https://doi.org/10.5194/essd-2024-233
Preprint. Discussion started: 18 July 2024

Open Access Earth System Science Data Discussions

Tracking spatiotemporal dynamics of crop-specific areas through machine learning and statistics disaggregating

Xiyu Li[1], Le Yu[1,2,3*], Zhenrong Du[4], Xiaoxuan Liu[5]
[1] Ministry of Education Key Laboratory for Earth System Modeling, Department of Earth System Science, Tsinghua University, Beijing, 100084, China
[2] Ministry of Education Ecological Field Station for East Asian Migratory Birds, Beijing 100084, China
[3] Tsinghua University (Department of Earth System Science)- Xi'an Institute of Surveying and Mapping Joint Research Center for Next-Generation Smart Mapping, Beijing 100084, China
[4] School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China
[5] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, 100190, China

Correspondence to: Le Yu (leyu@tsinghua.edu.cn)

Abstract. Mapping spatiotemporal dynamics of crop-specific areas is of great significance in addressing challenges faced by agricultural systems. But comparable multi-phase crop maps in year series have not yet been developed in most regions of the global. In this study, we developed a framework for updating annual crop-specific area maps at 10km resolution based on crop statistics disaggregating, multi-source data integrating and machine learning, taking factors related with crop distribution in different regions and complex agricultural systems into accounts. Experiments were conducted in three study areas (Africa, China, and USA) respectively corresponding to three conditions of the information coverage of crop distribution (low, median, and high). In our framework, we collected related spatial indicator used in previous studies and trained random forest regression models to predict spatiotemporal dynamics of crop-specific areas based on them. Annual crop statistics were further disaggregated based on probabilistic layer and harmonized based on multiple constraints. Our framework is a good attempt to integrate two strategies (top-down and bottom-up), creating more possibility for crop mapping to integrate statistic with remote sensing. Finally, our results include maps of crop-specific areas covering 42 types from 1961-2022 in Africa, maps of crop-specific areas covering 14 types from 1980-2022 in China and maps of crop-specific areas covering 15 types from 2008-2022 in USA. Results show that our products has a relatively good consistency with independent reference map or statistics. Our products provide approximate estimates for spatiotemporal dynamics of crop-specific areas in multiple regions over several decades, which could be used as data basis for food security and environmental impact assessments.

1 Introduction

As the largest land use in the world, agriculture is at the heart of many global problems, for it has greatly changed the biosphere in terms of land use change, freshwater use, nitrogen cycle and biodiversity (You and Sun, 2022; Siebert and Döll, 2010; Monfreda et al., 2008; Ellis et al., 2013). On the one hand, agricultural system is faced by higher food production demands coming with the continuous growth of population and higher proportion of high-protein foods in people's diet (Liu et al., 2022; Davis et al., 2016); On the other hand, agriculture is a driving factor behind many global environmental issues, such as

https://doi.org/10.5194/essd-2024-233
Preprint. Discussion started: 18 July 2024

deforestation and resulting carbon emissions (Song et al., 2021), climate change (Zhang et al., 2020), biodiversity loss (Hoang et al., 2023), etc. At the same time, agricultural production is also adversely affected by climate change (Vogel et al., 2019), epidemic outbreak (Laborde et al., 2020), war and conflict (Lin et al., 2023), and urbanization encroachment (Bren D'amour et al., 2017; Huang et al., 2020). Obtaining sufficient and detailed data basis is the prerequisite to deal with challenges and carry out decision analysis. Many studies have successfully mapped the distribution of global agricultural land in multiple time periods (Potapov et al., 2022; Ramankutty et al., 2008; Lu et al., 2020), but the diversity of global agricultural planting systems has not been fully detailed (You et al., 2014; Yu et al., 2020). Mapping spatiotemporal dynamics of crop-specific areas is of great significance in addressing the above challenges (You and Sun, 2022). Human needs for different crops vary according to their use, nutritional value and cultural factors. Food security is not only about providing sufficient calories, but also about meeting people's nutritional needs and dietary preferences (Nelson et al., 2018). Meanwhile, crop type mapping provides more information about the ecological and environmental impacts of agriculture and the threats it faces (Hoang et al., 2023; Song et al., 2021).

Global crop distribution information is uneven in time and space (Table S1, Fig S1). Despite operational monitoring and remote sensing mapping in some countries in recent years (D'andrimont et al., 2021; Boryan et al., 2011), there is still a lack of spatially explicit data sets for crop planting history. In most parts of the world, crop distribution information tends to come only from national or subnational scale statistics (Fig S1), which greatly limits the use of data (Alami Machichi et al., 2023). Therefore, many studies have developed crop maps by spatially allocating crop statistics based on cropland maps over the past two decades. These crop distribution maps often have large-scale spatial coverage (usually global coverage), but coarse spatial resolution (such as 5 arc minutes, about 10km). As early as 2004, Leff et al. (2004) allocated the statistics of 18 crops to the global 5-arc grids in 1990s by a simplified proportional disaggregating approach. Specifically, this method obtained a crop distribution map by calculating the proportion of the crop-specific harvested area in the total at the administrative unit level and multiplying it by the cropland proportion within the grid. By adopting this approach, Monfreda et al. (2008) developed the global harvest area and yield distribution map of 175 crops in 2000 (M3) with more detailed statistical data. On the basis of M3, Portmann et al. (2010) further combined additional data such as irrigated area, crop calendar to produce monthly irrigated and rainfed crop areas around the year 2000 (MIRCA).

However, in the first few global crop distribution maps, the harvest area of crops in administrative units is equally distributed proportionally, ignoring the diversity and difference of crop distribution in different regions and complex agricultural systems. By considering climatic and edaphic suitability of crops, GAEZ products (Global Agro-ecological Zones) provides new estimates of potential crop-specific areas of 23 crops (Fischer et al., 2021). Moreover, the Spatial Production Allocation Model (SPAM) is a new spatial allocation model of crop statistics by disaggregation at farming systems and optimization using a cross-entropy algorithm, also taking many related factors into account such as crop suitability, market accessibility and crop revenue, which greatly increases the complexity of model input and allocation process (You and Wood, 2006; You et al., 2009; You et al., 2014; Yu et al., 2020). However, the differences in crop-specific harvested area estimated by each model are significant, mostly resulting from differences in the input datasets and downscaling methodologies (Anderson et al., 2015).

https://doi.org/10.5194/essd-2024-233
Preprint. Discussion started: 18 July 2024

As multiple sets of crop-type mapping products based on remote sensing have been produced and widely used, crop distribution information in these regions has a more accurate and timely spatial representation. Some studies have adopted the scoring rule-based approach to integrate multi-source crop maps. For example, Backer-Reshef et al. (2023) built a scoring system based on five indicators and updated more accurate crop distribution information of 66 countries in the world based on SPAM 2010 map. Besides, Tang et al. (2024) updated the latest distribution information of crops into around 2020 on the basis of the M3 map in 2000, which integrated 27 sets of existing crop mapping products and new crop statistics by constructing scoring rules. Among the above global crop distribution mapping products, GAEZ and SPAM has multi-period map products (Yu et al., 2020; Grogan et al., 2022; Fischer et al., 2021). But the comparison of multi-period maps across time stages proved to be inappropriate. Although the research team ensured the spatial accuracy of each map as much as possible, system errors of various data sources were inevitably included in the integration process (Yu et al., 2020). In order to better understand temporal trends of crop distribution, many studies have focused on the development of time-series comparable mapping products. For example, Ray et al. (2012) plotted the planting area and yield change trend of maize, rice, wheat and soybean during 1961-2008 based on the of national or subnational crop statistics. However, collecting statistics at the subnational level takes a lot of manpower and time, especially when covering multiple crop types. Jackson et al. (2019) developed a new allocation algorithm called the Probabilistic Cropland Allocation Model (PCAM), which allocated crop statistic at national level based on suitability probability clusters and multiple Monte Carlo. Mapping result covers 17 crops from 1961 to 2014, but it is regarded more as only likely changes in the spatial distribution.

Inspired by previous studies, our study aims to develop a framework for updating annual crop-specific area maps based on crop statistics disaggregating, multi-source data integrating and machine learning, taking factors related to crop distribution in different regions and complex agricultural systems into account. We selected three regions as study areas, respectively Africa, China, and USA. They correspond to three conditions of the information coverage of crop distribution (low, median, and high) (Table S1, Fig S1). Updating annual crop-specific area maps is important especially for regions like Africa and China, where operational annual crop mapping and monitoring is not available and is largely limited by time-consuming sample collecting process and laborious field work. In our framework, we collected related spatial indicator used in previous studies and trained machine learning models to predict spatiotemporal dynamics of crop-specific areas based on them. Annual crop statistics were further disaggregated based on the probabilistic layer and harmonized based on multiple constraints. Finally, we produced maps of crop-specific areas covering 42 types from 1961-2022 in Africa, maps of crop-specific areas covering 14 types from 1980-2022 in China and maps of crop-specific areas covering 15 types from 2008-2022 in USA (for validation).

2 Data preparation

In this study, three regions were selected as study areas, respectively Africa, China, and USA. They correspond to three conditions of the information coverage of crop distribution (low, median, and high). In Africa, there are only crop statistics provided in mostly areas. While in China, except for statistics, there have been many studies producing single-type crop maps

https://doi.org/10.5194/essd-2024-233
Preprint. Discussion started: 18 July 2024

in recent years but not integrated multi-type maps. But in USA, Crop Data Layer product (CDL) provides annual-update ground truth layer covering the whole crop categories. Therefore, we respectively collected crop statistics and base maps for these regions. In addition, we also prepared cropland extent datasets and spatial indicators related with crop distribution for further spatiotemporal modeling.

2.1 Crop statistics

Firstly, we collected different sources of crop statistics in different study regions (Table 1), which provide the total amount of crop-specific areas. FAOSTAT, the biggest dataset in the field of food and agriculture, provides free access to food and agriculture data for over 245 countries and territories from 1961 to the most recent year available (https://www.fao.org/faostat/en/#data). Here, Crops and livestock products datasets in FAOSTAT's Production domain were collected, which contained information on harvest area of more than 143 crop types. According to the categories of SPAM products, it is integrated into 42 crop categories through FAO code (Table S2-1).

Table 1 Crop statistics data sources used in this study

Region	Source	Spatial Scale	Time period	Crop types
Africa	FAOSTAT	national (ADM0)	1961-2022	42
USA	USDA NASS	sub-national (ADM1)	2008-2022	15
China	CHN STAT	sub-national (ADM1)	1980-2022	14

In countries with large land area such as USA and China, national statistics are insufficient to capture the details of spatial changes. Therefore, we collected statistics at a sub-national level from national official reports in these two counties. As for USA, National Agricultural Statistics Service (NASS) from United States Department of Agriculture (USDA) provides access to census or survey of crop harvest area at the state level (https://quickstats.nass.usda.gov/). In China, statistics of crop-specific areas were acquired from China's economic and social big data research platform operated by China National Knowledge Infrastructure (https://data.cnki.net/valueSearch/index?ky=). Due to the differences in crop categories from multiple sources, we matched 15 and 14 crop types respectively in USA and China according to the categories of SPAM products (Table S2-2, Table S2-3).

In addition, it is also a vital step to relate statistics to georeferenced locations. The Global Administrative Unit Layers (GAUL), one of the most standard spatial datasets of administrative units, compiles and disseminates the best available information on different levels of administrative units for countries in the world. Here, we used GAUL at national levels (ADM0) and sub-national levels (ADM1) to form geo-referenced crop harvest datasets respectively in Africa and USA/China (Table 1).

https://doi.org/10.5194/essd-2024-233
Preprint. Discussion started: 18 July 2024

2.2 Base maps

Compared to statistics which represent total amount, base maps provide a detailed portrayal of crop-specific areas, which will be essential for spatiotemporal dynamic modeling. There are multiple maps of crop-specific areas with global coverage and detailed categories. And among which SPAM2010 product were chosen in this study for its sophisticated approaches and latest updates, providing key information about the distribution of crop-specific areas in regions in which remote sensing techniques have not been used widely in crop mapping such as Africa (Table 2).

Table 2 Crop-type base maps used in this study

Region	Source	Spatial Resolution	Time period	Crop types
Africa	SPAM2010	5 arcmin	2010	42
USA	CDL	30m	2008-2022	15
China	Multi sources	10m/30m/5 arcmin	Multi	14

However, more accurate and timely crop distribution information in countries like USA can be acquired by higher spatial resolution maps produced by remote sensing and field survey. Therefore, we adopted Crop Data Layer (CDL) as the base map in USA (Table 2). The correspondence between CDL's crop categories and SPAM is presented in the Supplement (Table S2-2). As for China, there is still not an integrated multi-type crop map with medium-high spatial resolutions (10-30m). Therefore, we collected multi-source base maps of 8 crop types to produce more accurate and timely references (Table S3-1), as for the rest crop types and the regions that are not covered, we still used SPAM2010 as the base map. To ensure uniform resolution, we calculate the proportion of the crop area within a 10 km (~5 arcmin) grid for the base map with higher resolution (10m-30m).

2.3 Cropland extent

Cropland extent determines where statistics of crop-specific areas can be allocated. In order to provide an annual-updated basis to track spatiotemporal dynamics, we integrated two datasets to produce annual cropland extent from history to the most recent year. It is an efficient and accurate way to identify cropland from classified land cover products. FROM-GLC Plus provides a framework for near real-time land cover mapping at multi-temporal (annual to daily) and multi-resolution (30 m to sub-meter) levels (Yu et al., 2022). Here, we used FROM-GLC Plus Global Land Cover Products (1982-2021, 1km subpixel) to extract cropland extent from 1982 to 2021. In periods where there were rarely remotely sensed images (before the 1980s), GCD (Global Cropland Dataset) which is produced by spatially allocating cropland statistics was used as extent (Cao et al., 2021) (Table S3-5). Given that crop production may take place over several seasons within a year, we also multiplied the cropland extent by the cropping intensity to get the annual maximum harvested area. Annual cropping intensity datasets cover periods from 2001 to 2019 (Liu et al., 2021). For the remaining years, data from the nearest year is used as a substitute (Table S3-6).

https://doi.org/10.5194/essd-2024-233
Preprint. Discussion started: 18 July 2024

To ensure uniform resolution, we resampled the cropland extent and maximum harvested area of all year periods to 10km (~5 arcmins).

2.4 Spatial indicators

To better model spatiotemporal dynamic of crop-specific areas, a total of 26 related spatial indicators were collected from 7 aspects, including climate, agro-system, suitability, potential yield, soil, terrain, and location (Table S3-10). These indicators are proved to be related to the distribution of crop-specific areas in previous studies. To ensure uniform resolution, we resampled the spatial resolution of all spatial indicators to 10km (~5 arcmin).

Climate determines the areas in which crops are suitable for cultivation, and climate change affects the potential yields and revenue, thus affecting farmer decisions. It is reported that climatic variables explain considerable portions of the variance in crop planted area (22–30%), harvestable fraction (15–28%) and yield (32–50%) (Wei et al., 2023). In this study, five climate variables were calculated from ERA5-Land including annual mean temperature (temp), total annual precipitation (prec), downwards surface solar radiation (radi_down), evaporation from vegetation transpiration (evap_veg) and growing degree day (gdd) (Table S3-10). We present a detailed description of these variables in the Supplement (Sect. S3-2).

In addition, terrain and soil properties are also essential variables of crop suitability. However, these variables are only available at single time period unlike climate variables that updated annually. More specifically, slope and elevation in the year 2010 were calculated from GMTED2010 as the terrain variables while aggregation of soil water content, PH, texture class, organic carbon content, sand content and clay content at different depths were calculated from OpenlandMap datasets as soil variables. We present a detailed description of terrain and soil variables in the Supplement (respectively Sect. S3-3 and Sect. S3-4).

We also selected the suitability assessment result of each crop as input for this prediction model. GAEZv4.0 produces a gridded suitability assessment for 48 major crops in two input levels (i.e., high, low), and two water supply regimes (i.e., irrigated or rainfed) at 5 arcmin resolution. The correspondence between GAEZ v4.0's crop categories and SPAM is presented in the Supplement (Table S2-4). Most of the SPAM2010 crops are included in GAEZ's crop categories, those not included are assigned values from similar crops. Potential yield has a greater impact on farmer decisions when multiple crops are suitable to cultivate, this variable could also be accessed by GAEZv4.0 product in two input levels and two water supply regimes. The suitability index and potential yield in three regimes (irrigated, rainfed and high input, rainfed and low input) were selected as input indicators. We present a detailed description of suitability and potential yield variables' processing progress in the Supplement (respectively Sect. S3-5 and Sect. S3-6).

Agro-system variables are directly related to crop distribution. Firstly, cropland maps determine where and to which extent crop could be cultivated (also in Section 2.3). Furthermore, crop cultivation suitability and potential yield vary greatly in irrigation or rainfed regimes. Therefore, we integrated two datasets (HID and SPAM) to extract irrigation area proportion in a long time series. HID (historical irrigation data set) provides estimates of the temporal development of the area equipped for irrigation since 1900 at 5 arcmin resolution (Siebert et al., 2015), while SPAM contains irrigation area proportion information in year 2000, 2005 and 2010. What's more, it is said that the proportions of crops being distributed on farms of different sizes

https://doi.org/10.5194/essd-2024-233
Preprint. Discussion started: 18 July 2024

vary greatly (Su et al., 2022). So we select a global field size map as an indicator which classifies farm size into 5 classes (Lesiv et al., 2019). Last but not least, rural population is closely related to agricultural production, which can be considered as a measure of agricultural labor and also market accessibility. Estimates of rural population from HYDE (History Database of the Global Environment) were selected as the last indicator of the agro-system group (Klein Goldewijk et al., 2017). We present a detailed description of agro-system variables' processing progress in the Supplement (respectively Sect. S3-7, S3-8, S3-9, S3-10).

3 Method

There are mainly two steps in annually updating spatiotemporal dynamics of crop-specific areas. The first step is to generate probabilistic spatiotemporal dynamics layers through machine learning based on spatial indicators and base maps, while the latter step is to allocate crop statistics based on the probabilistic layer and harmonize the result based on multiple constraints (Fig 1).

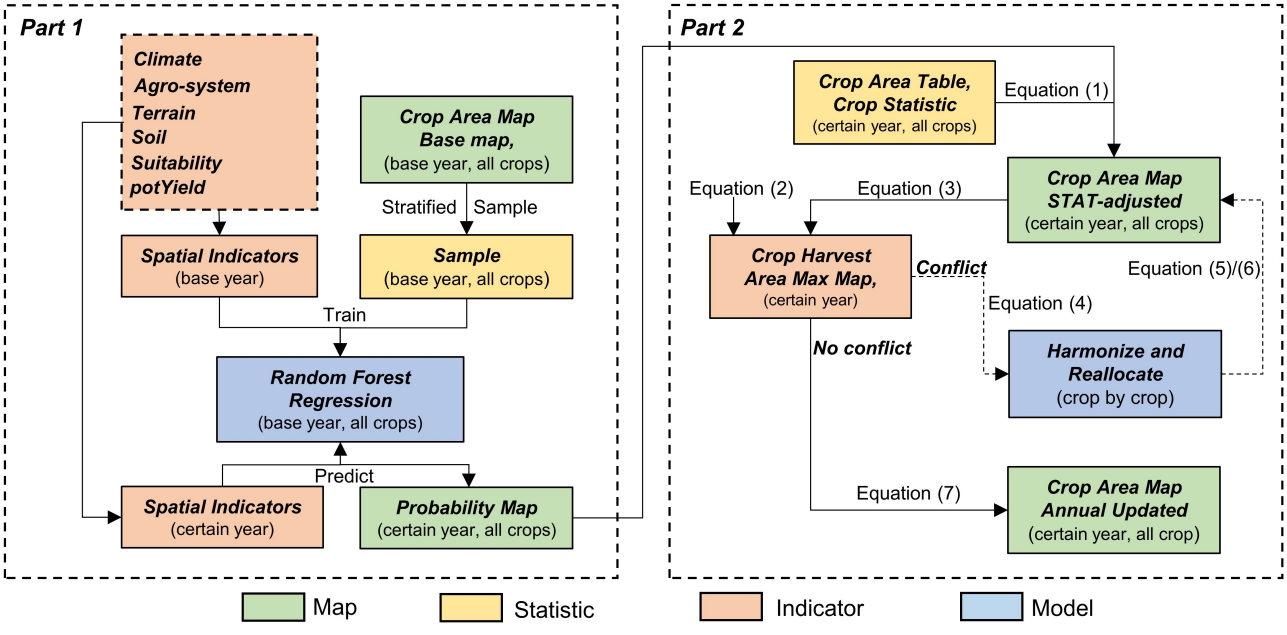


Figure 1. Workflow of this study.

3.1 Generating probabilistic spatiotemporal dynamics

In this part, we trained machine learning models for each crop in each region based on spatial indicators and base maps. To begin with, we collected samples on the base map using a stratified random strategy. The sample set of each crop is set to collect a maximum of 10000 points at the resolution of 10 km. We only collected samples at the pixel where crop-specific area

https://doi.org/10.5194/essd-2024-233
Preprint. Discussion started: 18 July 2024

exists. Combined with spatial indicators prepared in Section 2.4, we chose random forest (RF) regression model to learn the rules by which crop distribution is affected by 26 spatial indicators from sample sets. RF has been proven to be effective and efficient in dealing with multi-dimensional features (Breiman, 2001) and RF regression has been widely used in tasks of predicting the distribution of geospatial targets such as population, cropland, etc. (Cao et al., 2021; Sorichetta et al., 2015). More specifically, we use the random forest algorithm in Google Earth Engine to build the regression model. The number of decision trees is set to 100, and the number of features required for each node for splitting is set to the square root of the number of input features set by default.

3.2 Statistic allocation and harmonization

In this part, crop statistics were allocated based on probabilistic layers predicted by the RF regression model (Section 3.1) and harmonized according to multiple constraints. More specifically, we first calculate the sum of probabilistic values at the administrative unit level (excluding areas with no cropland or suitability index of 0). The sum value was used to compare with crop statistics at the same administrative unit and the rate was used to adjust probabilistic layers at this administrative unit (Equation (1)). The descriptions of variables in the equations are listed in Table 3 (the same below).

$$AdjProb_{ijy} = \frac{Stat_{jky}}{\sum_{i \in k} Prob_{ijy} \times PixelArea_i} \times Prob_{ijy} \quad (1)$$

Then, we defined the maximum crop harvest area in each grid as the limit that no more crop statistics could be allocated (Equation (2)). More specifically, the maximum crop harvest area is calculated by multiplying cropland and crop intensity (more data sources descriptions in Section 2.3).

$$MaxHarvArea_{iy} = Cropland_{iy} \times CropIntensity_{iy} \times PixelArea_i \quad (2)$$

Further, we used HarvAreaLeft_{iy} to represent the conflict between the maximum crop harvest area and statistics allocation (Equation (3)). In order to compare in the same unit, we transformed the adjusted proportion of crop-specific areas from proportion (0-1) into area (ha) by multiplying $PixelArea_i$.

$$\text{HarvAreaLeft}_{iy} = MaxHarvArea_{iy} - \sum_j AdjProb_{ijy} \times PixelArea_i \quad (3)$$

Statistic allocation was performed in a crop-by-crop order which is specified in each administrative unit. We classified the crop type list into several types: annual or perennial, specific name (e.g. wheat) or general name (e.g. other cereals). The priority of perennial crops is higher than annual ones, and then crops with specific name will be processed first compared with that of general names. We further ranked the crop type list in each sub-group according to the crop area statistics in the administrative unit in descending order. It's worth noting that when maximum crop harvest area is achieved in a certain grid i, the statistic of the crop type j being processed would not be allocated. The crop type j and crop types behind j in the ranked crop list would be saved in a grid-specific list $LeftCrop_{iy}$. We further summed the statistics still not be allocated yet (StatLeft_{jky}) in a grid where maximum of crop harvest area is achieved based on $LeftCrop_{iy}$ (Equation (4)).

$$\text{StatLeft}_{jky} = \sum_{i \in k} AdjProb_{ijy} \times PixelArea_i \ \ if\ \text{HarvAreaLeft}_{iy} < 0\ and\ j \in LeftCrop_{iy} \quad (4)$$

https://doi.org/10.5194/essd-2024-233
Preprint. Discussion started: 18 July 2024

Open Access Earth System Science Data Discussions

We adopted two strategies to deal with statistics still not being allocated yet ($\mathrm{StatLeft}_{jky}$). The first one is to allocate crop statistics to the grids where the maximum crop harvest area is not achieved according to the probability value predicted by the RF model (Equation (5)). If there are statistics still not allocated yet, the latter process will allocate crop statistics to the grids where the maximum crop harvest area is not achieved according to the rest space of crop harvest area (Equation (6)).

$$AdjProb_{ijy} += \frac{StatLeft_{jky}}{\sum_{i \in k,\ \mathrm{HarvAreaLeft}_{iy}>0} AdjProb_{ijy} \times PixelArea_i} \times AdjProb_{ijy}\ if\ \ \mathrm{HarvAreaLeft}_{iy} > 0 \quad (5)$$

$$AdjProb_{ijy} += \frac{StatLeft_{jky}}{\sum_{i \in k,\ \mathrm{HarvAreaLeft}_{iy}>0} \mathrm{HarvAreaLeft}_{iy}} \times \frac{\mathrm{HarvAreaLeft}_{iy}}{PixelArea_i}\ if\ \ \mathrm{HarvAreaLeft}_{iy} > 0 \quad (6)$$

At last, we got the adjusted proportions of crop-specific areas (after Equation (5) or (6)), and the final result $\mathrm{CropArea}_{ijy}$ represents the harvest area (ha) of crop type j in grid i in year y (Equation (7)).

$$\mathrm{CropArea}_{ijy} = AdjProb_{ijy} \times PixelArea_i \quad (7)$$

Table 3 Main input variables in the part of statistic allocation and harmonization

ID	Variables	Definition
1	i	10km grid cell
2	j	Crop type
3	k	Administrative unit
4	y	Year
5	$Prob_{ijy}$	Probability of crop type j in grid i in year y predict by RF.
6	$PixelArea_i$	Pixel area in grid i, factor that transforms proportion (0-1) into area (ha).
7	$Stat_{jky}$	Area Statistics (ha) of crop type j in administrative unit k in year y.
8	$AdjProb_{ijy}$	Adjusted probability of crop type j in grid i in year y in the process.
9	$Cropland_{iy}$	Cropland area proportion in grid i in year y.
10	$CropIntensity_{iy}$	Harvest frequency in grid i in year y.
11	$MaxHarvArea_{iy}$	Maximum of harvest area in grid i in year y.
12	$\mathrm{HarvAreaLeft}_{iy}$	Left harvest area space where crop statistics can be allocated in grid i in year y.
13	$LeftCrop_{iy}$	List of crop types whose statistics have not been allocated when the maximum harvest area is achieved in grid i in year y.
14	$\mathrm{StatLeft}_{jky}$	Statistics of crop type j which has not been allocated in administrative unit k in year y.
15	$\mathrm{CropArea}_{ijy}$	Harvest area of crop type j in grid i in year y

https://doi.org/10.5194/essd-2024-233
Preprint. Discussion started: 18 July 2024

4 Result

4.1 Allocated harvest area maps

Through the method proposed in Section 3, we allocated crop statistics into crop-specific area maps. To verify the effectiveness of this allocation process, we compared the crop statistics and the aggregation result of crop-specific area maps at the same administrative units. Results show that our method allocated almost all crop statistics into crop-specific areas at the corresponding administrative unit with R^2 over 0.8 and a slope close to 1 (Fig 2). The reason why crop statistic and mapping results are not completely consistent at the administrative unit level is that the pixels at the edge of the administrative boundary have not been uniformly processed. In addition, when the statistical data is too large, some of the data cannot be reasonably allocated to cropland and suitable land for specific crops, and these pixel values are replaced by the base map.

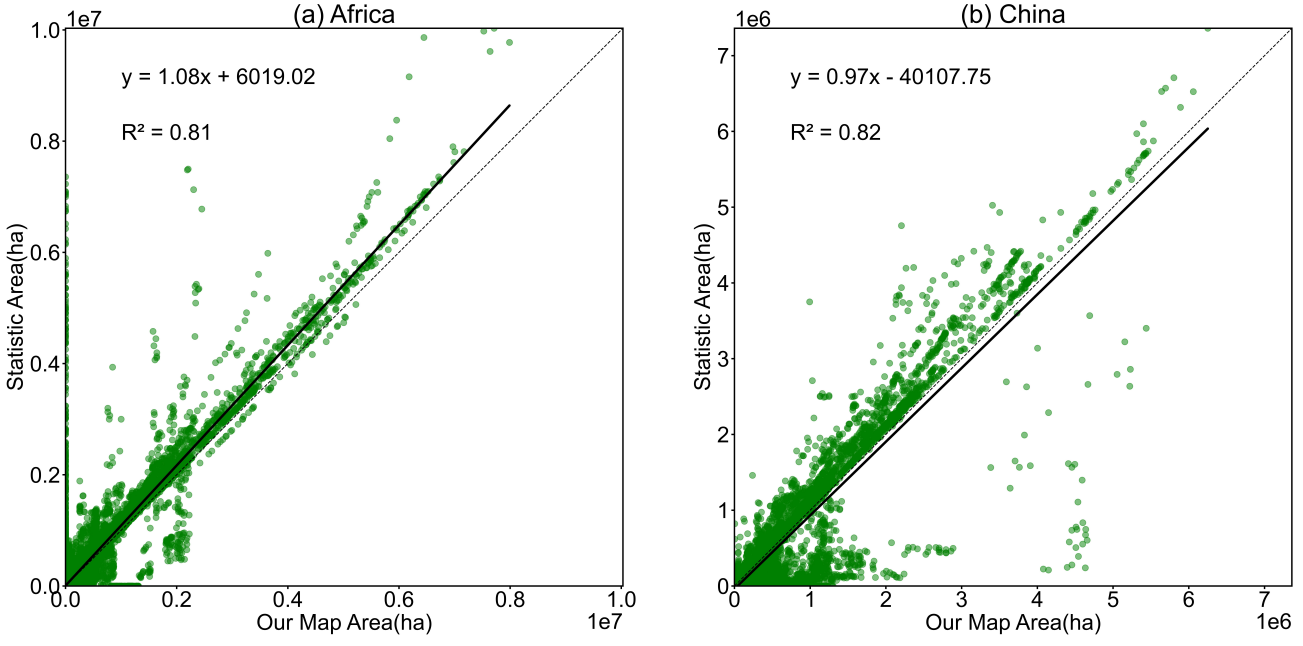


Figure 2. Consistency of crop-specific area maps and crop statistics (Fig 2a, Africa; Fig 2b China).

Here, we presented our mapping results of crop-specific areas in China (Fig 3) and Africa (Fig 4) in multiple time stages by selecting the representative crops for each category. It can be seen that our results can characterize the distribution of crop-specific areas over long periods of time. In the past few decades, the harvest area of many crops in both regions has increased substantially, and this increase has taken the form of expansion and intensification. Expansion refers to the emergence of new areas of crop cultivation, while intensification refers to the realization of more crop harvest areas within the original crop distribution range. For example, maize in the past 40 years has achieved intensification in the North China Plain and has achieved expansion in the Northeast Plain (Fig 3). The soybean planting area in China has both expanded and intensified to a large extent in the northeast plain (Fig 3). While Rice cultivation in China is relatively stable, distributed in the middle and

https://doi.org/10.5194/essd-2024-233
Preprint. Discussion started: 18 July 2024

lower reaches of the Yangtze River plain, northeast plain, southeast coast and other areas, without significant changes (Fig 3). Winter wheat in the North China Plain is the main wheat growing region in China, and the main distribution range has not changed, but the intensification degree has changed in different small regions (Fig 3).

In Africa, maize and sorghum are the cereals with the largest increases in harvested area. As for maize, the Niger River basin in West Africa, the Ethiopian Plateau in North Africa, and East Africa are regions with relatively rapid growth, while southern Africa is the main stable producing area (Fig 4). The Chad Basin and the middle Nile region are the main areas of increase in sorghum's harvest area. While wheat and rice harvests in Africa were relatively stable and did not increase significantly (Fig 4). Cassava, a tuber crop, is a traditional food for rural people in tropical Africa and has expanded and intensified along the coast of the Gulf of Guinea in West Africa (Fig 4). Other representative African crops have not changed as much relatively.

https://doi.org/10.5194/essd-2024-233
Preprint. Discussion started: 18 July 2024

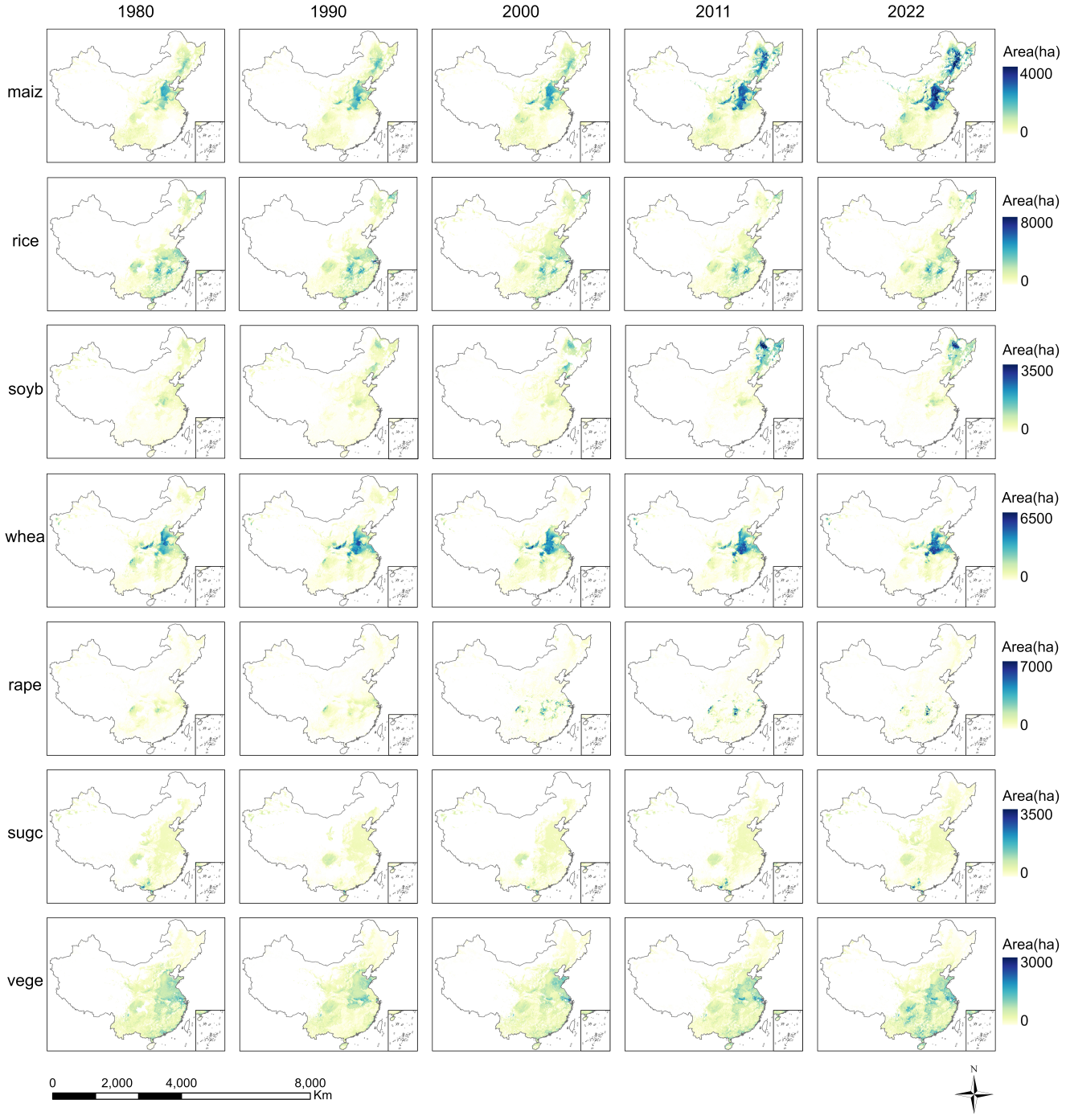
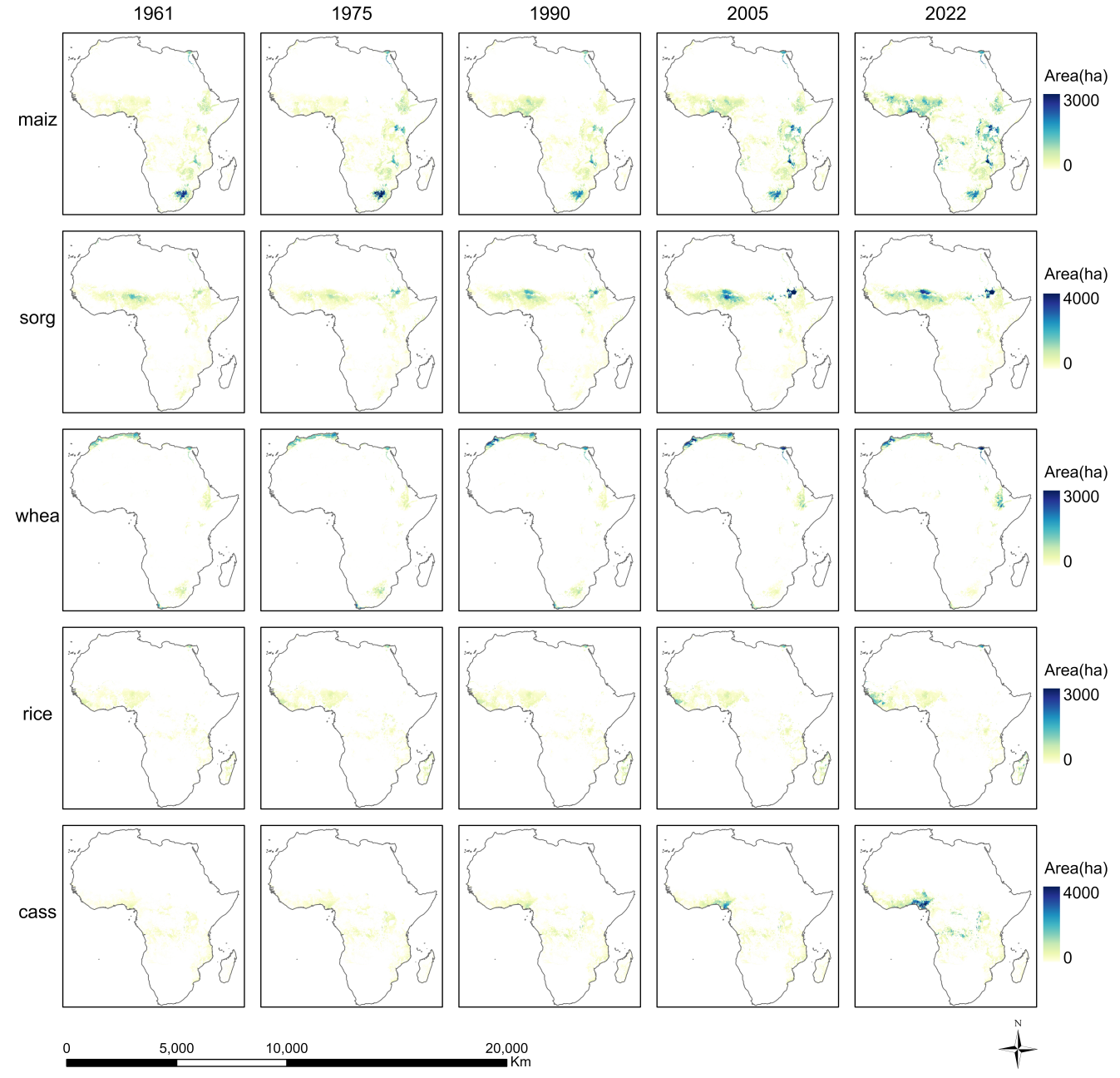


Figure 3. Mapping results of crop-specific harvest areas in China (covering 1980-2022, including maize, rice, soybean, wheat, rapeseed, sugarcane, vegetables, etc.).

https://doi.org/10.5194/essd-2024-233
Preprint. Discussion started: 18 July 2024



275

Figure 4. Mapping results of crop-specific harvest areas in Africa (covering 1961-2022, including maize, sorghum, wheat, rice, cassava etc.).

https://doi.org/10.5194/essd-2024-233
Preprint. Discussion started: 18 July 2024

4.2 Crop-specific changes in time and space

Our results provide annual distribution of crop-specific areas which could be further used to identify spatial patterns of change trends. In this part, we used the linear fitting method to calculate the variation trend of the harvested area of various crops at a 10km grid. In China, we noticed that the harvested area of maize increased significantly in almost all regions, and decreased only in the northern part of the North China Plain (Fig 5e). In contrast, there is a modest reduction of other cereals' harvested area in some planting regions. The rice harvest area increased significantly in the North China Plain and the southeast hills, but decreased in all other areas, especially in the southeast coastal area (Fig 5h). Harvest area of wheat increased only in the middle and southern part of the North China Plain, but decreased in all other areas, especially in the Sichuan Basin and the area north of the Qinling Mountains and south of the Loess Plateau (Fig 5n). As for oil crops, soybean planting area has expanded dramatically across the Northeast Plain but reduced in other regions over the past four decades (Fig 5i). The area of rapeseed harvested increased in the Yunnan-Guizhou Plateau and southeastern hills, but decreased in the middle and lower Yangtze River plain (Fig 5g). Moreover, vegetable harvest area increased in the southern area of China (Fig 5m) while bean harvest area increased in the middle and lower Yangtze River Plain and North China Plain (Fig 5a). Our spatial analysis of crop harvest area change trends can provide a database for food security and environmental impact analysis, which also implies changes in population dietary requirements for different crops.

https://doi.org/10.5194/essd-2024-233
Preprint. Discussion started: 18 July 2024

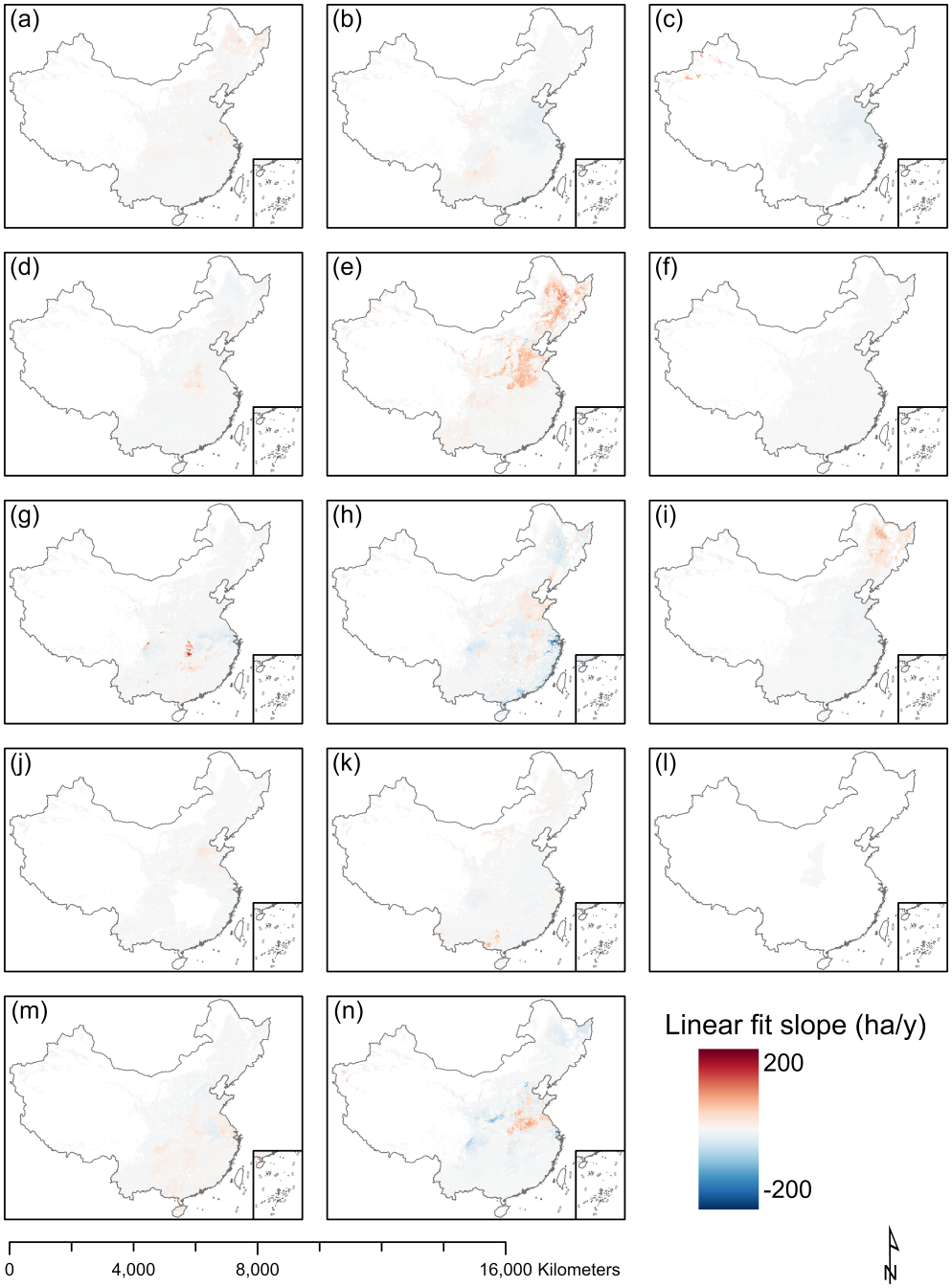


Figure 5. Linear fit slope of crop-specific areas in China (1980-2022, unit: ha/year), including a) bean; b) roots and tubers; c) cotton; d) groundnut; e) maize; f) bast fiber; g) rapeseed; h) rice; i) soybean; j) sugar beet; k) sugarcane; l) tobacco; m) vegetables; n) wheat.

https://doi.org/10.5194/essd-2024-233
Preprint. Discussion started: 18 July 2024

In Africa, we found that the harvested area of the major crop types in Africa is increasing, and there is little area with a clear trend of decreasing over the past six decades. Among them, cassava, a tuber crop, has an obvious growth trend in the Gulf of Guinea region of West Africa (Fig 6a). And cowpea harvest area increased in the middle Niger River basin (Fig 6c). The Chad Basin and the middle Nile region are the main areas of increase in sorghum harvest area (Fig 6g). Our maps and analysis can provide information on the underlying factors driving agricultural development in Africa and be incorporated into future decision-making.

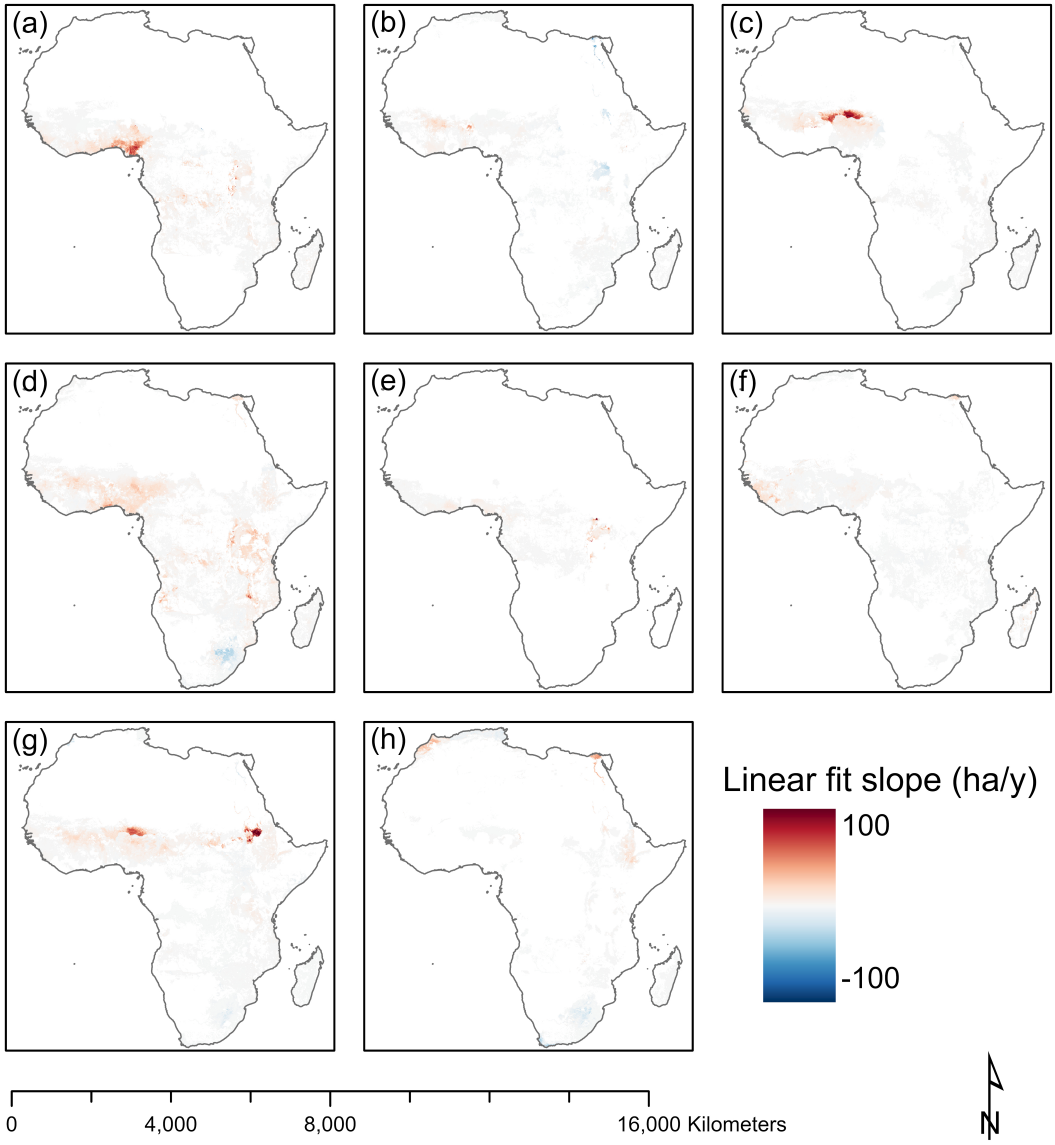


Figure 6. Linear fit slope of crop-specific areas in Africa (1960-2022, unit: ha/year), including a) cassava; b) cotton; c) cowpea; d) maize; e) plantations; f) rice; g) sorghum; h) wheat.

https://doi.org/10.5194/essd-2024-233
Preprint. Discussion started: 18 July 2024

4.3 Result validation

Crop-specific area maps produced in this study are estimates of crop harvest area distribution with various uncertainties. Therefore, we adopted multiple methods to validate the effectiveness of our method and the accuracy of our products. First of all, we selected independent crop maps which are not used in our mapping process to verify our products. In the part of data comparison, we calculated two accuracy indicators as a representation of consistency. The first one is the coefficient of determination (R^2) between the values of our products and others. A higher R^2 generally indicates a better performance. The second one is the root-mean-square-error (RMSE) between the values of our products and others. In contrast, a lower RMSE generally indicates a better consistency of two datasets.

SPAM2010 was used as base map in the process of our product especially in Africa. However, the SPAM model also produces global maps of crop harvest area in 2005 and 2000. The SPAM2005 was selected to validate our products for crop categories of SPAM2000 is not same as the other two, covering 20 crops rather than 42 crops. Besides for validation, the other reason why SPAM2005 was not used as the input in our model is that it is not recommended to cross-compare SPAM products over time for differences may contain more errors or inaccuracies than real changes in the ground (Yu et al., 2020).

Results show that crop map in 2005 updated by our method has a relatively good consistency with SPAM2005 at the grid level in Africa (Fig 7). Maize, cassava, groundnut, cowpea are the representative crops of cereals, roots & tubes, oil crops and pulses with the largest harvest area in each group in Africa. Among these crops, the values of R^2 are between 0.44 and 0.74, the values of RMSE are between 155 and 360 ha. We admit that there are crop types with a relatively low value of R^2, while these crop types are mainly crop aggregates (e.g. other fibers, rest crops) (Table S4-1).

https://doi.org/10.5194/essd-2024-233
Preprint. Discussion started: 18 July 2024

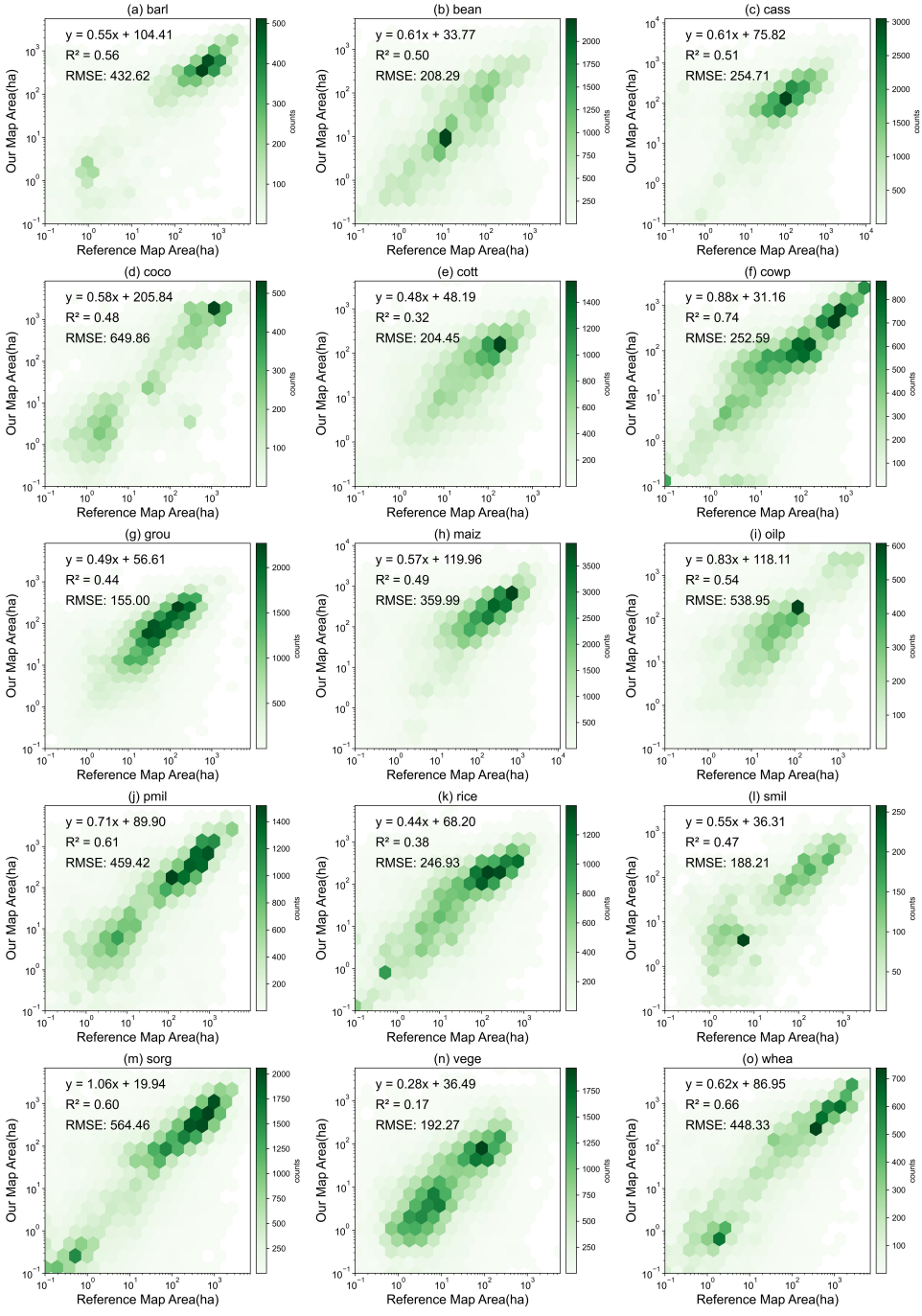


Figure 7. Comparison between our results (2005) and SPAM2005 in Africa at the grid level. A total of 15 crops with the large planting area in Africa were selected, including a) barley; b) bean; c) cassava; d) cocoa; e) cotton; f) cowpea; g) groundnut; h) maize; i) oil palm; j) pearl millet; k) rice; l) small millet; m) sorghum; n) vegetables; o) wheat. Comparison results of other crops could be found in Table S4-1.

https://doi.org/10.5194/essd-2024-233
Preprint. Discussion started: 18 July 2024


Open Access Earth System Science Data Discussions

As for China, we used integrated multi-source crop distribution maps as base map (not only SPAM2010), which will introduce uncertainty and also greater differences at the grid scale compared with the SPAM series products. Therefore, we aggregated the result of SPAM2005 at adm2 level to compare with ours at the same unit for SPAM series product collected as many adm2 level statistics data as possible, among which the adm2 level statistics coverage rate in SPAM2005 accounts for 54.6% (global average). Results show that crop map in 2005 updated by our method has a relatively good consistency with SPAM2005 at the adm2 level in China (Fig 8), especially maize (R^2=0.86), soybean (R^2=0.74), tobacco (R^2=0.84), vegetables (R^2=0.69) and wheat (R^2=0.89). We noticed that there are some crops that have a poor consistency such as beans, roots, bast fiber, sugarbeet, sugarcane, etc (Fig S4-1). It is caused by inconsistency of statistics from adm1 and adm2 for there is a missing value collected at the adm2 level (from SPAM2005) but the result collected at adm1 unit has a nonzero statistic value.

https://doi.org/10.5194/essd-2024-233
Preprint. Discussion started: 18 July 2024

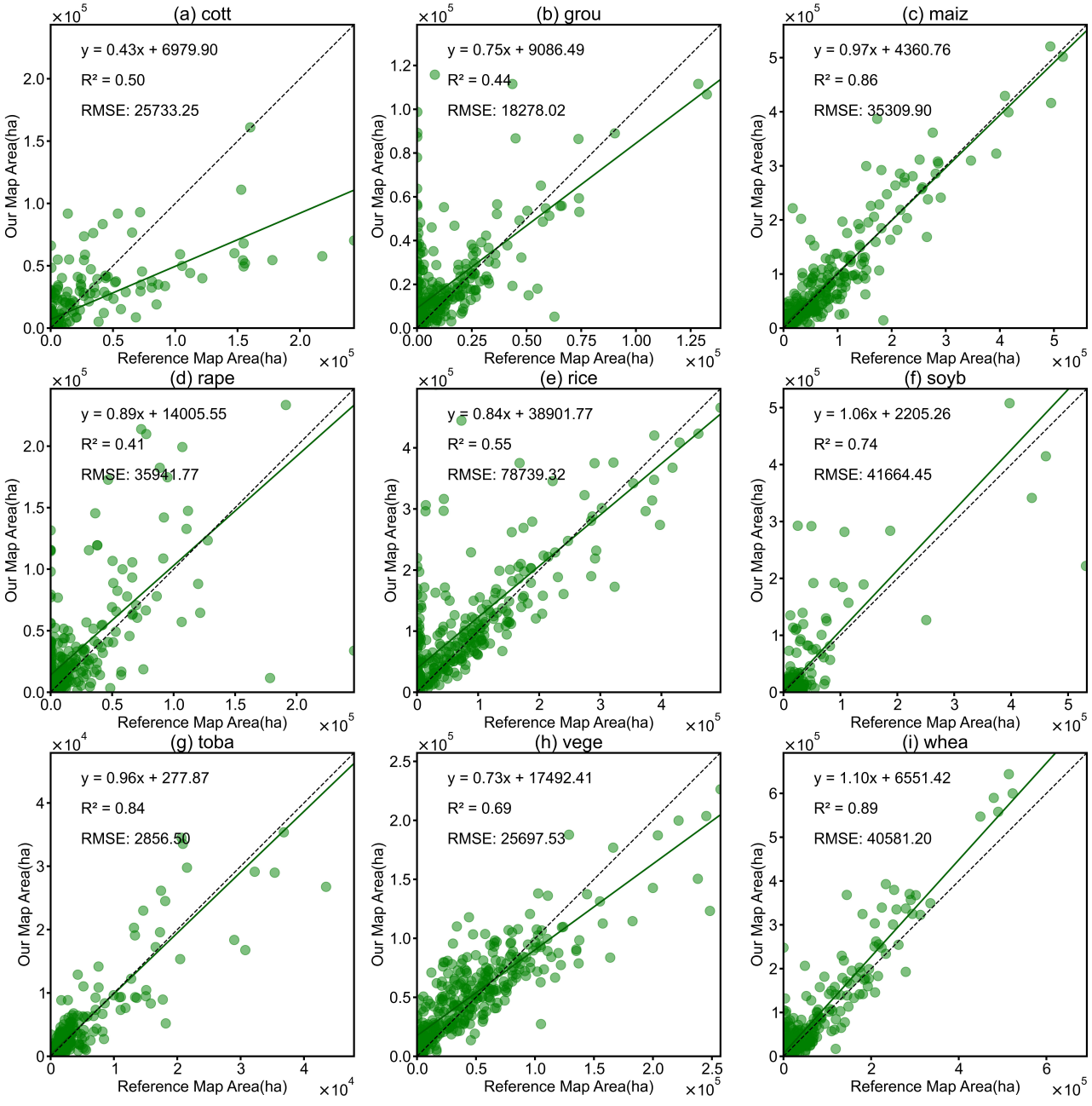


Figure 8. Comparison between our results (2005) and SPAM2005 in China at the Administrative unit (adm2 level), excluding crop types with strong inconsistency of statistics from adm1 and adm2, which includes a) cotton; b) groundnut; c) maize; d) rapeseed; e) rice; f) soybean; g) tobacco; h) vegetables; i) wheat. Full comparison results could be found in Fig S4-1.

To verify the products' accuracy in a time series, we compared our products with PCAM (Probabilistic Cropland Allocation Model). PCAM is produced based on randomly allocating statistics to the probability cluster of crop suitability through multiple Monte Carlo, which provides the crop-specific area harvested of 17 major crops in a global 0.5-degree grid from 1961

https://doi.org/10.5194/essd-2024-233
Preprint. Discussion started: 18 July 2024

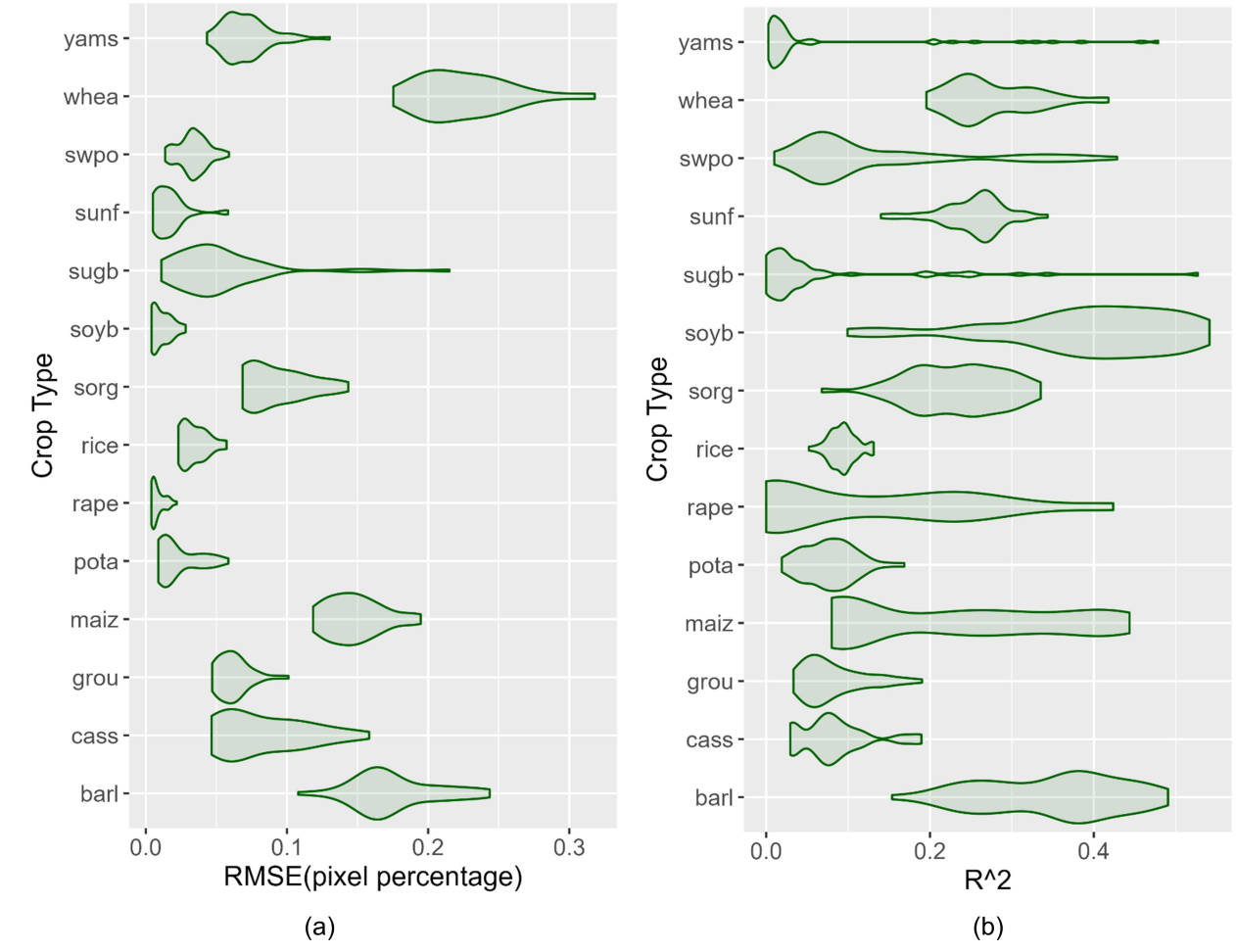


Figure 9. Comparison between our results (1961-2014) and PCAM in Africa at the grid level. (Fig 9a, RMSE (pixel percentage in a 10km grid), Fig 9b, R2)

to 2014 (Jackson et al., 2019). Other datasets such as M3, MIRCA, GAEZ, and SPAM only provide maps at certain years. We resampled the spatial resolution PCAM to 10km before comparison. Results showed that our products do not have a good consistency with PCAM, with relatively high RMSE (some crop types exceed 1000 ha) and low R^2 (mainly between 0.2 to 0.4) (Fig 9). This may relate to the reason that PCAM has a coarser spatial resolution, where crop-specific areas distributed in our product are not covered by PCAM (Fig S4-2).

Besides, we also carried out contrast experiments to verify the role of key parts of our model. Producing annual-updated probabilistic layers is one of the essential parts of this study, which were used as the basis of crop statistics allocation at certain years. Here, we conducted experiments to prove the effectiveness of this step. More specifically, we compared the accuracy differences between results produced by annual-updated probabilistic layers and only base-year layers (SPAM2010), other experiment settings (including method and input data) are the same. Results showed that the input of annual-updated

https://doi.org/10.5194/essd-2024-233
Preprint. Discussion started: 18 July 2024

Open Access Earth System Science Data Discussions

probabilistic layers significantly improved mapping accuracy in almost every crop type, with higher R^2 and lower RMSE (Fig 10). For example, compared with SPAM2005, R^2 of maize maps produced with updated probabilistic layers in 2005 is 0.487, which is higher than using only the base map (0.151), while RMSE is 360 ha, lower than using only the base map (524 ha).

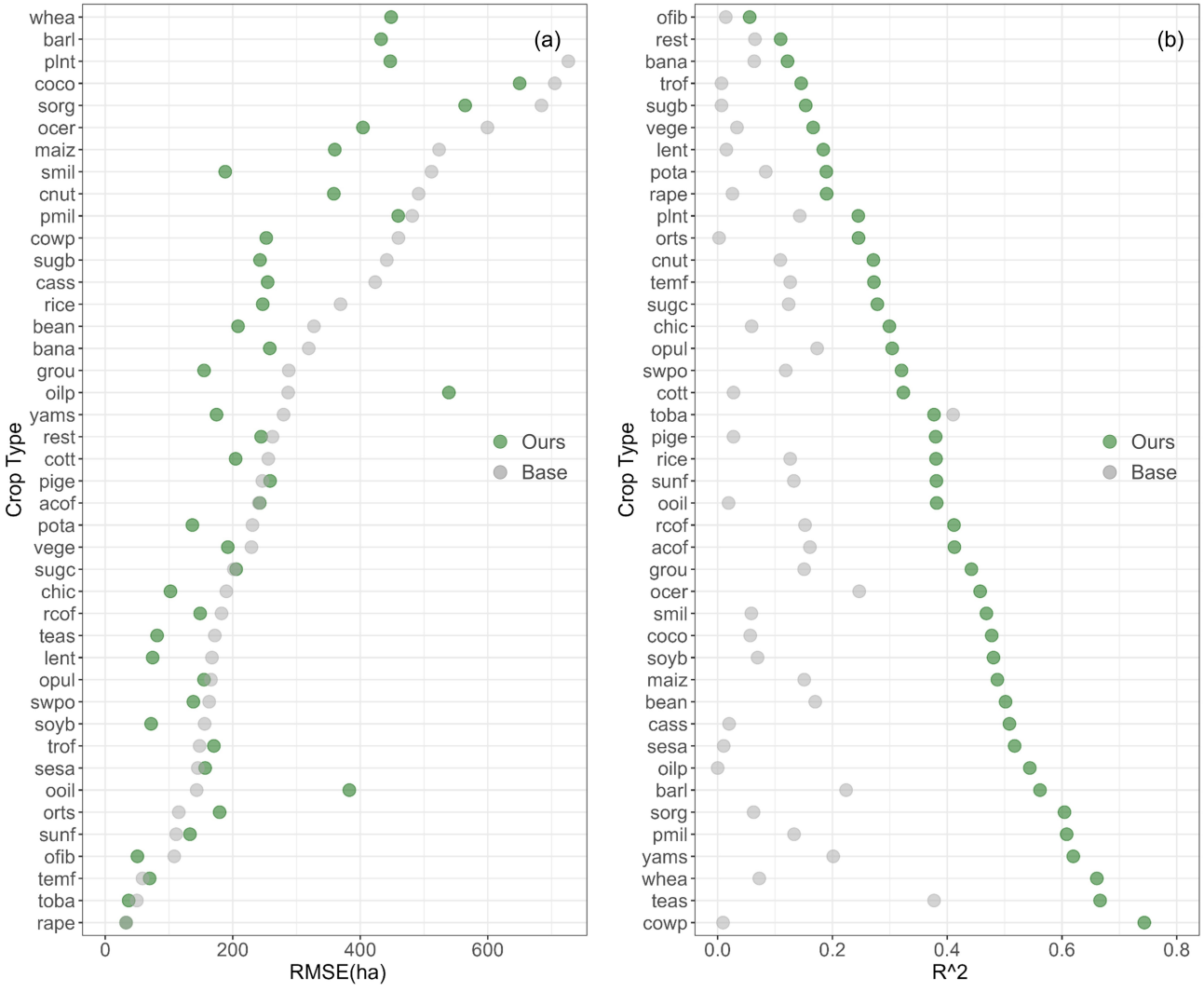
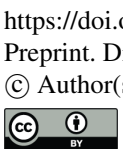


Figure 10. Accuracy comparison of mapping results using annual-updated probabilistic layers and only base-year layers (SPAM2010) (compared with SPAM2005 in Africa at the grid level, Fig 10a, unit: R2, Fig 10b, unit: RMSE (ha)).

Last but not least, we tested the whole model by transferring it into regions with rich data. USA, one of the regions with the most sufficient and accessible crop area references in the world, is selected as one of the study areas to conduct the same work flow to validate the effectiveness and feasibility, especially updating crop-specific area maps in annual series. Here, we selected CDL in the year 2010 as the base map (the same setting in Africa, using SPAM2010) and updated annual probabilistic layers based on the RF regression model trained based on CDL 2010. Final results were produced by allocating crop statistics from

https://doi.org/10.5194/essd-2024-233
Preprint. Discussion started: 18 July 2024



USDA based on annually updated layers. Results showed that our method could generate annual crop maps with relatively consistent accuracy (Fig 11). As for staple crop types such as maize, soybean and wheat, annually updated crop maps achieved relatively higher R^2 (between 0.53 to 0.78) and lower RMSE (between 492 to 727 ha) with less variation in time series, while we also noticed that mapping accuracy of some crops fluctuates greatly in a given year, especially crops of small harvest areas such as tobacco, sugarcane and groundnut in terms of RMSE. As for R^2, almost all of the rest crops except for staple ones have some fluctuation (Fig 11).

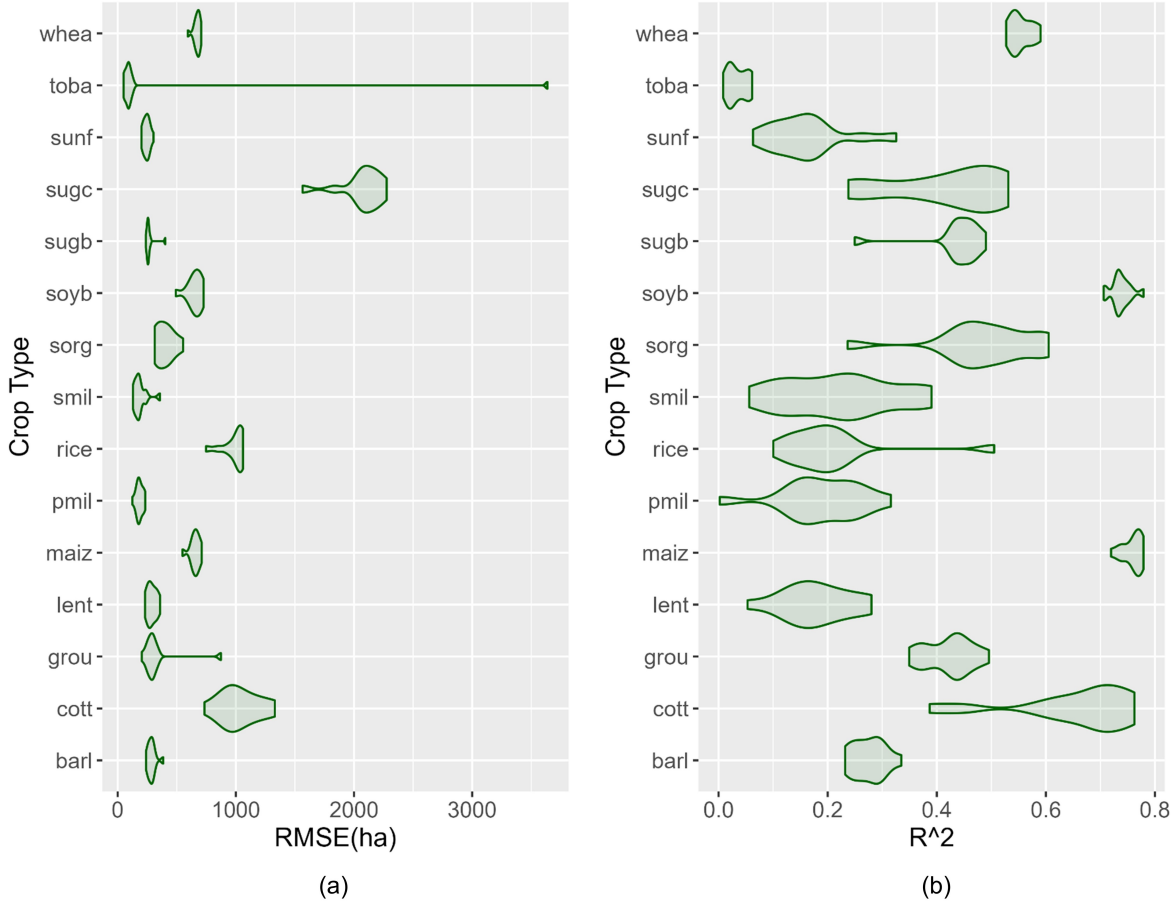


Figure 11. Comparison between our results (2008-2022) and CDL in USA at the grid level. (Fig 11a, RMSE, Fig 11b, R2)

5 Discussion

5.1 Input data uncertainty

First of all, our estimates of crop-specific areas are largely dependent on the quality of input data. In simple terms, crop statistics provide total amount on crop-specific areas, while the base map provides spatial details of crop distribution in a given base year. When the total is not accurate, especially if a number is too large, all suitable cropland in the administrative unit will

https://doi.org/10.5194/essd-2024-233
Preprint. Discussion started: 18 July 2024

exceed the maximum crop harvest area limit, which will make the allocation process impossible to continue. Under the circumstances, we use the results of the base year to replace such grid values. In addition, the uncertainty in the input map will directly affect our mapping results. For example, the differences in input cropland extent cause the different coverage of the crop-specific areas at the first stage, while the uncertainty and inaccuracy of the base map will increase the difficulty of model learning effective knowledge. The model performance in USA is better than that in Africa (Fig 10, Fig 7). Except for regional differences, one possible reason is that the base map used in USA is from CDL which is considered ground truth in many studies, while the base map used in Africa is from SPAM2010, which is the best approximation of crop distribution based on statistics allocation and still contain a lot of uncertainty.

5.2 Model uncertainty

Secondly, our model provides a framework for updating annual crop-specific area maps based on statistics, but large uncertainties remain. The results should be regarded more as approximate estimates rather than ground realities. Before statistic allocation started, we used RF regression models to generate annual-updated probabilistic layers of crop-specific areas. To compose the feature set, we selected 26 widely used and open-access spatial indicators that are related to crop distribution. They clearly can't represent all the potential indicators, additional related indicators published in the future may be beneficial to the model optimization. Compared with recent study that predicting crop suitability by temperature, precipitation and soil data in Africa (Chemura et al., 2024), the indicators included in our framework are more comprehensive. As for the machine learning task of predicting crop distribution area, the relationship between crop distribution area and spatial indicators on the 10km scale is implicit, which may not be fully captured by the model, unlike using medium-high resolution remote sensing images to construct the relationship between bands and crop characteristics (Li et al., 2021). In the similar task of decomposing population distribution, the population is directly proportional to the density of roads, built-up areas, and night lights (Sorichetta et al., 2015). Deep learning networks may be able to play a larger role, but this requires much larger amounts of training data as input, which is not sufficient for scarce crop distribution information (especially in regions like Africa).

Here, we further analyze the interpretability of the model. We found that there are some differences in the distribution prediction models of different crops, and the variables that play the key role are inconsistent. Cropland area, latitude and longitude, rural population, these variables have the highest importance in the distribution prediction model of almost all crops in Africa (Fig S4-5). While in models of USA, only cropland area still plays the most important role (Fig S4-6). It is worth noticing that location and terrain indicators both play a vital role in our models of China no matter whether using SPAM2010 or multi-year mapping results as training data (Fig S4-7,S4-8,S4-11,S4-12). We also calculated the average values of spatial indicators in each group (climate, agro-system, suitability, potential yield, soil, terrain, and location). The results showed that location variables were the most important for predicting crop distribution in Africa, followed by agricultural system, climate, terrain, soil, etc. (Fig S4-9). It explains that our model largely predicts crop distribution based on crop planting history (from the base map and cropland extent), while other related factors such as agricultural system and climate will be helpful to reflect fluctuations in the year series.

https://doi.org/10.5194/essd-2024-233
Preprint. Discussion started: 18 July 2024

5.3 Integrate statistics with remote sensing: Top-down and bottom-up

Last but not least, it will create more possibilities for crop mapping to integrate statistics with remote sensing, which means better integration of the two strategies (top-down and bottom-up). In many studies, statistical data are only used as validation or reference for remote sensing mapping of crop distribution, while in studies based on statistic spatial allocation, bottom-up strategies and utilization of remote sensing are often ignored. Our research only shows one kind of new possibility, which utilizes statistic decomposition and spatiotemporal prediction models, especially in studies conducted in China, which integrated multiple crop mapping results based on remote sensing, and the possible distribution in the remaining years was estimated by using a spatiotemporal prediction model and further harmonized when combined with statistical decomposition. Our study is only a simple fusion of the two strategies and data sources, and the in-depth integration of the two needs to be further explored in the future. Since statistical data can cover long time series and whole crop types, while remote sensing provides real-time monitoring of real conditions on the ground and sufficient spatial details. Giving play to the advantages of both will achieve mutually beneficial results. For example, the spatial allocation algorithm can be used to coordinate the sub-pixel results of remote sensing classification (Hu et al., 2021), the crop growing season information extracted from remote sensing can be incorporated into the spatial allocation model of statistics (Wang et al., 2022), or the deep learning algorithm can be trained for crop pattern recognition with the constraint of statistical data (Zhong et al., 2019).

6 Data availability

Datasets produced by this study are available to the public at https://doi.org/10.6084/m9.figshare.26028769 (Li et al., 2024) ,which include maps of crop-specific areas covering 42 types from 1961-2022 in Africa, maps of crop-specific areas covering 14 types from 1980-2022 in China and maps of crop-specific areas covering 15 types from 2008-2022 in USA.

7. Conclusion

In this study, we developed a framework for updating annual crop-specific area maps at 10km resolution based on crop statistics disaggregating, multi-source data integrating and machine learning, taking factors related to crop distribution in different regions and complex agricultural systems into account. Experiments were conducted in three study areas (Africa, China, and USA) respectively corresponding to three conditions of the information coverage of crop distribution (low, median, and high). In our framework, we collected related spatial indicators used in previous studies and trained random forest regression models to predict spatiotemporal dynamics of crop-specific areas based on them. Annual crop statistics were further disaggregated based on the probabilistic layer and harmonized based on multiple constraints. Our framework is a good attempt to integrate two strategies (top-down and bottom-up), creating more possibility for crop mapping to integrate statistics with remote sensing. Finally, we produced maps of crop-specific areas covering 42 types from 1961-2022 in Africa, maps of crop-specific areas covering 14 types from 1980-2022 in China and maps of crop-specific areas covering 16 types from 2008-2022

in USA (for validation). Results show that our products have a relatively good consistency with independent reference maps or statistics. Our products provide approximate estimates for spatiotemporal dynamics of crop-specific areas in multiple regions over several decades, which could be used as data basis for food security and environmental impact assessments.

Author contribution

XL and YL designed the framework and XL developed the datasets. XL prepared the manuscript with contributions from all co-authors.

Competing interests

The authors declare that they have no conflict of interest.

Financial support

This work was supported by the National Key R&D Program of China (2022YFE0209400).

References

Alami Machichi, M., mansouri, l. E., imani, y., Bourja, O., Lahlou, O., Zennayi, Y., Bourzeix, F., Hanadé Houmma, I., and Hadria, R.: Crop mapping using supervised machine learning and deep learning: a systematic literature review, International Journal of Remote Sensing, 44, 2717-2753, 10.1080/01431161.2023.2205984, 2023.

Anderson, W., You, L., Wood, S., Wood-Sichra, U., and Wu, W.: An analysis of methodological and spatial differences in global cropping systems models and maps, Global Ecology and Biogeography, 24, 180-191, https://doi.org/10.1111/geb.12243, 2015.

Becker-Reshef, I., Barker, B., Whitcraft, A., Oliva, P., Mobley, K., Justice, C., and Sahajpal, R.: Crop Type Maps for Operational Global Agricultural Monitoring, Scientific Data, 10, 172, 10.1038/s41597-023-02047-9, 2023.

Boryan, C., Yang, Z., Mueller, R., and Craig, M.: Monitoring US agriculture: the US Department of Agriculture, National Agricultural Statistics Service, Cropland Data Layer Program, Geocarto International, 26, 341-358, 10.1080/10106049.2011.562309, 2011.

Breiman, L.: Random Forests, Machine Learning, 45, 5-32, 10.1023/A:1010933404324, 2001.

Bren d'Amour, C., Reitsma, F., Baiocchi, G., Barthel, S., Güneralp, B., Erb, K.-H., Haberl, H., Creutzig, F., and Seto, K. C.: Future urban land expansion and implications for global croplands, Proceedings of the National Academy of Sciences, 114, 8939-8944, 10.1073/pnas.1606036114, 2017.

Cao, B., Yu, L., Li, X., Chen, M., Li, X., Hao, P., and Gong, P.: A 1km global cropland dataset from 10000BCE to 2100CE, Earth Syst. Sci. Data, 13, 5403-5421, 10.5194/essd-13-5403-2021, 2021.

Chemura, A., Gleixner, S., and Gornott, C.: Dataset of the suitability of major food crops in Africa under climate change, Scientific Data, 11, 294, 10.1038/s41597-024-03118-1, 2024.

d'Andrimont, R., Verhegghen, A., Lemoine, G., Kempeneers, P., Meroni, M., and van der Velde, M.: From parcel to continental scale – A first European crop type map based on Sentinel-1 and LUCAS Copernicus in-situ observations, Remote Sensing of Environment, 266, 112708, https://doi.org/10.1016/j.rse.2021.112708, 2021.

Davis, K. F., Gephart, J. A., Emery, K. A., Leach, A. M., Galloway, J. N., and D'Odorico, P.: Meeting future food demand with current agricultural resources, Global Environmental Change, 39, 125-132, https://doi.org/10.1016/j.gloenvcha.2016.05.004, 2016.

Ellis, E. C., Kaplan, J. O., Fuller, D. Q., Vavrus, S., Klein Goldewijk, K., and Verburg, P. H.: Used planet: A global history, Proceedings of the National Academy of Sciences, 110, 7978-7985, 10.1073/pnas.1217241110, 2013.

Fischer, G., Nachtergaele, F., Van Velthuizen, H., Chiozza, F., Franceschini, G., Henry, M., Muchoney, D., and Tramberend, S.: Global agro-ecological zones v4–model documentation, Food & Agriculture Org.2021.

Grogan, D., Frolking, S., Wisser, D., Prusevich, A., and Glidden, S.: Global gridded crop harvested area, production, yield, and monthly physical area data circa 2015, Scientific Data, 9, 15, 10.1038/s41597-021-01115-2, 2022.

Hoang, N. T., Taherzadeh, O., Ohashi, H., Yonekura, Y., Nishijima, S., Yamabe, M., Matsui, T., Matsuda, H., Moran, D., and Kanemoto, K.: Mapping potential conflicts between global agriculture and terrestrial conservation, Proceedings of the National Academy of Sciences, 120, e2208376120, 10.1073/pnas.2208376120, 2023.

Hu, Q., Yin, H., Friedl, M. A., You, L., Li, Z., Tang, H., and Wu, W.: Integrating coarse-resolution images and agricultural statistics to generate sub-pixel crop type maps and reconciled area estimates, Remote Sensing of Environment, 258, 112365, https://doi.org/10.1016/j.rse.2021.112365, 2021.

Huang, Q., Liu, Z., He, C., Gou, S., Bai, Y., Wang, Y., and Shen, M.: The occupation of cropland by global urban expansion from 1992 to 2016 and its implications, Environmental Research Letters, 15, 084037, 10.1088/1748-9326/ab858c, 2020.

Jackson, N. D., Konar, M., Debaere, P., and Estes, L.: Probabilistic global maps of crop-specific areas from 1961 to 2014, Environmental Research Letters, 14, 094023, 10.1088/1748-9326/ab3b93, 2019.

Klein Goldewijk, K., Beusen, A., Doelman, J., and Stehfest, E.: Anthropogenic land use estimates for the Holocene–HYDE 3.2, Earth System Science Data, 9, 927-953, 2017.

Laborde, D., Martin, W., Swinnen, J., and Vos, R.: COVID-19 risks to global food security, Science, 369, 500-502, 10.1126/science.abc4765, 2020.

Leff, B., Ramankutty, N., and Foley, J. A.: Geographic distribution of major crops across the world, Global Biogeochemical Cycles, 18, https://doi.org/10.1029/2003GB002108, 2004.

Lesiv, M., Laso Bayas, J. C., See, L., Duerauer, M., Dahlia, D., Durando, N., Hazarika, R., Kumar Sahariah, P., Vakolyuk, M. y., Blyshchyk, V., Bilous, A., Perez-Hoyos, A., Gengler, S., Prestele, R., Bilous, S., Akhtar, I. u. H., Singha, K., Choudhury,

https://doi.org/10.5194/essd-2024-233
Preprint. Discussion started: 18 July 2024

S. B., Chetri, T., Malek, Ž., Bungnamei, K., Saikia, A., Sahariah, D., Narzary, W., Danylo, O., Sturn, T., Karner, M., McCallum, I., Schepaschenko, D., Moltchanova, E., Fraisl, D., Moorthy, I., and Fritz, S.: Estimating the global distribution of field size using crowdsourcing, Global Change Biology, 25, 174-186, https://doi.org/10.1111/gcb.14492, 2019.

Li, X., Yu, L., Peng, D., and Gong, P.: A large-scale, long time-series (1984–2020) of soybean mapping with phenological features: Heilongjiang Province as a test case, International Journal of Remote Sensing, 42, 7332-7356, 10.1080/01431161.2021.1957177, 2021.

Li, X., Yu, L., Du, Z., and Liu, X.: Tracking spatiotemporal dynamics of crop-specific areas through machine learning and statistics disaggregating, 10.6084/m9.figshare.26028769, 2024.

Lin, F., Li, X., Jia, N., Feng, F., Huang, H., Huang, J., Fan, S., Ciais, P., and Song, X.-P.: The impact of Russia-Ukraine conflict on global food security, Global Food Security, 36, 100661, https://doi.org/10.1016/j.gfs.2022.100661, 2023.

Liu, X., Zheng, J., Yu, L., Hao, P., Chen, B., Xin, Q., Fu, H., and Gong, P.: Annual dynamic dataset of global cropping intensity from 2001 to 2019, Scientific Data, 8, 283, 10.1038/s41597-021-01065-9, 2021.

Liu, Y., Tan, Q., Chen, J., Pan, T., Penuelas, J., Zhang, J., and Ge, Q.: Dietary Transition Determining the Tradeoff Between Global Food Security and Sustainable Development Goals Varied in Regions, Earth's Future, 10, e2021EF002354, https://doi.org/10.1029/2021EF002354, 2022.

Lu, M., Wu, W., You, L., See, L., Fritz, S., Yu, Q., Wei, Y., Chen, D., Yang, P., and Xue, B.: A cultivated planet in 2010 – Part 1: The global synergy cropland map, Earth Syst. Sci. Data, 12, 1913-1928, 10.5194/essd-12-1913-2020, 2020.

Monfreda, C., Ramankutty, N., and 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, https://doi.org/10.1029/2007GB002947, 2008.

Nelson, G., Bogard, J., Lividini, K., Arsenault, J., Riley, M., Sulser, T. B., Mason-D'Croz, D., Power, B., Gustafson, D., Herrero, M., Wiebe, K., Cooper, K., Remans, R., and Rosegrant, M.: Income growth and climate change effects on global nutrition security to mid-century, Nature Sustainability, 1, 773-781, 10.1038/s41893-018-0192-z, 2018.

Portmann, F. T., Siebert, S., and Döll, P.: MIRCA2000—Global monthly irrigated and rainfed crop areas around the year 2000: A new high-resolution data set for agricultural and hydrological modeling, Global Biogeochemical Cycles, 24, https://doi.org/10.1029/2008GB003435, 2010.

Potapov, P., Turubanova, S., Hansen, M. C., Tyukavina, A., Zalles, V., Khan, A., Song, X.-P., Pickens, A., Shen, Q., and Cortez, J.: Global maps of cropland extent and change show accelerated cropland expansion in the twenty-first century, Nature Food, 3, 19-28, 10.1038/s43016-021-00429-z, 2022.

Ramankutty, N., Evan, A. T., Monfreda, C., and Foley, J. A.: Farming the planet: 1. Geographic distribution of global agricultural lands in the year 2000, Global Biogeochemical Cycles, 22, https://doi.org/10.1029/2007GB002952, 2008.

Ray, D. K., Ramankutty, N., Mueller, N. D., West, P. C., and Foley, J. A.: Recent patterns of crop yield growth and stagnation, Nature Communications, 3, 1293, 10.1038/ncomms2296, 2012.

Siebert, S. and Döll, P.: Quantifying blue and green virtual water contents in global crop production as well as potential production losses without irrigation, Journal of Hydrology, 384, 198-217, https://doi.org/10.1016/j.jhydrol.2009.07.031, 2010.

Siebert, S., Kummu, M., Porkka, M., Döll, P., Ramankutty, N., and Scanlon, B. R.: A global data set of the extent of irrigated land from 1900 to 2005, Hydrol. Earth Syst. Sci., 19, 1521-1545, 10.5194/hess-19-1521-2015, 2015.

Song, X.-P., Hansen, M. C., Potapov, P., Adusei, B., Pickering, J., Adami, M., Lima, A., Zalles, V., Stehman, S. V., Di Bella, C. M., Conde, M. C., Copati, E. J., Fernandes, L. B., Hernandez-Serna, A., Jantz, S. M., Pickens, A. H., Turubanova, S., and Tyukavina, A.: Massive soybean expansion in South America since 2000 and implications for conservation, Nature Sustainability, 4, 784-792, 10.1038/s41893-021-00729-z, 2021.

Sorichetta, A., Hornby, G. M., Stevens, F. R., Gaughan, A. E., Linard, C., and Tatem, A. J.: High-resolution gridded population datasets for Latin America and the Caribbean in 2010, 2015, and 2020, Scientific Data, 2, 150045, 10.1038/sdata.2015.45, 2015.

Su, H., Willaarts, B., Luna-Gonzalez, D., Krol, M. S., and Hogeboom, R. J.: Gridded 5 arcmin datasets for simultaneously farm-size-specific and crop-specific harvested areas in 56 countries, Earth Syst. Sci. Data, 14, 4397-4418, 10.5194/essd-14-4397-2022, 2022.

Tang, F. H. M., Nguyen, T. H., Conchedda, G., Casse, L., Tubiello, F. N., and Maggi, F.: CROPGRIDS: a global geo-referenced dataset of 173 crops, Scientific Data, 11, 413, 10.1038/s41597-024-03247-7, 2024.

Vogel, E., Donat, M. G., Alexander, L. V., Meinshausen, M., Ray, D. K., Karoly, D., Meinshausen, N., and Frieler, K.: The effects of climate extremes on global agricultural yields, Environmental Research Letters, 14, 054010, 10.1088/1748-9326/ab154b, 2019.

Wang, Y., Zhang, Z., Zuo, L., Wang, X., Zhao, X., and Sun, F.: Mapping Crop Distribution Patterns and Changes in China from 2000 to 2015 by Fusing Remote-Sensing, Statistics, and Knowledge-Based Crop Phenology, 10.3390/rs14081800, 2022.

Wei, D., Gephart, J. A., Iizumi, T., Ramankutty, N., and Davis, K. F.: Key role of planted and harvested area fluctuations in US crop production shocks, Nature Sustainability, 6, 1177-1185, 10.1038/s41893-023-01152-2, 2023.

You, L. and 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.

You, L. and 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.

You, L., Wood, S., and Wood-Sichra, U.: Generating plausible crop distribution maps for Sub-Saharan Africa using a spatially disaggregated data fusion and optimization approach, Agricultural Systems, 99, 126-140, https://doi.org/10.1016/j.agsy.2008.11.003, 2009.

You, L., Wood, S., Wood-Sichra, U., and 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.

https://doi.org/10.5194/essd-2024-233
Preprint. Discussion started: 18 July 2024

570 Yu, L., Du, Z., Dong, R., Zheng, J., Tu, Y., Chen, X., Hao, P., Zhong, B., Peng, D., Zhao, J., Li, X., Yang, J., Fu, H., Yang, G., and Gong, P.: FROM-GLC Plus: toward near real-time and multi-resolution land cover mapping, GIScience & Remote Sensing, 59, 1026-1047, 10.1080/15481603.2022.2096184, 2022.

Yu, Q., You, L., Wood-Sichra, U., Ru, Y., Joglekar, A. K. B., Fritz, S., Xiong, W., Lu, M., Wu, W., and Yang, P.: A cultivated planet in 2010 – Part 2: The global gridded agricultural-production maps, Earth Syst. Sci. Data, 12, 3545-3572, 10.5194/essd-
575 12-3545-2020, 2020.

Zhang, G., Xiao, X., Dong, J., Xin, F., Zhang, Y., Qin, Y., Doughty, R. B., and Moore, B.: Fingerprint of rice paddies in spatial–temporal dynamics of atmospheric methane concentration in monsoon Asia, Nature Communications, 11, 554, 10.1038/s41467-019-14155-5, 2020.

Zhong, L., Hu, L., Zhou, H., and Tao, X.: Deep learning based winter wheat mapping using statistical data as ground references
580 in Kansas and northern Texas, US, Remote Sensing of Environment, 233, 111411, https://doi.org/10.1016/j.rse.2019.111411, 2019.