Environmental Research Letters

LETTER • **OPEN ACCESS**

Probabilistic global maps of crop-specific areas from 1961 to 2014

To cite this article: Nicole D Jackson *et al* 2019 *Environ. Res. Lett.* **14** 094023

View the article online for updates and enhancements.

This content was downloaded from IP address 130.126.153.41 on 21/09/2019 at 04:34

Environmental Research Letters

LETTER

Probabilistic global maps of crop-specific areas from 1961 to 2014

Nicole D Jackson[1], Megan Konar[1], Peter Debaere[2] and Lyndon Estes[3]

[1] Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, Urbana IL, United States of America
[2] Darden School of Business, University of Virginia, Charlottesville, VA22903, United States of America
[3] Graduate School of Geography, Clark University Worcester, MA 01610, United States of America

E-mail: mkonar@illinois.edu

Keywords: algorithm development, global, gridded, probabilistic allocation, crop suitability, agricultural geography, time series

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
23 April 2019

REVISED
6 August 2019

ACCEPTED FOR PUBLICATION
15 August 2019

PUBLISHED
20 September 2019

Original content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence.

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

Abstract

Agriculture has substantial socioeconomic and environmental impacts that vary between crops. However, information on how the spatial distribution of specific crops has changed over time across the globe is relatively sparse. We introduce the Probabilistic Cropland Allocation Model (PCAM), a novel algorithm to estimate where specific crops have likely been grown over time. Specifically, PCAM downscales annual and national-scale data on the crop-specific area harvested of 17 major crops to a global 0.5-degree grid from 1961 to 2014. To do this, pixels are assigned into probability clusters based upon crop-specific pixel suitability (based on mean climate and soil characteristics) and gridded historical agricultural areas. PCAM maps compare relatively well with an existing gridded dataset of crop-specific areas circa 2000 (simple matching coefficient value >0.8 for all crops). PCAM estimates compare less well with time series county-level agricultural census data for the United States. Importantly, deviations between census data and PCAM benchmark estimates (driven by soil and climate suitability) can be used to infer the importance of other factors of agricultural production (e.g. labor, agricultural policy, extreme climate) in future work. Our results provide new insights into the likely changes in the spatial distribution of major crops over the past half-century.

1. Introduction

Agriculture is responsible for feeding the growing global population and is also one of the dominant ways in which humans impact the Earth system [1, 2]. Covering approximately 12% of the land surface as of 2000 [3], croplands now comprise one of the largest terrestrial biomes and continue to grow [4]. Among other impacts, croplands alter biogeochemical cycling [5], lead to forest clearing [6, 7], consume vast quantities of water that far exceed use by any other human activity [8], alter local, regional, and global climate [9–11], and degrade soil quality [12]. Specific crops impact natural resources and Earth surface processes in unique ways. For example, fertilized corn production in the US Midwest is responsible for increased nitrogen loads to the Gulf of Mexico [13], while wheat production in Northern India is rapidly depleting groundwater aquifers [14]. Similarly, rice production throughout Asia is a significant contributor to global atmospheric methane [15], while expanding soybean cultivation has converted large areas of South America's savannas and dry forests to cropland, and appears set to do the same in Southern Africa [16]. Estimating how the distribution of specific crops has likely evolved over time will enable better assessments of environmental system dynamics.

Agricultural management practices have changed substantially over the past several decades [17]. Although total global cropped area increased by about 18% since the mid-1900s, yields increased by 28% during the same time period [4]. This intensification of production was enabled by the increased use of fertilizers, pesticides, mechanization, irrigation, and cultivar improvements that are associated with the 'Green Revolution' [18, 19]. These productivity gains are substantial and have enabled production to outpace population growth. However, yields of the 'Big 4' (i.e. maize, rice, wheat, and soy) are increasing slower than the 2.4% per year rate required to double global

© 2019 The Author(s). Published by IOP Publishing Ltd

production by 2050 [20]. Additionally, extreme weather events have negatively impacted cereal production over the last several decades [21, 22]. Extreme events are projected to increasingly impact agriculture in the future, with different impacts depending on the crop and climate hazard. To determine historical and future exposure of crops to extreme events, it is important to identify where specific crops have been grown.

Despite their large consequences, understanding of these changes and their impacts is primarily confined to the national scale, as the only globally comprehensive, annual crop area dataset is the country-level statistics provided by the United Nations Food and Agriculture Organization (FAO) [4]. FAO provides an open-access source of global agricultural statistics which is widely utilized and provides a standard for national level agricultural statistics. FAO provides data on the area harvested [hectares] for each crop-country-year. Areas harvested multiple times in a single year are counted more than once in the national total. This means that the harvested area may exceed the physical area of the cropland that they are grown on. FAO also provides information on the production [tonnes] for each crop-country-year.

Previous studies have created sub-national, crop-specific distributions and production maps. Monfreda *et al* [23] developed a statistical disaggregation technique to downscale the FAO crop-specific areas to a 5 min resolution, using a gridded cropland area map as the disaggregation target [3]. Portmann *et al* [24] built on that effort by using crop calendars to temporally disaggregate the dataset to a monthly time step. However, both datasets represent the year 2000 crop distributions, and the disaggregation procedure does not account for distributional variations between crops below the scale of the administrative unit providing the crop area statistics [23]. You *et al* [25] developed the Spatial Production Allocation Model (SPAM), which is a crop allocation model that incorporates comparative advantage and potential economic value to spatially distribute crop production. SPAM uses a host of input data and a cross-entropy approach to make plausible estimates of the distribution of 42 crops under two production systems [25, 26]. As with the model presented by Monfreda *et al* [23], the SPAM dataset is also circa 2000 and does not vary in time, although model output for the years 2005 and 2010 has recently been generated [27, 28].

A number of datasets provide gridded agricultural land use histories, but not for specific crops. These include a modeled reconstruction of various land use distributions, including agriculture, for the past 300 [29, 30] and 12 000 years [31, 32]. These historical reconstructions were developed under Land Use Harmonization projects to create merged historical and future projected land use time series, which provided the forcings for Climate Model Inter-comparison Projects [33]. Additionally, Ramankutty and Foley [34] presented a historical reconstruction of total cropland for the period 1700–1992. Pongrantz *et al* [35] modeled cropland and pasture from 800 to 1992.

Despite this substantial progress in understanding agricultural geography, there is relatively sparse information on the spatial distribution of individual crops over time. The goal of this study is therefore to estimate crop-specific agricultural geography in time. Half degree spatial resolution maps of the annual distributions of 17 major crops for the period 1961–2014 are created. To generate these maps, a new cropland allocation algorithm was developed—the Probabilistic Cropland Allocation Model (PCAM). PCAM disaggregates the FAO's national census information [4] onto the global agricultural land use grids provided by the History Database of the Global Environment (HYDE) [32, 36]. The presented approach differs from previous disaggregation methods by using crop-specific land suitability surfaces provided by the Global Agro-Ecological Zones (GAEZ) database [37] to probabilistically allocate crop areas into grids. GAEZ provides gridded crop-specific suitabilities based upon mean climate and soil characteristics. Additionally, information on crop-specific calendars—including multi-cropping—is used to allocate crops in space.

The potential utility of these probabilistic maps is demonstrated by using them to answer two key questions: (1) how has the spatial distribution of key crops changed globally over the study period? (2) Which countries, crops, and time periods have experienced significant area changes? This dataset is expected to provide deeper insights into how environmental dynamics have changed when used as an input in Earth system models, which currently rely on temporally static or generic crop types in their land inputs (e.g., [38–40]).

2. Methods

The PCAM was developed to estimate gridded and annual crop-specific harvested area. Figure 1 outlines the framework. The approach relies on several key input data that are at different spatial resolutions (refer to table 1). In general, national census data is downscaled to geographic grids. To do this, gridded data on the crop-specific suitability of each pixel (based upon climate and soil characteristics) is incorporated with the fraction of each pixel that has historically been cropland. Then, national level information was probabilistically allocated to each pixel. The approach incorporates a Monte Carlo algorithm, with random selection across multiple trials to obtain a likelihood estimate. This novel algorithm enables researchers to estimate a global crop-specific, gridded product in time. The data sources, algorithm, and key assumptions are described in more detail below.

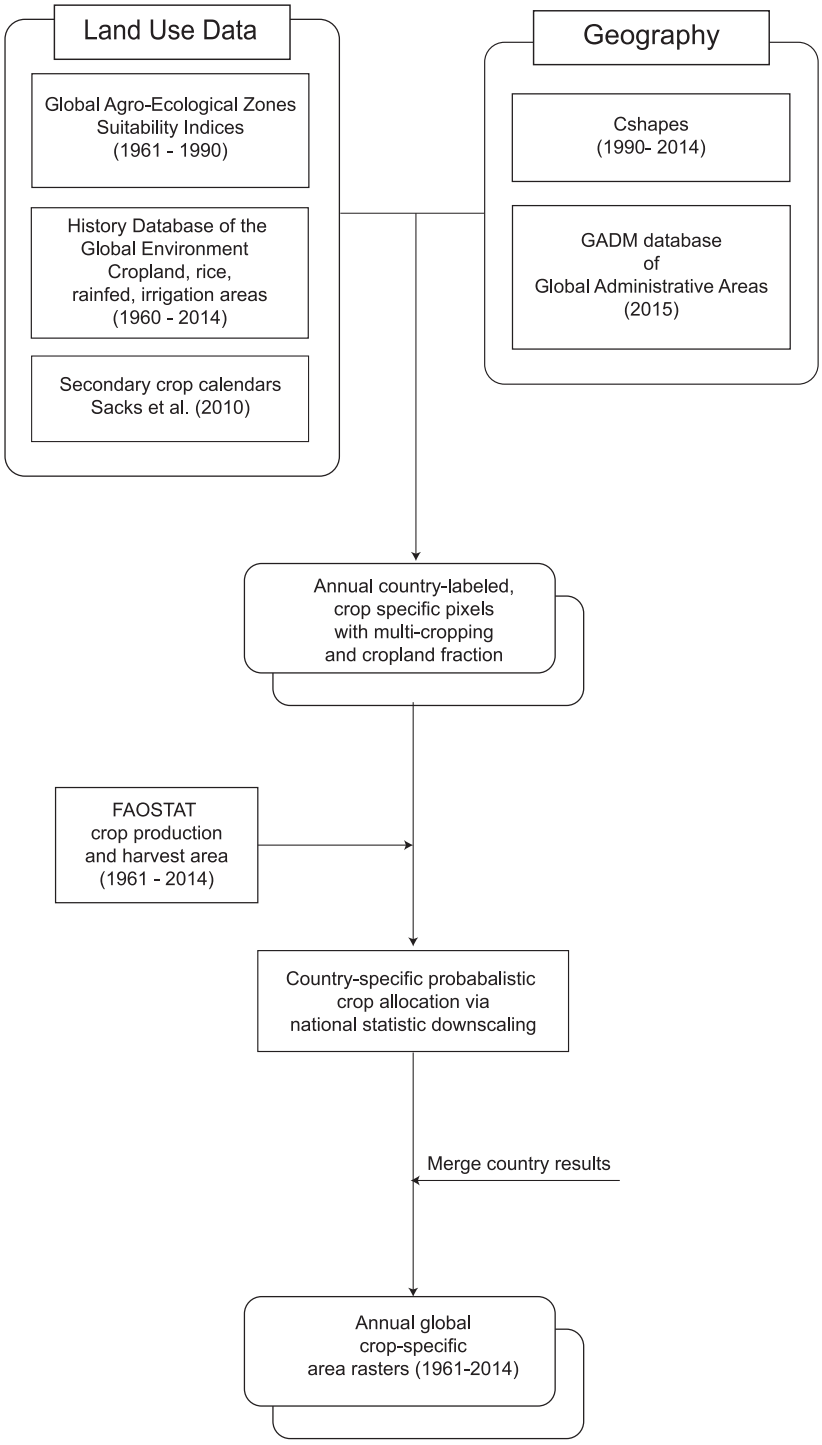

Figure 1. Schematic of the Probabilistic Cropland Allocation Model (PCAM) methodology.

Table 1. Summary of data used in this study.

Source	Spatial scale	Time period	Variable collected
FAO [4]	National	1961–2014	National census information on crop-specific production [tonnes] and harvested area [hectares]
GAEZ [37]	5 arc minute	1961–1990 (baseline)	Crop-specific suitability of each pixel (based on soil and climate)
HYDE [31]	5 arc minute	10 000 BC—2014	Fraction of each cell that is cropland, rice, rainfed or irrigated
Sacks *et al* [44]	5 arc minute	2000	Secondary crop calendars

2.1. Data sources and integration

Information from several key sources for this study was collected. Table 1 lists all of the (open source) data that were utilized. First, information on crop-specific production [tonnes] and area harvested [hectares] at the national level was collected from the FAO [4]. This information is available for 173 crops annually for the period 1961–2014 [4]. The FAO census data was used as the national allocation constraint in the PCAM algorithm (see next section).

Crop-specific suitability index (SI) rasters were obtained from GAEZ [37]. GAEZ is developed by the FAO and the International Institute for Applied Systems Analysis and provides information on the suitability of each pixel for specific crops based on mean climate and soil characteristics. Soil characteristics are principally based on the Harmonized World Soil Database with supplemental sub-national and regional information [41]. Mean climate information is derived from four general circulation models [42]. GAEZ has information for up to 280 crops/land utilization types under alternative input and management levels for historical, current and future climate conditions [37]. The SI values were selected based on intermediate agricultural inputs for both rainfed and irrigated water supply systems. The intermediate input level takes into account both subsistence and commercial production [42]. This GAEZ data provides an integral, gridded product for the PCAM allocation routine.

The History Database of the Global Environment (HYDE) provides 12 000 years (10 000 BC to 2017) of land use data at 5 arc minute spatial resolution [31, 32, 43]. HYDE provides gridded estimates of total cropland based on historical patterns of human populations. It does not specify the particular crops grown, with the exception of rice. From HYDE, time series information on the fraction of each grid cell that is cropland (rainfed and irrigated) as well as the fraction that contains rice was retrieved. This information is available decadally from 1960 to 2000 and annually from 2000 to 2014. Information on the total cropland of each pixel is used to constrain area accounting in the PCAM algorithm.

The FAO counts multi-cropped areas multiple times in the national statistic. This means that if the same land parcel is used twice in the same year, the area of this parcel will be counted twice [4]. For this reason, those pixels with multi-cropping were determined for better accounting in the algorithm. To incorporate pixel-scale information on multi-cropping, the secondary crop calendars provided by Sacks *et al* [44] were used for barley, maize, oats, rice, sorghum, and wheat. For example, if a pixel multi-crops rice in the data provided by Sacks *et al* [44], then the algorithm will count the pixel area twice if it is selected in the PCAM model.

The spatial scale, temporal domain, and crops considered were based upon the available input data. The study period was restricted to 1961–2014, since this is the time period for which FAO national agricultural statistics were available. The 0.5-degree spatial resolution was chosen for comparison with several existing global, gridded agricultural datasets (e.g. [23]). Finally, 17 crops were selected for consideration in this study (see SI for list). These 17 crops are selected because all of the required input data is available for them. These 17 crops represent 48%–62% of global agricultural production [tonnes] and 68%–75% of global harvested area for the years 1961–2014 (see figure S1 for details available online at stacks.iop.org/ERL/14/094023/mmedia).

The gridded data (i.e. GAEZ, HYDE, and crop calendars) are interpolated from a 5 arc minute grid to a 0.5-degree grid using a nearest neighbor approximation. A panel dataset was constructed from these three gridded data layers. To do this, information at the pixel level was combined by using each pixel's unique latitude-longitude pair. This enables researchers to construct a panel that has crop-specific information and varies in time. Each pixel with multi-cropping activities is assigned an indicator variable for the specific crops that are multi-cropped in that location.

In accordance with the UN, national boundaries from 1961 to 1990 were fixed to the 1990 boundaries. Dynamic country boundaries were obtained from the `Cshapes` v.0.6 package in R [45, 46]. These boundaries allow researchers to capture key events such as the dissolution of the Soviet Union. Pixels are paired with countries by using the boundaries active on 31 December for each year during the study time domain. For countries not found in `Cshapes`, the Database of Global Administrative Areas [47] was used. The pixel is labeled using the International Organization for Standardization (ISO) 3166-1 alpha-3, which provides each country with a unique 3-character code. These codes are used to match national statistics to pixels during the allocation process.

2.2. Algorithm development

The PCAM downscales national census data to a global grid. To do this, PCAM follows 5 major steps.

- Step 1: Assign pixels to probability clusters so that national information can be assigned to the most likely and suitable pixels.
- Step 2: Sort national crop-specific production (and corresponding harvested area) information for each year.
- Step 3: Assign the harvested area of crops to pixels in descending order of national crop production until the national harvested area statistic is met. This step is repeated for each produced crop.

Table 2. Global Agro-Ecological Zones (GAEZ) suitability index (SI) values and resulting probability clusters for use in the Probabilistic Cropland Allocation Model (PCAM). Clustering applies to pixels in countries where the HYDE cropland fraction is greater than zero for both rainfed and irrigated water supply systems across all crops. Final clustering is based on prioritizing rainfed over irrigation water types.

GAEZ Class	Suitability index threshold	Classification	Initial cluster	Final cluster by water type	
				Rainfed	Irrigation
1	SI $>$ 85	Very high	1	1	3
2	SI $>$ 70	High	1	1	3
3	SI $>$ 55	Good	1	1	3
4	SI $>$ 40	Medium	2	2	4
5	SI $>$ 25	Moderate	2	2	4
6	SI $>$ 10	Marginal	3	5	5
7	SI $>$ 0	Very Marginal	3	5	5
8	SI $=$ 0	Not suitable	4	6	6

- Step 4: Perform probabilistic downscaling many times using a Monte Carlo framework to obtain likely outcomes.
- Step 5: Repeat Steps 2–4 for all country-years in the study domain.

Additional details on these steps are provided here.

Step 1: Construct probability clusters. The goal of PCAM is to assign national information on the total area of cultivation for specific crops to pixels. Thus, it is essential to determine how to distribute national (lumped) information in space. A major novelty of the approach is the classification of pixels into probability clusters. These underlying probabilities are then used to probabilistically allocate national values. Gridded probabilities are based on a combination of GAEZ and HYDE data. First, pixels with HYDE agricultural areas $>$0 are available to allocate to. Then, pixels are assigned to probability clusters according to their underlying climate and soil SI derived from GAEZ data. SI values as provided by GAEZ are shown in table 2. Note that these values are pixel-crop specific. The SI categories were then clustered into six probability clusters in which the greatest probability (i.e. '1') is assigned to 'very high' quality pixels and the lowest probability (i.e. '6') is assigned to pixels that are 'not suitable'. Since FAO does not differentiate between varieties of rice (dryland versus wetland) and millet (foxtail versus pearl), generic rice and millet values were derived.

GAEZ information on the crop-specific SI of each pixel for rainfed and irrigated water supply systems is used to further refine probability clusters. Many locations that are not well-suited for a specific crop under rainfed conditions may be well-suited for that crop if irrigation is available. Unfortunately, time-varying irrigation infrastructure information is not available. Time-varying irrigation maps would enable researchers to determine where and when irrigation infrastructure was available to meet crop water demands. In the absence of such information, high quality rainfed pixels were assigned a more likely probability cluster than its irrigated counterpart, since it was not known if access to irrigation is available. In this scheme, moderate-very high suitability rainfed pixels were prioritized over comparable irrigation pixels. 'Marginal', 'very marginal', and 'not suitable' lands were treated as being equal regardless of water supply.

Figure 2 displays how probability clusters vary across crops within a country. In general, irrigation improves suitability as shown by an expansion of areas containing lower SI values. For example, the map of SI values for rainfed maize (see figure 2(A)) fairly closely resembles the cornbelt in the United States. Now, if irrigation was provided to maize, then many more locations throughout the US would be highly suitable for maize (figure 2(B)). The probability cluster assignment assumes that crops will be preferentially grown using rainfall (rather than irrigation) when the soil and climate characteristics are conducive to it. The right column in figure 2 presents an example of the resulting probability clusters for the four major crops in the United States. As figure 2(C) illustrates, the probability clusters for maize in the US will preferentially allocate maize to pixels that are most suitable under rainfed conditions (i.e. pixels with value of '1'), with less likelihood of allocation to pixels that are suitable only when irrigation is available (i.e. pixels with value of '3').

HYDE provides space and time-varying pixel-level information on rice harvest areas. However, the total harvest area as reported by HYDE is an order of magnitude less than what is reported by FAO during the study period (see SI). Thus, the HYDE rice information was used as a baseline to identify 'active' rice pixels. Then a nearest neighbor approximation is applied to identify the three nearest neighbors to these known active rice pixels. These pixels are also labeled as active rice pixels to accommodate the excess rice area in the

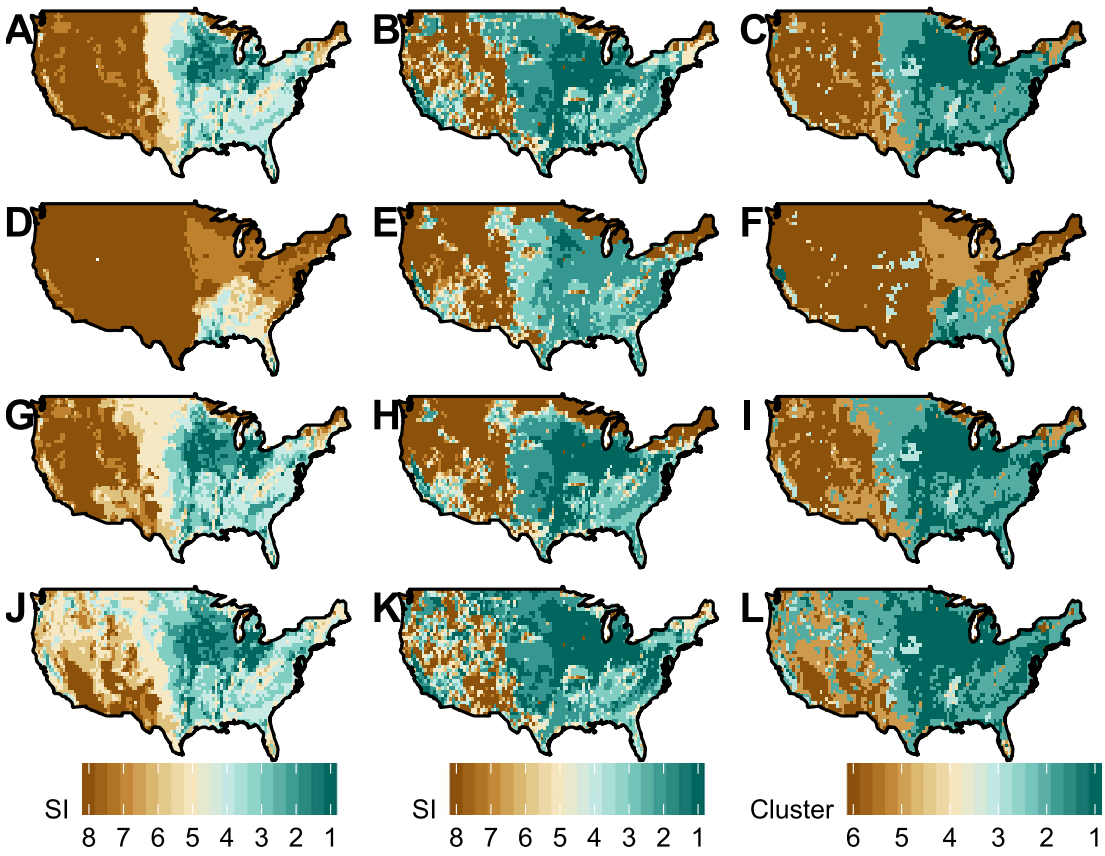

Figure 2. Example of the clustering approach employed in the Probabilistic Cropland Allocation Model (PCAM) for the United States. The left column presents suitability indices for rainfed crops from GAEZ. The middle column presents suitability indices for irrigated crops from GAEZ. The right column presents the probability clusters used in PCAM. The top row displays maize, the second row displays soy, the third row displays rice, and the bottom row displays wheat.

FAO database. For rice, the pixels that have been identified as active either from HYDE or the nearest neighbor identification are automatically placed in the first cluster (value of '1').

Step 2: National statistics analysis. The production [tonnes] and harvested area [hectares] for each country-year in the study domain were retrieved from FAO. First, the number of crops produced by each country for each year was obtained. The allocation algorithm begins by ranking the production information for all crops in a country-year in descending order. If a country has rice production, their rankings were adjusted to have rice be the first allocated crop. The remainder of the crops are then allocated in order of their production ranking. In other words, the algorithm is initialized with the known rice spatial locations and then sequentially moves through the rest of the crops in decreasing order of production. For example, maize is the most produced crop in the United States (in tonnage). However, the United States also produces rice; it is the seventh most produced crop in the year 2000. Rice would be allocated first, and then maize. This approach ensures that known gridded information is incorporated while still ensuring that the crops that are produced in the greatest quantity are prioritized for allocation.

The rice harvest area was adjusted if HYDE information is available. Rice cropland is aggregated for each country and compared to the data provided by FAO. If FAO indicates more rice harvest areas then HYDE at the country-level, the HYDE-based rice area was deducted from FAO. This net area becomes the rice harvest area constraint as outlined in Step 3.

Step 3: Harvest area assignment. National harvested area data is used as an adding-up constraint in the algorithm. Pixels are assigned harvested area until the total harvested area allocated reaches the national value reported by FAO or the HYDE adjusted value in the case of rice. The pixel-level cropland fractions provided by HYDE form a spatial distribution of generic crop-related activities across both space and time. This distribution was used as the random pixel selection probabilities during harvest area assignment. Thus, pixels are probabilistically selected until the following harvested area constraint is met:

$$\sum_{p=1}^{P} \text{Area}_{p,c,y} \approx \text{Harvested Area}_{c,y,i}, \qquad (1)$$

where p is pixel, c is crop, y is year, and i is country. Thus, the harvest area assigned to all pixels in a country-year should approximately sum to the FAO national harvested area statistic. An approximation is

used because the random accumulation of pixel areas will not necessarily be perfectly equal to the national harvested area.

The PCAM algorithm was constructed such that it can be consistently run for all countries in the entire time domain. There are countries in which the number of crops produced exceeds the number of pixels available to be allocated to. In these instances, each pixel was divided into equivalent area slices. These slices are determined by a time series of maximum crops produced by countries where the number of crops exceeds the number of country-labeled pixels. Each annual value is used to determine the number of slices a pixel is divided into for PCAM (see figure S2(a)). For example, in the year 2000, the maximum number of crops (from the PCAM crop list) produced across all countries is 11. Therefore, and regardless of country, each pixel will be divided into 11 slices, which provides a pixel 11 opportunities to be randomly selected. The slices available are adjusted if a pixel is identified by HYDE for rice. The area attributed to rice is converted into slices. These slices are then deducted from the total slices available. For example, a known rice pixel in the year 2000 may have a rice area that is equivalent to two slices. Therefore, it has a net of 9 slices available for random selection across all crops.

Two steps are taken when a pixel is randomly selected. First, the area of the pixel's slice is added to the accumulated harvest area (as in equation (1)). Second, the pixel's selection counter is reduced by one across all crops. However, the probability of selection was not altered as the selection counter is adjusted. When the pixel's selection counter equals zero, it is removed from the potential pool of pixels across all crops. Note that the slice area is doubled if the selected pixel has been designated as multi-cropped for a specific crop. With this approach, it was assumed that each pixel with multi-cropping capabilities will multi-crop in a given year. For this reason, the area in those pixels was double counted (since this is what FAO does).

In the event that the national adding-up constraint has not been satisfied and all clustered pixels have been allocated, pixels with a HYDE cropland fraction equal to zero are then considered. These pixels are clustered based on their original cluster assignment. For example, a pixel with zero cropland but an original cluster assignment of '1' would be placed in cluster 7. Since the selection probability is equal to the cropland fraction, these extra pixels are assigned a probability equivalent to the ratio of pixels in the cluster eligible for allocation to the total number of outstanding pixels. Thus, the probability for these zero cropland pixels is dynamically calculated each time a pixel is selected.

Step 4: Monte Carlo downscaling. Step 3 was run many times in a Monte Carlo framework. Since crops were probabilistically assigned to pixels, it is important to perform this operation multiple times in order to determine the most likely outcome across many trials. For this reason, the model was run 500 times and keep track of the allocation in each model run. A counter is used to track the number of times a pixel is selected for a given crop across all cycles. The number of crop-specific probabilistic selections was converted into an estimate of the crop-specific harvested area with the following equation:

$$f_{p,c,i,y} = \frac{N_{p,c,i,y}}{N_{total} N_{slice,y}} \cdot \text{Area}_{p,c,y}, \quad (2)$$

where f is fraction, p is pixel, c is crop, i is country, and y is year. $N_{p,c,i,y}$ is the number of crop-specific selections for a pixel in a country for a given year across all cycles. N_{total} is the number of total cycles run. $N_{slice,y}$ is the number of slices the pixel is divided into for a given year. $\text{Area}_{p,c,y}$ is the pixel area.

Step 5: Apply the framework to all countries. The core of the algorithm (Steps 2–4) is repeated for all country-years in the study domain. Upon completion, individual country results are bound together for each year to construct annual global PCAM results. The annual area is checked to determine if it is balanced with FAOSTAT data for each country in a given year. In the event that the areas do not balance, then the results were scaled so that each crop's total harvest area matches with the FAO (see SI for scaling factors used).

2.3. Model assumptions

The statistical downscaling algorithm makes several key assumptions. First, the algorithm relies on the national census data of FAO. Errors in this database would significantly alter the results. This is especially problematic for countries which may have political incentives to under or over-report their agricultural areas. For example, Seto *et al* [48] discuss China's propensity to over-report agricultural areas for political motives. Similarly, national standards of agricultural data collection remain poor in many regions, such as in sub-Saharan Africa [49]. Second, errors in the gridded data layers would carry over to the dataset. Third, a probabilistic downscaling approach was employed. The approach assumes that mean climate and soil characteristics are the most important factors in determining where specific crops are grown. This neglects several other important factors (e.g. extreme climate, agricultural labor, policies and regulations, etc). The probabilistic approach means that estimates of *likely* crop-specific areas for each year in the study period were produced. However, PCAM maps should not be misinterpreted as providing actual information on where each crop was grown in each year.

The GAEZ data on the suitability of each pixel was used for specific crops. This is a key gridded data layer that, unfortunately, does not vary in time. Ideally, the SI classes provided by GAEZ would vary in time. However, these SI values estimate the crop-specific suitability of each pixel based on average climate and soil conditions, with 1961–1990 as the baseline. It is likely that these underlying mean climate and soil

characteristics remain relatively stable over the study period. However, the suitability of some locations for specific crops has changed in this time domain; northern China has become increasingly suitable for maize cultivation since the 1980s [50]. This would mean that the approach which is based on these time-invariant SI classes is appropriate for characterizing mean suitability and allocation. However, these SI classes will become increasingly problematic to use as the climate diverges from the 1961–1990 base period for which the SI classes were developed.

Certain crops are more likely to be grown in rainfed and irrigated conditions. Unfortunately, the available data does not enable a determination of which crops are grown in rainfed versus irrigated conditions over time. Modeled crop-specific irrigated areas circa 2000 are available [51], but not at the annual temporal domain necessary for this study. Additionally, crop-specific pixels from Monfreda *et al* [23] were used to force the irrigation model of Portmann *et al* [51]. The agricultural maps that underpin the irrigation spatial model are fixed in time and thus inconsistent with the PCAM approach. For this reason, the circa 2000 irrigation maps were not used, but instead assume that crops are preferentially assigned to high-quality rainfed locations. However, it is likely that certain crops, such as high-value crops, will be preferentially grown where more certain irrigation water supplies are available. It was anticipated that this is particularly likely to be the case for high-value crops (e.g. produce, tree and vine crops, etc), and should not be as problematic for the major staples considered in this study, with the exception of rice [51].

2.4. Model assessment

Quantitative methods were used to assess PCAM performance against existing agricultural maps. The four types of metrics used are: (1) R^2; (2) mean absolute error (MAE); (3) the Jaccard coefficient (J); and (4) the simple matching coefficient (SMC). The first two metrics consider the 'intensity' of the pixels. These two metrics were considered to be the most strict in assessing model performance. The R^2 value is from a Type II linear regression of the estimated crop-specific pixel values of harvest area fractions on existing maps. The R^2 regression also includes pixels where the harvest fraction is zero. J and SMC are focused on estimating spatial similarity in terms of the presence and absence pixel-level harvest areas.

The Jaccard (J) [52] and SMC [53] are defined according to equations (3)–(4), respectively. J gives an approximation of how similar PCAM results are for identifying crop-specific areas. Similarly, SMC allows consideration of both the presence and absence of harvest areas. In other words, it is possible to quantify if PCAM is allocating crops to known harvest areas as well as not allocating to known non-harvest areas.

$$J_{crop} = \frac{M_{11}}{M_{01} + M_{10} + M_{11}} \tag{3}$$

$$\mathrm{SMC}_{crop} = \frac{M_{00} + M_{11}}{M_{00} + M_{01} + M_{10} + M_{11}}. \tag{4}$$

The results from PCAM and the values from the comparison datasets were converted to a binary indicator where '1' is assigned if a pixel is non-zero, and '0' elsewhere. For J and SMC, the parameters are identically defined as follows: M_{11} is the total number of pixels where both PCAM and the comparison datasets have values of 1; M_{01} is the total number of pixels where PCAM has a value of 0 but comparison datasets have a value of 1; M_{10} is the total number of pixels where PCAM has a value of 1 but comparison datasets have a value of 0; and the total number of pixels where both PCAM and comparison datasets have values of 0. J and SMC values closer to '1' indicate high spatial pattern matching between PCAM and these other datasets in terms of both the presence and absence of harvest areas.

The comparisons were performed by using identical spatial resolutions across datasets. For the global comparison, the results from Monfreda *et al* [23] were interpolated from its native 5 arc minute spatial resolution to the 0.5-degree resolution of PCAM. On the sub-national scale, PCAM was compared to the United States Department of Agriculture's (USDA) census for the years 1997 and 2012 [54]. County-level boundaries were obtained from the Newberry Library [55] for 1997 and the United States Census Bureau [56] for 2012. These boundaries and their associated Federal Information Processing Standard codes are used to match data from the USDA to its spatial corollary. Then, PCAM results were aggregated from the 0.5-degree resolution to the county-level to match the spatial resolution of the USDA census data. Cassava, groundnut, rapeseed, rye, and yam were excluded from sub-national assessment in the United States due to poor data availability for the assessment years.

2.5. Sensitivity analysis

PCAM heavily relies on the HYDE and GAEZ gridded datasets. In particular, the hierarchical clustering that boosts allocation towards the most suitable pixels is largely dependent on GAEZ. A combination of each crop's rainfed and irrigation values were used with both set for intermediate input levels. However, rainfed suitability is also available from GAEZ for low and high agricultural inputs, while irrigation data is also available for high agricultural inputs. GAEZ defines low agricultural inputs to be management practices aligned with subsistence farming, whereas high agricultural inputs use advanced management practices for complete commercial production [42]. To determine the sensitivity of PCAM estimates to GAEZ inputs, various combinations of probability clusters were run based on different GAEZ input scenarios. Specifically, the impact of changing rainfed input levels on PCAM results at the global scale for the

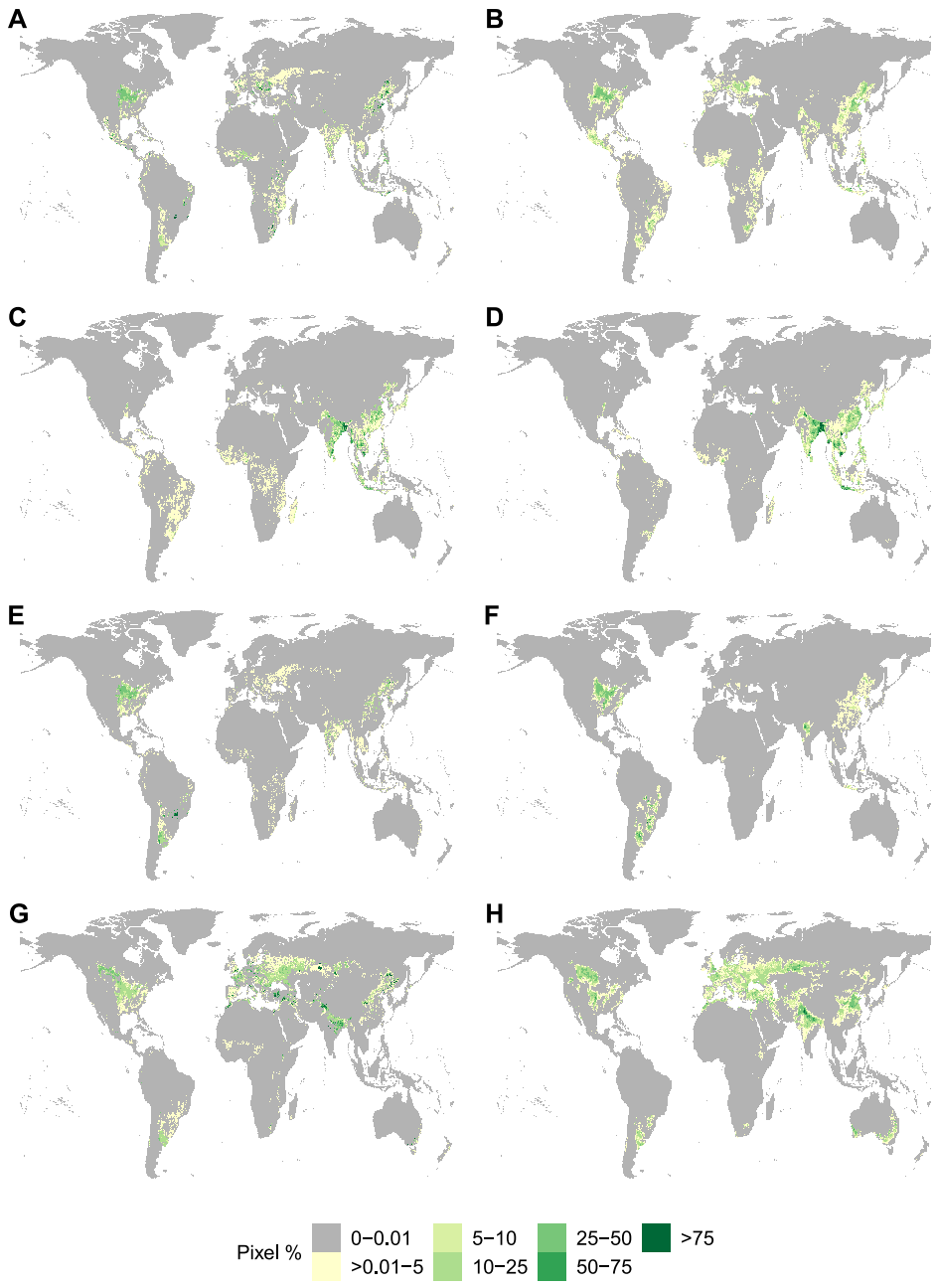

Figure 3. Comparison between PCAM and Monfreda *et al* [23] maps for major crops for the year 2000. PCAM maps (first column) compare reasonably well with Monfreda *et al* [23] maps (second column). The top row displays maize, second row displays rice, third row displays soybean, and bottom row shows wheat.

year 2000 was considered. The rainfed inputs' variability was the focus since rainfed cropland historically dominates our time domain (see figure S2(b)). The metrics defined in section 2.4 were used to understand how PCAM maps change relative to the Monfreda *et al* [23] dataset.

3. Results and discussion

The PCAM model results are presented here. First, the model results are compared with other crop-specific products at both the global and sub-national scale. Then, global crop-specific agricultural geography over time are assessed. Finally, the variability and uncertainty in crop-specific agricultural areas in time is evaluated.

3.1. Comparison with other datasets

3.1.1. Global scale

The comparisons between Monfreda *et al* [23] and PCAM for maize, soy, rice, and wheat are mapped in figure 3. The remaining crops are presented in figures S5–S7. From figure 3 and table 3 it is clear that PCAM captures similar global spatial trends for specific crops as Monfreda *et al* [23]. Additionally, PCAM approximates well the areas with and without crops as shown by the high SMC values across crops. So, PCAM

Table 3. Global comparison between PCAM and Monfreda *et al* [23] for the year 2000. Note that PCAM estimated areas match FAO data by design. R^2, MAE, *J* and SMC metrics are presented to quantify the similarity between PCAM and Monfreda *et al* [23]. Area (%) displays the percent difference in area between PCAM and Monfreda *et al* [23]. For example, PCAM has 12.7% more millet harvested area globally than does Monfreda *et al* [23], due to the underlying FAO data. Differences between FAO and Monfreda *et al* [23] harvested area suggest that PCAM results will never perfectly match Monfreda *et al* [23].

	Harvest area (10^6 hectares)							
Crop	FAO	PCAM	Monfreda *et al* [23]	Area (%)	R^2	MAE (%)	*J*	SMC
Barley	54.4	54.4	54	0.7	0.13	0.504	0.344	0.907
Cassava	16.9	16.9	15	12.7	0.044	0.131	0.16	0.95
Groundnut	23.2	23.2	22	5.5	0.043	0.19	0.15	0.94
Maize	136.8	136.8	136	0.6	0.177	1.138	0.274	0.869
Millet	37.1	37.1	33	12.4	0.144	0.285	0.153	0.936
Oats	12.6	12.6	13	−3.1	0.14	0.11	0.275	0.931
Potato	20.1	20.1	19	5.8	0.083	0.193	0.174	0.909
Rapeseed	25.8	25.8	24	7.5	0.092	0.263	0.139	0.919
Rice	128	128	154	−16.9	0.632	0.64	0.445	0.93
Rye	9.8	9.8	9	8.9	0.177	0.1	0.174	0.939
sorghum	41.2	41.2	39	5.6	0.091	0.336	0.159	0.91
Soybean	74.4	74.4	75	−0.8	0.191	0.583	0.212	0.913
Sugarbeet	6	6	6	0	0.038	0.065	0.13	0.956
Sunflower	21.2	21.2	21	1	0.152	0.184	0.182	0.923
Sweet potato	9.7	9.7	9	7.8	0.105	0.086	0.033	0.956
Wheat	215.1	215.1	209	2.9	0.094	2.049	0.373	0.869
Yam	3.9	3.9	4	−2.5	0.018	0.031	0.148	0.987

provides additional validation of the GAEZ-based approach to allocation, since PCAM does a good job at replicating sub-national patterns that are not incorporated in it. There are moderate differences in the global area of each crop between PCAM and Monfreda *et al* [23], driven by the underlying FAO data which serves as a constraint in PCAM (see table 3). For example, PCAM has 12.4% more millet harvested area globally than does Monfreda *et al* [23].

Differences between FAO and Monfreda *et al* [23] harvested area mean that PCAM results should not be expected to perfectly match Monfreda *et al* [23]. This is because PCAM total agricultural area is equivalent to FAO harvested area, by design. This is confirmed in tables 3 and 4, where the FAO and PCAM values are the same. Figure S2(c) shows a time series of Goldewijk [31], FAO [4], and PCAM area. Note that Goldewijk [31] reports much less cropland area than does FAO [4]. PCAM areas have been calibrated to match FAO [4] which is why PCAM and FAO values align in tables 3 and 4. Surprisingly, Monfreda *et al* [23] and FAO [4] do not contain the same quantity of harvested area.

Table 4 provides country-level analysis for the top 10 countries by harvest area for the four major crops. The starting national area from FAO deviates from the total area in Monfreda *et al* [23], as at the global scale. In some countries, the performance of PCAM is not as good as at the global scale. For example, the global SMC for maize is 0.87 (see table 3), but 9 of the top 10 countries do not approach a value this high at the national scale (see table 4). The United States has the highest SMC for maize with a value of 0.85. However, PCAM demonstrates higher R^2 values at the country scale than it does globally. For example, the global R^2 value for maize is 0.18 versus 0.62 and 0.46 for the United States and Argentina, respectively.

Harvest area discrepancies between FAO and Monfreda *et al* [23] exist across a supermajority of country-crop pairs in a non-uniform manner. Thirteen of the 17 crops analyzed showed more reported area by FAO than by Monfreda *et al* [23] at the global scale. However, when country-level harvest areas are considered, larger area discrepancies were observed. For example, FAO/PCAM has 33.3% more maize harvested area in South Africa than does Monfreda *et al* [23]. These differences in national harvested area contribute to the differences between PCAM and Monfreda *et al* [23] (see figure 3).

3.1.2. National scale

The United States has agricultural census data available at the sub-national scale for several points in time from the USDA. The harvest fraction estimates provided by equation (2) are compared to county-level information provided by the USDA. Cassava and yams were not compared as they are not grown in the United States. Census information was obtained for 1997 and 2012 for 12 of our 17 assessed crops [54]. These years were chosen as 1997 is the first readily available electronic census with county-level information and 2012 is the most recent census available in the time domain.

Table 5 compares PCAM with USDA data. PCAM provides the best pixel-level intensity matching for maize and rice as demonstrated by their high R^2

Table 4. Comparison of national harvest areas from FAO, PCAM, and Monfreda *et al* [23] for the year 2000 for the top 10 countries by harvest area. R^2, MAE, *J*, and SMC metrics are presented to quantify the similarity between PCAM and Monfreda *et al* [23]. Area (%) displays the percent difference in area between PCAM and Monfreda *et al* [23]. For example, PCAM has 19.8% less rice harvested area in India than does Monfreda *et al* [23].

		Harvest area (10^6hectares)							
Country	Global fraction	FAO	PCAM	Monfreda *et al* [23]	Area (%)	R^2	MAE	*J*	SMC
A. Maize									
USA	0.214	29.3	29.3	30	−2.3	0.617	1.979	0.441 8	0.8541
China	0.168	23.1	23.1	25	−7.6	0.115	3.542	0.390 4	0.78
Brazil	0.085	11.6	11.6	11	5.5	0.001	2.551	0.087 2	0.7588
Mexico	0.052	7.1	7.1	8	−11.3	0.024	5.769	0.352 2	0.6074
India	0.048	6.6	6.6	6	10	0.006	3.053	0.287 3	0.5466
South Africa	0.029	4	4	3	33.3	0.028	5.163	0.258 7	0.7618
Indonesia	0.026	3.5	3.5	3	16.7	0.017	2.698	0.254 1	0.7754
Argentina	0.023	3.1	3.1	3	3.3	0.462	0.831	0.469 5	0.8516
Nigeria	0.023	3.2	3.2	4	−20	0.015	5.593	0.270 1	0.3399
Romania	0.022	3	3	3	0	0.273	12.499	0.436 2	0.5268
B. Rice									
India	0.288	36.9	36.9	46	−19.8	0.463	8.801	0.753 4	0.819
China	0.197	25.3	25.3	31	−18.4	0.489	2.214	0.720 1	0.9143
Indonesia	0.078	10	10	11	−9.1	0.491	4.52	0.666 7	0.8003
Bangladesh	0.073	9.4	9.4	10	−6	0.401	24.196	0.936 2	0.9362
Thailand	0.065	8.4	8.4	10	−16	0.632	9.785	0.786 6	0.7941
Viet Nam	0.051	6.5	6.5	7	−7.1	0.644	13.674	0.909 9	0.9099
Myanmar	0.042	5.4	5.4	6	−10	0.635	6.246	0.560 6	0.6234
Philippines	0.027	3.5	3.5	4	−12.5	0.413	9.98	0.914 6	0.9157
Brazil	0.022	2.8	2.8	3	−6.7	0.188	0.4	0.181 1	0.7322
Nigeria	0.013	1.7	1.7	2	−15	0.029	2.687	0.370 2	0.5677
C. Soybean									
USA	0.394	29.3	29.3	30	−2.3	0.556	2.005	0.466 6	0.8612
Brazil	0.183	13.6	13.6	14	−2.9	0	3.214	0.057 9	0.8038
China	0.125	9.3	9.3	9	3.3	0.086	1.422	0.263 2	0.7534
Argentina	0.116	8.6	8.6	9	−4.4	0.352	2.65	0.569	0.8651
India	0.086	6.4	6.4	6	6.7	0	3.307	0.164	0.5842
Paraguay	0.016	1.2	1.2	1	20	0.004	5.879	0	0.6338
Canada	0.014	1.1	1.1	1	10	0.237	0.085	0.488 9	0.993
Indonesia	0.011	0.8	0.8	1	−20	0	0.738	0.075	0.8153
Bolivia	0.008	0.6	0.6	1	−40	0.026	0.92	0.219	0.776
Nigeria	0.007	0.5	0.5	1	−50	0	0.835	0.136 8	0.7294
D. Wheat									
India	0.128	27.5	27.5	26	5.8	0.029	11.938	0.34	0.5878
China	0.124	26.7	26.7	27	−1.1	0.001	5.115	0.281 1	0.7207
USA	0.1	21.5	21.5	22	−2.3	0.109	2.889	0.289 8	0.7338
Russian Federation	0.099	21.3	21.3	20	6.5	0.397	0.886	0.483 9	0.9195
Australia	0.056	12.1	12.1	12	0.8	0.003	3.274	0.041 7	0.8416
Canada	0.05	10.9	10.9	11	−0.9	0.599	0.608	0.560 6	0.969
Kazakhstan	0.047	10.1	10.1	9	12.2	0.129	6.272	0.536 5	0.8497
Turkey	0.044	9.4	9.4	9	4.4	0.137	16.404	0.233 6	0.2719
Pakistan	0.039	8.5	8.5	9	−5.6	0.191	12.721	0.375	0.6451
Argentina	0.029	6.2	6.2	6	3.3	0.358	1.902	0.5	0.8462

values. These results are unsurprising. PCAM's allocation of rice is bolstered by pre-identification of active areas from HYDE. The rice allocation in the US for both years exceed the global rice performance in the year 2000 (SMC = 0.97 in 1997 and 2012 versus 0.93 globally in the year 2000). Maize, which is the second crop to be allocated in the United States, is often grown in areas where rice is not. Therefore, there is minimal competition to allocate maize. The SMC values between PCAM and USDA indicate that the model is capturing the geography of specific crops reasonably well. Again, SMC values are not as good as they are at the global scale, but all crops across both years have SMC values above 0.5. Additional high performing crops across both years are millet, soybeans, sugarbeets, and sweet potatoes. These spatial statistics

Table 5. Comparison between PCAM and USDA county harvest area for the years 1997 and 2012.

	Harvest area (10^4 hectares)							
Crop	FAO	PCAM	USDA	Area (%)	R^2	MAE (%)	J	SMC
A. 1997								
Barley	250.8	250.8	244.7	2.5	0	0.898	0.2594	0.5673
Maize	2940.9	2940.9	2872.2	2.4	0.536	7.248	0.3839	0.5479
Millet	16.3	16.3	13.7	19.1	0	0.063	0.0202	0.7591
Oats	113.8	113.8	108.3	5.1	0.081	0.312	0.4604	0.6244
Potato	54.4	54.4	50.2	8.4	0.002	0.234	0.1196	0.6453
Rice	125.6	125.6	127	−1.1	0.545	0.243	0.3796	0.9667
Sorghum	370.6	370.6	345.7	7.2	0.014	1.178	0.3036	0.6319
Soybean	2796.7	2796.7	2739.2	2.1	0.492	6.427	0.5216	0.7134
Sugarbeet	57.8	57.8	55.3	4.5	0	0.257	0.0233	0.7705
Sunflower	113	113	101.3	11.5	0.002	0.418	0.1012	0.5941
Sweet potato	3.3	3.3	2.9	13.8	0.004	0.0121	0.0435	0.9344
Wheat	2541.4	2541.4	2508.5	1.3	0.048	7.667	0.473	0.5787
B. 2012								
Barley	131.3	131.3	130.6	0.5	0	0.483	0.2102	0.5688
Maize	3535.9	3535.9	3533.2	0.1	0.54	8.840	0.3767	0.5315
Millet	8.3	8.3	8.1	2.6	0.001	0.0323	0.0026	0.8107
Oats	42.3	42.3	41.8	1.2	0.03	0.132	0.4234	0.6672
Potato	45.8	45.8	34.7	31.9	0.001	0.194	0.2424	0.5559
Rice	108.4	108.4	107.8	0.6	0.624	0.179	0.3462	0.9662
Sorghum	200.5	200.5	205.3	−2.3	0.002	0.697	0.2595	0.6145
Soybean	3081.5	3081.5	3076.8	0.2	0.458	7.511	0.51	0.6965
Sugarbeet	48.7	48.7	48.8	0	0	0.218	0.0286	0.7804
Sunflower	74.5	74.5	73.8	0.9	0.006	0.282	0.0819	0.5822
Sweet potato	5.1	5.1	3.9	32.4	0	0.0195	0.0345	0.9165
Wheat	1979.8	1979.8	1981	−0.1	0.022	6.174	0.4592	0.5688

indicate a reasonable performance, particularly given that FAO and USDA national data do not match one another (see table 5), which means the PCAM algorithm is being run with a different national constraint.

Maps of changes in harvest area for maize, soybeans, rice, and wheat from 1997 to 2012 are presented in figure 4 for the United States. PCAM spatial trends are disappointing and typically estimate gains in harvest area that the USDA data does not support. This disparity highlights a potential shortcoming of the approach, which is that PCAM allocated to suitable pixels based on mean climate and soil characteristics. This means that PCAM may be capturing crop *planted* area spatial trends better than *harvested* area. This is because the planted area decision will largely be driven by mean suitability, but extreme weather may impact the eventual locations of crop harvest. This is what would have happened in the year 2012 in the US, which was an extreme drought year in the US corn and soy belt [57]. This extreme event will be captured by the USDA 2012 census data on harvested area. However, the drought would not show up in the stationary pixel-scale suitability driving PCAM spatial allocation. The national harvested area statistic should capture this information, yet FAO and USDA data diverge for the US for this year.

Note that PCAM tends to underestimate declining areas compared to USDA data when considering whole country pixel trends. For example, PCAM projects that 2.1% of maize-related pixels will decline in intensity versus 14.6% based on USDA. Similarly, PCAM estimates that 5.4% soy-related pixels will decline in intensity versus 10.5% for USDA. There is also an underperformance of wheat in estimating the decline of wheat-related pixel (19.4% for PCAM versus 33.6% from USDA). However, PCAM is comparable to USDA when examining areas where pixels increase in intensity. For rice, PCAM and USDA are extremely close approximations (0.34% versus 0.67%). Thus, PCAM is able to closely capture the loss of rice harvest area in the Mississippi Embayment region. For soybeans, 13.8% of pixels are projected to increase according to PCAM compared to the 17.0% based on USDA.

Figure 4 highlights that area harvested in the United States are driven by factors other than mean climate and soil. Many factors are likely important when determining where to grow crops, particularly in countries with advanced agricultural systems, such as the United States. However, only gridded information on crop-specific suitability driven by mean climate and soil was used. Other factors—such as labor, access to machinery, agricultural policies [58]—influence where specific crops are grown as much, or even more than, the physical variables that were used to guide the downscaling. Thus, divergences between PCAM estimates and census

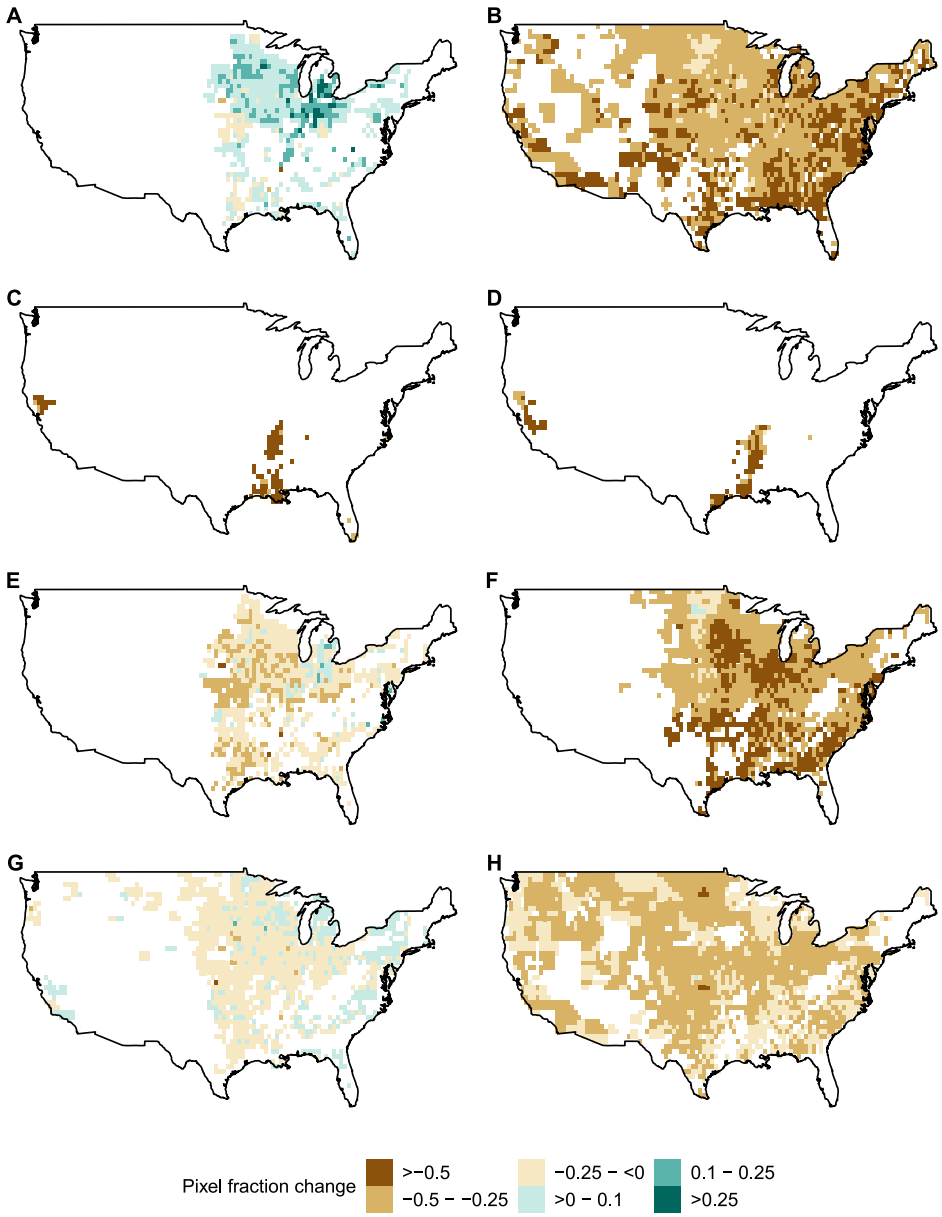

Figure 4. Comparison between PCAM and USDA changes in harvest area fraction in the continental United States. Changes from 1997 to 2012 are mapped. The first column shows PCAM maps and the second column shows USDA NASS [55] maps. The top row displays maize, second row rice, third row soybeans, and the last row is wheat. White areas show where PCAM or USDA indicate that no crop is grown.

information, such as those provide in figure 4 may actually prove useful in future research that aims to identify other important factors guiding agricultural geography. Determining these factors is unfortunately beyond the scope of this project, but future work may aim to analyze these deviations and compare them with maps of other potentially important factors of agricultural production.

3.2. Crop-specific changes in time and space

First, changes in total cropland over time are examined. Figure 5 presents PCAM estimates of total cropland for the beginning, middle, and end of our study period for the 17 crops considered. The common scale across sub-figures makes it clear that both extensification and intensification of crop area have occurred over time. The intensities in total crop area increase significantly in 2010, driven by reported gains in crop area from FAO (see figure S2(c)). In 1970, 28.2% of the global land area was estimated to be cultivating at least one of the 17 crops with a mean harvest fraction of 0.09. By 2010, PCAM estimates a slight expansions of the global area that is harvested to 30.2%, but the mean cropland fraction has increased to 0.11. The observed trends in both modest intensification and extensification is driven largely by differences in crop area reported by FAO and HYDE (as shown in figure S2(c)).

Table 6 summarizes the PCAM area estimates at the beginning and end of the study period by crop (see

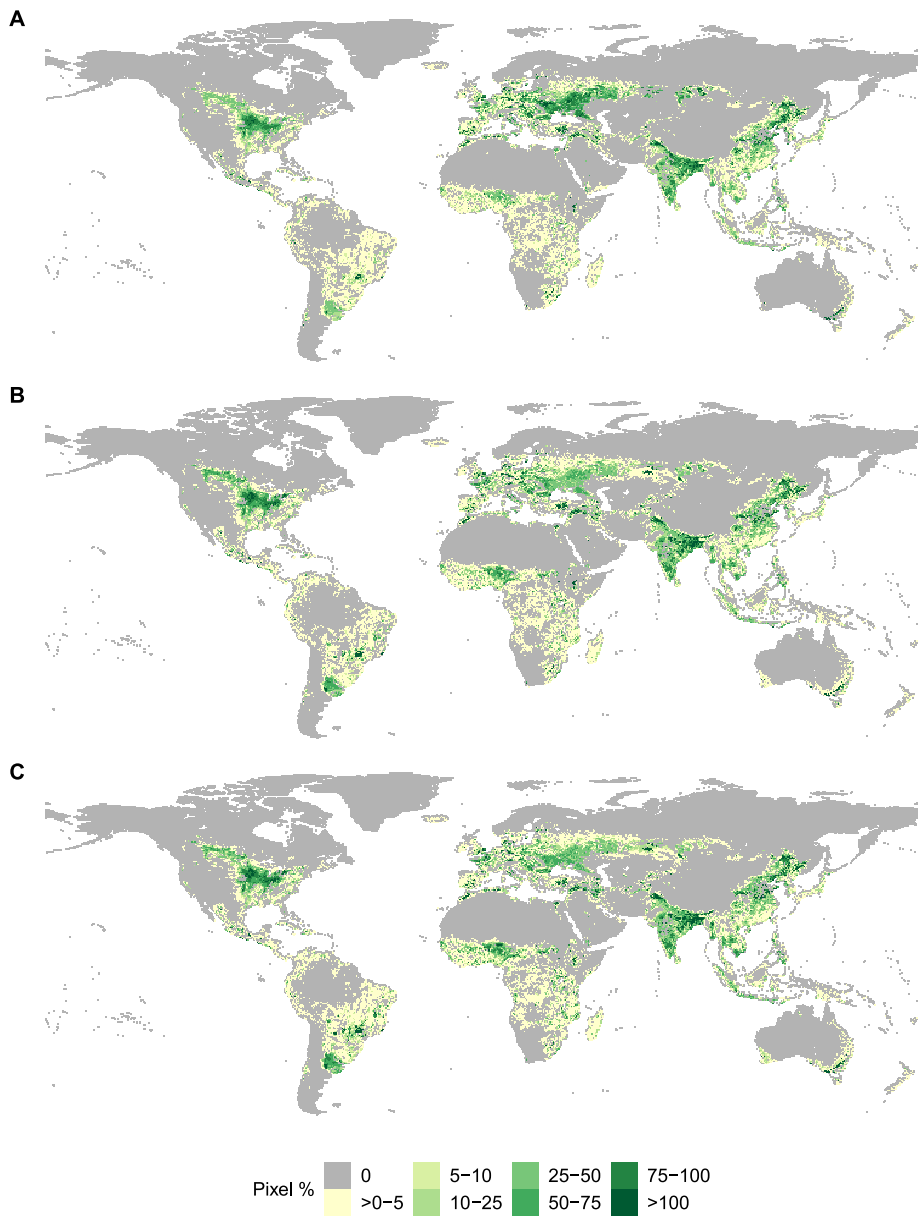

Figure 5. Maps of total cropland area over time. The global distribution of the 17 crops considered in this study is shown for (A) 1970, (B) 2000, and (C) 2010. Note that these maps highlight the occurrence of both intensification and extensification over time.

figure S11 for the broader time series). To further understand these changes, two specific sub-domains are considered to understand how crops have changed in both time and space: 1970 versus 2000, and 2000 versus 2010. The year 2000 was used as the 'benchmark' year, since global comparison data was available for this year (i.e. Monfreda *et al* [23] as discussed above). PCAM's output of pixel-level harvest fractions was treated as proxies for the likelihood of a crop being present in a given pixel. By extension, PCAM enables estimates of projected changes of a crop's likelihood over time to be determined as the relative change in these harvest fractions. As a result, it is possible to further identify two extremes of crop-specific pixel behavior. The first extreme is if a pixel converts from having a non-zero to a zero likelihood. These pixels are then considered inactive. The other extreme is where a pixel is estimated to increase its likelihood by several orders of magnitude. This latter case is a surrogate of pixel-level crop intensification. These extreme pixels are referred to as 'super pixels'.

Figure 6 shows changes in time for each of the four major crops across both domains (see figures S12–S14 for the rest of the crops). Table 7 quantifies these spatial changes by presenting the fraction of pixels that have demonstrated crop-related increases and decreases over each sub-domain. A majority of pixels show no change over either period. In other words, these pixels have likely remained identically zero over time. Between 1970 and 2000, a majority of the crops studied (nine of 17) have more pixels that increased their likelihood than decreased. The remaining crops show a greater percentage of pixels that declined. Similarly, for the second period, 10 crops were estimated to have

Table 6. Global cropland area changes in time. The global area is provided by crop for the start (1961) and end (2014) of the study time period as estimated by PCAM. The percent change in area over this time period is also shown. The percent of total cropland is shown for each crop for the start and end of the study time period. Note that soybean increases from 3.2% to 12% of all harvested area during this period.

	Area (10^6 hectares)			Crop contribution (%)	
Crop	1961	2014	Area change (%)	1961	2014
Barley	54.5	49.4	−9.3	7.3	5.1
Cassava	9.6	23.9	148.1	1.3	2.4
Groundnut	16.6	26.5	59.8	2.2	2.7
Maize	105.5	184.7	75.1	14.2	18.9
Millet	43.4	31.3	−27.7	5.8	3.2
Oats	38.3	9.6	−74.9	5.1	1
Potato	22.1	19.1	−13.8	3	2
Rapeseed	6.3	36.1	475.4	0.8	3.7
Rice	115.3	162.6	41	15.5	16.6
Rye	30.3	5.3	−82.5	4.1	0.5
Sorghum	46	44.9	−2.3	6.2	4.6
Soybean	23.8	117.5	393.5	3.2	12
Sugarbeet	6.9	4.5	−35.4	0.9	0.5
Sunflower	6.7	25.2	278	0.9	2.6
Sweet potato	13.4	8.3	−37.5	1.8	0.9
Wheat	204.2	220.4	7.9	27.4	22.6
Yam	1.1	7.7	579.7	0.2	0.8
Total	743	977			

a net gain in active pixels while the remaining seven will have a net decline. Sweet potatoes is the additional crop demonstrating these gains in pixel activation.

The number of pixels activated for maize is estimated to increase by 6.95% by the year 2000. This increase is being driven by the 35% of pixels who significantly increase their likelihood. China (21%), Mozambique (8.4%), India (8.1%) are drivers of this increase. In particular, China and Mozambique see their maize likelihoods increase by 54% and 169%, respectively. For the 2000–2010 period, maize is projected to have 53% more pixels that increase their likelihood, rather than decrease. China is the estimated leading driver of this change. Its maize pixels increase their likelihood by 30.5% and it is home to 25% of the super gaining pixels.

For soybeans, there are 53% more pixels that are estimated to have increased than decreased the likelihood with a mean relative change of 283% for 1970–2000. This large increase in likelihood is being driven by 35% of its pixels being super gaining. The United States (22.3%), Brasil (12.5%), and Argentina (9.1%) are where most of this extreme extensification is occurring for 1970–2000. For Argentina, these projections highlight a case of both extensification and intensification as their mean likelihood increased by 3000%. Paraguay also illustrates substantial extensification of soy (i.e. gaining pixels), which indicates that PCAM is able to capture areal expansion of crops. For 2000–2010, Brasil (27.6%), the United States (13.6%), and Russia (7%) lead the global expansion of soybean pixels.

Wheat is another crop that is anticipated to have a continuous pixel-level expansion over both domains. It has 7.65% and 8.6% more pixels that increase rather than decrease in likelihood for 1970–2000 and 2000–2010, respectively. Nearly 50% of the pixels with the largest increases in likelihood for 1970–2000 are found in China, the United States, and Canada. For the 2000–2010 period, Kazakhstan (26.9%), Brasil (13.8%) and China (10.4%) are the leading countries for extreme intensification.

Rice is crop of net expansion across the study period, with increases in expansion slowing during the 2000–2010 period. For this latter period, Tanzania, Brazil and Guinea are projected to contain 24% of the super gaining pixels. India is the leading global producer of rice by harvest area. Their rice pixels were approximated to have a mean likelihood change of −2.0% during 2000–2010, with 32.7% of its pixels declining in likelihood while 20.7% increased their likelihood.

Barley is estimated to have 28.4% more of its pixels decline than increase in likelihood when comparing 1970–2000. Over this period, there is an average 8.11% increase in pixel likelihood globally. Of those estimated to decline, 17% of them are projected to become completely inactive in 2000. The United States contains 20% of these lost pixels, which coincides with United States shifting from the third to eighth largest barley producer. Russia and China lose barley pixels from 1970 to 2000. Russian barley losses continue in 2000–2010 as their average barley pixel has an estimated mean likelihood reduction of −3.50%.

Cassava is projected to have 19.7% pixels increase in likelihood than decrease, while the mean pixel increases by 5.15% from 1970 to 2000. Nearly 25% of cassava pixels are expected to become inactive. Brasil (18.6%), Congo (15.1%), and Indonesia (10.4%) are the leading countries for pixel loss. Brasil sees a modest increase in pixel likelihood of 3.23%, which coincides with a dip in global ranking from sixth to seventh. For 2000–2010, cassava experiences a nominal 2.4% increase in mean likelihood change with 20.1% more pixels estimated to increase their likelihood than decrease. These latter changes are being driven by Brazil (15.3%), Congo (10.1%), and Nigeria (7.0%).

Groundnuts are considered to have 14.98% more pixels gain likelihood than lose by 2000. The global mean change in likelihood improves by 7.48%. Approximately 33% and 31% of pixels experience extreme gain and loss, respectively. India (21.3%), Brasil (8.4%), and the United States (7.5%) are home to most of these losses. Note that Brasil is not a top 10 country by area for groundnuts between 1970 and 2000. However, India contained the most harvest area globally during this time. While India lost pixels, the remaining active pixels are projected to have a mean

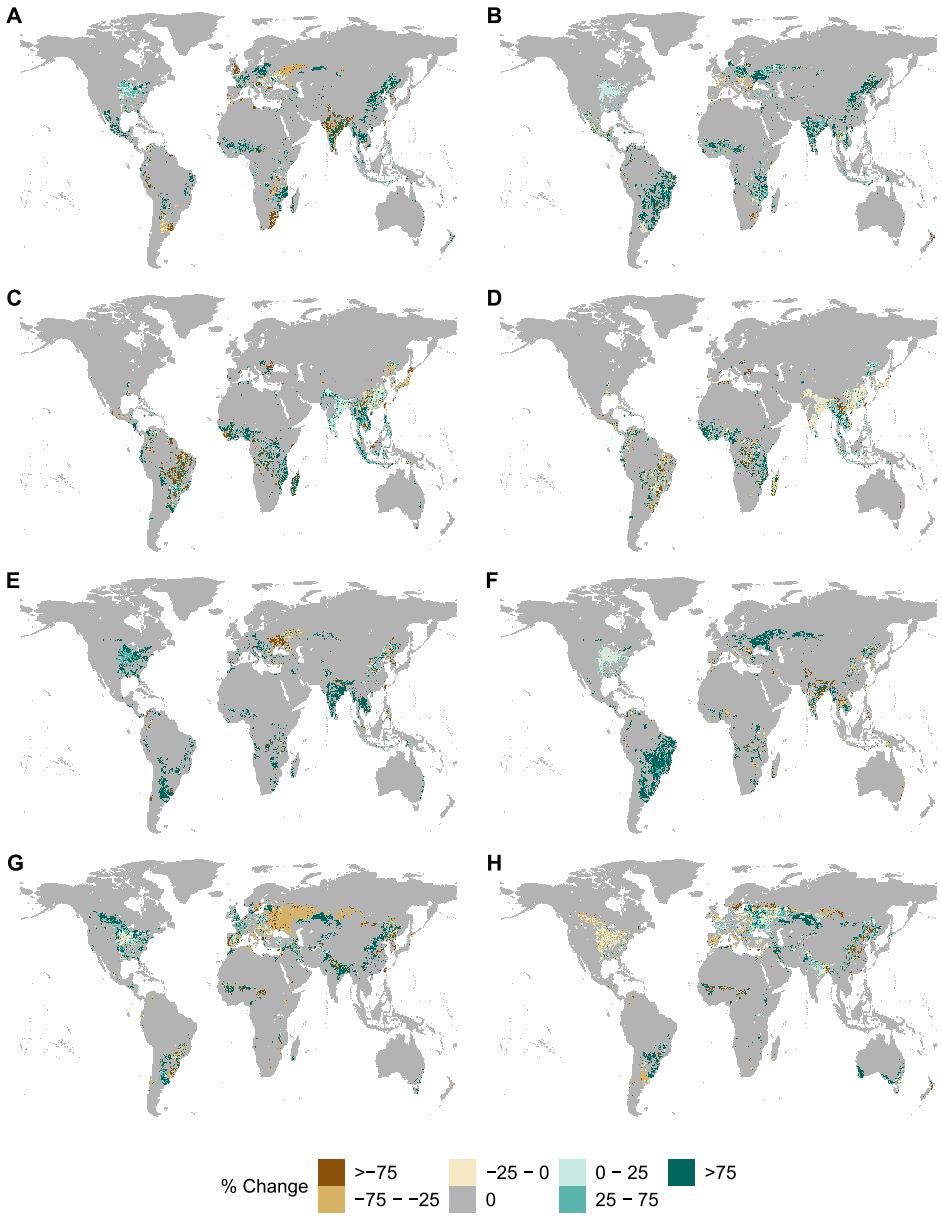

Figure 6. Maps of crop-specific changes over time. The first column shows changes between 1970 and 2000; the second column shows changes between 2000 and 2010. The top row displays maize, second row rice, third row soybeans, and the last row is wheat.

likelihood increase of 59.4%. For the 2000–2010 period, the global likelihood is projected to increase by 3.91% with a difference of 23.7% between pixels that increased versus decreased their likelihood. Groundnut's expansion during this period is lead by India (14.7), Nigeria (9.5%), and Argentina (6.3%). Nigeria sees its groundnut likelihood increase by 34.8%. India is a more complex case as 16.1% of its pixels are projected to become inactive while the 27.7% of its remaining pixels become extreme intensifiers.

For rapeseed, there are 80% more pixels that are estimated to have increased than decreased their likelihood in 1970–2000 with a global mean relative change of 185%. This large increase is being driven by 31.6% of its pixels classified as super gaining in their likelihood. China (28.3%), Canada (16.2%), and India (11.8%) are where most of this extreme extensification is occurring for 1970–2000. For example, the mean percent increase likelihood is 92.4% in China. Russia (21.2%), the United States (33.3%), and China (32.3%) are the leading drivers of expansion and contraction of rapeseed pixels for 2000–2010, respectively.

Sunflowers demonstrate 72.1% more pixels with increases than decreases in likelihood for 1970–2000 with a mean likelihood change of 64.9%. Nearly 38% of its pixels significantly increase their likelihood. The United States (19.2%), Argentina (13.4%) and Ukraine (8.3%) are where most of this extreme extensification is occurring for 1970–2000. Argentina and Ukraine are also projected to have extreme intensification as their likelihoods increase by 51.9% and 91.4%, respectively. For 2000–2010,

Table 7. Estimated changes in pixel activation by crop over two time domains: 1970 versus 2000, and 2000 versus 2010. Pixels without change were identically equal to 0 across these time periods.

	1970–2000			2000–2010		
Crop	% Increase	% Decrease	% No change	% Increase	% Decrease	% No change
Barley	3.95	7.09	88.96	3.92	6.75	89.33
Cassava	2.81	1.88	95.31	2.74	1.82	95.44
Groundnut	3.46	2.56	93.97	3.63	2.24	94.13
Maize	6.95	4.1	88.96	9.06	2.77	88.17
Millet	3.18	4.09	92.74	3.14	3.62	93.24
Oats	2.11	7.97	89.92	2.76	6.18	91.06
Potato	4.81	4.88	90.31	4.19	5.33	90.48
Rapeseed	6.78	0.73	92.49	6.49	2.61	90.91
Rice	7.44	4.23	88.32	6.09	5.06	88.84
Rye	1.45	7.5	91.05	2.25	5.57	92.18
Sorghum	4.47	5.13	90.4	4.27	4.84	90.89
Soybean	6.18	1.89	91.93	7.63	2.68	89.69
Sugarbeet	1.91	3.12	94.97	2.11	2.67	95.22
Sunflower	7.13	1.15	91.72	4.8	3.92	91.27
Sweet potato	2.21	2.25	95.54	2.58	1.75	95.68
Wheat	7.11	6.1	86.79	7.25	6.1	86.65
Yam	1.12	0.5	98.38	1.19	0.61	98.2

Russia is the leading driver of heterogeneous change in the global sunflower distribution as it is home to 23.9% of pixels that alter their sunflower activity level. Conversely, the United States (27.6%) is the leading driver of sunflower pixel loss.

Yams are estimated to have 38% and 32.5% more of its pixels increase than decline in likelihood with changes of mean likelihood increasing by 3.67% and 1.03% for 1970–2000 and 2000–2010, respectively. These are considered to be negligible changes over the course of the study domain. Nigeria is the leading country over the entire domain.

It is important to note that PCAM spatial patterns are likely most representative of planted area. This is because PCAM uses SIs to probabilistically allocate crops to pixels. This is most likely capturing the locations that a crop is suitable to be planted in. However, the algorithm downscaled harvest area information. This means that there is a potential mismatch in spatial locations. These differences are likely to be small in locations and time periods in which crops were both planted and harvested in the most suitable locations. However, in time periods and places that experienced significant divergence—say, due to drought—these spatial estimates may be problematic. Recent work has shown that extreme weather events have reduced cereal production in recent decades, with drought events impacting both harvested area and yield, whereas temperature extrema have only impacted grain yield [22]. As such, this distinction between planted and harvested area estimates is important to keep in mind when using PCAM estimates, particularly in times of drought. Future work could bring PCAM area estimates together with information on climate extremes to better estimate harvested area.

3.3. PCAM sensitivity

The sensitivity of PCAM model output was analyzed using different scenarios from GAEZ. Specifically, the GAEZ inputs used were varied to evaluate the pixel-scale suitability of specific crops. All PCAM results presented thus far are based upon the 'intermediate' input scenario produced by GAEZ. Here, PCAM was run with the 'low' and 'high' input scenarios for rainfed agriculture. Again, PCAM model output was compared with Monfreda *et al* [23]. Both the low and high rainfed inputs produce PCAM outputs that are similar (SMC $>$ 0.8) to the Monfreda *et al* [23] data (see table 8). Note that the SMC of each crop was weighted by its global harvest area fraction for the year 2000 to obtain an overall weighted SMC. The low versus intermediate and high versus intermediate scenarios differ by 0.12% and 1.24%, respectively. Therefore, there is limited impact on spatial similarity as a result of altering agricultural input scenarios.

Each crop's statistical distribution was examined to further elucidate the effect of changing GAEZ inputs on PCAM estimates. Table S4 shows how these properties and the number of contributed pixels fluctuate. Figure S4 presents box plots of harvest fraction for maize, rice, soybeans, and wheat as a function of these GAEZ inputs (see figure S15 for the remainder crops). A suppression in pixel intensity was observed for the high rainfed results due to significant increases in the number of pixels selected. Similarly, the low inputs scenario produces the largest pixel intensities and is generally based on having selected the fewest number of pixels. This makes sense, as more pixels are suitable for selection when many inputs are available in agriculture; fewer pixels are suitable when fewer inputs are used. There is a comparable number of outliers (i.e. excessively high harvest fractions) across the

Table 8. Sensitivity results comparing PCAM to Monfreda *et al* [23] across 'low', 'intermediate', and 'high' agricultural input scenarios provided by GAEZ. 'Weight' is the global harvested area fraction for the year 2000 as given by FAO [4].

		Low		Intermediate		High	
Crop	Weight	R^2	SMC	R^2	SMC	R^2	SMC
Barley	0.06	0.139	0.91	0.13	0.9074	0.206	0.8895
Cassava	0.02	0.05	0.9571	0.044	0.956	0.128	0.9423
Groundnut	0.03	0.037	0.9432	0.043	0.9397	0.038	0.9308
Maize	0.16	0.197	0.8686	0.177	0.8688	0.22	0.8557
Millet	0.04	0.12	0.9363	0.144	0.936	0.167	0.9287
Oats	0.01	0.145	0.9286	0.14	0.931	0.178	0.9197
Potato	0.02	0.061	0.9084	0.083	0.9094	0.126	0.9069
Rapeseed	0.03	0.082	0.9169	0.092	0.9194	0.088	0.9146
Rice	0.18	0.636	0.9301	0.632	0.9304	0.631	0.9121
Rye	0.01	0.134	0.9365	0.177	0.9391	0.253	0.9382
Sorghum	0.05	0.047	0.9091	0.091	0.9103	0.263	0.8986
Soybean	0.09	0.224	0.9079	0.191	0.9134	0.4	0.9121
Sugarbeet	0.01	0.061	0.9475	0.038	0.9555	0.058	0.9476
Sunflower	0.02	0.205	0.9227	0.152	0.9229	0.163	0.9121
Sweet potato	0.01	0.119	0.9525	0.105	0.9557	0.065	0.9444
Wheat	0.25	0.162	0.8669	0.094	0.8688	0.108	0.8594
Yam	0	0.039	0.9862	0.018	0.9873	0.119	0.9855
Weighted value			0.89		0.89		0.88

four dominant crops and inputs scenarios (see figure S4). Overall, the intermediate scenario is nestled between the two inputs extrema. Thus, the intermediate rainfed PCAM outputs was treated as baseline values, whereas the low and high form the upper and lower bounds, respectively.

Many additional sensitivity analyses can be performed by future researchers. Future work can evaluate the 'truthfullness' of national harvested area and run PCAM with multiple national constraints and weight the most truthful statistic. For example, FAO was relied upon for the truth in agricultural statistics. However, national values presented by [23] or USDA may be better to use. Future research could determine the best national statistic for each country-year and then re-run PCAM to obtain better results. Additionally, future work could vary the probability clusters that underpin PCAM for each country-year. This could be done using an optimization algorithm that would determine the probability cluster assignment that most closely represents a 'target' database (e.g. [23] or USDA) and then use this cluster for the algorithm. More sophisticated versions of this would involved varying the probability clusters in time in a way that reflects the underlying suitability of pixels. Other factors that influence spatial allocation could be added, such as infrastructure (e.g. irrigation and/or road network) and proximity to urban centers. These extensions would build upon and improve the results presented here.

4. Conclusions

The PCAM model was introduced to downscale national agricultural census information to a geographic grid. This model enabled the creation of estimated gridded areas of individual crops at the annual time scale. Remarkably, this probabilistic approach performs reasonably well against a global dataset assembled with sub-national agricultural census information. PCAM global gridded maps of specific crops for each year from 1961 to 2014 are provided in the supplementary information, available at https://doi.org/10.13012/B2IDB-7439710_V1 in order to ensure transparency and enable future research.

PCAM employs a probabilistic framework that enables estimation of *likely* locations of specific crops for each year for the last half century. There are many potential applications of both the algorithm and the maps that have been produced. In its current form, PCAM provides a unique opportunity to determine the other factors of production that are important in the spatial distribution of agriculture. This can be done by analyzing deviations between census data and PCAM estimates, which are based on the underlying suitability of the mean climate and soil of pixels. This means that PCAM provides a useful benchmark of where crops were likely grown taking only climate and soil into consideration. PCAM has the additional advantage of requiring relatively few data inputs. The PCAM algorithm may also be useful in allocating national variables to a global grid in other settings.

It is important to note that the methodology is designed to provide a probabilistic assessment of crop distributions. For this reason, PCAM's maps should not be considered to be actual data. Rather, they should be used to determine likely locations of major crops in time. For locations in which sub-national census data exists, such as it does in the United States, then the census data is the preferred source of

information. However, agricultural census information is often not available for many locations and time periods. This means that PCAM estimates may be suitable for certain uses in these instances.

Future work could improve upon the approach, especially as new datasets and computational techniques emerge. For example, machine learning shows much promise to leverage satellite data for estimating crop-specific areas [59, 60]. However, this approach would be restricted to the time domain of necessary satellite data. Future work could assemble available sub-national census data and fuse it with PCAM estimates for locations without data. Insights into additional determinants of crop-specific suitability could be used to improve the suitability mesh. Additionally, future work could integrate information on extreme climate events to better determine locations of harvest as opposed to planted areas.

Acknowledgments

This material is based upon work supported by the National Science Foundation Grant No. ACI-1639529 ('INFEWS/T1: Mesoscale Data Fusion to Map and Model the US Food, Energy, and Water (FEW) System'), EAR-1534544 ('Hazards SEES: Understanding Cross-Scale Interactions of Trade and Food Policy to Improve Resilience to Drought Risk'), and CBET-1844773 ('CAREER: A National Strategy for a Resilient Food Supply Chain'). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The authors thank Qian Dang, Xiaowen Lin, and Landon Marston for their valuable feedback. NDJ is thankful for the support from the Diversifying Faculty in Illinois fellowship. All data sources are detailed in table 1 and are publicly accessible. The authors gratefully acknowledge these sources, without which this work would not be possible.

Data availability statement

Any data that support the findings of this study are included within the supplementary information, or are openly available at https://doi.org/10.13012/B2IDB-7439710_V1.

ORCID iDs

Nicole D Jackson https://orcid.org/0000-0002-3814-9906
Megan Konar https://orcid.org/0000-0003-0540-8438
Lyndon Estes https://orcid.org/0000-0002-9358-816X

References

[1] Vitousek P M, Mooney H A, Lubchenko J and Melillo J M 1997 Human domination of earth as ecosystems *Science* **277** 494–9
[2] Tilman D 1999 Global environmental impacts of agricultural expansion: the need for sustainable and efficient practices *PNAS* **96** 5995–6000
[3] Ramankutty N, Evan A T, Monfreda C and Foley J A 2008 Farming the planet: 1. Geographic distribution of global agricultural lands in the year 2000 *Glob. Biogeochem. Cycles* **22** GB1003
[4] FAO 2016 Food and Agriculture Organization of the United Nations *FAOSTAT* (http://www.fao.org/faostat/en/#home)
[5] Bouwman L, Klein Goldewijk K, Van Der Hoek K W, Beusen A H W, Van Vuuren D P, Willems J, Rufino M C and Stehfest E 2013 Exploring global changes in nitrogen and phosphorus cycles in agriculture induced by livestock production over the 1900–2050 period *PNAS* **110** 20882–7
[6] Gibbs H K, Ruesch A S, Achard F, Clayton M K, Holmgren P, Ramankutty N and Foley J A 2010 Tropical forests were the primary sources of new agricultural land in the 1980s and 1990s *PNAS* **107** 16732–7
[7] Hansen M C, Stehman S V and Potapov P V 2010 Quantification of global gross forest cover loss *PNAS* **107** 8650–5
[8] Elliott J *et al* 2014 Constraints and potentials of future irrigation water availability on agricultural production under climate change *PNAS* **111** 3239–44
[9] Bounoua L, DeFries R, Collatz G J, Sellers P and Khan H 2002 Effects of land conversion on surface climate *Clim. Change* **52** 29–64
[10] Foley J A, Delire C, Ramakutty N and Snyder P 2003 Green surprise? How terrestrial ecosystems could affect earth's climate *Frontiers Ecol. Environ.* **1** 38–44
[11] Mahmood R, Foster S A, Keeling T, Hubbard K G, Carlson C and Leeper R 2006 Impacts of irrigation on 20th century temperature in the northern Great Plains *Glob. Planet. Change* **54** 1–18
[12] Sanderman J, Hengl T and Fiske G J 2017 Soil carbon debt of 12 000 years of human land use *PNAS* **114** 9575–80
[13] Donner S D and Kucharik C J 2008 Corn-based ethanol production compromises goal of reducing nitrogen export by the mississippi river *PNAS* **105** 4513–8
[14] Rodell M, Velicogna I and Famiglietti J S 2009 Satellite-based estimates of groundwater depletion in india *Nature* **460** 999–1002
[15] Lansing J S and Kremer J N 2011 Rice, fish, and the planet *PNAS* **108** 19841–2
[16] Gasparri N I, Kuemmerle T, Meyfroidt P, le Polain de Waroux Y and Kreft H 2016 The emerging soybean production frontier in Southern Africa: conservation challenges and the role of South-South telecouplings *Conservation Lett.* **9** 21–31
[17] Foley J A *et al* 2005 Global consequences of land use *Science* **309** 570–4
[18] Lobell D B, Cassman K G and Field C B 2009 Crop yield gaps: their importance, magnitudes, and causes *Annu. Rev. Environ. Resour.* **34** 179–204
[19] Lobell D B, Schlenker W and Costa-Roberts J 2011 Climate trends and global crop production since 1980 *Science* **333** 616–20
[20] Ray D K, Mueller N D, West P C and Foley J A 2013 Yield trends are insufficient to double global crop production by 2050 *PLoS One* **8** e66428
[21] Ray D K, Ramankutty N, Mueller N D, West P C and Foley J A 2012 Recent patterns of crop yield growth and stagnation *Nat. Commun.* **3** 1293
[22] Lesk C, Rowhani P and Ramankutty N 2016 Influence of extreme weather disasters on global crop production *Nature* **84** 84–7
[23] Monfreda C, Ramankutty N and Foley J A 2008 Farming the planet: 2. Geographic distribution of crop areas, yields,

physiological types, and net primary production in the year 2000 *Glob. Biogeochem. Cycles* **22** GB1022
[24] Portmann F T, Siebert S and Döll P 2010 MIRCA2000-Global monthly irrigated and rainfed crop areas around the year 2000: a new high-resolution data set for agricultural and hydrological modeling *Glob. Biogeochem. Cycles* **24** 24 PP
[25] You L, Wood S, Wood-Sichra U and Wu W 2014 Generating global crop distribution maps: from census to grid *Agric. Syst.* **127** 53–60
[26] Anderson W, You L, Wood S, Wood-Sichra U and Wu W 2014 An analysis of methodological and spatial differences in global cropping systems models and maps *Glob. Ecol. Biogeogr.* **24** 180–91
[27] Wood-Sichra U, Joglekar A B and You L 2016 Global spatially-disaggregated crop production statistics data for 2005 version 3.2 *Technical Report* International Food Policy Research Institute (IFPRI) and International Institute for Applied Systems Analysis (IIASA) (https://doi.org/10.7910/DVN/DHXBJX)
[28] Wood-Sichra U, Joglekar A B and You L 2019 Global spatially-disaggregated crop production statistics data for 2010 version 1.0 *Technical Report* International Food Policy Research Institute (IFPRI) and International Science and Technology Practice and Policy (InSTePP) Center, University of Minnesota (https://doi.org/10.7910/DVN/PRFF8V)
[29] Klein Goldewijk K 2004 Estimating global land use change over the past 300 years: the HYDE database *GeoJournal* **61** 335–44
[30] Klein Goldewijk K and Ramankutty N 2004 Land cover change over the last three centuries due to human activities: the availability of new global data sets *GeoJournal* **61** 335–44
[31] Klein Goldewijk K 2016 A historical land use data set for the holocene; HYDE 3.2 (replaced) (https://doi.org/10.17026/dans-znk-cfy3)
[32] Klein Goldewijk K, Beusen A, Van Drecht G and De Vos M 2011 The HYDE 3.1 spatially explicit database of human-induced global land-use change over the past 12 000 years *Glob. Ecol. Biogeogr.* **20** 73–86
[33] Hurtt G C *et al* 2011 Harmonization of land-use scenarios for the period 1500–2100: 600 years of global gridded annual land-use transitions, wood harvest, and resulting secondary lands *Clim. Change* **109** 117
[34] Ramankutty N and Foley J A 1999 Estimating historical changes in global land cover: croplands from 1700 to 1992 *Glob. Biogeochem. Cycles* **13** 997–1027
[35] Pongratz J, Reick C, Raddatz T and Claussen M 2008 A reconstruction of global agricultural areas and land cover for the last millennium *Glob. Biogeochem. Cycles* **22** GB3018
[36] Klein Goldewijk K, Van Drecht G and Bouwman A F 2007 Mapping contemporary global cropland and grassland distributions on a 5 × 5 min resolution *J. Land Use Sci.* **2** 167–90
[37] GAEZ 2016 Food and agriculture organization of the united nations and the international institute for applied systems analysis, global agro-ecological zones (http://www.fao.org/nr/gaez/en/)
[38] Sitch S *et al* 2003 Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the lpj dynamic global vegetation model *Glob. Change Biol.* **9** 161–85
[39] New M, Lister D, Hulme M and Makin I 2002 A high-resolution data set of surface climate over global land areas *Clim. Res.* **21** 1–25
[40] Lawrence D M *et al* 2011 Parameterization improvements and functional and structural advances in version 4 of the community land model *J. Adv. Model. Earth Syst.* **3** 1–27
[41] FAO/IIASA/ISRIC/ISS-CAS/JRC 2009 Harmonized world soil database (version 1.1) *Technical Report* FAO and IIASA, Rome, Italy and Laxenburg, Austria
[42] Fischer G, Nachtergaele F O, Prieler S, Teixeira E, Toth G, van Velthuizen H, Verelst L and Wiberg D 2012 *GAEZ ver 3.0 Global Agro-ecological Zones Model Documentation* (Food and Agriculture Organization of the United Nations and the International Institute for Applied Systems Analysis) (http://fao.org/fileadmin/user_upload/gaez/docs/GAEZ_Model_Documentation.pdf)
[43] Klein Goldewijk K 2001 Estimating global land use change over the past 300 years: the hyde database *Glob. Biogeochem. Cycles* **15** 417–33
[44] Sacks W J, Deryng D, Foley J A and Ramankutty N 2010 Crop planting dates: an analysis of global patterns *Glob. Ecol. Biogeogr.* **19** 607–20
[45] Weidmann N B, Kuse D and Gleditsch K S 2010 The geography of the international system: the cshapes dataset *Int. Interact.* **36** 86–106
[46] Weidmann N B and Gleditsch K S 2010 Mapping and measuring country shapes the cshapes package *R J.* **2** 18–24
[47] Hijmans R, Garcia N and Wieczorek J 2015 GADM: database of global administrative areas. Version 2.8 (https://gadm.org)
[48] Seto K C, Kaufmann R K and Woodcock C E 2000 Landsat reveals chinaas farmland reserves, but they are vanishing fast *Nature* **406** 121
[49] Carletto C, Jolliffe D and Banerjee R 2013 The emperor has no data! agricultural statistics in Sub-Saharan Africa *World Bank Working Paper*
[50] Meng Q, Hou P, Lobell D B, Wang H, Cui Z, Zhang F and Chen X 2014 The benefits of recent warming for maize production in high latitude china *Clim. Change* **122** 341–9
[51] Portmann F T, Siebert S and Döll P 2010 Mirca2000-global monthly irrigated and rainfed crop areas around the year 2000: a new high-resolution data set for agricultural and hydrological modeling *Glob. Biogeochem. Cycles* **24** GB1011
[52] Urbani C B 1979 A statistical table for the degree of coexistence between two species *Oecologia* **44** 287–9
[53] Sokal R R and Michener C 1958 A statistical method for evaluating systematic relationship *Univ. Kansas Sci. Bull.* **38** 1409–38
[54] USDA NASS 2017 United States Department of Agriculture (USDA), National Agricultural Statistics Service (NASS) (https://nass.usda.gov/)
[55] Newberry Library 2017 Atlas of historical county boundaries (https://publications.newberry.org/ahcbp/index.html)
[56] United States Census Bureau 2012 TIGER/Line Shapefiles (https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.2012.html)
[57] Mazdiyasni O and AghaKouchak A 2015 Substantial increase in concurrent droughts and heatwaves in the united states *PNAS* **112** 11484–9
[58] Lark T J, Salmon J M and Gibbs H K 2015 Cropland expansion outpaces agricultural and biofuel policies in the united states *Environ. Res. Lett.* **10** 044003
[59] Defourny P *et al* 2019 Near real-time agriculture monitoring at national scale at parcel resolution: performance assessment of the sen2-agri automated system in various cropping systems around the world *Remote Sens. Environ.* **221** 551–68
[60] Wang S, Azzari G and Lobell D B 2019 Crop type mapping without field-level labels: random forest transfer and unsupervised clustering techniques *Remote Sens. Environ.* **222** 303–17