



OPEN

DATA DESCRIPTOR

A 10 m maize, rice and soybean yield dataset from 2016 to 2021 in Northeast China

Fei Teng^{1,2}, Minglei Wang^{1,3}, Wenjiao Shi^{1,4}✉, Li Pan⁵, Jinghan Guo^{1,4} & Xiangming Xiao⁵

Accurate mapping of crop yields is essential for informed agricultural decision-making and optimal allocation of resources. Current crop yield datasets are deficient in large-scale, high-resolution information regarding the long-term spatial and temporal distribution of crop yields. To address this challenge, we developed a method of vegetation photosynthesis model combined with transition coefficient, producing a detailed dataset with 10 m resolution, covering major regions of maize, rice, and soybean in Northeast China from 2016 to 2021. The method introduces a dynamic observation index (APAR_{ε_g}) and a composite yield-conversion coefficient (a), which presents an innovative method for estimating crop yields without field measurements. Validation results show that, for maize, rice, and soybean, the model achieves r values of 0.39, 0.51, and 0.52; MREs of 12.14%, 11.96%, and 14.06%; and rRMSEs of 16.97%, 16.12%, and 17.26%, respectively. The dataset offers valuable insights into crop yield distribution, supporting better agricultural decision-making and resource optimization.

Background & Summary

Accurate estimation of crop yields is crucial for maintaining agricultural stability and ensuring national food security, especially for staple crops such as maize, rice and soybean¹. These crops play a critical role in meeting the basic needs of billions of people worldwide. However, global climate change has brought unprecedented challenges to food security, including extreme weather events, water scarcity, and land degradation. Concurrently, external factors like geopolitical tensions, the COVID-19 pandemic, and economic volatility further complicate the global agricultural and food security landscape. Given these dynamic challenges, monitoring crop growth and estimating crop yields are vital for making informed import and export decisions, efficiently allocating agricultural resources, and formulating robust national food security strategies². Crop yield estimates not only play a pivotal role in accurately forecasting food production and assessing essential supply need, but also foster economic stability and drive agricultural progress³.

In recent years, significant progress has been made in crop yield estimation, especially through remote sensing technology⁴. This has enabled the development of crop yield datasets at various spatial resolutions (Table 1), such as 550 km⁵, 125 km⁶, 43 km⁷, 10 km^{8,9}, 4 km^{10,11}, 1 km¹², 500 m¹³ and 30 m¹⁴, which together constitute an important research foundation in this field. However, although these existing global and regional scale crop yield datasets provide valuable scientific basis for effectively monitoring and analyzing the dynamic changes of agricultural productivity at the provincial, river basin and even county levels, these datasets are mainly generated by spatial allocation or downscaling of statistical yield data of national or provincial administrative units¹⁴. This downscaling approach, which relies on top-level aggregated statistics of crop production or yield, may fail to accurately capture the true fluctuations and driving factors of field-level crop yields in regions with high spatial heterogeneity. This challenge is particularly acute in China, where a decentralized farming model, fragmented landholdings, diverse cropping systems, and varying management practices are the norm^{10–14}. The main reason is that the analyses with coarse resolution data contain a mixed pixel effect, blurring and interfering with the unique yield signal characteristics of small-scale plots.

¹Key Laboratory of Land Surface Pattern and Simulation, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, 100101, China. ²Shanghai Surveying and Mapping Institute, Shanghai, 200063, China. ³Collage of Geographical Science, Shanxi Normal University, Taiyuan, 030031, China. ⁴College of Resources and Environment, University of Chinese Academy of Sciences, Beijing, 100049, China. ⁵School of Biological Sciences, University of Oklahoma, Norman, OK, 73019, USA. ✉e-mail: shiwj@reis.ac.cn

References	Resolution	Spatial coverage	Span	Crop types
You <i>et al.</i> ⁵	550 km	Global	2000	20 crops
Grogan <i>et al.</i> ⁶	550 km	Global	2015	160 crops
Iizumi <i>et al.</i> ⁷	125 km	Global	1982–2006	4 crops
Qin <i>et al.</i> ⁸	10 km	Global	2010–2020	4 crops
Monfreda <i>et al.</i> ⁹	10 km	Global	2000	175 crops
Wu <i>et al.</i> ¹⁰	4 km	Asian	1995–2015	Rice
Zhang <i>et al.</i> ¹¹	4 km	Global	1982–2020	Wheat
Cheng <i>et al.</i> ¹²	1 km	China	2001–2015	Maize and wheat
Wu <i>et al.</i> ¹³	500 m	United States	2008–2018	Maize and soybean
Zhao <i>et al.</i> ¹⁴	30 m	China	2016–2021	Wheat

Table 1. Summary of studies on crop yield datasets.

Methodologies for crop yield estimation are broadly categorized into four main approaches: traditional methods, process-based models, spatial statistical and artificial intelligence (AI) models, and satellite-based crop photosynthesis-yield models. Traditional methods, based on resource-intensive field surveys, provide high accuracy at the cost of significant labor and financial investment, limiting their scalability¹⁵. Process-based models, such as Decision Support System for Agrotechnology Transfer (DSSAT), use climate and soil data as inputs to simulate crop growth and estimate grain yields¹⁶. By modeling crop development in relation to environmental factors, they provide a comprehensive approach to yield prediction¹⁷. However, these models often require detailed, localized input data, which can be difficult to obtain for large-scale applications¹⁸. In contrast, spatial statistical and AI models, which include those utilizing remote sensing data (such as vegetation indices) and crop yield data for training, focus on identifying spatial patterns to predict yields¹⁹. Machine learning models have been widely used to enhance prediction accuracy²⁰. Nevertheless, they depend heavily on large training datasets and high computational demands, which may limit their application in real-time field settings²⁰. Additionally, these models face challenges related to generalizability and environmental heterogeneity, especially when down-scaling estimates to smaller areas²¹. Satellite-based crop photosynthesis-yield models, merging the strengths of the aforementioned approaches, enhance prediction reliability by simulating dry matter dynamics^{22,23}.

Currently, the satellite-based crop photosynthesis-yield models such as the Carnegie Ames Stanford Approach (CASA)²⁴, Photosynthesis (PSN)²⁵, and Vegetation Photosynthesis Model (VPM)²⁶ have played pivotal roles in estimating terrestrial ecosystem Gross Primary Productivity (GPP) and Net Primary Productivity (NPP) through remote sensing methods²⁷. Among these, the VPM is widely used worldwide for its integration of the Enhanced Vegetation Index (EVI) and Land Surface Water Index (LSWI), developed using satellite remote sensing and flux observation data²⁸, especially in estimating C₃ and C₄ cropland yields^{29–31}. However, to estimate crop yield, the yield estimated by NPP from VPM combined with other constant parameters requires field measurements of parameters like maximum light energy utilization, harvest index, and plant carbon content. The acquisition of this data not only requires calculations from multiple sample sources but is also susceptible to various factors, including theories related to crop growth conditions³², which hinder low-cost and rapid estimation of crop yields³³.

To address this challenge, we developed a method (named yield_{APAR_{ε_g}} & *a*) of yield estimated by the dynamic observation index according to the model principle of VPM (APAR_{ε_g}) combined with transition coefficient (*a*) to estimate the spatial distribution of crop yields with 10 m spatial resolution in Northeast China from 2016 to 2021. As a major grain-producing region, Northeast China contributes more than 20% of China's total grain production annually. Maize, rice, and soybean are the primary crops in this region³⁴. Specifically, combining meteorological information and remotely sensed images generated by the Sentinel-2 satellite, we improve the previous method (combining NPP from VPM with some empirical parameters, yield_{NPP&EP}) by calculating and linearly regressing them against the yield statistics, developing a yield estimation method (yield_{APAR_{ε_g}} & *a*) that does not require direct field measurements, while ensuring high accuracy and cost-effectiveness. Finally, the accuracy of the staple crop yield dataset of Northeast China was assessed by using both field observation data and official crop yield statistics. Our dataset provides a detailed description of yield patterns for maize, rice, and soybean in Northeast China, thereby supporting strategic agricultural planning and precise management. A key advantage is the ability to reveal subtle field-management differences that regional analyses miss, providing vital support for precision services such as agricultural insurance underwriting and yield-efficiency evaluations for smallholders.

Methods

Study area. The study area, located in Northeast China, encompasses Heilongjiang (HLJ), Jilin (JL), and Liaoning (LN) provinces (38°71′–53°44′N; 118°52′–134°17′E) (Fig. 1). This region features a cold-temperate continental monsoon climate, characterized by an average annual temperature ranging from −3.8 °C to 11.3 °C and annual precipitation between 298 mm and 880 mm. Northeast China experiences four distinct seasons, with dry, cold winters and rainy, warm summers. Benefits from well-balanced light, temperature, and water availability, as well as fertile black soil, this area has become a major grain-producing region. Maize, rice, and soybean are the primary crops, and the planting period is from May to September. In 2022, the production of maize, rice and soybean in Northeast China accounted for 33.39%, 18.34%, and 51.79% of their national production respectively, and the planting areas of these crops accounted for 30.64%, 16.81%, and 52.29% of their national areas, respectively³⁵.

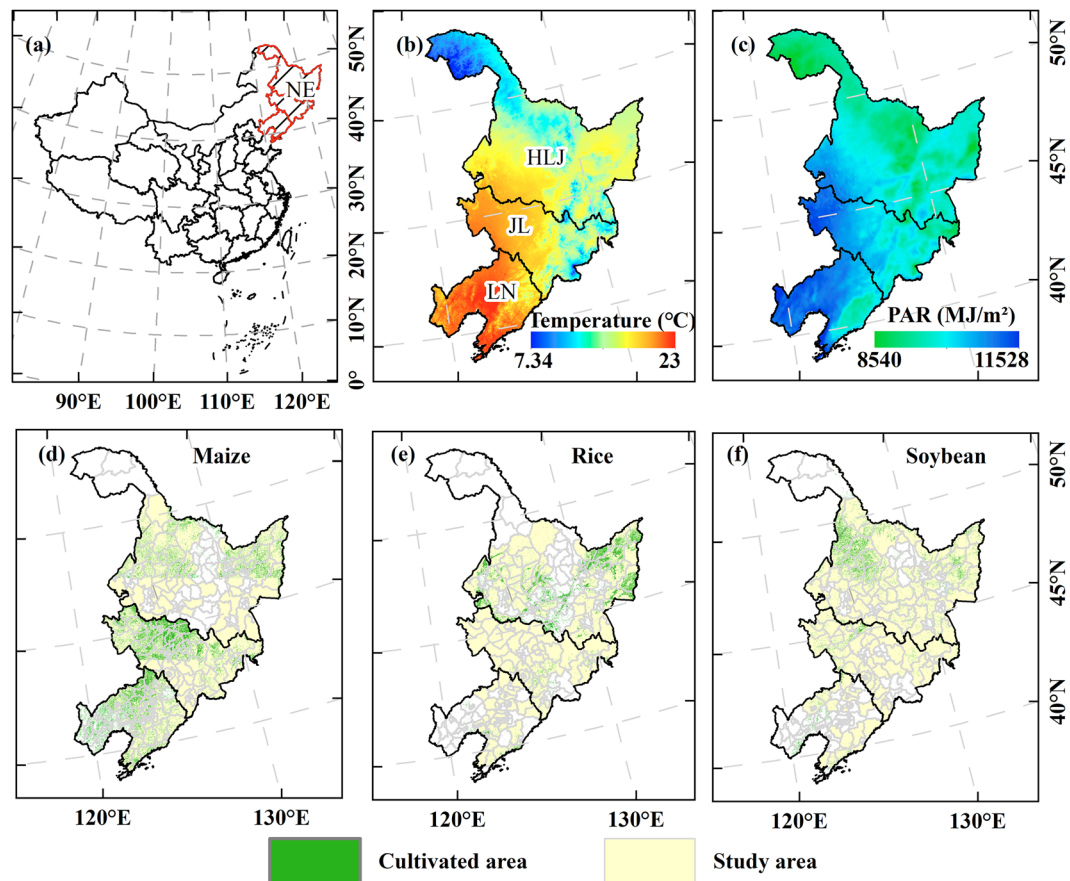


Fig. 1 The geographical location of the study area and the spatial distribution of key parameters. (a) the location of the Northeast China; (b) the average temperature from 2016 to 2021; (c) the average Photosynthetically Active Radiation (PAR) from 2016 to 2021; the study area and distribution of maize (d), rice (e) and soybean (f), respectively.

Data collection, preprocessing and parameter calculation. The data used for the model in this study are listed in Table 2. Full technical details of the model input and output, data preprocessing, quality control, and mechanism for handling missing data are provided in Supplementary information.

Sentinel-2 Multi-Spectral Instrument (MSI) Level-2A imagery, provided by the European Space Agency (ESA) at a 10 m spatial resolution, was acquired for the growing seasons of maize, rice, and soybean in the study area from 2016 to 2021. The data acquisition and all subsequent preprocessing steps were performed on the Google Earth Engine (GEE) platform. To ensure data quality, a series of preprocessing procedures were implemented. Initially, cloud and cloud-shadow masking were applied using the QA60 quality assessment band. The resulting clear-sky observations were then subjected to spatio-temporal filtering to reduce noise. Our preprocessing workflow also included the standardization of auxiliary meteorological data and the calculation of a custom moisture index, culminating in a multi-source fused dataset. Finally, using the processed dataset, we computed two remote sensing indices including Enhanced Vegetation Index (EVI) and Land Surface Water Index (LSWI)³⁶.

The Photosynthetically Active Radiation (PAR) data are derived from the Global Land Surface Characteristic Parameter Data Product dataset GLASS (Global Land Surface Satellite), with a spatial resolution of 5 km and a temporal resolution of 1 day³⁷. Since the PAR data in this dataset are only available up to 2020, the average value of the PAR from 2016 to 2020 was adopted for the data in 2021 in this study.

The monthly mean temperature, maximum and minimum temperature data from May to September were obtained from the National Tibetan Plateau Data Centre (<https://data.tpdc.ac.cn/>)³⁸, with a spatial resolution of 1 km.

The spatial distribution dataset of maize, sourced from the National Ecological Science Data Centre (<http://www.nesdc.org.cn/>), has a spatial resolution of 30 m and a temporal resolution of one year³⁹. The spatial distribution dataset of rice is from the National Ecological Science Data Centre (<http://www.nesdc.org.cn/>), with corresponding a spatial resolution of 20 m and a temporal resolution of one year⁴⁰. In addition, the spatial distribution data of rice for 2017 were used for 2016. The spatial distribution data of soybean were obtained from the National Earth System Science Data Centre (<http://www.geodata.cn/>), with a spatial resolution of 30 m and a temporal resolution of one year⁴¹. Similarly, the distribution for soybean for 2017 were used in place of the data for 2016.

Data types	Contents	Resolution	Span	Sources
Sentinel-2 data	EVI and LSWI	10 m	2016–2021	Google Earth Engine (GEE)
Meteorological data	Temperature	1 km	2016–2021	Peng <i>et al.</i> ³⁸
Global Land Surface Characteristic Parameter Data Product dataset	PAR	5 km	2016–2020	Liang <i>et al.</i> ³⁷
Maize cultivation distribution data	Classification of Maize	30 m	2016–2021	Shen <i>et al.</i> ³⁹
Rice cultivation distribution data	Classification of Rice	30 m	2017–2021	Li <i>et al.</i> ⁴⁰
Soybean cultivation distribution data	Classification of Soybean	30 m	2016–2021	Di <i>et al.</i> ⁴¹
Statistical data	Crop yields	County	2016–2021	National Bureau of Statistics ⁴²
Statistical data	Crop yields in the state farms of Heilongjiang Province	State farm	2016–2018	https://www.chinabdh.com/
Site-based yield data	Crop yields	Site	2016–2021	National Ecosystem Research Network of China

Table 2. Details on the datasets used in this study.

Site-based yield data of different crops in the study area were obtained from the National Field Scientific Observation Research Stations in Hailun City and Shenyang City, covering the period from 2016 to 2021 (<https://www.cnern.ac.cn/>). These data were used to verify the accuracy of the estimated output. The statistical data at the city and county scales are derived from statistical yearbooks for the same period (<http://www.stats.gov.cn>)³⁵. It is worth noting that due to the reform of the agricultural reclamation system in Heilongjiang Province in 2019, the statistical yearbooks for the period of 2016–2018 lack production measurements for the state farms in reclamation districts. Consequently, the yield data in Heilongjiang Province used in this study were derived by combining data from the relevant reclamation areas with county-level statistics⁴² (Table S1).

VPM for crop yield estimation. The schematic of data collection and preprocessing, model construction and validation are shown in Fig. 2.

The formula for crop yield converting from NPP is as follows^{43,44}:

$$Y = \frac{NPP * (1 - b) * HI}{c * (1 - \omega)} \quad (1)$$

where Y is the crop yield; b is the proportion of biomass in the underground part of the crop compared to the whole plant; HI is the harvest index; c is the plant carbon content; ω is the water content coefficient of the crop in the post-harvest storage period. According to the previous study, b, HI, c, and ω are taken as 0.1, 0.49, 0.45, and 14%, respectively⁴⁴.

NPP can be estimated using VPM. The absorption of PAR by the vegetation canopy can be classified into chlorophyll absorbing fraction ($FPAR_{chl}$) and non-photosynthetic vegetation absorbing fraction ($FPAR_{NPV}$) in VPM. Among them, the chlorophyll absorption part ($FPAR_{chl}$) is involved in photosynthesis, while the non-photosynthetic vegetation absorption part ($FPAR_{NPV}$) is not involved in photosynthesis^{25,45}. NPP can be estimated according to the following equations:

$$NPP = GPP * CUE \quad (2)$$

$$GPP = LUE * APAR_{chl} \quad (3)$$

$$APAR_{chl} = FPAR_{chl} * PAR \quad (4)$$

$$LUE = LUE_0 * T_{scalar} * P_{scalar} * W_{scalar} \quad (5)$$

where NPP stands for net primary productivity; GPP stands for gross primary productivity; CUE stands for carbon use efficiency, which is the conversion coefficient of GPP to NPP^{46–48}; LUE stands for light use efficiency (in gC/MJ); PAR stands for Photosynthetically Active Radiation (in MJ/m²); $APAR_{chl}$ stands for PAR absorbed by chlorophyll in the canopy; $FPAR_{chl}$ stands for the proportion of PAR absorbed by the vegetation canopy accounted for by the absorbing fraction of chlorophyll; LUE_0 represents the maximum light energy use efficiency (in gC/MJ)^{13,49}; T_{scalar} , P_{scalar} and W_{scalar} represent the stress coefficients of temperature, phenology and water on the maximum light energy utilization at the canopy scale⁵⁰.

T_{scalar} characterizes the effect of temperature on photosynthesis and can be calculated using the following equation⁵¹:

$$T_{scalar} = \frac{(T - T_{min})(T - T_{max})}{(T - T_{min})(T - T_{max}) - (T - T_{opt})^2} \quad (6)$$

where T_{min} , T_{max} , T_{opt} are the minimum, maximum and optimum temperatures required for vegetation photosynthesis, respectively. The look-up table method was used to determine the triple base temperatures of different

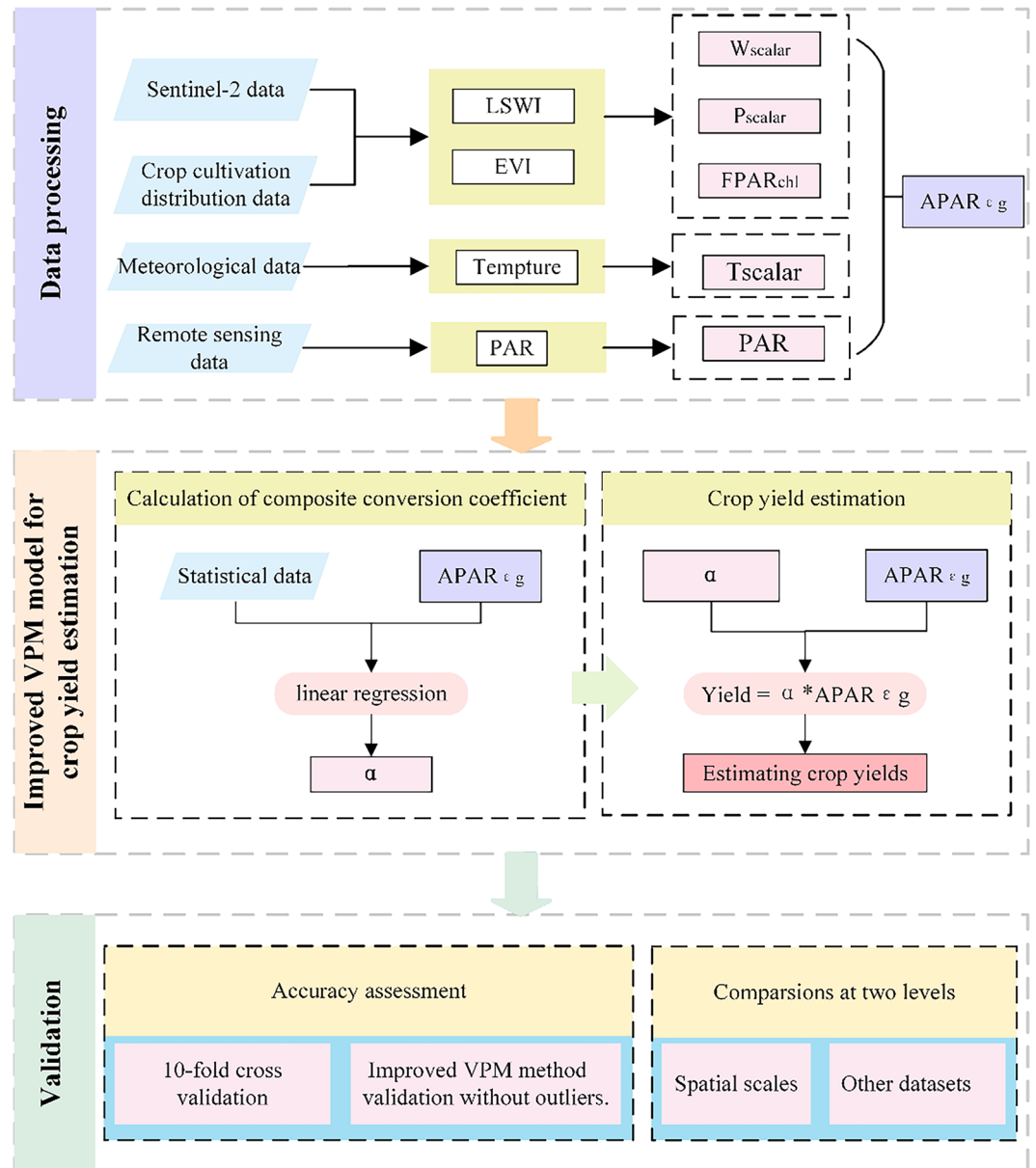


Fig. 2 Schematic of data preprocessing, model construction and dataset generation for estimating yields of jade, rice and soybean using yield_{APAR_{εg}} & *a*.

vegetation types⁵¹. We distinguished between T_{min} , T_{max} and T_{opt} for C3 and C4 crops, which can be obtained from the relevant literature⁵². T_{scalar} is set to 0 when the air temperature is less than the minimum photosynthetic temperature.

P_{scalar} characterizes the effect of changes in leaf phenology on photosynthesis at the crop canopy scale and is calculated depending on the type of crop. Crops like rice, soybean, and maize undergo distinct phases-from leaf emergence to full spreading. During the emergence to spreading phase, the calculation is defined by formula (7). Once leaves reach full spreading (denoted by $LSWI_{max}$), P_{scalar} reaches its peak value of 1 for the remainder of the growing season.

$$P_{scalar} = \frac{1 + LSWI}{2} \tag{7}$$

W_{scalar} represents the vegetation moisture factor, which quantifies the influence of moisture on photosynthesis, typically determined using $LSWI$ ⁵³. The calculation of the vegetation moisture factor is expressed as:

$$W_{scalar} = \frac{1 + LSWI}{1 + LSWI_{max}} \tag{8}$$

The $FPAR_{chl}$ is approximated by a linear function of EVI:

$$FPAR_{chl} = \alpha * EVI \quad (9)$$

where α is an empirical coefficient, which takes the value of 1⁵³.

Yield_ $APAR_{\epsilon_g}$ & a method for crop yield estimation. In the VPM, relatively fixed regionalized parameters such as LUE_0 , HI, CUE, c , b and ω are mainly affected by factors such as crop growth, environment, climatic conditions, crop varieties, and agricultural management conditions in the study area. In this study, we proposed a dynamic observation index $APAR_{\epsilon_g}$ and a composite yield-conversion coefficient a to replace multiple dynamic observation variables and multiple regionalized fixed parameters to reduce and integrate the key parameters of the original equation. According to equations ((1)–(9)), the crop yield can be estimated as follows:

$$Y = \frac{T_{scalar} * P_{scalar} * W_{scalar} * FPAR_{chl} * PAR * CUE * (1 - b) * HI * \epsilon_0}{c * (1 - \omega)} \quad (10)$$

We propose a dynamic observational index termed $APAR_{\epsilon_g}$. From a physical and biological perspective, $APAR_{\epsilon_g}$ represents the Photosynthetically Active Radiation that is jointly constrained by temperature (T_{scalar}), phenology (P_{scalar}), water availability (W_{scalar}), and canopy light absorption capacity ($FPAR_{chl}$), quantifying the effective photosynthetic energy input under real-world environmental constraints (in MJ/m²). Its mathematical expression is:

$$APAR_{\epsilon_g} = T_{scalar} * P_{scalar} * W_{scalar} * (FPAR_{chl} * PAR) \quad (11)$$

Correspondingly, we propose a yield conversion coefficient a that consolidates all the empirical parameters LUE_0 , HI, CUE, c , b and ω in the comprehensive yield equation. They collectively define the overall efficiency of converting the captured effective energy ($APAR_{\epsilon_g}$) into final yield. Therefore, we consolidate these empirical factors into a single yield conversion coefficient a :

$$a = \frac{CUE * (1 - b) * HI * LUE_0}{c * (1 - \omega)} \quad (12)$$

The $APAR_{\epsilon_g}$ can be calculated by the parameters we have mentioned in the section of “Data collection, pre-processing and parameter calculation”. The coefficient a is affected by a variety of factors such as crop type, environmental conditions of the study area, management practices, and crop varieties. Rather than prescribing fixed values for each component of a , we calibrate it empirically. Specifically, a for each city was derived using a zero-intercept linear regression model that relates the average $APAR_{\epsilon_g}$ in the specific crop distribution area to the statistical yields of the counties within the city for the period 2016–2021 (Fig. S1). Finally, the crop yield in each grid can be calculated as follows:

$$Y = a * \sum_{t_0}^{t_1} (APAR_{\epsilon_g t} * D_t) \quad (13)$$

where t_0 and t_1 are the starting and ending months of the crop growing season, $APAR_{\epsilon_g t}$ is the monthly average $APAR_{\epsilon_g}$ for month t and D_t is the number of days in month t .

Validation. The validation of our simulation results was evaluated through three types of data including statistical data, field data, and existing data products. For statistical data, we implemented a 10-fold cross-validation procedure to prevent overfitting and obtain an unbiased performance estimate because our model relies on statistical datasets for calibrating key parameters (specifically, parameter a). Specifically, based on the county-level statistical yield data and $APAR_{\epsilon_g}$ values for each crop across Northeast China from 2016 to 2021, we divided the dataset into 10 subsets. During each iteration, 90% of the data is used to train the linear regression model, and the remaining 10% is retained as an independent validation set to calculate the prediction error. For field data and existing data products, we conducted ground-truth validation by comparing the field data from Hailun and Shenyang stations during 2016–2021 with our results, the 10 km Global Gridded Crop Production (GGCP) dataset⁸, and the Spatial Production Allocation Model (SPAM) yield dataset⁵⁴. The above verification processes were all evaluated in this study using correlation coefficient (r), mean relative error (MRE), and relative root mean square error (rRMSE)² for evaluation.

Data Records

The dataset is available on ref. ⁵⁵. We provide the maize, rice and soybean yield distribution maps and datasets for Northeast China from 2016 to 2021. The crop yield distribution maps and datasets for Northeast China from 2016 to 2021 are named according to the respective years and are provided in Geo TIFF format with a spatial resolution of 10 m. Pixel values in these maps and datasets represent crop yield values, expressed as crop yield per unit area, in tonne per hectare (t/ha). Specific datasets detailing the yield for each crop are available in the provided link (<https://doi.org/10.6084/m9.figshare.27717624.v3>)⁵⁵.

Technical Validation

Comparing the performance of different methods using statistical data. The result shows that the yield estimation accuracy of the yield_{APAR_{ε_g} & *a* method is superior to that of the yield_{NPP&EP}. Notably, the yield_{APAR_{ε_g} & *a* method has achieved significant enhancement in *r*, MRE and rRMSE indexes, especially for rice, which *r* has improved by 0.41, MRE by 9.15% and rRMSE by 10.04%. Overall, crop yield estimation in the improved model shows better consistency with statical data at county level (Fig. 3). The advantage of the yield_{APAR_{ε_g} & *a* method lies in that it uses of APAR_{ε_g} and *a* to replace numerous regionally fixed parameters and dynamic observation variables in the original model, enabling the capture of key parameters without the need for direct measurement. This approach effectively captures interaction effects among diverse geographic factors, thereby optimizing and adjusting many of the discrepancies inherent in the yield_{NPP&EP}.}}}

Comparing the performance of our model for estimating yields across different yield levels using statistical data.

By comparing the estimated yields of maize, rice, and soybean in Northeast China from 2016 to 2021 with statistical data, we observed large deviations from estimated norms in several counties, especially in areas with low statistical anomalies. To improve the accuracy of yield predictions, we screened the average statistical yields of different crops separately by crop type: for maize, we excluded yields below 4 t/ha, 5 t/ha, and 6 t/ha; for rice, yields below 5 t/ha, 6 t/ha, and 7 t/ha; and for soybean, yields lower than 1.5 t/ha or greater than 3.5 t/ha. Some counties exhibited significant deviations from the expected low-value statistical anomalies (Fig. 3a,d,g). Meanwhile, maize, rice and soybean performed well under the yield_{APAR_{ε_g} & *a* method algorithm in districts with average yields of more than 6 t/ha, 7 t/ha and 3.5 t/ha, resulting in MRE of 11.78%, 11.46% and 18.75% and rRMSE of 16.73%, 15.39% and 23.33%, respectively. However, for maize, rice and soybean using the dataset within the optimal accuracy range mentioned above led to significant disruptions in temporal continuity across some counties, resulting in poor performance at the time-series scale. To tackle this issue, we chose datasets that closely aligned with the optimal accuracy criteria, focusing on districts with average statistical yields surpassing 5 t/ha for maize, 6 t/ha for rice and between 1.5 t/ha and 3.5 t/ha for soybean. This approach resulted in MRE of 12.14%, 11.93% and 14.06% and rRMSE of 16.97%, 15.97% and 17.26% for maize, rice and soybean, respectively. Such meticulous selection not only ensures the accuracy of the model predictions and maintains temporal consistency across counties, but also illustrates the particular applicability of the yield_{APAR_{ε_g} & *a* method to yield estimation in areas with high crop yields. Post-screening, prediction accuracy was assessed separately for each crop (Fig. S2).}}

In areas where maize yields averaged over 5 t/ha, LN exhibited poor cross-validation results with MRE and rRMSE values of 15.76% and 21.41%, respectively. Conversely, other regions demonstrated high consistency in measured and validated data, with JL achieving the best cross-validation results (MRE = 9.67%, rRMSE = 12.99%) followed by HLJ (MRE = 11.74%, rRMSE = 15.89%). Similarly, for rice, LN's cross-validation results were subpar, with MRE and rRMSE values of 13.30% and 17.09%, respectively. JL led with superior performance (MRE = 10.93%, rRMSE = 14.61%), followed closely by HLJ (MRE = 11.96%, rRMSE = 16.12%). Regarding soybean, LN exhibited poorer cross-validation results (MRE = 16.29%, rRMSE = 20.47%), whereas HLJ demonstrated the best performance (MRE = 13.97%, rRMSE = 18.32%), and JL also showed strong results (MRE = 14.06%, rRMSE = 17.26%).

Validation using field data and comparison with other datasets in Northeast China. To verify the accuracy and advancement of the dataset in this study, we compared the results of this study which aggregated to the 10 km spatial resolution (Fig. 4a), the GGCP dataset (Fig. 4b), this study at 10 m resolution (Fig. 4c) and the SPAM dataset (Fig. 4d) with the site-based yield data from Hailun and Shenyang for the 2016–2021. Specifically, the 10 m yield estimates from this study first aggregated to the 10 km spatial resolution prior to quantitative comparison. Compared with the GGCP dataset, our results show lower MRE and rRMSE (Fig. 4a,b). Furthermore, a comparison with maize and soybean from the 2020 SPAM dataset (with no matching sample points for rice) (Fig. 4c,d) confirmed the superior accuracy of our dataset in yield estimation. Specifically, our dataset exhibited a higher *r* value (0.09 higher), lower MRE (6.6% lower), and lower rRMSE (7.0% lower) compared to the SPAM dataset, with tighter clustering around the 1:1 line, further indicating improved accuracy in both *r* value and error metrics.

Patterns of crop yields in Northeast China from 2016 to 2021. We evaluated the annual yield maps by comparing their spatiotemporal dynamics against government-reported county-scale yields for maize, rice, and soybean. Yield distributions aligned closely across sources: maize predominantly ranged between 5–10 t/ha, rice between 5–10 t/ha, and soybean between 1–3 t/ha (Figs. S3–5). Spatial patterns from 2016 to 2021 consistently reflected statistical yield trends at this scale (Fig. 5): maize yields decreased east to west, rice declined west to east, and soybean diminished south-to-north (Figs. S6–8). However, these spatial relationships exhibited inconsistencies in select years, likely attributable to anomalous events such as localized disasters.

Uncertainties and limitations. The input-dependency analysis of the maize yield model in Jilin Province indicates that uncertainties associated with spatial representation constitute the dominant and explicitly quantified factor influencing dataset reliability. A comparison between simulations driven by 30 m and 1 km resolution inputs (Fig. S9) demonstrates improved performance when higher-resolution spatial data are used, highlighting the critical role of spatial scale in resolving yield variability. The spatially heterogeneous model performance observed across regions can therefore be partly attributed to these scale-related effects¹³. Even with limited observation points, this method can effectively assess spatial heterogeneity. It remains capable of supporting a range of

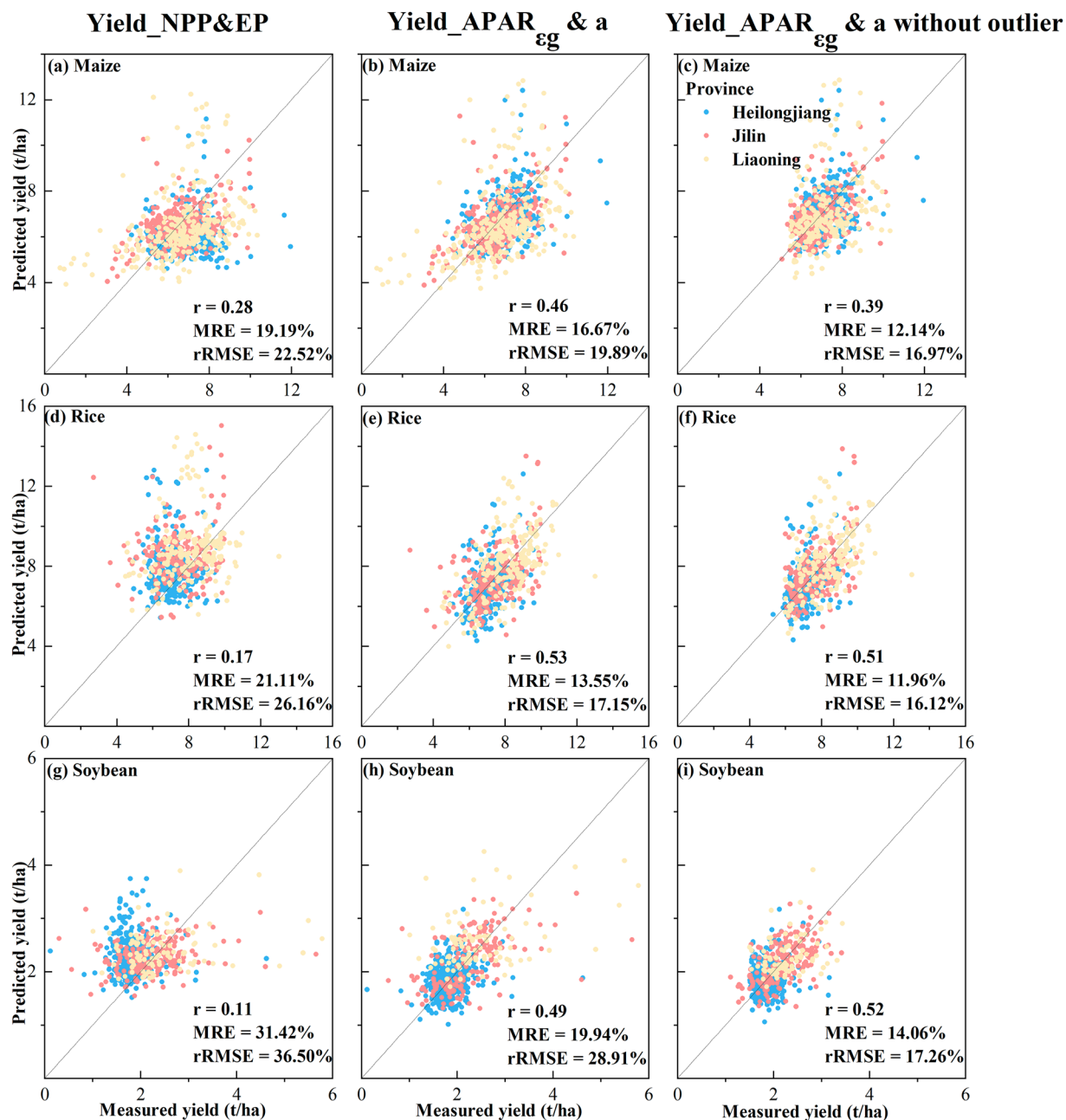


Fig. 3 The prediction performance of maize yield from 2016 to 2021 simulated by the yield_NPP&EP method (a), the yield_APAR_{εg} & a method (b) and the yield_APAR_{εg} & a method (c) without outliers; the prediction performance of rice yield from 2016 to 2021 simulated by the yield_NPP&EP method (d), the yield_APAR_{εg} & a method (e) and the yield_APAR_{εg} & a method (f) without outliers; the prediction performance of soybean yield from 2016 to 2021 simulated by the yield_NPP&EP method (g), the yield_APAR_{εg} & a method (h) and the yield_APAR_{εg} & a method (i) without outliers.

applications, including rapid and reliable regional yield assessment, analysis of yield spatial patterns, comparison of inter-county differences, and study of long-term yield trends. At the county level, the dataset reliably captures overall yield levels and inter-county differences, and cross-validation among counties supports its regional applicability. However, within individual counties, yield variability is strongly influenced by fine-scale heterogeneity in management practices, environmental conditions, and stress factors that are not fully represented in the model inputs. In high-yield areas, relatively homogeneous cropping systems and favorable growing conditions reduce the sensitivity of yield estimates to input uncertainties, resulting in more stable model performance. In contrast, low-yield or anomalous regions often exhibit greater variability in crop management, pest and disease impacts, and episodic stress events, which can amplify spatial and temporal uncertainties and lead to reduced stability of yield estimates.

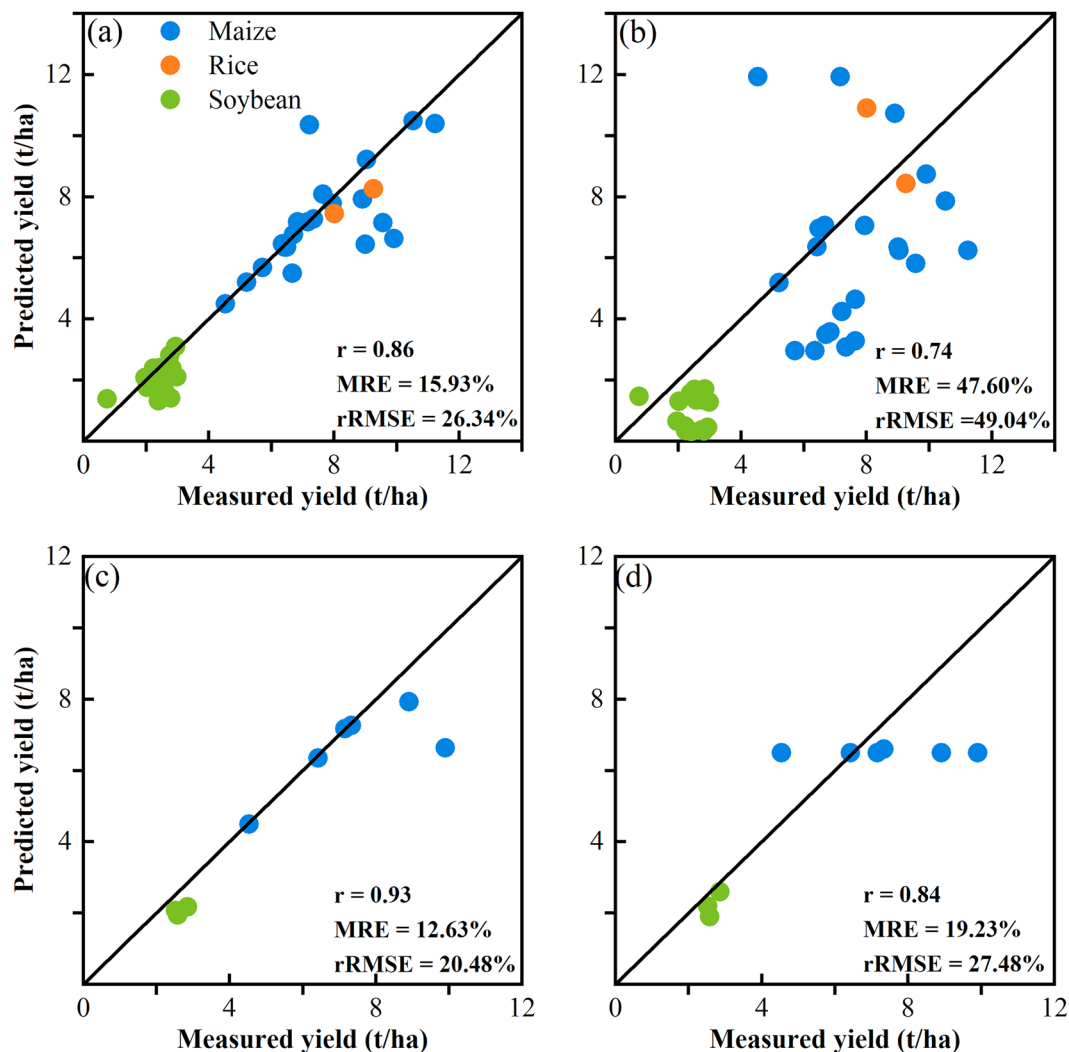


Fig. 4 Comparison of simulated yields from this study which aggregated to the 10 km spatial resolution (a), the GGCP dataset (b), this study at 10 m resolution (c) and the SPAM dataset (d) with the measured yields.

Beyond spatial representation, several sources of uncertainty further constrain model performance¹³. Temporal gaps and cloud contamination in optical remote sensing data primarily introduce noise into vegetation signal retrieval during periods critical for yield formation, affecting interannual variability rather than systematic yield levels²⁸. Uncertainties in phenology detection may propagate into seasonal productivity estimates, particularly in regions characterized by heterogeneous cropping practices. In addition, the reliance on monthly temperature data limits the representation of short-term thermal stress and daytime photosynthetic conditions, which may influence yield estimates in anomalous years. Finally, the absence of long-term, plot-level yield observations does not directly contribute to model error, but constrains independent validation at sub-county or field scales, thereby limiting the direct applicability of the dataset for field-scale or management-level applications⁵⁶.

Future improvements to the model could address these limitations. These limitations reflect common constraints in large-scale yield datasets derived from remote sensing and reanalysis products. Future extensions of the dataset may reduce some of these uncertainties through multi-sensor data integration to mitigate cloud effects, the incorporation of pixel-level uncertainty characterization⁵⁷, and the use of refined temperature metrics to better approximate daytime crop growth conditions.

Usage Notes

The 10 m annual crop yield maps produced in this study represent a crucial regional-scale dataset for agricultural research. This dataset is most suitable for crop yield monitoring and trend analysis in medium or high yield, large-scale cultivation areas in Northeast China and similar agro-ecological zones, and is particularly well suited for regional-scale analyses, spatial pattern assessment, inter-county comparisons, and long-term yield variability studies. The yield_{APAR_c} & *a* method is suitable for using in the estimation of relatively high-yield areas for maize, rice and soybean. This dataset not only holds direct significance for improving agricultural yield efficiency and ensuring food security but also has far-reaching impacts on environmental protection, disaster mitigation, and decision-making.

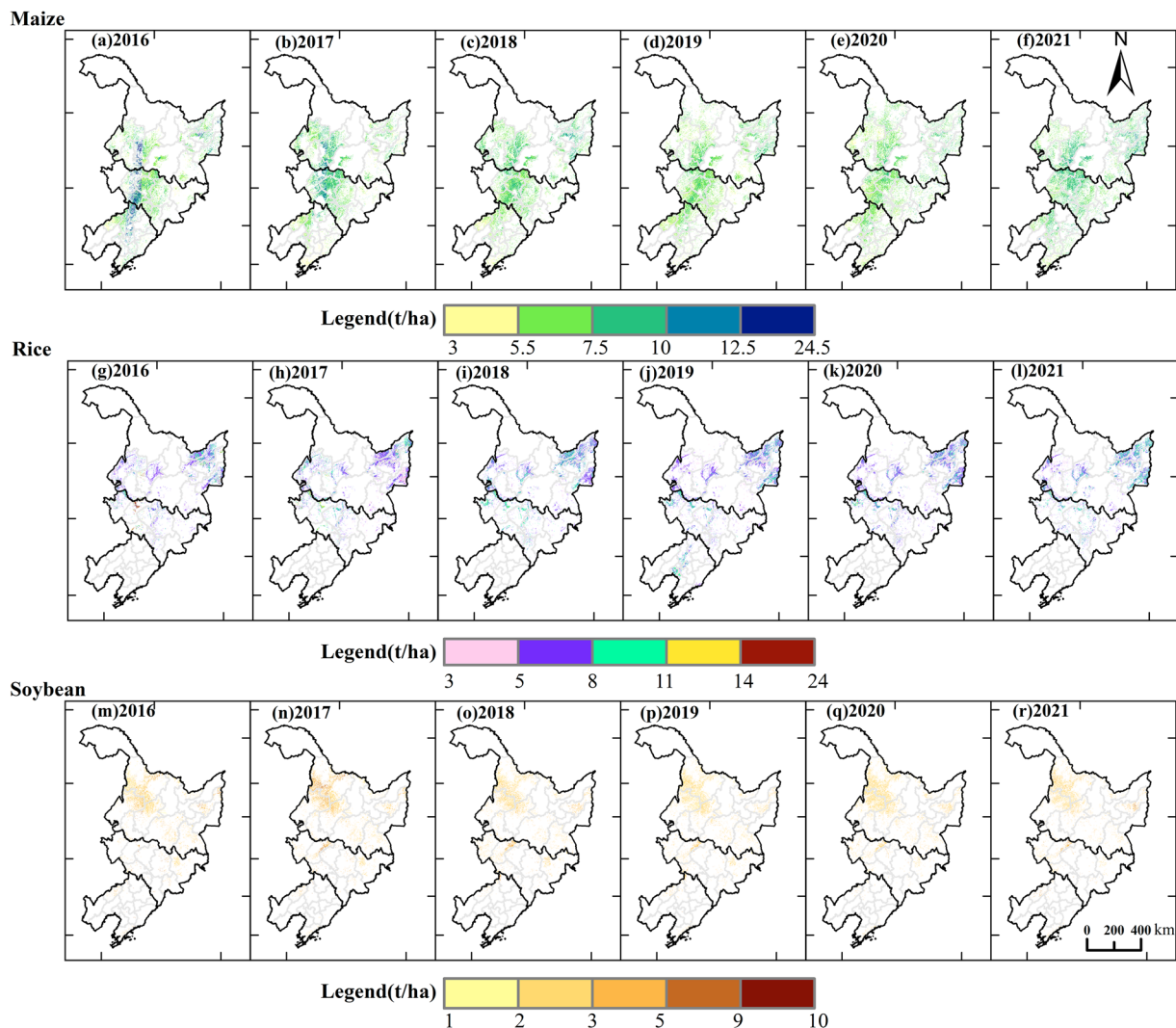


Fig. 5 The yield distribution of maize (a–f), rice (g–l) and soybean (m–r) with 10 m spatial resolution in the Northeast China from 2016 to 2021.

Data availability

The dataset generated during the current study is available in the figshare under the <https://doi.org/10.6084/m9.figshare.27717624.v3>.

Code availability

The codes we developed for crop yield computation and crop yield dataset generation are available at <https://doi.org/10.6084/m9.figshare.27717624.v3>. In this code, we use GEE to call Sentinel-2 data for the 2016–2021 crop growth period. In addition, we used the ArcGIS 10.7 mosaics data tool to generate a crop yield dataset for the Northeast China region.

Received: 21 January 2025; Accepted: 26 January 2026;

Published online: 03 February 2026

References

1. Liu, Y. *et al.* Analysis of spatio-temporal variation of crop yield in China using stepwise multiple linear regression. *Field Crops Research* **264**, 108098 (2021).
2. Yan, H. *et al.* Satellite-based evidences to improve cropland productivity on the high-standard farmland project regions in Henan Province, China. *Remote Sensing* **14**, 1724 (2022).
3. Karthikeyan, L., Chawla, I. & Mishra, A. K. A review of remote sensing applications in agriculture for food security: Crop growth and yield, irrigation, and crop losses. *Journal of Hydrology* **586**, 124905 (2020).
4. Jing, X. *et al.* Comparison of machine learning algorithms for remote sensing monitoring of rice yields (in Chinese). *Spectroscopy and Spectral Analysis* **42**, 1620–1627 (2022).
5. You, L. *et al.* Generating global crop distribution maps: From census to grid. *Agricultural Systems* **127**, 53–60 (2014).
6. Grogan, D. *et al.* Global gridded crop harvested area, production, yield, and monthly physical area data circa 2015. *Scientific Data* **9**, 15 (2022).

7. Iizumi, T. *et al.* Historical changes in global yields: major cereal and legume crops from 1982 to 2006. *Global Ecology and Biogeography* **23**, 346–357 (2014).
8. Qin, X. *et al.* Global Gridded Crop Production Dataset at 10 km Resolution from 2010 to 2020. *Scientific Data* **11**, 1377 (2024).
9. Monfreda, C., Ramankutty, N. & Foley, J. A. Farming the planet: 2. Geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000. *Global Biogeochemical Cycles* **22**, 1–19 (2008).
10. Wu, H. *et al.* AsiaRiceYield4km: seasonal rice yield in Asia from 1995 to 2015. *Earth System Science Data* **15**, 791–808 (2023).
11. Zhang, Z. *et al.* Estimating global wheat yields at 4 km resolution during 1982–2020 by a spatiotemporal transferable method. *Remote Sensing* **16**, 2342 (2024).
12. Cheng, M. *et al.* High-resolution crop yield and water productivity dataset generated using random forest and remote sensing. *Scientific Data* **9**, 641 (2022).
13. Wu, X. *et al.* Spatial-temporal dynamics of maize and soybean planted area, harvested area, gross primary production, and grain production in the Contiguous United States during 2008–2018. *Agricultural and Forest Meteorology* **297**, 108240 (2021).
14. Zhao, Y. *et al.* ChinaWheatYield30m: a 30 m annual winter wheat yield dataset from 2016 to 2021 in China. *Earth System Science Data* **15**, 4047–4063 (2023).
15. Gao, Z., Xu, X., Wang, J., Jin, H. & Yang, H. Cotton yield estimation based on similarity analysis of time-series NDVI (in Chinese). *Transactions of the Chinese Society of Agricultural Engineering* **28**, 148–153 (2012).
16. Meng, J. H., Wang, Y. N., Lin, Z. X. & Fang, H. T. Progress and perspective of crop growth models (in Chinese). *Transactions of the Chinese Society for Agricultural Machinery* **55**, 1–15 (2024).
17. Brisson, N. *et al.* STICS: a generic model for the simulation of crops and their water and nitrogen balances. I. Theory and parameterization applied to wheat and maize. *Agronomy* **18**, 311–346 (1998).
18. Huang, H. *et al.* 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 Transactions on Geoscience and Remote Sensing* **61**, 1–18 (2023).
19. Dela Torre, D. M. G., Gao, J. & Macinnis-Ng, C. Remote sensing-based estimation of rice yields using various models: A critical review. *Geo-Spatial Information Science* **24**, 580–603 (2021).
20. Zhou, X. *et al.* Predicting grain yield in rice using multi-temporal vegetation indices from UAV-based multispectral and digital imagery. *ISPRS Journal of Photogrammetry and Remote Sensing* **130**, 246–255 (2017).
21. Jiang, H. *et al.* A deep learning approach to conflating heterogeneous geospatial data for maize yield estimation: A case study of the US Maize Belt at the county level. *Global Change Biology* **26**, 1754–1766 (2020).
22. Wang, X., Zhang, F. & Johnson, V. C. New methods for improving the remote sensing estimation of soil organic matter content (SOMC) in the Ebinur Lake Wetland National Nature Reserve (ELWNNR) in northwest China. *Remote Sensing of Environment* **218**, 104–118 (2018).
23. Kayad, A. *et al.* Radiative transfer model inversion using high-resolution hyperspectral airborne imagery-Retrieving maize LAI to access biomass and grain yield. *Field Crops Research* **282**, 108449 (2022).
24. Potter, C. S. *et al.* Terrestrial ecosystem production: a process model based on global satellite and surface data. *Global Biogeochemical Cycles* **7**, 811–841 (1993).
25. Running, S. W. *et al.* A continuous satellite-derived measure of global terrestrial primary production. *BioScience* **54**, 547–560 (2004).
26. Niu, Z. E., Yan, H. M., Chen, J. Q., Huang, M. & Wang, S. Q. Comparison of crop gross primary productivity estimated with VPM model and MOD17 product in field ecosystem of China (in Chinese). *Transactions of the Chinese Society of Agricultural Engineering* **32**, 191–198 (2016).
27. Ji, Z., Pan, Y., Zhu, X., Zhang, D. & Wang, J. A generalized model to predict large-scale crop yields integrating satellite-based vegetation index time series and phenology metrics. *Ecological Indicators* **137**, 108759 (2022).
28. Wang Y. H. *et al.* Estimating maize yield in Jilin Province of China using VPM model combined with conversion coefficient (in Chinese). *Transactions of the Chinese Society of Agricultural Engineering* **40**, 195–201 (2024).
29. Wu, X. *et al.* Spatiotemporal consistency of four gross primary production products and solar-induced chlorophyll fluorescence in response to climate extremes across CONUS in 2012. *Journal of Geophysical Research: Biogeosciences* **123**, 3140–3161 (2018).
30. He, M. *et al.* Regional crop gross primary productivity and yield estimation using fused Landsat-MODIS data. *Remote Sensing* **10**, 372 (2018).
31. Wu, X. *et al.* Spatiotemporal changes of winter wheat planted and harvested areas, photosynthesis and grain production in the contiguous United States from 2008–2018. *Remote Sensing* **13**, 1735 (2021).
32. World Health Organization. *The state of food security and nutrition in the world 2023: urbanization, agrifood systems, transformation and healthy diets across the rural-urban continuum.* (Food and Agriculture Organization of the United Nations Press, 2023).
33. Dong, J. *et al.* The obstacles and breakthrough paths of security assurance of grain production capacity in Northeast China (in Chinese). *Research of Agricultural Modernization* **44**, 755–764 (2023).
34. Zhang, Z., Luo, Y., Han, J., Xu, J. & Tao, F. Estimating global wheat yields at 4 km resolution during 1982–2020 by a spatiotemporal transferable method. *Remote Sensing*, **16**, 2342 (2024).
35. National Bureau of Statistics of China. *China Statistical Yearbook 2023* (China Statistics Press, 2023).
36. Xiao, X. M. *et al.* Mapping paddy rice agriculture in southern China using multi-temporal MODIS images. *Remote Sensing of Environment* **95**, 480–492 (2005).
37. Liang, S. L. *et al.* The global land surface satellite (GLASS) product suite. *Bulletin of the American Meteorological Society* **102**, E323–E337 (2021).
38. Peng, S. Z., Ding, Y. X., Liu, W. Z. & Li, Z. 1 km monthly temperature and precipitation dataset for China from 1901 to 2017. *Earth System Science Data* **11**, 1931–1946 (2019).
39. Shen, R. Q. *et al.* A 30 m resolution distribution map of maize for china based on landsat and sentinel images. *Journal of Remote Sensing* **5**, 1–12 (2022).
40. Li, S. *et al.* A long-term paddy rice distribution dataset in Asia at a 30 m spatial resolution. *Scientific Data* **12**, 1052 (2025).
41. Di, Y. *et al.* Recent soybean subsidy policy did not revitalize but stabilize the soybean planting areas in Northeast China. *European Journal of Agronomy* **147**, 126841 (2023).
42. National Bureau of Statistics of China. *Statistical Yearbook of Heilongjiang State Farms.* (China Statistics Press, 2022).
43. Liu, G. *et al.* On the accuracy of official Chinese crop production data: Evidence from biophysical indexes of net primary production. *Proceedings of the National Academy of Sciences* **117**, 25434–25444 (2020).
44. Xiao, X. M. *et al.* Modeling gross primary production of temperate deciduous broadleaf forest using satellite images and climate data. *Remote Sensing of Environment* **91**, 256–270 (2004).
45. Xiao, X. M. *et al.* Satellite-based modeling of gross primary production in an evergreen needleleaf forest. *Remote Sensing of Environment* **89**, 519–534 (2004).
46. Delucia, E. H. *et al.* Forest carbon use efficiency: is respiration a constant fraction of gross primary production? *Global Change Biology* **13**, 1157–1167 (2007).
47. Peng, D. *et al.* Modelling paddy rice yield using MODIS data. *Agricultural and Forest Meteorology* **184**, 107–116 (2014).
48. Albrizio, R. & Steduto, P. Photosynthesis, respiration and conservative carbon use efficiency of four field grown crops. *Agricultural and Forest Meteorology* **116**, 19–36 (2003).
49. Pan, L. *et al.* Interannual variations and trends of gross primary production and transpiration of four mature deciduous broadleaf forest sites during 2000–2020. *Remote Sensing of Environment* **304**, 114042 (2024).

50. Raich, J. W. *et al.* Potential net primary productivity in South America: application of a global model. *Ecological Applications* **1**, 399–429 (1991).
51. Chen, J. *et al.* Estimation of gross primary productivity in Chinese terrestrial ecosystems by using VPM model (in Chinese). *Quaternary Sciences* **34**, 732–742 (2014).
52. Xiao, X. M. *et al.* Characterization of forest types in Northeastern China, using multi-temporal SPOT-4 VEGETATION sensor data. *Remote Sensing of Environment* **82**, 335–348 (2002).
53. Yan, H. M. *et al.* Modeling gross primary productivity for winter wheat-maize double cropping system using MODIS time series and CO₂ eddy flux tower data. *Agriculture, Ecosystems & Environment* **129**, 391–400 (2009).
54. Yu, Q. *et al.* A cultivated planet in 2010-Part 2: the global gridded agricultural production maps. *Earth System Science Data* **12**, 3545–3572 (2020).
55. Teng, F. Data underlying the publication: A 10 m maize, rice and soybean yield dataset from 2016 to 2021 in Northeast China. *figshare* <https://doi.org/10.6084/m9.figshare.27717624.v3> (2025).
56. Zhang, Y. *et al.* A global moderate resolution dataset of gross primary production of vegetation for 2000–2016. *Scientific Data* **4**, 170165 (2017).
57. Azzari, G. & Lobell, D. B. Landsat-based classification in the cloud: An opportunity for a paradigm shift in land cover monitoring. *Remote Sensing of Environment* **202**, 64–74 (2017).

Acknowledgements

This research was supported by the National Key Research and Development Program of China (2022YFB3903504) and the National Natural Science Foundation of China (72221002 and 42330707).

Author contributions

Fei Teng: Data curation, methodology, software, visualization, validation, writing-original draft. Minglei Wang: writing-review & editing. Wenjiao Shi: Resources, methodology, writing-review & editing, funding acquisition. Li Pan: writing-review & editing. Jinghan Guo: writing-review & editing. Xiangming Xiao: writing-review & editing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41597-026-06719-0>.

Correspondence and requests for materials should be addressed to W.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026