scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Mapping global yields of four major crops at 5-minute resolution from 1982 to 2015 using multi-source data and machine learning

Juan Cao[1,2], Zhao Zhang[1]✉, Xiangzhong Luo[3], Yuchuan Luo[1], Jialu Xu[1], Jun Xie[1], Jichong Han[1] & Fulu Tao[2,4]✉

Accurate, historical, and continuous global crop yield data are essential for assessing risks to the global food system. However, existing datasets often have limited spatial and temporal resolution. Here, we introduce GlobalCropYield5min, a novel gridded dataset providing crop yield data for major crops — including maize, rice, wheat, and soybean — from 1982 to 2015, with a spatial resolution of 5 arc-minutes. We developed three machine learning (ML) models for each country and crop, using crop statistics from approximately 12,000 administrative units, along with satellite data, climate variables, soil properties, agricultural practices, and climate modes. The optimal predictors and ML model were selected to estimate annual crop yield for each 5 × 5 arc-minute grid cell. Results show good model performance, with R^2 ranging from 0.70 to 0.95, and RMSE (NRMSE) from 0.16 t/ha (5%) to 1.1 t/ha (20%). GlobalCropYield5min outperforms other global yield datasets in spatial resolution, temporal coverage, and accuracy. This dataset is crucial for investigating climate-crop yield interactions and managing agricultural disaster risks.

Background & Summary

The increasing frequency of extreme climate events, coupled with increasing global volatility — such as the Russian-Ukrainian war and food trade restrictions — has had a dramatic impact on global food security and agricultural trade liberalization in recent years[1–3]. Since the 1990s, crop production has significantly increased both locally and globally[4,5], primarily due to higher crop yields (production per harvested area), rather than the expansion of harvested areas. However, year-to-year variability remains substantial due to climate fluctuations[6,7]. As the world's population continues to grow and environmental pressures increase, research on crop yields and their spatiotemporal changes has gained increasing prominence[6,8]. Therefore, a high-quality, spatially explicit, gridded crop yield dataset spanning several decades would be invaluable for addressing the risks posed by climate change, identifying yield gap, maintaining stability in international markets, and ensuring food security[9–11].

Several global historical crop yield datasets, covering recent decades and derived from census or satellite data, are publicly available. These datasets have greatly supported studies on global food security, sustainable development, and climate change impacts and adaptation[12–14]. For example, M3Crops assigned uniform statistical crop yields to downscaled, crop-specific areas and generated the first global yield mapping circa 2000[14]. The Global Agro-Ecological Zones (GAEZ) dataset provided the potential yields circa 2000, 2010 and 2015, respectively, through statistical downscaling[12,15]. Ray2012, based on approximately 13,500 crop censuses, allocated four major crop yields to each grid within each political unit and provided three 5-year average maps (1995, 2000 and 2005)[16]. The Spatial Production Allocation Model (SPAM) used cross entropy to downscale crop statistics to grid cells[17], producing global crop yields data around 2000, 2005, and 2010 [13,18,19]. The four datasets primarily rely on downscaling agricultural census data and other supplementary information to a resolution of 5 arc-minutes.

[1]School of National Safety and Emergency Management, Beijing Normal University, Beijing, 100875, China. [2]Key Laboratory of Land Surface Pattern and Simulation, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, 100101, China. [3]Department of Geography, National University of Singapore, Singapore, Singapore. [4]College of Resources and Environment, University of Chinese Academy of Sciences, Beijing, China. ✉e-mail: zhangzhao@bnu.edu.cn; taofl@igsnrr.ac.cn

However, these datasets are limited to specific years, and as a result, lack temporal continuity. The GDHY (Global Dataset of Historical Yields) and GGCMI (Global Gridded Crop Model Intercomparison) are the only two datasets that provide continuous temporal coverage, supporting studies on the inter-annual variations in crop yields[10,20–22]. Unfortunately, these two crop yield datasets use national-level crop statistics or data from limited experimental sites as inputs, and have a coarser spatial resolution of 0.5°[2,23]. Therefore, they may overlook localized spatial variations at the sub-national or smaller scale, especially in major crop-producing countries across larger areas. Additionally, capturing interannual fluctuations and upward trends becomes challenging with these datasets[24]. Therefore, current global crop yield datasets are lack of detailed time and spatial information, warranting the necessary to develop a new, high-resolution, and long-term global crop yield dataset[2].

The current global crop yield estimations primarily rely on process-oriented crop models (e.g., GGCMI) and downscaling methods (including M3Crops, GAEZ, SPAM and GDHY). However, the limited availability of detailed and precise input data on a global scale — such as spatially heterogeneous soil, cultivars, and management practices — poses challenges in calibrating process-oriented crop models. As a result, they often rely on national statistics or limited experimental data, and in some cases, remain uncalibrated. These limits them to accurately simulate crop yields over larger areas[24–26]. The downscaling methods depend on current agricultural statistics and other supplementary information, limiting the scalability of these methods across multiple crops, regions, and years[10,23]. Encouragingly, in recent years, Machine Learning (ML) algorithms — known for their ability to capture complex, higher-order, and nonlinear relationships between predictors and target variable — have been successfully employed for yield estimation in some countries[27,28]. However, to the best of our knowledge, global yields mapping for four major crops based on ML methods has not yet been conducted.

The objectives of this study are to: (1) develop ML models and select the optimal one for each country and crop by integrating satellite data, climate data, soil properties, agricultural management practices, and detailed census records for approximately 12,000 administrative units; (2) producing yield maps for maize, soybean, rice, and wheat, with a 5 arc-minute spatial resolution for the period 1982–2015 (GlobalCropYield5min), using the optimal ML model; and (3) comprehensively evaluate the data products. These four crops, as the primary cereal and legume sources for global population, account for nearly two-thirds of total calorie production worldwide[16].

Methods

Research framework. The study's flowchart is presented as Fig. 1. First, we compared three commonly used ML models for yield estimation, and selected the optimal models for each country and crop. Next, we applied the selected model to estimate annual crop yield for each 5×5 arc-minute grid cell from 1982 to 2015, producing a global crop yield dataset GlobalCropYield5min. To assess the accuracy of the dataset, we conducted a comprehensive evaluation from multiple perspectives. The evaluation included comparing simulation accuracy, analyzing the spatial patterns of the coefficient of variation (CV) and mean annual yield, as well as examining yield temporal trends and variations between GlobalCropYield5min and recorded data. Finally, the study compared the accuracy of GlobalCropYield5min with the SPAM and GDHY crop yield data for the years 2000, 2005, and 2010. The selected datasets are shown in Table 1.

Model development. The performance of ML models varies depending on factors such as data structure, distribution, and volume. It is important to note that a single model may not be universally applicable[29,30]. Consequently, three commonly used ML models — Random Forest (RF), eXtreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LightGBM) were employed for crop yield estimation. The LightGBM model supports efficient parallel training and offers notable advantages, including faster training, lower memory consumption, higher accuracy, and support for distributed, fast processing of large datasets. The XGBoost model utilizes optimization algorithms to reduce computational complexity and effectively mitigates overfitting through regularization, exhibiting significant advantages in handing large-scale datasets. Compared to XGBoost, the RF model is better suited for managing high-dimensional and noisy data. Further details about these models are described in previous studies[31–33].

Crop planting and harvesting months were determined according to Vogel *et al.*[34] (Fig. S4). All spatial data were aggregated to the administration unit level using the Google Earth Engine (GEE) platform. To ensure consistency, we re-gridded all datasets — including climate, NDVI, irrigation, fertilizer application rates, and soil data — onto a 5×5 arc-minute resolution grid over all months of the growing season, aligning with the harvested area dataset[14]. Cropped areas were masked using grid cells with a crop-specific harvested area fraction of $\geq$1% for each crop[35]. Before applying the ML models, all variables were standardized to have mean of 0 and standard deviation of 1. To evaluate and compare the accuracy of the three models, the datasets for all years and administrative units from 1982 to 2015 were randomly divided into two parts: 70% for predictor selection, model calibration/training, and through determination of optimal hyperparameters, and 30% for evaluating model performance.

For maize and rice, only the primary seasons were considered in this study, as they contribute most to national production and economic impacts[36]. For wheat, winter wheat and spring wheat differ in variety, growth habits, cropping regions, climate preferences, and harvest times[37]. For example, winter wheat requires vernalization to flower, so it must be sown before winter, typically in temperate climates. Spring wheat, on the other hand, does not need vernalization and is grown in warmer climates. Therefore, when modeling yields or studying the impacts of climate change, it is important to treat spring and winter wheat separately. Winter wheat and Spring wheat were determined following Vogel *et al.*[34] (Fig. S5).

Model predictor selection and parameter optimization. We designed an automatic selection process for the optimal combination of predictors and the parameters in a modular and extensible manner, allowing for the selection of the optimal models for different crops and countries. First, Recursive Feature Elimination

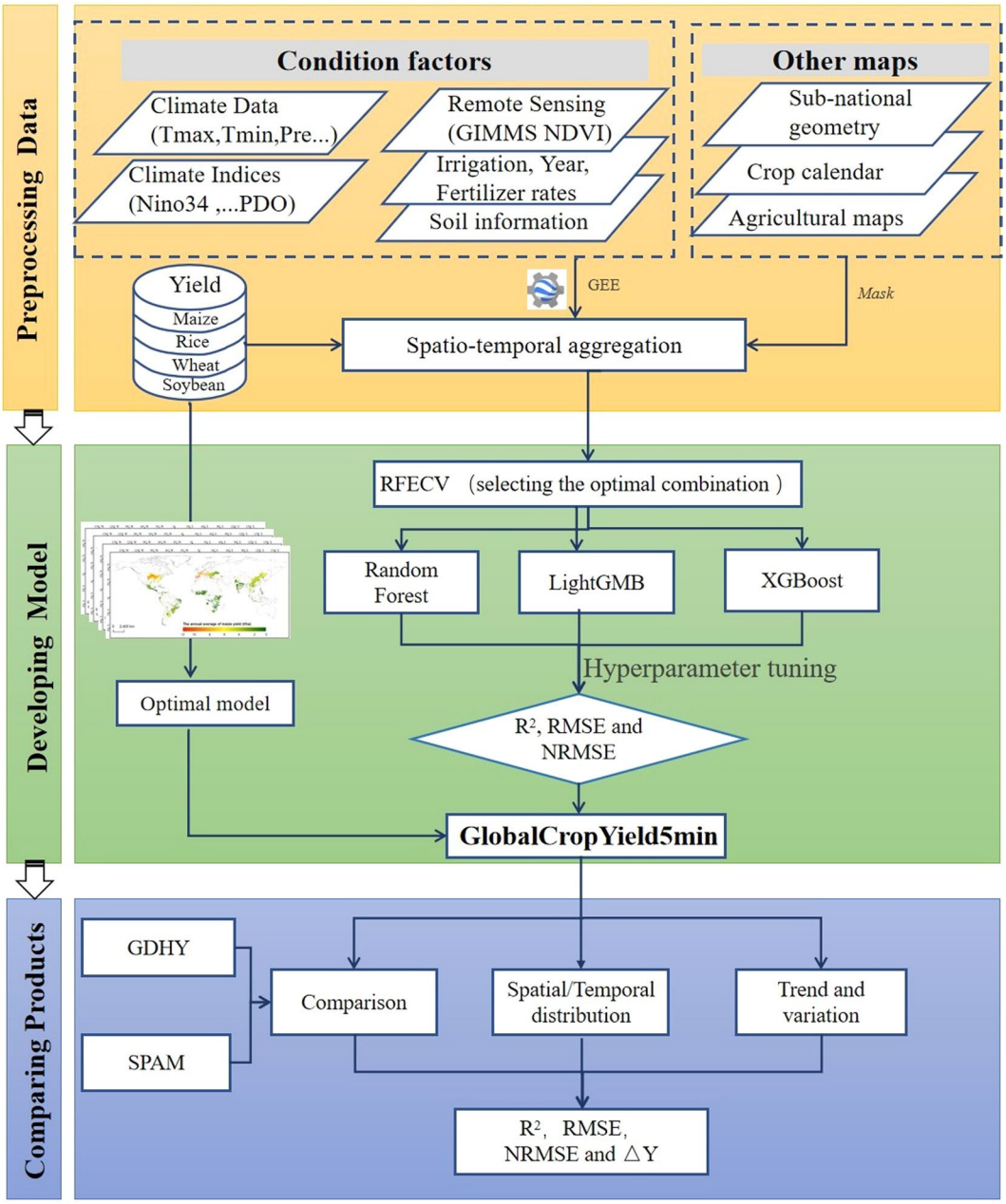

Fig. 1 The flowchart of this study.

cross-validation (RFECV)[38] was used to automatically select potential predictors with the best mean score across each country and crop, reducing predictors redundancy[39] and minimizing the risk of overfitting and collinearity. For parameter optimization, we first identified the range of parameters values requiring tuning. Subsequently, Bayesian optimization was utilized to select the optimal parameters[40,41]. The selection of models, predictors, and the model parameters varied by country and crops (Table S4).

Comparison and selection of the optimal yield estimation models. To evaluate the simulation accuracy of the GlobalCropYield5min product, this study first aggregated the dataset by administrative unit and then compared it with recorded data. The coefficient of determination (R^2), root mean square error (RMSE), and normalized root mean square error (NRMSE) were used to assess accuracy[42].

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_{i,j}^{True} - y_{i,j}^{Pred}\right)^2}{\sum_{i=1}^{n}\left(y_{i,j}^{True} - \bar{y}_{i,j}^{True}\right)^2} \quad (1)$$

Data	Data sources	Variables	Spatial/temporal resolution	Time coverage	Reference/data access
Yield data	~12,000 globally agricultural statistics	Yield	Administrative units, annual	1982–2015	See Table S1
Growing season dates	AgMIP harmonized crop calendar v1.25	Planting and harvesting months	0.5 × 0.5°, static	—	https://zenodo.org/record/3773820#.Y9e9AXZBzyS
Area harvested	Cropland Area fraction	Area harvested fraction	5 × 5 minute, static	circa 2000	http://www.earthstat.org/
Irrigation Areas	MIRCA2000	Irrigation Areas fraction	5 × 5 minute, static	circa 2000	https://www.uni-frankfurt.de/
Fertilizer	N application rates	Fertilizer application rates	0.5 × 0.5 °, annual	1982–2015	https://zenodo.org/record/4954582
Climate factors	TerraClimate	Tmin, Tmax, Tmean, Pre, Vap, Vpd, Srad	2.5 × 2.5 minute, monthly	1982–2015	https://doi.org/10.7923/G43J3B0R
	Climate indices	EA, IOD, Nino34, PDO, NAO, TSA	—, monthly	1982–2015	See Table S3
Soil information	HWSD	T_CLAY, T_GRAVEL, T_OC, T_PH_H2O, T_REF_BULK, T_SAND, T_SILT	1 × 1 km, static	—	https://www.fao.org/soils-portal
Remote sensing	GIMMS NDVI	NDVI	5 × 5 minute, 15-day	1982–2015	https://gee-community-catalog.org/projects/gimms_ndvi/#license

Table 1. Information on the selected datasets.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(y_{i,j}^{True} - y_{i,j}^{Pred}\right)^2}{n}} \tag{2}$$

$$NRMSE = \frac{RMSE}{\overline{y}_i} \tag{3}$$

Here, i represents the index of the administrative unit, n denotes the total number of administrative units, and j corresponds to the year. $y_{i,j}^{True}$ refers to the observed crop yield obtained from governmental or FAO websites for the i-th administrative unit in year j, $\overline{y}_{i,j}^{True}$ represents the average observed crop yield for the i-th administrative unit in year j, and $y_{i,j}^{Pred}$ indicates the GlobalCropYield5min yield estimate for the i-th administrative unit in year j. To ensure results stability, we ran each model for 50 times and calculated the mean predicted R^2, RMSE, and NRMSE, and the standard deviation (SD) as a measure of model performance. Therefore, all the R^2, RMSE, and NRMSE hereafter refer to the mean predicted R^2, RMSE, and NRMSE. The model with the highest predicted R^2 and lowest RMSE is selected as the optimal model for each crop and country.

Development of the GlobalCropYield5min product. For each crop and country, predictors at the gridded scale were aligned with those selected at the administrative scale level. These predictors were then input into the selected optimal model to generate gridded annual crop yield maps from 1982 to 2015 with a 5 arc-minute spatial resolution. To ensure the robustness of the results, 50 simulations were performed, and their outputs were averaged to create a global gridded long-term yield dataset. This process was consistently applied across all countries for each crop, and the resulting datasets were combined to form the GlobalCropYield5min product.

Verification of the GlobalCropYield5min product. The gridded crop yields were initially averaged and aggregated at the administrative unit level and then compared with recorded data to evaluate its accuracy. To further assess the product's ability to capture interannual yield variations, we compared the temporal patterns of actual yield and yield anomalies (ΔY) at the national level. This comparison utilized data from GlobalCropYield5min and reported yields, calculated as average values across all administrative units within each country. The ΔY was defined as follows:

$$\Delta Y_{i,j} = \frac{y_{i,j}^{True} - \hat{y}_{i,j}^{True}}{\overline{Y}_{t,g}} \times 100 \tag{4}$$

Here, $\hat{y}_{i,j}^{True}$ is calculated by the 5-year moving average method. The $\Delta Y_{i,j}$ (yield anomalies) represents the percent crop yield anomalies, which are widely used to quantify the changes in yields caused by climatic variability by removing the trend of the yield caused by non-climatic and time-dependent factors (i.e., demand, prices, technology, and other factors)[43,44]. It is worth mentioning that the percentage yield anomalies for the first two years and the last two years were not available because we used a time window of t−2 to t+2 to get moving average means.

Comparison with existing global crop yield products. To assess the accuracy of our product and compare it with two widely used global products — SPAM and GDHY — across various crops, countries, and globally, we followed a two-step process. First, we computed the average values for each administrative unit in 2000, 2005, and 2010, as these are the only three years publicly accessible for SPAM. Then, we compared the three sets of global gridded yield products with our collected yield data at the administrative unit level, using R^2 and RMSE between the reported yield and estimates.

Spatial uncertainty assessment of globalCropYield5min products. To evaluate the spatial uncertainty of the GlobalCropYield5min Products, we calculated the NRMSE following previous similar researches[45,46]. The NRMSE for each administrative unit was allocated to its centroid, and the kriging interpolation method was applied to spatially distribute the uncertainty.

Data

Crop system data. We collected crop yield, harvested area and production from various public sources, including national and regional statistical bureaus and agricultural agencies (Table S1, all websites are available and accessed before April 2020). Data availability varied across regions and time periods. To ensure a consistent global database for the four crops across all spatial levels, we conducted a preliminary assessment of data quality, excluding potential outliers that exceeded biophysically attainable yields for each crop type. Additionally, we excluded administrative units with missing yield data for more than one-third of the data collection period. In cases where crop yield data was entirely absent for a political unit during the study period or in a specific year, we examined the harvested areas and production data in the upper-level administrative units where the crop was harvested. If confirmed, we used an interpolation method to estimate the production and harvested area for the individual administrative unit. This involved selecting the five closest years of available data preceding the missing year and calculating the average harvested area and production for each administrative unit. Using the average values, we determined the proportion of the total harvested and production for each administrative unit. Subsequently, we used these proportion values to estimate the missing harvested area and production for the specific administrative unit, based on the corresponding harvested area and production at its upper administrative unit. For years when we have data for the administrative unit, we summed the harvested area and production up to the national level and compared it to the FAO-reported national data. In cases that discrepancies arose, we scaled the sub-national sum proportionately to match the FAO-reported data at the national level, assuming the FAO-reported data to be accurate. To ensure consistency, we converted all data for the period of 1982–2015 to the same units, specifically hectares for harvest areas, tons for production, and tons per hectare (ton/ha) for yield, as some countries used different units.

Regional climate and large-scale climate modes. In this study, we used the TerraClimate gridded monthly dataset, with a spatial resolution of 1/24° (~4 km)[47]. To verify the TerraClimate dataset in representing regional climate (RC) conditions, we initially examined the correspondence between monthly temperature and precipitation values from TerraClimate and national weather stations, specifically the China Meteorological Administration (CMA). For example, in the cultivation areas of winter wheat in mainland China (Fig. S1), the results showed the TerraClimate dataset effectively captured the spatial and temporal variations in regional climate, exhibiting high correlation coefficients (R^2) of 0.83 for temperature and 0.99 for precipitation (Fig. S2). The primary climate variables considered in the study include precipitation (Pre), minimum temperature (Tmin), maximum temperature (Tmax), mean temperature (Tmean), Vapor pressure (Vap), vapor pressure deficit (Vpd), and Srad (Surface shortwave radiation).

In addition to regional climate variables, large-scale climate (LC) oscillations, such as El Niño–Southern Oscillation (ENSO), Indian Ocean Dipole (IOD) and Pacific Decadal Oscillation (PDO), have been reported to cause extreme events like floods, droughts, and storms, with significant impacts on harvested area and productions[9,43,44,48–52]. Hence, we selected the monthly values of six major climate mode indices as potential predictors of crop yields. These indices include Nino34 (Niño 3.4 SST Index), EA (Eastern Atlantic pattern of the 500-hPa height), PDO, North Atlantic Oscillation (NAO), Atlantic Multidecadal Oscillation (AMO) and IOD (Table S3).

Satellite dataset, environmental stressor factors, and technology advancement. The Normal Difference Vegetation Index (NDVI) has been widely used for estimating crop yields and detecting crop drought stress[53,54]. The PKU GIMMS provides NDVI values with a temporal resolution of 15 days and a spatial resolution of 5 minutes, covering the entire globe, starting from 1982. The GIMMS NDVI product stands out due to its extended temporal coverage, enabling yield simulations before 2000 and surpassing other available satellite NDVI datasets in this regard[55]. In this study, we utilized NDVI data from PKU GIMMS Normalized Difference Vegetation Index dataset as a predictor for crop yield.

In addition, to capture the spatial and time-dependent improvement of agriculture technology information (TI), we incorporated additional predicting variables into our analysis[56–59]. Specifically, we used the irrigation area ratio (Fig. S3)[60], dynamic fertilizer application rates, and year[46,61]. These variables serve as indicators of evolving field management and techniques in agriculture. Recognizing the substantial impact of soil characteristics on crop growth and yield[62,63], we also considered six soil parameters for each 1 km × 1 km grid. These parameters, obtained from the Harmonized World Soil Database[31], include topsoil sand fraction (T_SAND), soil texture (T_TEXTURE), organic carbon content (T_OC), pH (T_PH_H2O), cation exchange capacity (T_CEC_SOIL) and bulk density (T_REF_BULK). Incorporating these soil parameters and spatial information (longitude, latitude, and elevation) as yield prediction variables enhances our understanding of the influence of constant environmental conditions (CEC) on crop performance.

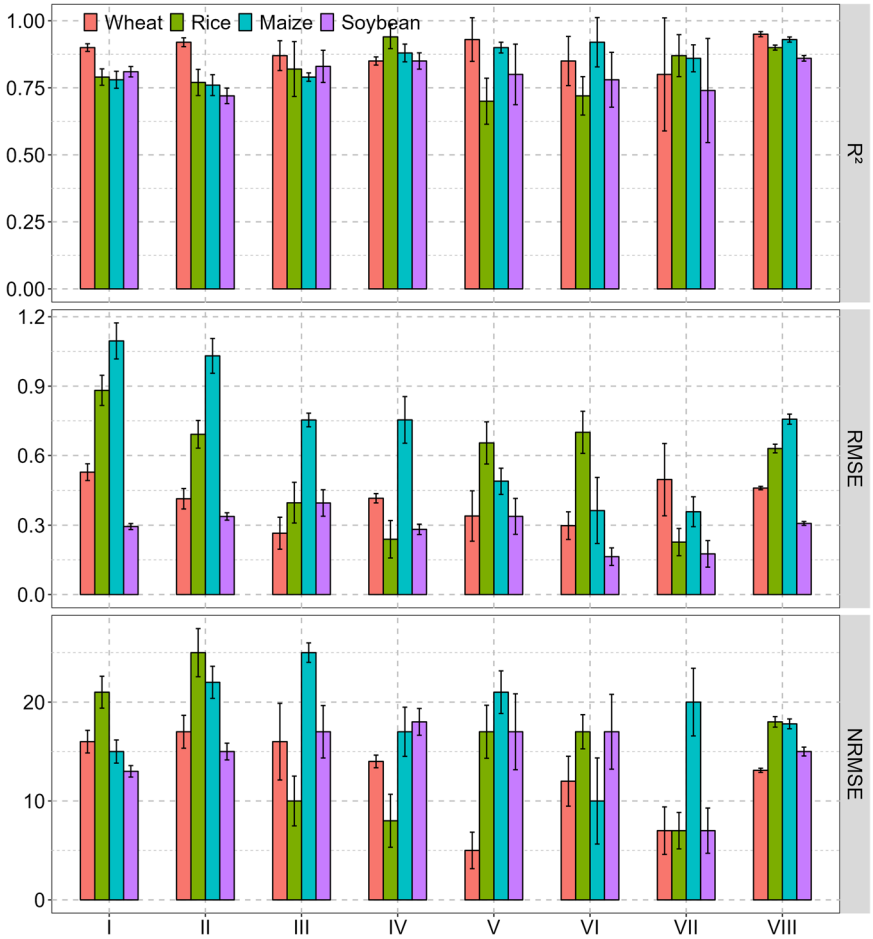

Fig. 2 The R^2 values of the three ML models for the top seven production countries and the globe using the testing dataset. The error bars are one standard deviation of R^2 from 50 ensemble simulations by randomly separating training and testing datasets. Note: I-VII represent the top seven production countries for each crop respectively and VIII represents the globe (see Table S2).

Data Records

The resultant GlobalCropYield5min dataset[64] in this paper is freely available online at https://doi.org/10.17632/hg8wzgx4yp.3. The dataset contains global gridded annual yield for the four major crops with 5-minute resolution from 1982 to 2015. Each crop dataset is in standard NetCDF4 format with the georeferenced information embedded. The unit of crop yield is t/ha the dataset.

Technical Validation

Performance of the ML models in estimating crop yield. We trained the three ML models separately for each country and crop. Figure 2 presents the R^2, RMSE (t/ha), and NRMSE (%) values for the top seven countries and globally, using the selected optimal ML models. Overall, the results demonstrate a high accuracy, although the optimal model type and predictor combinations varied across different crop and countries. At the global scale, wheat yield estimation exhibits the highest accuracy among the four crops, with a R^2 of 0.95 and a RMSE (NRMSE) of 0.46 t/ha (13.1%). Maize follows closely, with a R^2 of 0.93 and a RMSE (NRMSE) of 0.76 t/ha (17.8%). Rice yield estimation achieved a R^2 of 0.90 and a RMSE (NRMSE) of 0.63 t/ha (18.3%). The model for soybean yield estimation performed comparatively worse, with a R^2 of 0.86 and a RMSE (NRMSE) of 0.31 t/ha (15.3%).

The model skill is also the best for wheat among the major producing countries, with a R^2 ranging from 0.84 to 0.91 and a NRMSE (RMSE) from 5.6% (0.27 t/ha) to 16.7% (0.53 t/ha). For maize, the models perform best in Mexico and Ukraine, with a R^2 from 0.90 to 0.92, and a RMSE (NRMSE) from 0.50 t/ha (21.2%) to 0.39 t/ha (10.7%), respectively. In other countries, R^2 values range from 0.76 to 0.87, RMSE ranges from 0.36 to 1.1 t/ha, and NRMSE ranges from 15.3% to 25.1%. Regarding rice, the model achieved the highest accuracy in Bangladesh, with a R^2 of 0.90 and a RMSE (NRMSE) of 0.3 t/ha (9.8%). In other countries, the models had a R^2 ranging from 0.69 to 0.87 and a RMSE from 0.23 (7.6%) to 0.88 t/ha (24.6%). For soybean, the models exhibited a R^2 ranging from 0.72 to 0.85 and a RMSE (NRMSE) from 0.17 t/ha (7.2%) to 0.40 t/ha (18.3%). Furthermore, the error bars show relatively low deviations, suggesting the robustness of our models. The recorded and estimated yields for countries and the globe closely align along the 1: 1 line (Figs. S6-9). We further found that

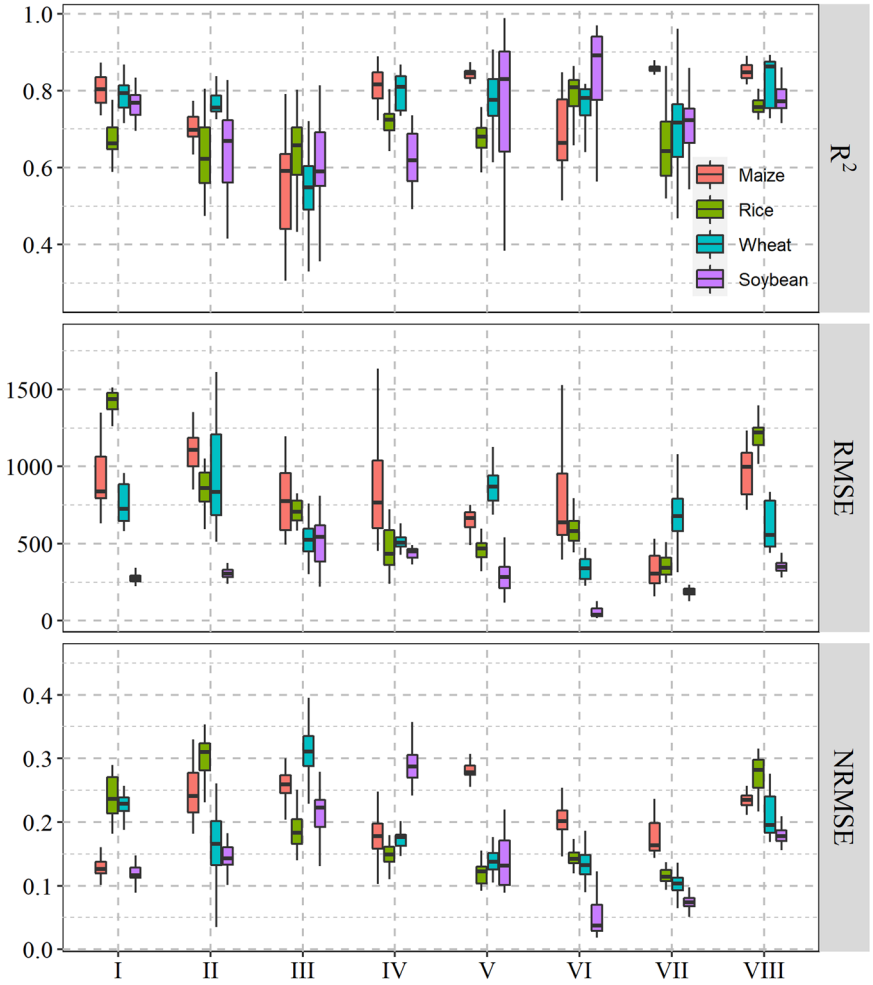

Fig. 3 Evaluation of the simulation accuracy using R^2, RMSE(t/ha), and NRMSE (%) of the four crops across the top seven production countries and the globe from 1982 to 2015. The black lines within the box indicate the medians in 34 years, while red dots represent means. The boundaries of boxes are the 25th and 75th percentiles and the whiskers below and above the boxes represent the 10th and 90th percentiles.

year and NDVI were the consistently important factors for four crop yield estimations, highlighting the role of long-term agronomic advancements and vegetation health in yield determination.

Comparison of GlobalCropYield5min products with observed records. Because global-scale field crop yield measurements were not available, we aggregated the GlobalCropYield5min data to the administrative unit level to compare the estimated yields with recorded data, enabling an analysis of yield estimation accuracy at the grid scale. At the global scale (Fig. 3), the mean R^2 between the estimates and recorded data for the top seven production countries was 0.85 for maize, 0.82 for wheat, 0.76 for rice and 0.78 for soybean. The mean RMSE (NRMSE) was 0.97 t/ha (24%), 0.62 t/ha (21%), 1.19 t/ha (28%), and 0.36 t/ha (18%) for maize, wheat, rice, and soybean, respectively. Overall, the simulation accuracy was significantly higher for maize and wheat than rice and soybean. For both maize and wheat, the average R^2 was greater than 0.7 in all major producing countries, with RMSE ranging from 0.32 t/ha to 1.1 t/ha and NRMSE from 10.6% to 28.1%, except for maize in Brazil and wheat in Russia. For rice, the R^2 ranged from 0.62 to 0.79, RMSE mostly ranged from 0.34 to 0.86 t/ha, and NRMSE ranged from 11.1 to 24.1%. Regarding soybean, the R^2 ranged from 0.61 to 0.76, RMSE ranged from 0.05 t/ha to 0.51 t/ha, and NRMSE ranged mostly from 5% to 21.8%. Note that a higher R^2 did not necessarily correspond to a lower RMSE (NRMSE) when comparing different crop types and countries, due to the substantial variation in recorded yields.

Regionally, the performance of the selected crop estimation model at the grid level varied by country and crop type. For the top seven production countries, the model performance was generally better for maize and wheat, with mean R^2 exceeding 0.7, except for maize in Brazil and wheat in Russia. The RMSE ranged from 0.33 t/ha to 1.1 t/ha, and NRMSE from 11% to 28%. The model performance was moderate for rice (mean R^2 ranging from 0.62 to 0.79 and RMSE from 0.34 t/ha to 1.45 t/ha) and soybean (mean R^2 ranging from 0.61 to 0.76 and RMSE from 0.19 t/ha to 0.51 t/ha, except for India).

In addition, we examined the prediction skill over time (Fig. S15). The average R^2 at the grid level ranged from 0.65 to 0.81, with RMSE (NRMSE) from 0.24 t/ha (13%) to 1.1 t/ha (38%) during the period of 1982–2015.

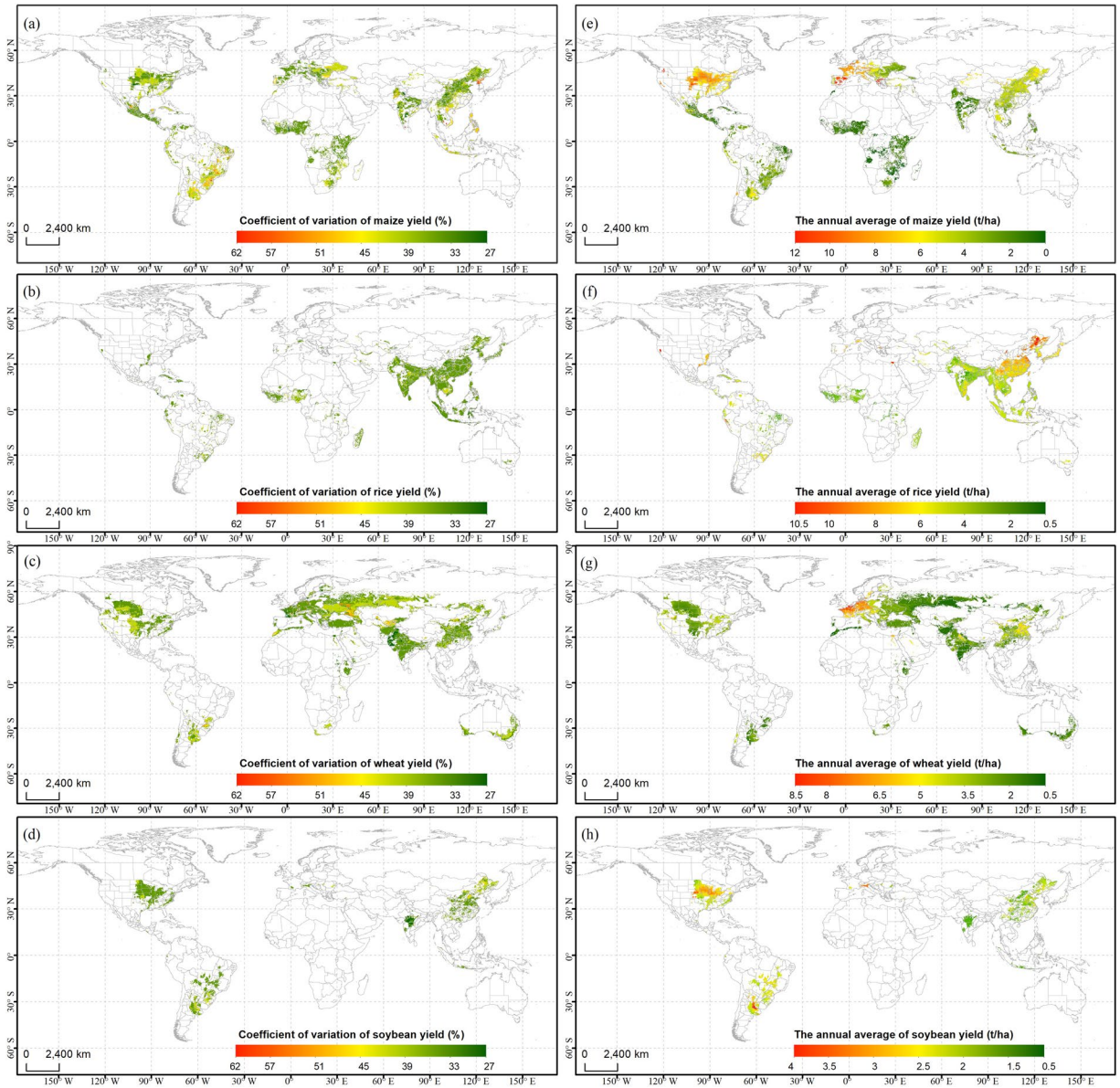

Fig. 4 Spatial pattern of CV and mean crop yields over the entire study period based on GlobalCropYield5min product for maize (**a, e**), rice (**b, f**), wheat (**c, g**) and soybean (**d, h**). The sample size is approximately 12,000 administrative units × 34 years for each crop. White represents the areas where crop was not harvested or analyzed.

The highest R^2 was achieved for maize in 2010, while the lowest R^2 was observed for rice in 2002. The prediction skill for maize (with R^2 ranging from 0.7 to 0.81 and mean R^2 being 0.76) and wheat (with R^2 ranging from 0.66 to 0.80 and mean R^2 being 0.74) was higher than that for rice (R^2 from 0.65 to 0.73 and mean R^2 being 0.70) and soybean (R^2 from 0.68 to 0.78 and mean R^2 being 0.72). These results are consistent with the earlier sub-national analysis. The time series analysis indicated a relatively high skill of these models even at the grid level (all $R^2 \geq 0.65$), although the skill varied across different years.

The spatial patterns of CV and mean crop yield of GlobalCropYield5min over the past three decades closely align with the reported yields (Fig. 4 and Fig. S16). The CV and mean of crop yields exhibit distinct patterns, particularly in large countries. Regions with high crop production such as the United States tend to have relatively low CV. Over the study period, the global average CV of maize yield was approximately 0.39 tons/ha/year (Fig. 4a). The CV of maize was higher than the global average in Brazil, Argentina, India, parts of China and Africa, and the Southeast Asia. The global average CV for rice yield was relatively low, except in southern China, northeastern Brazil, central India, and northwestern Africa. The global average CV for wheat yield was 0.38 tons/ha/year. The highest CV values were observed in the Australian wheat belt, northeastern China, Argentina, and southeastern Brazil. Conversely, in major soybean-producing countries such as the Midwestern U.S., India, Southeast and Central Asia, and parts of Latin America, the CV of soybean yield was low. However, it was higher

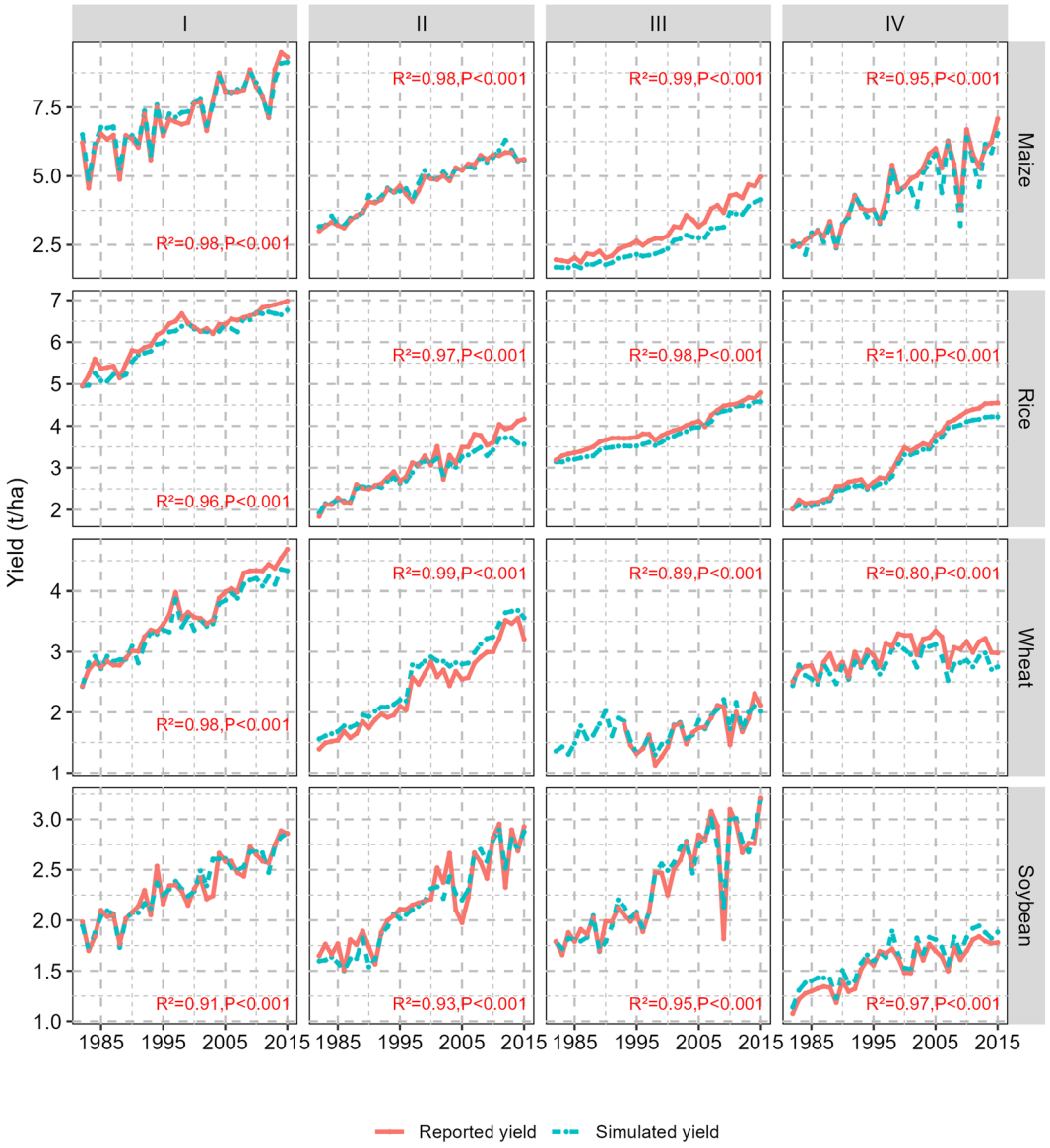

Fig. 5 The yield change at the nationally aggregated level for the top four production countries according the Reported yield (red dashed line) and Simulated yield (green dashed-dotted line) during 1982–2015 for global maize, rice, wheat, and soybean, respectively.

in northeastern China, marginal soybean-producing areas of the United States, Argentina, and Nigeria. Overall, the spatial patterns of crop yield CV and mean exhibit substantial spatial variability at the global level.

We also compared the yield time series for the top four producing countries to assess GlobalCropYield5min's ability to capture yield trends (Fig. 5) and year-to-year variations (Fig. S17). The simulated and reported yields exhibit closely aligned increasing trends, with R^2 values ranging from 0.80 to 1 (Fig. 5). Notably, a clear increasing yield trend is observed for each crop and major country from 1982 to 2015, although the magnitude varies. Additionally, interannual year-to-year yield variations are well captured, with R^2 values spanning from 0.36 to 0.96. Both show statistically significant positive correlations ($P < 0.001$). Overall, these results indicate that GlobalCropYield5min captures both the spatial heterogeneity of yield and its year-to-year variation fairly well.

Comparison with existing global crop yield products. We compared GlobalCropYield5min, SPAM, and GDHY with sub-national yield records from approximately 12,000 administrative units (Figs. 6, 7 and Figs. S18–20) across the top seven crop-producing countries and globally. Their performance was evaluated using the Taylor Diagram (Fig. 6 and Figs. S18–20) for 2000, 2005 and 2010, respectively. Notably, GlobalCropYield5min aligns more closely with the observed data (purple dot) on the x-axis than SPAM and GDHY. For all crop types in the top seven production countries, GlobalCropYield5min generally shows the highest correlation (grey lines) and the lowest RMSE (yellowish dashed lines), with the predicted yields closely matching the observed data (black dashed line, Fig. 6). SPAM exhibits a lower correlation with the observed data, moderate RMSE values, and

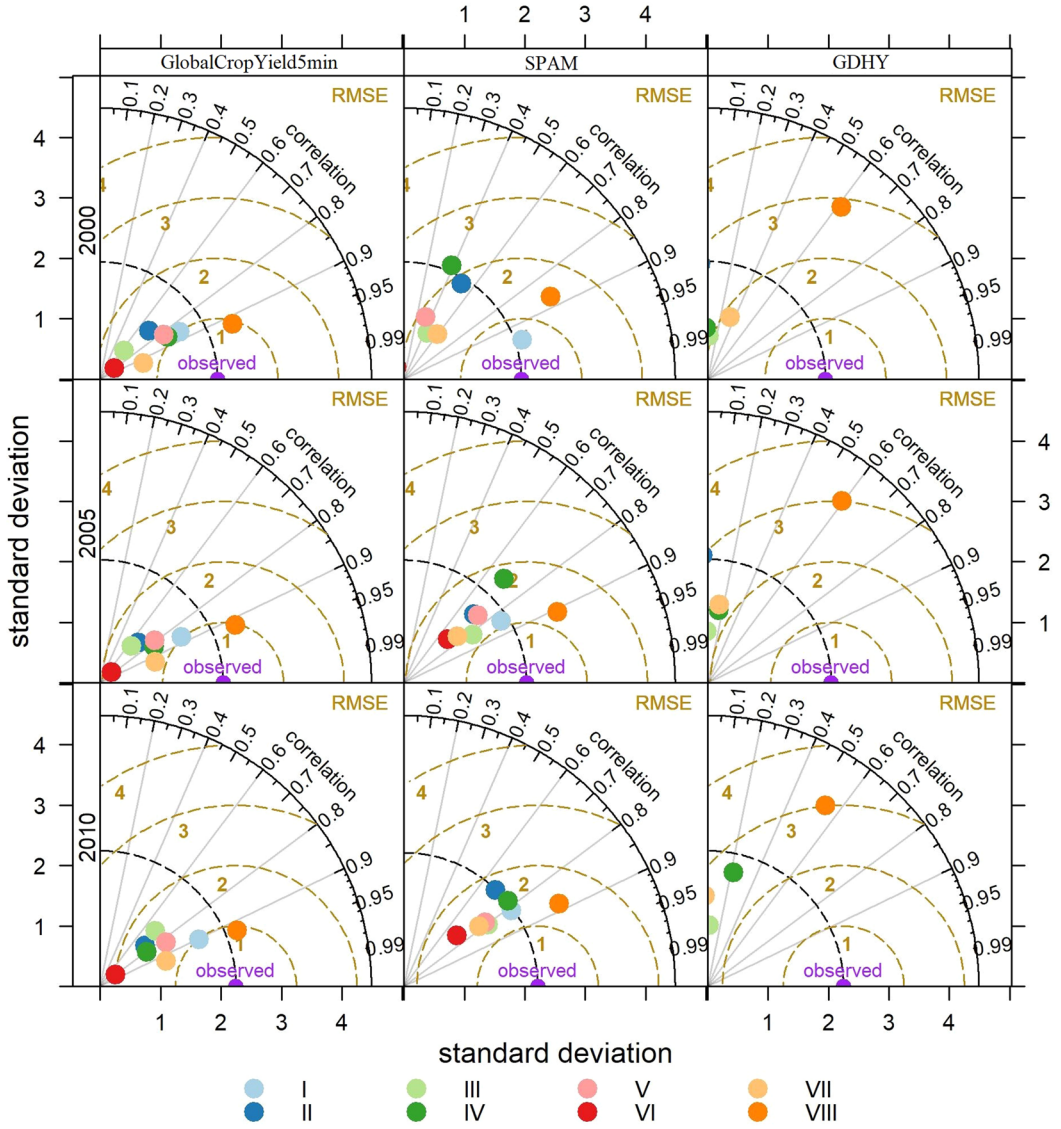

Fig. 6 Taylor diagram comparing the existing three global maize yield data products with the records (i.e., observed in the figure) in 2000, 2005, and 2010, respectively.

greater spatiotemporal variations. In contrast, GDHY performs the worst at the sub-national level, showing the weakest correlation with the observed data (correlation coefficient <0.1 in many countries, which were excluded from the figures), and a high RMSE compared to observed data.

At the global level (Fig. 7), similar to previous results, the GlobalCropYield5min dataset shows the highest performance, with R^2 ranging from 0.73 to 0.86. This is followed by SPAM, with R^2 ranging from 0.44 to 0.82, and GDHY, with R^2 from 0.01 to 0.39. Clearly, the accuracy of GlobalCropYield5min and SPAM is significantly higher than that of GDHY. This is likely because the yield datasets used to generate GlobalCropYield5min and SPAM include data from all sub-national administrative units, whereas GDHY relies solely on FAO statistics and has a spatial resolution of 0.5°. Additionally, GlobalCropYield5min provides continuous coverage from 1982 to 2015, whereas SPAM offers crop yield data only for 2000, 2005, and 2010, though its accuracy continues to improve.

Collecting records yield data at the 5-minute resolution for model verification purposes is indeed challenging. However, to address this limitation, we collected actual maize yields from agro-meteorological stations in China for the years 2000, 2005, and 2010 (Figs. S21–S22), and compared them with three existing products for crop yield prediction. Our analysis revealed a significant correlation between the GlobalCropYield5min and agro-meteorological stations yield data ($P < 0.001$), with an average R^2 of 0.69 and an NRMSE of 16.2%. Importantly, the GlobalCropYield5min product consistently demonstrated the lowest NRMSE. Notably, the GlobalCropYield5min product consistently exhibited the lowest RRMSE, whether analyzed for all three years collectively or separately. In comparison, the SPAM dataset exhibited slightly lower accuracy, with an average R^2 of 0.61 and an NRMSE of 21.15%. The GDHY dataset performed the worst, aligning with our validation

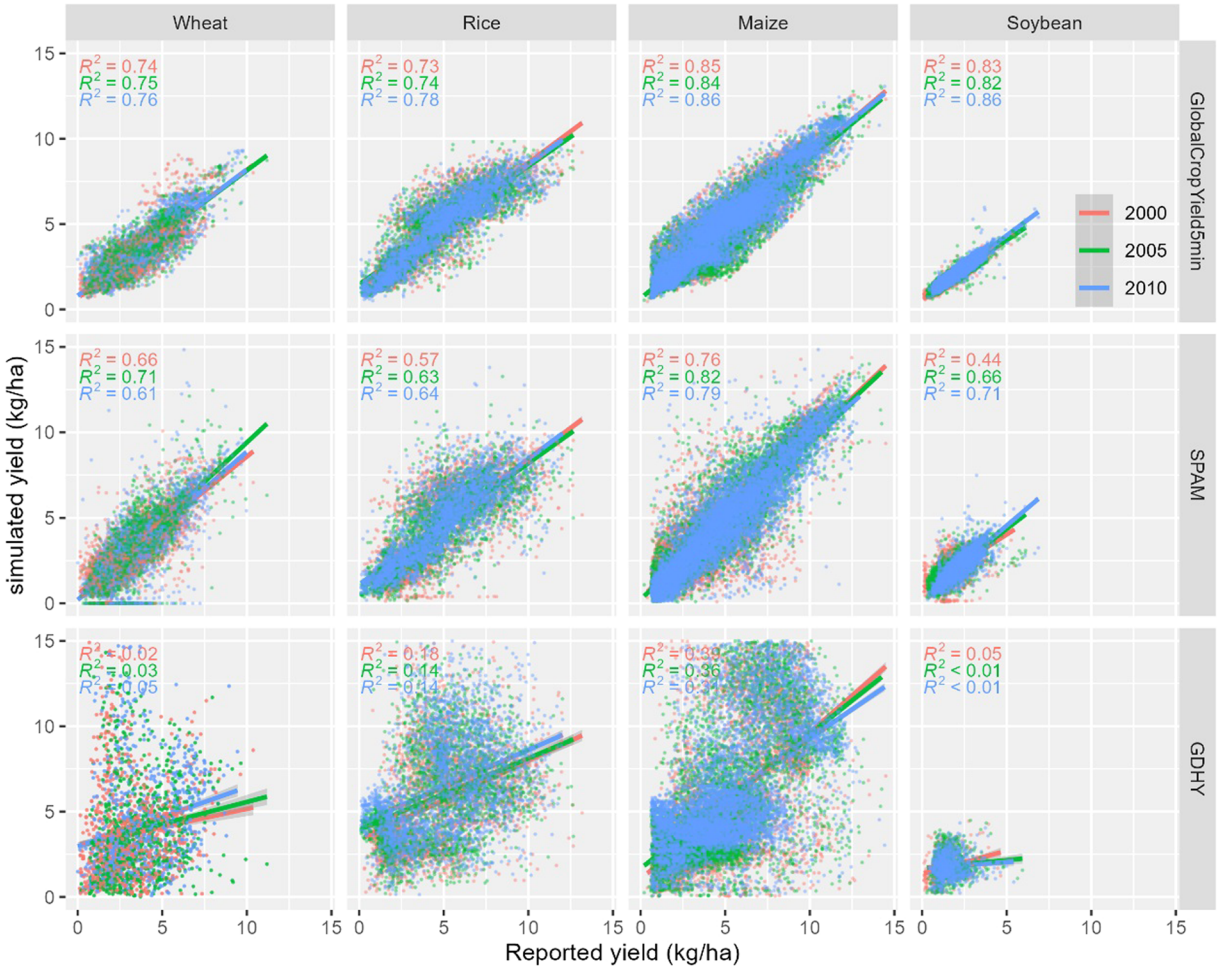

Fig. 7 The scatter plots between the simulated yields from the GlobalCropYield5min, SPAM and GDHY and the reported yield (~12,000 administrative units) in 2000, 2005, and 2010, respectively. The solid and gray dotted lines represent the fitted linear and 1:1 line, respectively.

results at the administrative scale (Figs. 6, 7, S13–14). Consequently, GDHY consistently shows the lowest accuracy in both administrative and field-scale validations. Although SPAM has slightly lower accuracy than GlobalCropYield5min, its temporal discontinuity, with updates only every five years, is a major limitation. In contrast, our dataset offers continuous time-series data, supporting both spatial and temporal analyses. Overall, GlobalCropYield5min outperforms the other two global yield datasets in terms of simulation accuracy, spatial resolution, and temporal coverage.

Spatial distribution of uncertainty in GlobalCropYield5min. Regarding spatial uncertainty, the mean NRMSE for maize, rice, wheat, and soybean was 25.7%, 22.1%, 23.7%, and 22.1%, respectively. Notably, 83.6%, 72.9%, 74.6% and 78.4% of the grids for maize, rice, wheat, and soybean, respectively, had a NRMSE below 30%, indicating low uncertainty in the GlobalCropYield5min dataset. Specifically, uncertainty was low in the Midwest region of the United States, most of Europe, India, Thailand and Pakistan for maize. Uncertainty was low in Myanmar, Thailand, Lao People's Democratic Republic, Cambodia, Viet Nam, Indonesia, and southern Brazil for rice. It was low in eastern United States, eastern Argentina, Europe, and parts of Asia including Afghanistan, Pakistan, and the North China Plain for wheat; and in the Midwest United States and central-western Brazil for soybean. However, uncertainty was high (NRMSE > 40%) in 12.4%, 1.8%, 4.0%, and 5.8% of grids for maize, rice, wheat, and soybean, respectively. These regions with higher uncertainty were primarily located in northeastern Brazil, northern Argentina, northeastern China, and the Philippines for maize; southwestern China, central Pakistan, and northwestern Brazil for rice (Fig. 8); northeastern China and western Australia for wheat; and southern Brazil and northern China for soybean.

Uncertainties and caveats. We applied ML models optimized for crop yield prediction at global scales, demonstrating their notable performance compared to previous studies (See Supplementary Text 2 for details). While the GlobalCropYield5min dataset provides a high-resolution global crop yield coverage, we acknowledge the uncertainties in its production. Firstly, the input datasets from multiple sources[47,65] may introduce biases in the crop yield estimates. For example, variations in planting and harvesting dates over time due to climate,

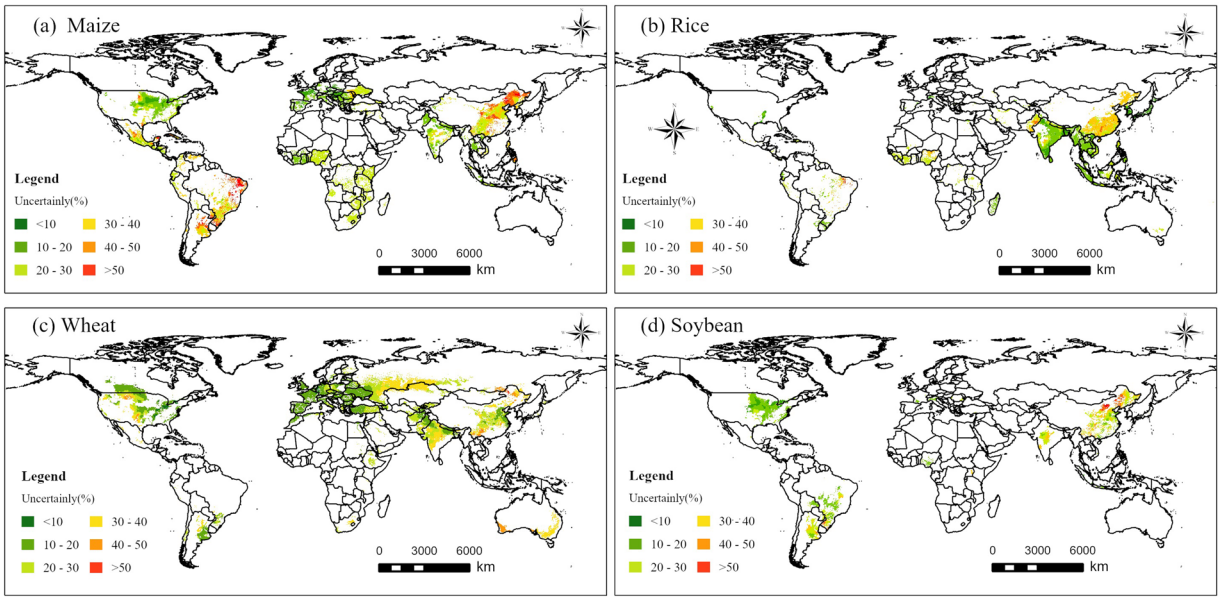

Fig. 8 Spatial pattern of uncertainty (NRMSE in %) in GlobalCropYield5min.

technology, and socioeconomic changes can affect the estimation results. As some previous studies[50,58,66], although we converted the fixed crop calendar to a monthly time step to reduce sensitivity, it is important to consider the dynamic nature of planting and harvesting dates for accurate yield mapping. Additionally, the use of a dynamic crop harvested area map would enhance long-time yield mapping. In this study, we employed crop harvested area-weighted gridded data for the administrative units based on some previous studies[25,67,68], but future studies should address the challenge of obtaining globally continuous coverage, high-resolution and temporal crop harvested area dataset. The uncertainties mentioned above are difficult to reduce unless substantial improvements in data quality[28,65,69–71]. Secondly, due to the difference in time length of collected sub-national statistics across regions, we interpolated data from high-level units to low-level units by using their proportions based on 5 years of average proportions[7,50,72]. By assuming the fluctuations in the proportion and spatial distributions conserved over time, minor errors may arise if these assumptions do not hold true. Moreover, misreporting of the collected sub-national agricultural statistics, spatial interpolating methods, various data availability, and imperfect modeling can contribute to uncertainty in developing the crop yield products. Additionally, the large discrepancies in crop yield gaps between two adjacent countries may lead to apparent spatial edges in the GlobalCropYield5min dataset because the model was built for the country, particularly for wheat and maize in Europe. Finally, it is important note that while the GlobalCropYield5min dataset offers high-resolution and long-time coverage, it may not be suitable for the regions with microclimate features, complex terrains, or heterogeneous land cover, as these factors were not explicitly considered in the simulations.

Code availability

The modeling code used to generate the GlobalCropYield5min dataset is implemented on Python 3.7 but is potentially applicable to other Python versions. The resultant figures are plotted on R. Codes in this paper is freely available online at https://github.com/caojuanLove/GlobalCropYield5min.git.

Received: 29 March 2024; Accepted: 14 February 2025;
Published online: 28 February 2025

References

1. Pereira, P., Bašić, F., Bogunovic, I. & Barcelo, D. Russian-Ukrainian war impacts the total environment. *Science of The Total Environment* **837**, 155865, https://doi.org/10.1016/j.scitotenv.2022.155865 (2022).
2. Kim, K.-H., Doi, Y., Ramankutty, N. & Iizumi, T. A review of global gridded cropping system data products. *Environmental Research Letters* **16**, 093005, https://doi.org/10.1088/1748-9326/ac20f4 (2021).
3. Iizumi, T. *et al.* Prediction of seasonal climate-induced variations in global food production. *Nature Climate Change* **3**, 904–908, https://doi.org/10.1038/nclimate1945 (2013).
4. Gianessi, L. P. The increasing importance of herbicides in worldwide crop production. *Pest Manag Sci* **69**, 1099–1105, https://doi.org/10.1002/ps.3598 (2013).
5. Miflin, B. Crop improvement in the 21st century. *Journal of Experimental Botany* **51**, 1–8, https://doi.org/10.1093/jexbot/51.342.1 (2000).
6. Blomqvist, L., Yates, L. & Brook, B. W. Drivers of increasing global crop production: A decomposition analysis. *Environmental Research Letters* **15**, https://doi.org/10.1088/1748-9326/ab9e9c (2020).
7. Ray, D. K., Gerber, J. S., MacDonald, G. K. & West, P. C. Climate variation explains a third of global crop yield variability. *Nature communications* **6**, 5989–5989, https://doi.org/10.1038/ncomms6989 (2015).
8. Fróna, D., Szenderák, J. & Harangi-Rákos, M. The Challenge of Feeding the World. *Sustainability* **11**, https://doi.org/10.3390/su11205816 (2019).

9. Iizumi, T., Takaya, Y., Kim, W., Nakaegawa, T. & Maeda, S. Global Within-Season Yield Anomaly Prediction for Major Crops Derived Using Seasonal Forecasts of Large-Scale Climate Indices and Regional Temperature and Precipitation. *Weather and Forecasting* **36**, 285–299, https://doi.org/10.1175/waf-d-20-0097.1 (2021).
10. Iizumi, T. *et al*. Impacts of El Niño Southern Oscillation on the global yields of major crops. *Nat Commun* **5**, 3712, https://doi.org/10.1038/ncomms4712 (2014).
11. Yuan, S. *et al*. Southeast Asia must narrow down the yield gap to continue to be a major rice bowl. *Nature Food* **3**, 217–226, https://doi.org/10.1038/s43016-022-00477-z (2022).
12. Fischer, G. *et al*. Global agro-ecological zones (gaez v4)-model documentation. (2021).
13. Yu, Q. *et al*. A cultivated planet in 2010 – Part 2: The global gridded agricultural-production maps. *Earth System Science Data* **12**, 3545–3572, https://doi.org/10.5194/essd-12-3545-2020 (2020).
14. 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**, n/a-n/a, https://doi.org/10.1029/2007gb002947 (2008).
15. Grogan, D., Frolking, S., Wisser, D., Prusevich, A. & Glidden, S. Global gridded crop harvested area, production, yield, and monthly physical area data circa 2015. *Sci Data* **9**, 15–15, https://doi.org/10.1038/s41597-021-01115-2 (2022).
16. Ray, D. K., Ramankutty, N., Mueller, N. D., West, P. C. & Foley, J. A. Recent patterns of crop yield growth and stagnation. *Nature Communications* **3** https://doi.org/10.1038/ncomms2296 (2012).
17. You, L. & Wood, S. An entropy approach to spatial disaggregation of agricultural production. *Agricultural Systems* **90**, 329–347, https://doi.org/10.1016/j.agsy.2006.01.008 (2006).
18. Wood-Sichra, U., Joglekar, A. & You, L. Spatial production allocation model (SPAM) 2005: Technical documentation. Washington, DC: International Food Policy Research Institute (IFPRI) and St. Paul: International Science and Technology Practice and Policy (InSTePP) Center, University of Minnesota; 2016. (HarvestChoice Working Paper.[Google Scholar], 2021).
19. You, L., Wood, S., Wood-Sichra, U. & Wu, W. Generating global crop distribution maps: From census to grid. *Agricultural Systems* **127**, 53–60, https://doi.org/10.1016/j.agsy.2014.01.002 (2014).
20. Vesco, P., Kovacic, M., Mistry, M. & Croicu, M. Climate variability, crop and conflict: Exploring the impacts of spatial concentration in agricultural production. *Journal of Peace Research* **58**, 98–113, https://doi.org/10.1177/0022343320971020 (2021).
21. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* **26** (2013).
22. Brás, T. A., Seixas, J., Carvalhais, N. & Jägermeyr, J. Severity of drought and heatwave crop losses tripled over the last five decades in Europe. *Environmental Research Letters* **16**, https://doi.org/10.1088/1748-9326/abf004 (2021).
23. You, L. & Sun, Z. Mapping global cropping system: Challenges, opportunities, and future perspectives. *Crop and Environment* **1**, 68–73, https://doi.org/10.1016/j.crope.2022.03.006 (2022).
24. Müller, C. *et al*. Global gridded crop model evaluation: benchmarking, skills, deficiencies and implications. *Geoscientific Model Development* **10**, 1403–1422, https://doi.org/10.5194/gmd-10-1403-2017 (2017).
25. Lobell, D. B. & Burke, M. B. On the use of statistical models to predict crop yield responses to climate change. *Agricultural and Forest Meteorology* **150**, 1443–1452, https://doi.org/10.1016/j.agrformet.2010.07.008 (2010).
26. Lobell, D. B. & Burke, M. *Climate change and food security: adapting agriculture to a warmer world*. Vol. 37 (Springer Science & Business Media, 2009).
27. Palanivel, K. & Surianarayanan, C. An approach for prediction of crop yield using machine learning and big data techniques. *International Journal of Computer Engineering and Technology* **10**, 110–118 (2019).
28. Cao, J. *et al*. Wheat yield predictions at a county and field scale with deep learning, machine learning, and google earth engine. *European Journal of Agronomy* **123**, https://doi.org/10.1016/j.eja.2020.126204 (2021).
29. Lamichhane, S., Adhikari, K. & Kumar, L. National soil organic carbon map of agricultural lands in Nepal. *Geoderma Regional* **30** https://doi.org/10.1016/j.geodrs.2022.e00568 (2022).
30. Guevara, M. *et al*. No silver bullet for digital soil mapping: country-specific soil organic carbon estimates across Latin America. *SOIL* **4**, 173–193, https://doi.org/10.5194/soil-4-173-2018 (2018).
31. Cao, J. *et al*. Integrating Multi-Source Data for Rice Yield Prediction across China using Machine Learning and Deep Learning Approaches. *Agricultural and Forest Meteorology* **297**, https://doi.org/10.1016/j.agrformet.2020.108275 (2021).
32. Ji, F., Meng, J., Cheng, Z., Fang, H. & Wang, Y. Crop Yield Estimation at Field Scales by Assimilating Time Series of Sentinel-2 Data Into a Modified CASA-WOFOST Coupled Model. *IEEE Transactions on Geoscience and Remote Sensing* **60**, 1–14, https://doi.org/10.1109/tgrs.2020.3047102 (2022).
33. van Klompenburg, T., Kassahun, A. & Catal, C. Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture* **177**, https://doi.org/10.1016/j.compag.2020.105709 (2020).
34. Vogel, E. *et al*. The effects of climate extremes on global agricultural yields. *Environmental Research Letters* **14.5**, 054010 (2019).
35. Anderson, W. B., Seager, R., Baethgen, W., Cane, M. & You, L. Synchronous crop failures and climate-forced production variability. *Sci Adv* **5**, eaaw1976–eaaw1976, https://doi.org/10.1126/sciadv.aaw1976 (2019).
36. Lobell, D. B., Bänziger, M., Magorokosho, C. & Vivek, B. Nonlinear heat effects on African maize as evidenced by historical yield trials. *Nature Climate Change* **1**, 42–45, https://doi.org/10.1038/nclimate1043 (2011).
37. Sherman, J. D. & Talbert, L. E. Vernalization-induced changes of the DNA methylation pattern in winter wheat. *Genome* **45**, 253–260 (2002).
38. Buitinck, L. *et al*. API design for machine learning software: experiences from the scikit-learn project. In: *ECML PKDD workshop: languages for data mining and machine learning*, pp. 108–122, https://doi.org/10.48550/ARXIV.1309.0238 (2013).
39. Brownlee, J. *Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python*. (Machine Learning Mastery, 2020).
40. Snoek, J., Larochelle, H. & Adams, R. P. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems* **25** (2012).
41. Brochu, E., Cora, V. M. & De Freitas, N. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *Technical Report TR-2009-23. University of British Columbia, Computer Science*, https://doi.org/10.48550/arXiv.1012.2599 (2009).
42. Cai, Y. *et al*. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. *Agricultural and Forest Meteorology* **274**, 144–159, https://doi.org/10.1016/j.agrformet.2019.03.010 (2019).
43. Wang, B. *et al*. Quantifying the impacts of pre-occurred ENSO signals on wheat yield variation using machine learning in Australia. *Agricultural and Forest Meteorology* **291** https://doi.org/10.1016/j.agrformet.2020.108043 (2020).
44. Anderson, W., Seager, R., Baethgen, W., Cane, M. & You, L. Synchronous crop failures and climate-forced production variability. *Science advances* **5**, eaaw1976 (2019).
45. 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**, https://doi.org/10.3390/rs16132342 (2024).
46. Wu, H. *et al*. AsiaRiceYield4km: seasonal rice yield in Asia from 1995 to 2015. *Earth System Science Data* **15**, 791–808, https://doi.org/10.5194/essd-15-791-2023 (2023).
47. Abatzoglou, J. T., Dobrowski, S. Z., Parks, S. A. & Hegewisch, K. C. TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958-2015. *Scientific data* **5**, 170191, https://doi.org/10.1038/sdata.2017.191 (2018).

48. Heino, M. *et al*. Two-thirds of global cropland area impacted by climate oscillations. *Nat Commun* **9**, 1257, https://doi.org/10.1038/s41467-017-02071-5 (2018).
49. Najafi, E., Pal, I. & Khanbilvardi, R. Climate drives variability and joint variability of global crop yields. *Sci Total Environ* **662**, 361–372, https://doi.org/10.1016/j.scitotenv.2019.01.172 (2019).
50. Ray, D. K. *et al*. Climate change has likely already affected global food production. *PLoS One* **14**, e0217148, https://doi.org/10.1371/journal.pone.0217148 (2019).
51. Schillerberg, T. A., Tian, D. & Miao, R. Spatiotemporal patterns of maize and winter wheat yields in the United States: Predictability and impact from climate oscillations. *Agricultural and Forest Meteorology* **275**, 208–222, https://doi.org/10.1016/j.agrformet.2019.05.019 (2019).
52. Tian, D. *et al*. Does decadal climate variation influence wheat and maize production in the southeast USA? *Agricultural and Forest Meteorology* **204**, 1–9, https://doi.org/10.1016/j.agrformet.2015.01.013 (2015).
53. Ren, J., Chen, Z., Zhou, Q. & Tang, H. Regional yield estimation for winter wheat with MODIS-NDVI data in Shandong, China. *International Journal of Applied Earth Observation and Geoinformation* **10**, 403–413 (2008).
54. Roznik, M., Boyd, M. & Porth, L. Improving crop yield estimation by applying higher resolution satellite NDVI imagery and high-resolution cropland masks. *Remote Sensing Applications: Society and Environment* **25**, 100693 (2022).
55. Pinzon, J. & Tucker, C. A Non-Stationary 1981–2012 AVHRR NDVI3g Time Series. *Remote Sensing* **6**, 6929–6960, https://doi.org/10.3390/rs6086929 (2014).
56. Lobell, D. B. & Gourdji, S. M. The influence of climate change on global crop productivity. *Plant Physiol* **160**, 1686–1697, https://doi.org/10.1104/pp.112.208298 (2012).
57. Aggarwal, P., Vyas, S., Thornton, P., Campbell, B. M. & Kropff, M. Importance of considering technology growth in impact assessments of climate change on agriculture. *Global Food Security* **23**, 41–48, https://doi.org/10.1016/j.gfs.2019.04.002 (2019).
58. Lobell, D. B., Schlenker, W. & Costa-Roberts, J. Climate trends and global crop production since 1980. *Science* **333**, 616–620 (2011).
59. Burchfield, E. K., Nelson, K. S. & Spangler, K. The impact of agricultural landscape diversification on U.S. crop production. *Agriculture, Ecosystems & Environment* **285**, https://doi.org/10.1016/j.agee.2019.106615 (2019).
60. Portmann, F. T. *Global dataset of monthly growing areas of 26 irrigated crops: version 1.0*. (Univ.-Bibliothek Frankfurt am Main, 2008).
61. Lu, C. *et al*. In-season maize yield prediction in Northeast China: The phase-dependent benefits of assimilating climate forecast and satellite observations. *Agricultural and Forest Meteorology* **358**, https://doi.org/10.1016/j.agrformet.2024.110242 (2024).
62. Shi, H. & Xingguo, M. Interpreting spatial heterogeneity of crop yield with a process model and remote sensing. *Ecological Modelling* **222**, 2530–2541, https://doi.org/10.1016/j.ecolmodel.2010.11.011 (2011).
63. Stadler, A. *et al*. Quantifying the effects of soil variability on crop growth using apparent soil electrical conductivity measurements. *European Journal of Agronomy* **64**, 8–20, https://doi.org/10.1016/j.eja.2014.12.004 (2015).
64. Juan, C. *et al*. GlobalCropYield5min: A global gridded annual major crops yield dataset at 5-minute resolution during 1982-2015. *Mendeley Data, V3*, https://doi.org/10.17632/hg8wzgx4yp.3 (2024).
65. Li, X. & Xiao, J. A Global, 0.05-Degree Product of Solar-Induced Chlorophyll Fluorescence Derived from OCO-2, MODIS, and Reanalysis Data. *Remote Sensing* **11** https://doi.org/10.3390/rs11050517 (2019).
66. Franke, J. A. *et al*. The GGCMI Phase 2 experiment: global gridded crop model simulations under uniform changes in CO₂, temperature, water, and nitrogen levels (protocol version 1.0). *Geoscientific Model Development* **13**, 2315–2336, https://doi.org/10.5194/gmd-13-2315-2020 (2020).
67. Zhang, Z., Song, X., Tao, F., Zhang, S. & Shi, W. Climate trends and crop production in China at county scale, 1980 to 2008. *Theoretical and Applied Climatology* **123**, 291–302, https://doi.org/10.1007/s00704-014-1343-4 (2015).
68. Zhang, T., Zhu, J. & Wassmann, R. Responses of rice yields to recent climate change in China: An empirical assessment based on long-term observations at different spatial scales (1981–2005). *Agricultural and Forest Meteorology* **150**, 1128–1137, https://doi.org/10.1016/j.agrformet.2010.04.013 (2010).
69. Fan, J. *et al*. Light Gradient Boosting Machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. *Agricultural Water Management* **225**, 105758, https://doi.org/10.1016/j.agwat.2019.105758 (2019).
70. Jeong, J. H. *et al*. Random Forests for Global and Regional Crop Yield Predictions. *PLoS One* **11**, e0156571, https://doi.org/10.1371/journal.pone.0156571 (2016).
71. Luo, Y. *et al*. Accurately mapping global wheat production system using deep learning algorithms. *International Journal of Applied Earth Observation and Geoinformation* **110**, https://doi.org/10.1016/j.jag.2022.102823 (2022).
72. Ray, D. K., Mueller, N. D., West, P. C. & Foley, J. A. Yield Trends Are Insufficient to Double Global Crop Production by 2050. *Plos One* **8**, e66428, https://doi.org/10.1371/journal.pone.0066428 (2013).

Acknowledgements

This study was supported by the China Postdoctoral Science Foundation (Grant No. 2023M743450 and GZC20232614) and National Natural Science Foundation of China (Project Nos. 42061144003, 41977405).

Author contributions

Zhao Zhang: Conceptualization, Methodology, Writing- Reviewing and Editing. Juan Cao: Data curation, Writing-Original draft preparation. Xiangyun Luo: Visualization, Writing- Reviewing and Investigation. Jun Xie: Reviewing. Jichong Han: Reviewing. Yuchuan Luo: Reviewing. Jialu Xu: Reviewing. Fulu Tao: Conceptualization, Methodology, Writing- Reviewing and Editing.

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.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-04650-4.

Correspondence and requests for materials should be addressed to Z.Z. or F.T.