5

Gridded 5-arcmin, simultaneously farm-size- and crop-specific harvested area for 56 countries

10 Han Su[1, 2], Bárbara Willaarts [2], Diana Luna-Gonzalez [2], Maarten S. Krol [1], Rick J. Hogeboom [1 ,3]

[1] Multidisciplinary Water Management group, Faculty of Engineering Technology, University of Twente, Enschede, 7500AE, the Netherlands
[2] Water Security group, International Institute for Applied Systems Analysis (IIASA), Laxenburg, 2361, Austria
15 [3] Water Footprint Network, Enschede, 7522NB, the Netherlands

Correspondence to: Han Su (h.su@utwente.nl)

Abstract.

Farms are not homogeneous. Smaller farms generally have different planted crops, yields, agricultural inputs, and ~~irrigations~~irrigation applications compared to larger farms. ~~Mapping~~However, gridded farm-size ~~could facilitate studies~~specific data—that is moreover crop specific—is currently lacking. This obscures our understanding of differences between small- and large-scale farms, e.g. with respect to ~~quantify how water availability and~~ climate change ~~affect small and large farms, respectively. Given the lack of gridded farm size specific data, this~~adaptation and mitigation strategies, contribution to (local) food security, and water consumption patterns. This study ~~aims to develop a global~~fills a significant part of the current data gap, by developing high-resolution gridded, simultaneously farm-size- and crop-specific datasets of harvested area ~~We achieved it by downscaling a best available dataset, which collected~~ for 56 countries (i.e., covering about half the global cropland). Hereto, we downscaled the most complete global direct measurements ~~on crop and farm size, using~~of farm size and crop type by compiling state-of-the-art datasets, including crop maps, cropland extent maps, and dominant field size ~~distributions for~~distribution, representative of the year 2010. ~~Uncertainties in crop maps were explicitly considered by using two crop maps separately during downscaling. Due to data availability, our downscaled maps cover 56 countries accounting for half of the global cropland. Based on the~~ Using two different crop ~~maps, we have~~map sources, we were able to produce two new 5-arcmin gridded~~,~~ datasets on simultaneously ~~derived~~ farm-size- and crop-specific ~~dataset of~~harvested areas~~,~~: one for 11 farm sizes, 27 crops, and 2 farming systems~~,~~ and ~~another~~ one for 11 farm sizes, 42 crops, and 4 farming systems. ~~The downscaled maps~~In line with previous findings, our resulting datasets show major ~~differences in~~ planted crops and ~~irrigation change along with~~irrigated area (%) between farm sizes~~, which support previous findings. Validations show well consistencies with~~. Consistency between our resulting datasets and i) observations ~~on farm size specific oil palm~~from satellite images~~,~~ on farm-size-specific ~~irrigation from~~oil palm, ii) household surveys~~, and~~ on the farm-size-specific irrigated area (%), and iii) previous studies that ~~map farm size but are not~~mapped non-crop-specific ~~We observed~~ farm sizes, support the validity of our datasets. Although at grid level some uncertainties ~~at the grid cell level and found conclusions~~remain to be overcome, particularly those stemming from uncertainties in crop maps, results at ~~the~~ country level ~~are~~seem robust ~~to these uncertainties including the uncertainties from the crop maps. Our downscaled maps will help to explicitly include farm size into global agriculture modeling. The source~~. Source data, code, and ~~downscaled maps~~resulting datasets are open-access and freely available at https://doi.org/10.5281/zenodo.~~5747616~~6976249 (Su et al., 2022).

1 Introduction

There are over 608 million farms around the world~~.~~, which highly vary in their characteristics (Lowder et al., 2016; Lowder et al., 2021). ~~Land and water resources are not equally distributed among these farms. More~~ For example, more than 80% of the~~se~~ farms are smaller than 2 hectares and ~~they only~~ utilize only around 20% of global farmland area ~~(~~of 2.5 billion hectares ~~(Bosc et al., 2013;~~ Lowder et al., 2021~~; Bosc et al., 2013). The~~). In contrast, the largest ~~one percent~~1% of the farms ~~utilize~~occupy 70% of global farmland area (Lowder et al., 2021). Smaller farms ~~also insufficiently~~typically apply less irrigation ~~to adapt to water scarcity~~ in low- and middle-income countries~~,~~, making them more vulnerable to water scarcity than larger farms (Ricciardi et al., 2020).

~~In addition to water and land resources, the characteristics of agricultural production differ across farm sizes, which may be country-dependent. For example, in terms of~~ In terms of crops and mindful of national differences, smaller farms ~~tend to~~ plant more fruits, pulses, and roots and tubers, while larger farms plant more vegetables, nuts, and oil crops (~~Ricciardi et al., 2018b;~~ Herrero et al., 2017~~). In terms of agricultural practices used to increase agricultural productivity,~~; Ricciardi et al., 2018a, b). Furthermore, farmers who operate smaller farms tend to increase the use of non-fixed inputs~~,~~ to increase their productivity, such as fertilizers and pesticides, ~~while~~whereas larger farms ~~tend to~~rather increase fixed inputs~~,~~ such as machinery (Ren et al., 2019). ~~Smaller~~Whether smaller farms also ~~have a greater biodiversity on average (Ricciardi et al.,~~ generate~~2021; Noack et al., 2021). Though whether smaller farms have a~~ higher yield~~s~~ has long been debated, although it appears that yield~~s~~ often ~~correlates~~correlate positively with farm size (see Rudra (1968); Savastano and Scandizzo (2017); Gollin (2019); Ricciardi et al. (2021)). What seems undisputed, however, is that smaller farms on average display greater biodiversity than their larger counterparts (Ricciardi et al., 2021; Noack et al., 2021).~~(2021)).~~

~~These above-mentioned~~Since characteristics ~~stimulate~~vary widely between farms, many studies ~~to explicitly~~ set out to map the differences, particularly along the dimension of their size to discern small- and large-~~-~~scale farms ~~in agriculture studies and map farm sizes (Meyfroidt, 2017;~~ (Riesgo et al., 2016; Meyfroidt, 2017). At the global level, farm size mapping ~~farm sizes can be traced back to the studies of~~was pioneered by Lowder et al. (2016), Samberg et al. (2016), and Fritz et al. (2015). Lowder et al. (2016) estimated the country-~~-~~level distribution of farm size, based on multiple agricultural censuses. Samberg et al. (2016) used the Mean Agricultural Area (MAA) to assign each ~~subnational~~sub-national administrative unit with a farm size. ~~This~~A limitation of this approach is that it may overestimate the area of small farms ~~because not all farms are small, even if they are,~~ since being located in ~~the~~an administrative unit dominated by small farms does not necessarily mean that all farms within that unit are indeed small (Ricciardi et al., ~~2018b~~2018a, b). Fritz et al. (2015) ~~developed~~mapped a gridded global dominant field size ~~map~~distribution, using manually labeled field size data on ~~the~~ satellite images and spatial interpolation. The dominant field size ~~map~~distribution by Fritz et al. (2015) was updated by Lesiv et al. (2019). ~~When~~A consequence of interpreting fields as farms, ~~the~~however, is that small farm ~~area will also~~areas may be overestimated~~ as,~~ since large farms can include small-sized fields~~.~~ as well.

3

Further developments ensued through Herrero et al. (2017~~)~~, who used the country~~-~~level farm size data from Lowder et al. (2016) and Fritz et al. (2015) to develop a dominant farm size map ~~which~~. This map, in turn, was ~~later~~ updated by Mehrabi et al. (2020) using the field size ~~map~~distribution from Lesiv et al. (2019). ~~Given that dominant farm size only~~ However, despite its improvements, the method employed by Mehrabi et al. (2020) still assigns only one (i.e., a dominant) farm size to each grid cell (~~usually 10 km by 10 km), dominant farm size may over/underestimate some kinds of farm sizes when it is used to estimate~~5 × 5 arcmin), which reduces its usefulness in estimating the number and area distribution of different farm sizes.

~~In~~ Another important shortcoming in previous studies~~,~~ is that current farm size ~~mapping is~~maps are not crop~~-~~ specific. ~~One way~~A potential solution to estimat~~e~~ing the planted crops for different farm sizes is to overlap the farm size map with crop maps, e.g.~~, Monfreda et al. (2008) in~~. Samberg et al. (2016~~) and Mehrabi et al. (2020), Ray et al. (2013) in~~), Herrero et al. (2017~~). Overlays with crop maps~~), and Mehrabi et al. (2020). Yet still, such overlays may lead to biases in the ~~allocation~~assigning of crop-specific ~~cropping~~ areas to farm sizes ~~(Ricciardi et al., 2018b),~~, because of differences between farm size and MAA, field sizes, and dominant farm sizes~~,~~ and potentially also due to possible structural differences in crop choices between farm sizes~~.~~

~~One way to avoid such biases is to develop a simultaneously (Ricciardi et al., 2018a, b).~~ In order to address these limitations, farm~~-~~size- and crop-specific ~~map~~datasets would need to be developed simultaneously, which is what Ricciardi et al. ~~(2018b); Ricciardi et al.~~ (2018a~~) established an~~, b) attempted. Arguably the most complete empirical global ~~database using~~dataset to day, they collated data from agriculture census~~es~~ and household surveys that directly measured crop production or areas in combination with farm size. The~~i~~r dataset covers about half of the global cropland, including data for 56 countries[1]~~-~~, with subnational data for 46 countries. ~~Ricciardi's dataset, however, does not have gridded maps, so it has limited~~ Still, being defined at administrative unit level, the dataset by Ricciardi et al. (2018a, b) lacks a high-resolution grid-level representation of the data. This resolution gap limits the capability to fulfill the needs of ~~global~~e.g. climate~~-change,~~ agricultural and ~~water resources studies, where the~~ hydrological ~~model and climate~~ models ~~which~~ commonly ~~use grided maps~~need gridded data as input~~.~~ ~~Lacking gridded farm size- and crop-specific maps limits the evaluations of how water scarcity and climate change affect~~, which, in turn, obscures our understanding of differences between small- and large-~~scale~~ farms, ~~respectively~~e.g. with respect to climate change adaptation and mitigation strategies, contribution to local food security, and water consumption patterns.

This study ~~aims to develop~~fills a ~~global~~significant part of the current data gap, by developing high-resolution gridded, simultaneously farm~~-~~size- and crop-specific dataset~~s~~ of harvested ~~areas with~~area for 56 countries, representative of the year 2010. The datasets, moreover, provide additional information on farming systems. ~~Considering the data availability, the baseline year is 2010 with data covering 56 countries. We compiled multiple datasets~~ To obtain the datasets, we developed and applied a downscaling procedure, in which we used state-of-the-art datasets on field size and crop type, including crop
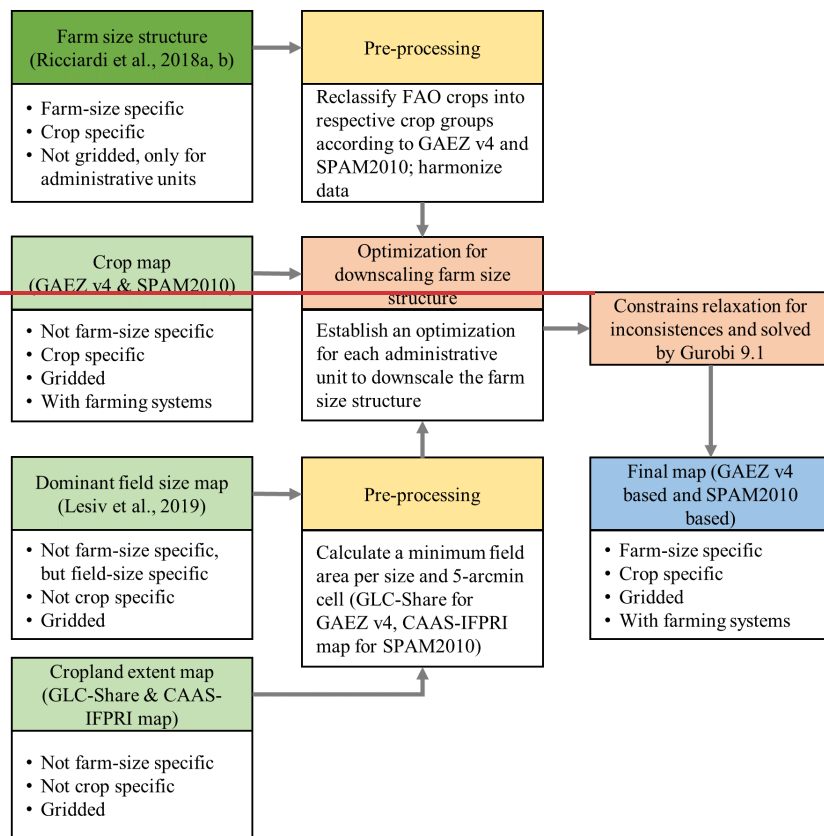
---

[1] ~~In their~~Their paper~~, they claim to have states~~ data is available for 55 countries~~. In,~~ but the associated dataset ~~they published, it~~actually contains ~~the 56th country, the~~ 56 (Czech Republic~~.~~ seems to be added).

maps (Yu et al., 2020; FAO and IIASA, 2021; Fischer et al., 2021), cropland extent, (Latham et al., 2014; Lu et al., 2020), and

110 dominant field size distribution, as well as crop distribution and farming systems and used them (Lesiv et al., 2019), to downscale the most complete empirical global farm-size- and crop-specific datasets developeddataset by Ricciardi et al. (2018b); Ricciardi et al. (2018a)., b) from the level of administrative units intounit to a 5 arcmin grid cell level. We alsogridded spatial resolution. Two crop maps were used to explicitly considered the consider uncertainties in crop distributions by using two crop maps. The . We validated our resulting downscaled maps were validated with datasets using empirical data and

115 comparedisons with previous studies.

2 Methods

2.1 Overview

Imagine that we know the crop area of small and large farms within an administrative unit, to downscale it, if we get a high spatial resolution map of crop area, we may have some idea on where the small and large farms may locate because some

120 crops are planted more by small farms and some crops are planted more by large farms. In addition, if we have the field size distribution within the administrative unit, we could know more about the location of small and large farms because large fields only belong to large farms and small farms can only be located in small fields. When we combine the information from the crop map and field size distribution, even though we could not precisely locate small and large farms, we can estimation their distributions in this administrative unit with some extent of uncertainties. This is how we develop the gridded,

125 simultaneously farm size- and crop-specific dataset of harvested areas. Theoretically, we could estimate multiple distributions of small and large farms that are consistent with all the administrative level and grid cell level data. Practically, however, these distributions may not exist because of the background inconsistencies in the datasets. To deal with the background inconsistencies, we assume the best estimation of the farm size- and crop-specific distributions are the distributions that could maximize consistencies with datasets. In those cases, we tried to find multiple distributions that meet the same level of

130 consistency with datasets and averaged the multiple distributions to get the final estimation.

Farm size structure (Ricciardi et al., 2018a, b)	Pre-processing
• Farm-size specific • Crop specific • Not gridded, only for administrative units	Reclassify FAO crops into respective crop groups according to GAEZ v4 and SPAM2010; harmonize data

Crop map (GAEZ v4 & SPAM2010)	Optimization for downscaling farm size structure	Constrains relaxation for inconsistences and solved by Gurobi 9.1
• Not farm-size specific • Crop specific • Gridded • With farming systems	Establish an optimization for each administrative unit to downscale the farm size structure	

Dominant field size map (Lesiv et al., 2019)	Pre-processing	Final map (GAEZ v4 based and SPAM2010 based)
• Not farm-size specific, but field-size specific • Not crop specific • Gridded	Calculate a minimum field area per size and 5-arcmin cell (GLC-Share for GAEZ v4, CAAS-IFPRI map for SPAM2010)	• Farm-size specific • Crop specific • Gridded • With farming systems

Cropland extent map (GLC-Share & CAAS-IFPRI map)
• Not farm-size specific • Not crop specific • Gridded

The gridded, simultaneously farm-size- and crop-specific dataset of harvested areas can be achieved by downscaling the administrative unit level crop-specific farm size structure using gridded crop distribution and gridded dominant field size distribution (Fig. 1). Since certain crops are more prevalent on small farms and others on larger farms as indicated by crop-specific farm size structure, the gridded crop distribution primarily indicate where small and large farms are located. Gridded dominant field size distribution further helps specify the location of small and large farms because, by definition, large fields only belong to large farms and small farms can only be located in small fields. We assumed the best estimation of the farm-size- and crop-specific harvested area distribution is the one that maximizes consistencies with the underlying administrative unit farm-size and grid cell level data.

6

140 **The dataset**~~Figure 1. Diagram of map development processors.~~

~~The map~~ development involved pre-processing of multiple datasets, establishing optimization for downscaling, and constraints relaxation and solving optimization problems (Fig. 1). The pre-processing included two parts: i) reclassifying crops to accommodate differences in crop classification used in the underlying datasets and harmonizing ~~Ricciardi's~~the dataset by Ricciardi et al. (2018a, b) and ii) converting the dominant field size ~~map~~distribution into a minimum field area per field size

145 and 5-arcmin grid cell (Sect. 2.2). The downscaling was achieved by maximizing consistencies with multiple datasets that provide information on the location of each farm/field size and planted crops. Specifically, we ~~established~~formulated an optimization for each administrative unit (Sect. 2.3) and solved it via constraints relaxations (Sect. 2.4). Priorities in achieving consistency with the various underlying datasets were considered during these processes (Sect. 2.3 and 2.4). The spatial crop distribution affects both crop location and farm size location during downscaling and is ~~usually uncertain. To consider~~

150 ~~the~~associated with considerable uncertainties ~~in~~. To consider propagation of such uncertainties, we used two different crop maps, ~~we used two crop maps~~i.e. GAEZv4 (FAO and IIASA, 2021; Fischer et al., 2021) and SPAM2010 (Yu et al., 2020). Doing so allowed us to develop two alternative versions of the final downscaled ~~map~~dataset separately.

7
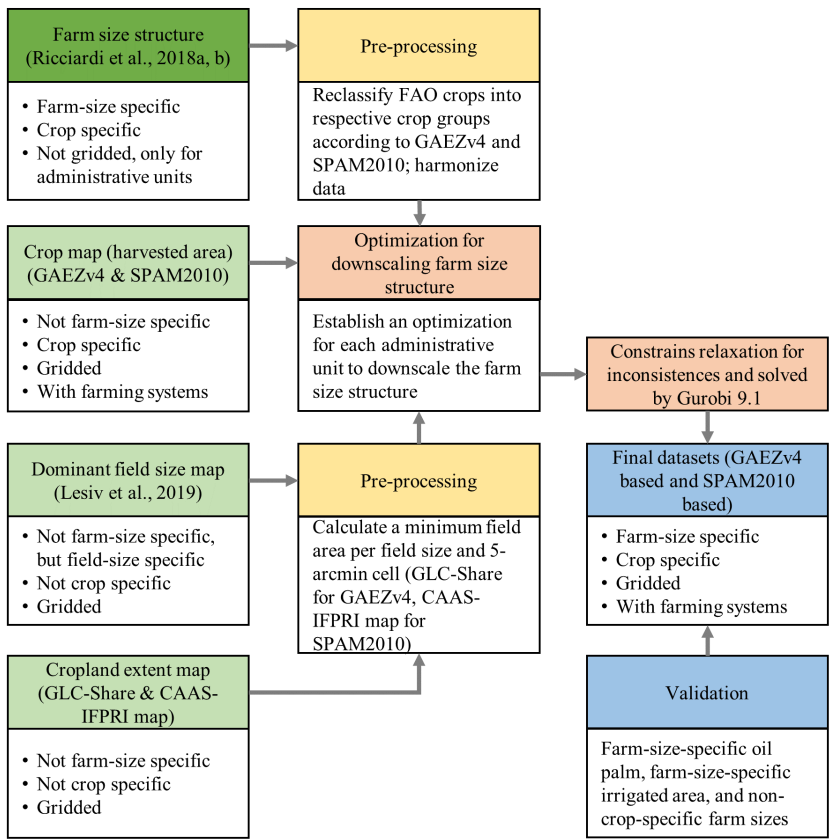
Figure 1. Diagram of dataset development processors.

2.2 Datasets and pre-processing

~~All the datasets used in this study can be found in Table 1. Ricciardi's~~The main dataset by Ricciardi et al. (2018a, b) provides the farm ~~size~~-size- and crop-specific cropping area for 56 countries at the administrative unit level (~~(~~see [S1] for ~~the~~a list of the 56 countries included). The eleven farm sizes in this dataset are based on the classification from the World Census of Agriculture (WCA) (FAO, ~~2020b~~2015; Ricciardi et al., 2018a, b; FAO, 2022): 0–1 ha, 1–2 ha, 2–5 ha, 5–10 ha, 10–20 ha, 20–50 ha, 50–100 ha, 100–200 ha, 200–500 ha, 500–1000 ha, and >1000 ha. The cropping area in this dataset ~~means~~indicates

8

either crop area, planted area, harvested area, or cultivated area. Because ~~the~~ data quality varies from country to country and because this dataset was not harmonized ~~over~~in time, we chose to downscale its crop-specific farm size structure~~,~~ (i.e., the ~~crop-specific~~ percentage of harvested area per farm size~~,~~ for each crop) instead of the absolute area. ~~The crop~~

165 Crop-specific harvested area is taken from two separate crop ~~maps: GAEZ v4 (Fischer et al., 2021;~~ map sources: GAEZv4 (FAO and IIASA, 2021; Fischer et al., 2021) and SPAM2010 (Yu et al., 2020). These are the ~~only two~~most comprehensive crop maps available, containing harvested area of ~~tens~~dozens of crops for the year 2010 at 5 arcmin spatial resolution (Kim et al., 2021). ~~GAEZ v4~~GAEZv4 and SPAM2010 have their own crop classification systems ~~(~~[, which are given in [S2, S3~~] for details). GAEZ v4~~]. Furthermore, GAEZv4 distinguishes two farming systems~~:~~, namely irrigated and rainfed~~.~~, while

170 SPAM2010 further ~~distinguishes~~specifies rainfed into low- and high-input rainfed and subsistence rainfed~~.~~ (in addition to irrigated farming systems).

The dominant field size distribution (Lesiv et al., 2019) indicates where larger farms may ~~locate. It provides~~be located and contains the spatial distribution for five field sizes: < 0.64 ha, 0.64–2.56 ha, 2.56–16 ha, 16–100 ha, and >100 ha. For pre-processing the dominant field size distribution, cropland extent maps were also included (detailed steps could be found below).

175 All datasets used in this study are listed in Table 1.

Table 1. Datasets that were used to develop the gridded, farm‑size ~~specific,~~ and crop-specific dataset~~s~~ of harvested area.

Dataset	Indicator	Spatial coverage and resolution	Time	Crop	Note
Ricciardi et al. (2018b); Ricciardi et al. (2018a)	Farm size structure*	56 countries; (sub)national administrative unit	Varies from 2001 to 2015	154 FAO crops	11 farm sizes
GAEZ v4 (Fischer et al., 2021; FAO and IIASA, 2021)	Harvested area	Global; gridded, 5 arcmin (10 km)	2010	27 GAEZ crops**	2 farming systems (irrigated and rainfed)
SPAM2010 (Yu et al., 2020)	Harvested area	Global; gridded, 5 arcmin (10 km)	2010	42 SPAM crops	4 farming systems (irrigated, low- and high-input rainfed and subsistence rainfed)
Dominant field size distribution (Lesiv et al., 2019)	Dominant field size	Global; gridded, 30 arcsec (1 km)	Varies from 2000 to 2017	Not crop-specific	5 field sizes
GLC-Share (Latham et al., 2014)	Cropland extent	Global; gridded, 30 arcsec (1 km)	Around 2010	Not crop-specific	The based map of GAEZ v4

~~CAAS-IFPRI cropland extent map (Lu et al., 2020)~~	~~Cropland extent~~	~~Global; gridded, 15 arcsec (0.5 km)~~	~~2010~~	~~Not crop-specific~~	~~The base map of SPAM2010~~

* Here ~~we mean~~ the crop-specific percentage of harvested area per farm size within an administrative unit **is meant. ** ** The 27th crop is Fruits and Nuts which is not listed in the document but available in their dataset.

Dataset	Indicator	Spatial coverage and resolution	Time	Crop	Note
Ricciardi et al. (2018a, b)	Farm size structure*	56 countries; (sub)national administrative unit	Varies from 2001 to 2015	154 FAO crops	11 farm sizes
GAEZv4 (FAO and IIASA, 2021; Fischer et al., 2021)	Harvested area (crop map)	Global; gridded, 5 arcmin (10 km)	2010	27 GAEZv4 crops**	2 farming systems (irrigated and rainfed)
SPAM2010 (Yu et al., 2020)	Harvested area (crop map)	Global; gridded, 5 arcmin (10 km)	2010	42 SPAM2010 crops	4 farming systems (irrigated, low- and high-input rainfed, and subsistence rainfed)
Dominant field size distribution (Lesiv et al., 2019)	Dominant field size	Global; gridded, 30 arcsec (1 km)	Varies from 2000 to 2017	Not crop specific	5 field sizes
GLC-Share (Latham et al., 2014)	Cropland extent	Global; gridded, 30 arcsec (1 km)	Around 2010	Not crop specific	The based map of GAEZv4
CAAS-IFPRI cropland extent map (Lu et al., 2020)	Cropland extent	Global; gridded, 15 arcsec (0.5 km)	2010	Not crop specific	The base map of SPAM2010

180 To pre-process ~~Ricciardi's~~the dataset~~,~~ by Ricciardi et al. (2018a, b), we first reclassified their crops (who followed the FAO ~~crops in this dataset~~classification) into 27 GAEZv4 crops and 42 SPAM2010 crops, respectively. ~~Detailed criteria~~Crop reclassification details can be found in [S2, S3]. We used the cropping area to ~~get~~obtain the crop-specific farm size structure. In this dataset, the cropping area is crop- specific and includes four items: crop area, planted area, harvested area, and cultivated

area. These variables were identified by ~~Ricciardi's~~the dataset by Ricciardi et al. (2018a, b) from the local agriculture census. There is no worldwide standard definition for these items (FAO, 2015~~). Local~~), meaning local agriculture censuses ~~have~~can apply their own ~~preference to use one of them for specific crops.~~preferred definitions. In ~~generally~~general, however, *planted area* is used for temporary crops; *cultivated area* for temporary crops and permanent crops; *crop area* for temporary crops, permanent crops, fallow fields, meadows, and pastures; and *harvested area* ~~is~~for the cultivated area excluding the area ~~destroyed~~rendered unsuitable for cultivation by natural disasters or other reasons (FAO, ~~2020a,~~2015, 2020). In terms of data availability, one or two of these items are available for most countries~~. To harmonize data, when~~ at the least. If more than one item ~~is~~was available, we ~~used~~harmonized the data by taking the item with ~~a larger~~the largest overall area (after crop reclassification) to estimate farm size structure ~~because~~, since a larger overall area typically means that more farm size classes have available data ~~in most cases.~~. If none of the four items ~~were~~was available, we used crop production data provided by ~~Ricciardi's~~the dataset ~~to get~~by Ricciardi et al. (2018a, b) as a proxy for the crop-specific farm size structure~~. In this case, we assumed~~, assuming constant yield~~s~~ across farm sizes.

~~We~~During pre-processing we also converted the 1 ~~*~~× 1 km dominant field size distribution map into a minimum field area per field size and 5-arcmin grid cell ~~during pre-processing.~~to align with the spatial resolution of crop maps. We interpreted *dominant field size* as ~~a field~~that fields of that size accounting for at least 50% of cropland in the grid cell. For each field size, we calculated the minimum field area for each 1-km cell by using the 50% of cropland extent ~~that is dominated by the respective field size.~~. We then summed ~~and scaled~~ the minimum field area ~~to cover all croplands of~~from 1-km to 5-arcmin cells. ~~To keep cropland extent consistent with crop map during downscaling, GLC-Share is used when the crop map is GAEZ v4; CAAS-IFPRI cropland extent map is used when the crop map is SPAM2010.~~ and scaled the summed area to cover 50% of croplands in 5-arcmin cells. The minimum field area of field size 16–100 ha is 120 ha in ~~the~~a 5-arcmin cell~~#23~~ which means, for example, farms larger than 16 ha should occupy at least 120 ha in the cell~~#23~~. To keep cropland extent consistent with the crop maps during downscaling, GLC-Share was used with the GAEZv4 crop map, while we used CAAS-IFPRI cropland extent map with the SPAM2010 crop map.

2.3 Optimization for downscaling

For each administrative unit defined in ~~Ricciardi's~~the dataset~~.~~ by Ricciardi et al. (2018a, b), we established the following optimization problem for our downscaling~~:~~ procedure. Note that the dataset by Ricciardi et al. (2018a, b) identifies eleven farm sizes and the dominant field size distribution (Lesiv et al., 2019) identifies five field sizes.

Sets:

c, Crops

f, Farm size, labelled by the lower bound of the eleven farm sizes

e, Field size, labelled by the lower bound of the five field sizes

s, Farming system

a, Administrative unit

g, Grid cell

Parameters:

$ha.R_{c,f,a}$, Crop-specific farm size structure, percentage of the harvested area of farm size f that plant crop c in the administrative unit a, (from ~~Ricciardi's~~the dataset by Ricciardi et al. (2018a, b))

$ha.S_{c,s,g}$, Harvested area of crop c under farming system s at grid cell g, (from crop map, either SPAM2010 or GAEZv4)

$ha.L_{e,g}$, Minimum field area of field size e at grid cell g, (from dominant field size ~~map~~distribution by Lesiv et al. (2019) and crop extent map by Latham et al., (2014) and Lu et al., (2020)

p_f, The minimum farm area of farm size f in any ~~gir~~grid cell when the farm size f exists~~; it is~~, i.e., the lower bound of the farm size class f

l, Elastic factor

Variables:

$ha_{c,f,s,g}$ Harvested area of crop c, farm size f, farming system s ~~at~~in grid cell g, (estimated by this study)

Objective function:

Since we aim to downscale ~~Ricciardi's~~the dataset, by Ricciardi et al. (2018a, b), we ~~wanted to maximize~~maximized—within the constraints— consistencies with ~~Ricciardi's~~the dataset ~~when constraints allow:~~by Ricciardi et al. (2018a, b):

$$min \sum_{c,f} abs\left(ha.R_{c,f,a} \sum_{s,g\in a} ha.S_{c,s,g} - \sum_{s,g\in a} ha_{c,f,s,g} \right) \qquad (1)$$

Constraints:

The first constraint ~~ensures consistencies~~is meant to ensure consistency with the respective crop ~~map:~~maps we used and states that the total harvested area per crop per farming system per grid cell in our ~~map equals~~datasets must be equal to the harvested area per crop per farming system per grid cell in the respective crop map.

$$\sum_f ha_{c,f,s,g} = ha.S_{c,s,g}, \forall c, s, g \qquad (2)$$

The second constraint ~~ensures~~requires a minimum ~~consistencies~~level of consistency with ~~Ricciardi's~~the dataset. ~~The~~ by Ricciardi et al. (2018a, b) and states that the relative difference in farm size structure between our estimation and ~~Ricciardi's~~the dataset ~~would~~by Ricciardi et al. (2018a, b) cannot be ~~less~~more than 10%. This constraint ensures that ~~we do not diverge far from Ricciardi's dataset,~~ even when other constraints are hard to meet. ~~In this case,~~, we do not diverge too far from the dataset by Ricciardi et al. (2018a, b). This constraint takes priority over the following constraints, meaning we would relax other constraints to ~~ensure these minimum consistencies with Ricciardi's dataset. The arbitrary~~ meet this one. The 10% relative difference ~~considers~~mark is an educated guess based on timestamp differences in ~~Ricciardi's~~the dataset by Ricciardi et al. (2018a, b) and overall assumed uncertainties underlying each of the datasets.

$$90\% * ha.R_{c,f,a} \sum_{s,g\in a} ha.S_{c,s,g} \le \sum_{s,g\in a} ha_{c,f,s,g} \le 110\% * ha.R_{c,f,a} \sum_{s,g\in a} ha.S_{c,s,g}, \forall c,f \tag{3}$$

~~Third, we also applied~~The third constraint sets a minimum allocated area for each farm size ~~at~~in each grid cell. ~~The~~ if the farm size exists in the cell. This minimum allocated area is not necessarily required by the definition of farm size ~~since the farm size is defined based on the total operated or cultivated area that does not need to be a single crop area and single farming system. It~~, yet we reckoned it is still reasonable to include ~~it~~ because ~~we applied it at the~~ a 5-arcmin ~~(~10 km)~~ grid cell ~~level. Considering the~~is mostly much larger than a single farm. Given both the presence of uncertainties in these constraints and inconsistencies among datasets used, we ~~consider~~incorporated this constraint in a hard form and soft form. ~~We~~ during the optimization: we used the hard form by default. ~~We consider relaxing these constraints using the~~, but transitioned to the more relaxed soft form when the optimization ~~is~~was infeasible (see also Sect. 2.4). ~~The~~Note that the soft form does not strictly require ~~the~~a minimum allocation area for each farming system.

Hard form:

$$ha_{c,f,s,g} \ge p_f, \forall c,f,s,g, if\ ha_{c,f,s,g} > 0 \tag{4}$$

Soft form:

$$\sum_s ha_{c,f,s,g} \ge l \times p_f, \forall c,f,g, if\ ha_{c,f,s,g} > 0 \tag{5}$$

~~Fourth, we applied~~The fourth constraint sets a minimum area ~~constraint~~ for ~~some~~certain farm sizes according to the spatial distribution of dominant field size ~~distribution. This constraint follows the logic~~. The rationale is that a field ~~could~~can only belong to ~~an~~a farm equal or larger than its own size ~~of farm~~. We assumed a ~~linear~~uniform distribution of area within each farm size, like Ricciardi et al. (2018a, b), to accommodate the different classifications of size in farms and fields. ~~Given~~For example, 40% of the area ~~of~~in the farm size 10–20 ha was assumed to be in 16–20 ha class in Eq. (7).

For field ~~larger than 100 ha, for~~areas and farms larger than 100 ha:

$$\sum_{c,s,f\ge100} ha_{c,f,s,g} \ge ha.L_{100,g}, \forall g \tag{6}$$

~~Given the area of~~For field areas larger than 16 ha, ~~for~~and farms larger than 10 ha:

$$\sum_{c,s,f\ge20} ha_{c,f,s,g} + \frac{20-16}{20-10}\sum_{c,s} ha_{c,10,s,g} \ge ha.L_{100,g} + ha.L_{16,g}, \forall g \tag{7}$$

~~Given the area of~~For field areas larger than 2.56 ha, ~~for~~and farms larger than 2 ha:

$$\sum_{c,s,f\ge5} ha_{c,f,s,g} + \frac{5-2.56}{5-2}\sum_{c,s} ha_{c,2,s,g} \ge ha.L_{100,g} + ha.L_{16,g} + ha.L_{2.56,g}, \forall g \tag{8}$$

~~Given the area of~~For field areas larger than 0.64 ha, ~~for all farms~~ and any farm size:

$$\sum_{c,s,f\ge1} ha_{c,f,s,g} + \frac{1-0.64}{1-0}\sum_{c,s} ha_{c,0,s,g} \ge ha.L_{100,g} + ha.L_{16,g} + ha.L_{2.56,g} + ha.L_{0.64,g}, \forall g \tag{9}$$

~~Last but~~Since areas should not ~~least~~assume negative values, we ~~have~~also include non-negative area constraints:

13

$$ha_{c,f,s,g} \geq 0, \forall c,f,s,g \tag{10}$$

2.4 Constraints relaxation and solving procedures

When the above optimization (Eq. (1)–(10)) proved infeasible, we first replaced the hard form of minimum allocated area (i.e., the third constraint) (Eq. (4)) for all farm sizes with the soft form (Eq. (5)) and applied the elastic factor with the following values in order: 1, 1/2, 1/4, 1/8, 1/16, 1/32, 1/64, and 0. If optimization was still infeasible, we relaxed the minimum area constraint required by the dominant field size distribution (i.e., the fourth constraint) by removing the constraints from large to small farms until the optimization was feasible. Relaxing the minimum area constraint did not happen often during downscaling.

Once the above optimization became feasible, we did not necessarily strike a unique global optimum. Therefore, we calculated up to 80 (sub)optimal solutions with the same level of consistencies and averaged these to obtain the final solution. Since the number and quality of solutions depend on the searching process of the solving toolbox, this procedure may still leave some bias in the final averaged solution.

Each optimization was solved by Gurobi v9.1, a fast commercial optimization solver, using the dual simplex method (Gurobi Optimization, 2021). Optimization was taken as infeasible by the solver's initial evaluation or if it is computationally unsolvable (cannot be solved within 150 seconds). Most of the optimal solutions were obtained within 60 seconds when feasible. For those administrative units that contained more than 300 5-arcmin grid cells, the optimization becomes highly complex. This posed a challenge for the solver, with the number of decision variables increasing to over half a million. As a workaround, we applied a two-tiered optimization, where we first divided all grid cells randomly into several groups. Each group included ~100 grid cells (except for Russia, where groups were set to contain ~200 grid cells to keep the total number of groups below 300). Next, we solved the optimization at group level, followed by solving it at the cell level within each group. Out of 3,421 administrative units, 244 units underwent this workaround procedure, collectively covering 89.4% of grid cells in this study. The entire optimization was performed on a desktop computer (Intel(R) Core(TM) i7-8700 CPU @ 3.20GHz, RAM 16 GB) taking 9 days.

Finally, we masked the crop-specific farm size as unknown if their crops are not covered by the dataset by Ricciardi et al. (2018a, b). For these crops, the optimization would still estimate their farm size structures only based on the distribution of crops and dominant field size. Since the overall farm size structure is absent and dominant field size is not sufficient to estimate all farm sizes, uncertainties of these crops are significantly larger than those associated with the dataset by Ricciardi et al. (2018a, b).

14

2.5 Validation ~~of downscaled maps~~ and comparison with previous studies ~~mapping farm sizes~~

~~The ideal way to validate~~Ideally, we would have validated our ~~downscaled simultaneously~~ farm-size- and crop-specific ~~dataset is to compare it~~datasets with observations. However, ~~most of the~~ there are limited empirical datasets available ~~datasets are not farm size specific. This creates challenges for validating our downscaled maps for all crops and farming systems. We searched for validation datasets that~~, and if there ~~are~~ ~~global focused,~~, most are not farm-size specific ~~with additional information on crop or farming systems. Limited by data availability, we were able to validate our downscaled maps with~~. Given these limitations, we validated our datasets using two empirical datasets ~~and we compared them with previous studies to assess the reliability of our downscaled maps. More validations are expected when more validation datasets become available.~~

~~For the first validation with empirical datasets, we compared our downscaled map with~~. The first is by Descals et al. (2020~~)~~ ~~on oil palm map. Descals et al. (2020)~~), who developed a global gridded farm-size-specific oil palm map using deep learning and satellite images for the year 2019. We validated our datasets for five countries that are covered by both our datasets and the dataset by Descals et al. (2020) (Fig. A1). ~~With satellite images, they classified oil palm areas into small farms and large farms based on landscape features. In order to~~To interpret th~~eir~~sr size classification, we adopted the definition of small oil palm farms by Indonesia (the world's largest palm oil producer and exporter) ~~and used 25 ha as the threshold for the two scales (mentioned by~~ Descals et al~~.,~~., (2020~~). The validation was in five countries because only the five countries are covered by both our dataset and validation dataset(Fig. A1). The crop *Oil palm* in GAEZ v4 and SPAM2010 based map was used for validation separately.~~), who apply a 25 ha threshold to distinguish small from large farms, i.e., between the two scales as included in Descals et al. (2020). We calculated the Pearson correlation coefficient ~~between our downscaled map and Descals et al. (2020)~~ at grid cell level ~~on three~~ (i.e, 5 arcmin) and two additional spatial scales ~~using spatial moving average, 5 arcmin,~~, i.e., 15 arcmin~~,~~ and 25 arcmin, using a spatial moving average. We validated our GAEZv4 and SPAM2010 crop map based datasets, separately.

~~For the~~The second ~~validation with~~ empirical ~~datasets,~~dataset to which we compared our ~~downscaled maps with~~datasets is that of farm-size-specific ~~irrigation~~ percentage of irrigated area at the country level ~~using~~from the FAO RuLIS (Rural Livelihoods Information System) database (FAO, 2021). ~~Eleven of 56 countries'~~RuLIS includes micro-level household survey data ~~around~~representative of the year 2010. Eleven out of 56 countries included in our study are ~~available [S4].~~also listed in RuLIS (see an overview in [S4]). Based on the~~se~~ household surveys, we calculated the percentage of ~~the~~total irrigated area (irrigated area divided by cultivated area) for each farm size (classified by crop area) where at least ~~5~~five survey samples are available. ~~We~~Once more, we calculated the correlations between our estimati~~on~~es and those derived from the household surveys. ~~This~~ Although this validation considers farm-size-specific farming systems, ~~wi~~the data is aggregated over crops.

~~We also compared our downscaled map with previous studies, Lowder et al. (2016) and Mehrabi et al. (2020), which mapped the geographic distribution of farm sizes but were not crop-specific and not farm-system specific. Lowder et al. (2016) provides the percentage of harvested area operated by each farm size at the~~To further validate our datasets, we compared our datasets to two other studies. The first is by the FAO and has just been published (FAO, 2022). This dataset contains structural data

15

obtained through agricultural censuses, including total crop areas per farm size, at country level. ~~Mehrabi's dataset keeps,~~ for the years 1990, 2000, and 2010. We compared our datasets with the structural data of 2010 (the year our datasets are most representative of), and complementary with data of the year 2000 as well. The reason to include data on 2000 too is that data does not rely so heavily on interpolation as does 2010 (FAO, 2022), making the comparison more robust although temporal representativeness is less appropriate. Another advantage of including FAOSTAT structural data of 2000 is that it allows for the comparison with the widely used dataset by Lowder et al. (2016) at the same time since the dataset by Lowder et al. (2016) is largely the same as FAOSTAT structural data of 2000 [S5].

The second study to which we compared our datasets is by Mehrabi et al. (2020), who mapped geographic distributions of farm sizes. The dataset by Mehrabi et al. (2020) uses the same farm size distribution as ~~Lowder's~~the dataset by Lowder et al. (2016) at the country level, but ~~provides~~adds the dominant farm size ~~per~~at 5-arcmin grid cell. ~~We~~ level. For our comparison, we calculated ~~—~~at grid cell level—the dominant farm size from our ~~downscaled map~~datasets with the farm size that operates the largest total harvested area per grid cell, for ~~GAEZ based downscaled map~~our GAEZv4 and ~~SPAM based downscaled map,~~ respectively. ~~The comparison was pixel-to-pixel by counting the number of cells that have similar, larger, and smaller dominant farm size in our maps compared with Mehrabi's dataset. Similar dominant farm size means the farm size in our downscaled map are the same or next to the farm size in Mehrabi's dataset.~~SPAM2010 crop map based datasets, separately.

3 Results ~~and analysis~~

3.1 ~~The crop type~~Dataset statistics

3.1.1 Crop types and farm size~~s~~

~~With the crop map from GAEZ v4 (SPAM2010), we~~ We identified ~~the 5-arcmin~~ gridded harvested area for 56 countries, 11 farm sizes, 27 crops ~~(42 crops for SPAM based map),~~ and 2 farming systems ~~(based on the GAEZv4 crop map, and for 42 crops and~~ 4 farming systems ~~for SPAM~~ based on the SPAM2010 crop map~~)~~. ~~One example can be found in~~, both at 5-arcmin spatial resolution. Fig. 2~~, where we illustrate~~ illustrates the harvested area of rainfed maize belonging to two farm sizes (2–5 ha and 500–1000 ha)~~. Overall,~~) according to our ~~results~~farm-size- and crop-specific harvested area dataset based on the GAEZv4 crop map. Statistics of crop type and farm size show the ~~preference for~~prevalence of certain crop groups for ~~eleven~~certain farm sizes ~~(~~(see [S2] for the crop group~~ing~~s of ~~the~~ 27 GAEZv4 crops). ~~Fig. As~~3(a) shows that, as farm size increases, oil crops and fodder crops become more ~~popular;~~prevalent, while fruits and nuts, pulses, and roots and tubers become less ~~popular (~~widespread. ~~Our dataset~~Fig. 3(a)). ~~Larger farms (>20 ha) dominate the planting of fodder crops, sugar crops and oil crops; smaller farms (< 20 ha) dominate the planting of vegetables, stimulates, roots and tubers, pulses, fruits, nuts, and cotton (Fig. 3(b)). The SPAM~~ based on the SPAM2010 crop map shows comparable results ~~(Fig. A2 and~~to that based on GAEZv4 (see Fig. A2 for crop groupings as per [S3]). These ~~results~~statistics are consistent with ~~our datasets~~earlier findings

16

360 by Ricciardi et al. (~~2018b~~2018a, b) and ~~previous studies~~ Herrero et al. (2017~~), which indicate that the optimization resulted in modest remaining inconsistency.~~).
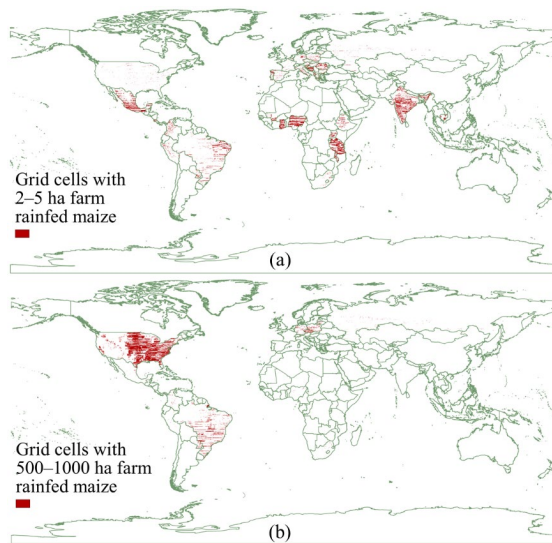
Figure 2. ~~The gird~~Grid cells with a harvested area of rainfed maize ~~belonging to the~~on farm size 2–5 ha (a) and farm size 500–1000 ha (b),

365 according to ~~the GAEZ~~our farm-size- and crop-specific dataset based ~~downscaled~~on the GAEZv4 crop map.
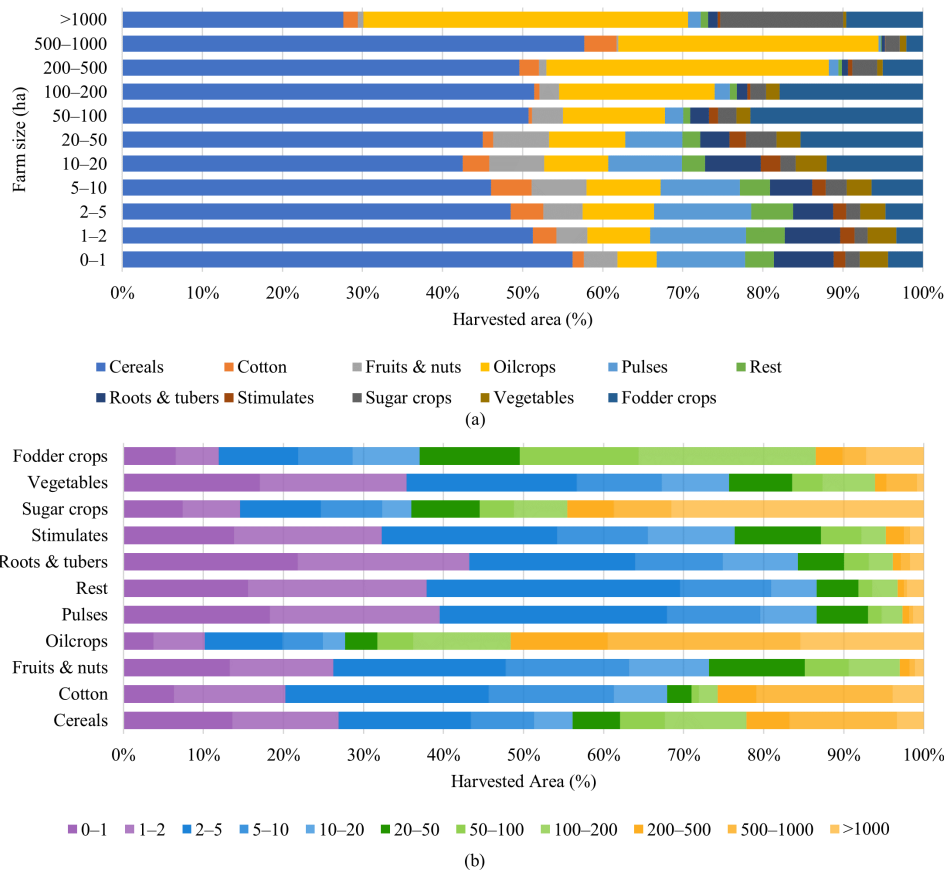
(a)

(b)

Figure 3. Harvested area of crop groups within each farm size (a) and harvested area of crop groups by farm size (b) according to ~~GAEZ based downscaled map~~our farm-size- and crop-specific harvested area dataset based on the GAEZv4 crop map. The alternative version based on SPAM2010 crop map is given in Fig. A2.

370 ### 3.1.2 Farming systems and farm size

~~Comparing between irrigated~~Besides providing farm-size- and ~~rainfed harvested area, overall, our results show~~ crop-specific harvested areas, we added information on farming systems inherited from crop maps. Statistics of farming system and farm size derived from our dataset reveal that small farms irrigate a larger relative share of their harvested area than large farms

18

(Fig. 4~~(a))~~,, Fig. 5), which ~~supports the observations of~~ aligns well with earlier ones by Ricciardi et al. (2020). ~~Plausible thresholds to differentiate~~Here, the finding is not sensitive to the threshold used to set apart small ~~and~~from large farms, whose possible values can ~~be country specific and~~ range from 1~~–~~ ha to 42 ha ~~for most countries (~~as suggested by Khalil et al~~.,~~. (2017~~:~~) and FAO~~,~~ (2017, ~~2019b). With any threshold within~~ 2019). Note, however, that this ~~range, our dataset supports previous observations. The same observation can also be found in the SPAM based downscaled map (Fig. A3). The overall observations may not~~alignment does not hold for some countries (~~see~~ Sect. 3.4~~2.2~~ for further details).

Our dataset based on the SPAM2010 crop map further divides rainfed farming systems into low-input, high-input, and subsistence rainfed systems (Fig. 4(b)). Associated statistics show a clear correlation between low-input and subsistence rainfed farming systems and smaller farm sizes. At the same time, smaller farms do not consist exclusively of low-input and subsistence rainfed farming systems, since these smaller farms also operate a sizable portion of the irrigated and high-input rainfed area (see Fig. 4(b)). Similarly, the predominant farming system type of larger farms is high-input rainfed, but high-input rainfed systems are not solely employed at larger farms.
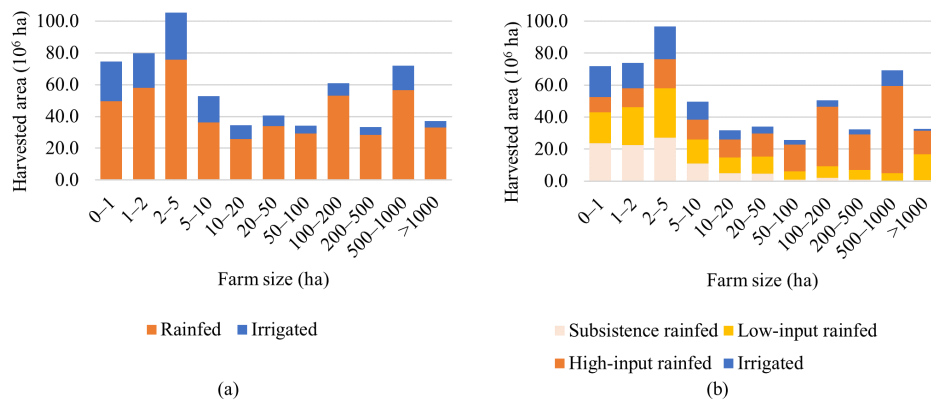

(a)

Rainfed ■ Irrigated

(b)

Subsistence rainfed ■ Low-input rainfed
■ High-input rainfed ■ Irrigated

Figure 4. ~~The overall higher~~The distribution of irrigated and rainfed farming systems per farm size according to our farm-size- and crop-specific harvested area datasets based on the GAEZv4 crop map (a) and the SPAM2010 crop map (b). Note, SPAM2010 further divides rainfed farming system into low-put, high-input, and subsistence rainfed farming systems.

To further explore irrigation ~~of small farms may be because most of small farms are in the severe water scarce regions (Fig. practices~~4(b)). ~~Here, to get water scarcity information,~~ we overlapped our ~~downscaled map~~datasets with the annual average blue water scarcity map ~~where~~ by Mekonnen and Hoekstra (2016), who classified water scarcity ~~is classified as~~in four ~~levels:~~ categories, i.e., low, moderate, significant, and severe water scarcity ~~(Mekonnen and Hoekstra, 2016; Hoekstra et al., 2012). It remains unknown whether small farms adapt to water scarcity via irrigation or irrigation of small farms.~~ This analysis also confirms an earlier finding by Ricciardi et al. (2020) that even though small farms irrigate a larger relative share of their area than ~~increase water scarcity (Grafton et al., 2018).~~Another explanation for the overall higher irrigation of small farms is the

19

~~country coverage. In our dataset, a~~ large ~~number of small~~ farms ~~are from Asia. Previous studies show, on~~ *on average*, ~~an~~
~~independent of regional water scarcity, with the percentage of irrigated area in Asian small farms being high: over 50% when~~
~~water is scarce and over 20% when water is not scarce (Ricciardi et al., 2020). This percentage is much higher than that in~~
400 ~~Europe, Central Asia, Latin America, and Sub-Saharan Africa (Ricciardi et al., 2020). Thus, the overall portion of irrigated~~
~~areas in small farms is high.~~
~~With water scarcity (moderate, significant, and severe), we observed that~~ large farms irrigate ~~to~~ a larger ~~extent~~relative share
than small farms when water is scarce (Fig. 5). Fig. ~~, which still supports the observations of Ricciardi et al. (2020) with most~~
~~thresholds to differentiate small and~~ 5 shows a relatively low irrigation share for farms >1000 ha which would undermine this
405 finding. However, the total relative irrigation share of large farms ~~within 1–42 ha. This observation does not depend on the~~
~~relatively low irrigation extent of >1000 ha farm size since the farm size >1000 ha only contributes to~~is still larger than that
of small farms, because this farm size makes up less than 4.5% of ~~water scarce area of~~ large farms. ~~The same observation can~~
~~also be found in the SPAM-based downscaled map (Fig. A3). The reason is that the water scarce area of the >1000 ha farm~~
~~size is mainly contributed by limited crops from a few regions in our dataset. In this case, the characteristics of these crops and~~
410 ~~regions have more impact on the overall relationship between water scarcity and irrigation. For example, sugarcane in São~~
~~Paulo, Brazil, is one of the main contributors to the significant and severe water scarce area of >1000 ha farm size. However,~~
~~water scarcity is not present all year round. The level of water scarcity is low from January to June, which is the tillering phase~~
~~for sugarcane. During the dry season, sugarcane is usually harvested, during which moisture in sugarcane is relatively low and~~
~~the sugar is highly concentrated (Kavats et al., 2020). This may help to explain why the large farms in this area are rainfed~~
415 ~~even though under a certain level of water scarcity. Note,~~ located in water scarce areas. Note, that the main aim of Fig. ~~4 and~~
~~Fig. A3~~5 is to compare statistics of our dataset~~s~~ with previous ~~observations~~studies instead of drawing conclusions on irrigation
levels for specific farm sizes, which may need further investigation on influencing factors and uncertainties.

(a)   (b)

Figure   5.

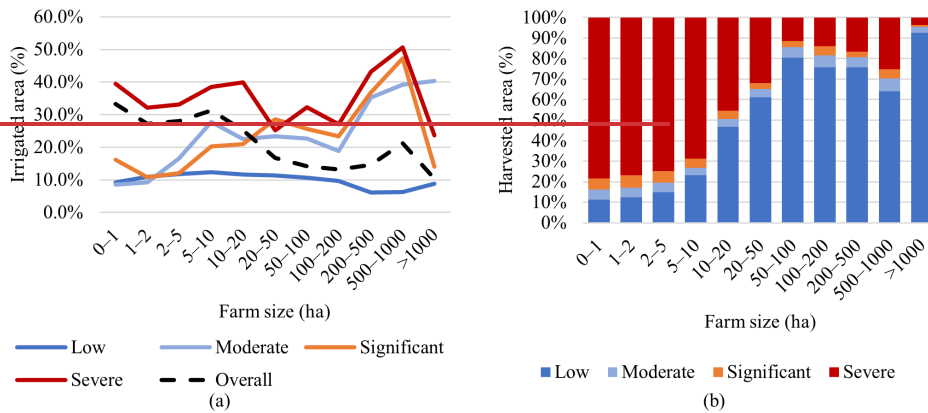(a)   (b)
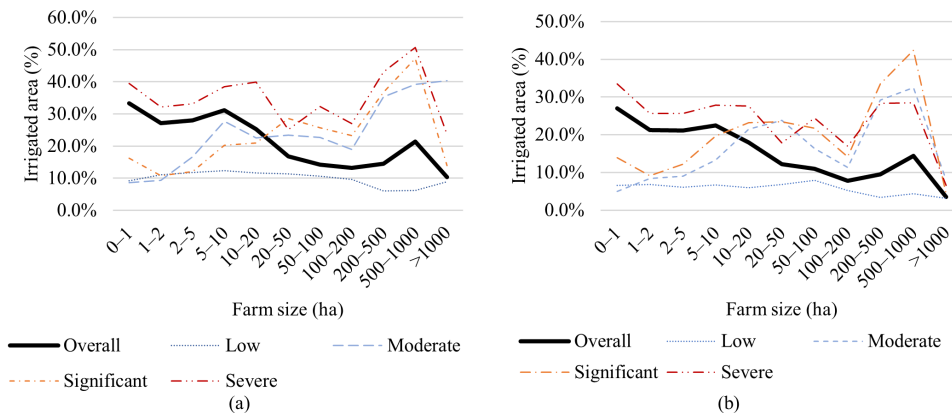
Figure 4. ~~The percentage of~~ the irrigated area by farm size according to our farm-size- and crop-specific harvested area datasets based on the GAEZv4 crop map (a) and the SPAM2010 crop map (b) under ~~each~~four blue water scarcity ~~level (a) and~~ levels ~~of water scarcity within each farm size (b) according to GAEZ based downscaled map.~~

~~SPAM2010 further divided the rainfed farming system into low- and high-input rainfed~~(WSL) by Mekonnen and ~~subsistence rainfed. With SPAM based downscaled map, our dataset~~Hoekstra (2016). Low blue WSL indicates ~~the subsistence~~blue water consumption does not exceed blue water availability; moderate WSL indicates blue water consumption is 100–150% of blue water availability; significant WSL indicates blue water consumption is 150–200% of blue water availability; ~~and~~ ~~low-input rainfed farming system is mainly operated at smaller farms, but the smaller farms do not exclusively consist of subsistence and low-input rainfed farming system: they also operate a sizable portion of the irrigated and high-input rainfed area (~~severe WSL indicates blue water consumption is ~~Fig. 5). Similarly, the main type of farming system of~~ larger ~~farms is high-input rainfed, but the high-input rainfed is far from being limited to larger farms (~~than 200% of blue water availability.~~Fig. 5).~~

420

425

430
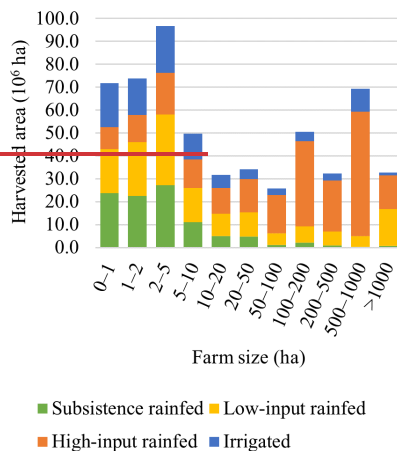
Formatted: Caption

3.2 Validation

3 ~~Validated.~~2.1 Validation with **empirical data on** farm-size-specific oil palm ~~from satellite images~~harvested area

~~Validations~~Table 2 shows that validation with farm-size-specific oil palm data ~~show~~yields a significant positive correlation ~~between our downscaled maps and the validation dataset on oil palm from satellite images~~ in most countries, for both small and large farms ~~(Table 2)~~. At larger spatial scales, the correlation becomes stronger~~. This means~~, indicating that the s~~ub~~national distributions of oil palm harvested area in our ~~downscaled maps and~~datasets are similar to those of Descals et al. (2020~~) are similar.~~). Besides the threshold of 25 ha ~~for~~to set apart small ~~and~~from large farms, we ~~also tried~~repeated the comparison with 10 ha and 50 ha ~~as~~ thresholds ~~and conducted the same comparison. We found the above conclusions on~~which resulted in similar correlations (see [S6, S7] for detailed results of these comparisons). This indicates that, at least for oil palm ~~comparison,~~ found relations are not sensitive to the choice of threshold.

~~Still, there are~~ Despite strong overall correlations, we observed some differences ~~especially in the case of~~for certain regions, particularly Costa Rica and the United Republic of Tanzania. ~~Part~~Some of ~~the above~~these differences ~~results from the~~ can be attributed to inconsistencies between harvested area according to the crop maps we used and the validation dataset. We compared total oil palm area according to the crop maps we used and the validation dataset. ~~We compared all farms area between,~~ and found that if the oil palm locations in the crop maps ~~and validation dataset, i.e., the total area of small and large farms (Table 2). We noticed that if the cropland location in crop maps differs~~differed from the validation map (no~~t~~ significant positive correlation), ~~the~~ farm-size-specific validation ~~will be~~was poor~~.~~ as well (Table 2). This ~~means~~implies that the accuracy

of our ~~estimations are~~estimates of farm-size- and crop-specific harvested area is limited by the accuracy of oil palm locations in crop maps. The (minor) differences between validation results for the GAEZv4 based dataset and the SPAM2010 based dataset can also largely be attributed to the same reason.

Table 2. ~~accuracies of farm location~~Pearson correlation coefficient between the harvested area of oil palm estimated by satellite images from Descals et al. (2020) and i) GAEZv4 crop map based farm-size- and crop-specific dataset (Gb) and ii) SPAM2010 crop map based farm-size- and crop-specific dataset (Sb), respectively, for small farms (<25 ha), large farms (≥25 ha), and all farms at various spatial resolutions. All farms compared the oil palm area in GAEZv4 and SPAM2010 crop map, whose results imply the accuracy of our estimates of farm-size- and crop-specific harvested area is limited by the accuracy of oil palm locations **in crop maps.** ~~The differences between validations results for the GAEZ based map and the SPAM based map can also be attributed to the same reason, the differences in farm location between GAEZ v4 and SPAM2010~~* p<0.005. ** p<0.001.

~~Table 2. Pearson correlation coefficient of the harvested area between oil palms from satellite images Descals et al. (2020) and GAEZ based downscaled map and SPAM based downscaled map respectively for small farms, large farms and all farms. Since all farms results do not distinguish farm size, they indicate the differences in oil palm spatial distribution between Descals et al. (2020) and crop map datasets (GAEZ v4 and SPAM2010).~~

		Small farms			Large farms			All farms		
		5 arcmin	15 arcmin	25 arcmin	5 arcmin	15 arcmin	25 arcmin	5 arcmin	15 arcmin	25 arcmin
Colombia	Gb~~GA EZ based~~	0.177*	0.313**	0.397**	0.112**	0.238**	0.334**	0.232**	0.374**	0.465**
	Sb~~SP AM based~~	0.218**	0.547**	0.684**	0.385**	0.620**	0.701**	0.409**	0.652**	0.729**
Costa Rica	Gb~~GA EZ based~~	0.086	0.183**	0.215**	-0.012	-0.074	-0.144**	0.032	0.001	-0.043
	Sb~~SP AM based~~	0.836**	0.944**	0.971**	0.771**	0.891**	0.925**	0.877**	0.925**	0.929**
Brazil	Gb~~GA EZ based~~	0.245**	0.396**	0.483**	0.177**	0.258**	0.271**	0.326**	0.398**	0.423**
	Sb~~SP AM based~~	0.133**	0.190**	0.248**	0.087**	0.091**	0.084**	0.148**	0.154**	0.156**

United republic of Tanzania	GbGA EZ based	0.01	-0.109*	-0.202**	-0.011	-0.039	-0.063	0.022	-0.115*	-0.218**
	SbSP AM based	0.024	0.025	0.069				0.022	0.014	0.065
Peru	GbGA EZ based	0.172**	0.350**	0.438**	0.024	0.139**	0.237**	0.111**	0.263**	0.363**
	SbSP AM based	0.367**	0.389**	0.429**	0.141**	0.216**	0.240**	0.302**	0.395**	0.436**

$* p<0.005$
$** p<0.001$

3.4 Validated 2.2 Validation with empirical data on farm-size-specific irrigation from household surveys estimates

Our results also have positive correlations with household surveys in farm size specific irrigation for the GAEZ based map
(Fig. 6(a)) and the SPAM based map (shows Fig. 6(b)) respectively. This means that our downscaled maps datasets are quite consistent with validation data empirical data on farm-size-specific irrigation estimates in terms of country-level farm-size-specific irrigation. Detailed results show that the maps could percentage of irrigated area. More detailed results in [S8] further illustrate how our datasets capture the higher percentage of irrigated areas in as indicated by the household surveys in both small or and large farms in most countries along the indications of household surveys [S5].
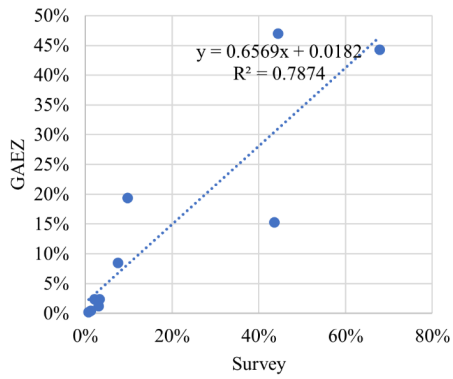
From the validations. However, we noticed also found that our downscaled maps datasets systematically underestimate the extent percentage of the irrigated area compared with respect to these same household surveys, both for the GAEZ based map in our GAEZv4 and the SPAM SPAM2010 based datasets of harvested areas. Fig. map. If 6(c) and 6(d) show that these underestimations are still present if we compare the percentages of irrigated area for all farms from the datasets, we find these underestimations are still there (Fig. 6(c), (d)). This means the crop maps. This systematic underestimation may come from the therefore be explained by different measurements of irrigated area and cultivated area in the validation dataset and datasets of compared to the crop maps.
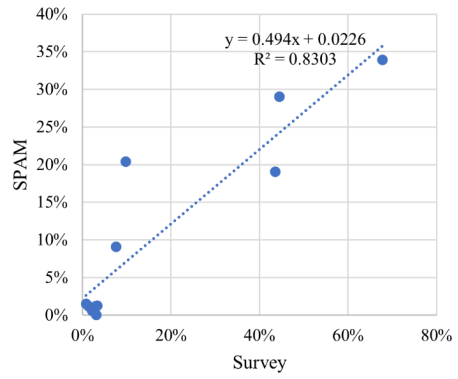
y = 0.6704x + 0.0246
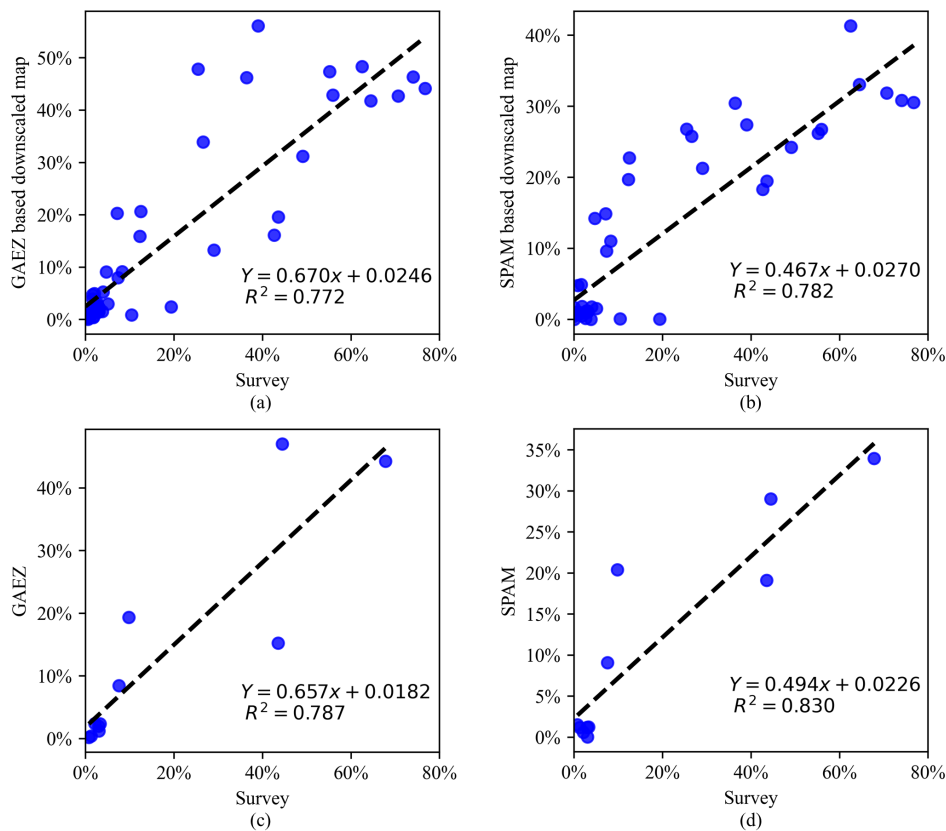R² = 0.7717

(a)

y = 0.4672x + 0.027
R² = 0.7819

(b)

y = 0.6569x + 0.0182
R² = 0.7874

(c)

y = 0.494x + 0.0226
R² = 0.8303
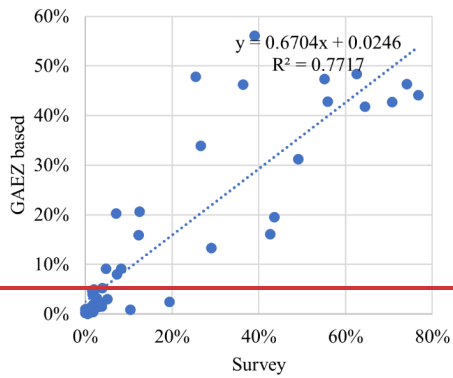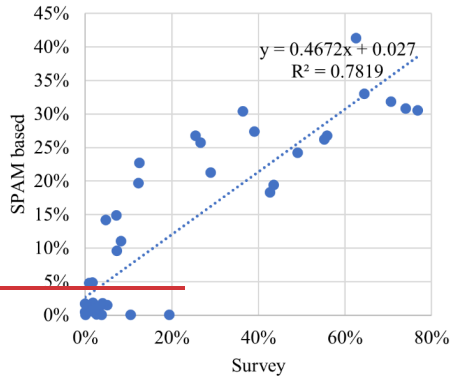
(d)

Figure 6. Correlations on the farm-~~size-~~specific irrigated area (% of total harvested area per farm size) between household survey ~~and GAEZ based downscaled map~~ data from FAO RuLIS (Rural Livelihoods Information System) database (FAO, 2021) and our GAEZv4 based (a) and ~~SPAM based downscaled map (b) for eleven countries. The~~SPAM2010 based, farm-size- and crop-specific datasets of harvested area (b), and correlations on the irrigated area of all farms (% of the total harvested area) between household survey data from FAO RuLIS (Rural Livelihoods Information System) database (FAO, 2021) and ~~GAEZ v4~~GAEZv4 (c) and SPAM2010 (d~~) are also provided.~~), all for eleven countries.

3.2.3 Validation through comparison with other studies

Finally, we compared our high-resolution farm-size- and crop-type specific harvested area datasets with FAOSTAT, whose structured data contains farm size structures of 44 overlapping countries for the years 2000 and 2010 (FAO,

26

2022). Results show that (non-crop-specific) farm size structures of our datasets are similar to FAOSTAT structure data for most countries. Fig. 3.5 Compared with previous studies mapping farm sizes

495 ~~Compared with Lowder's dataset (Lowder et al., 2016) on the percentage of harvested area operated by each farm size, we observed positive correlations for GAEZ based map (Fig. 7(a)) and SPAM based map (Fig. 7(b)). This means at the country level, the number of farms for each farm size is similar to Lowder's dataset ([S6] for details). There are still differences between our downscaled map and Lowder's dataset. For example, Lowder's dataset estimate 78.5% of harvested area is under the farm size 50–100 ha in Bulgaria while our downscaled maps give around 5%. However, our downscaled maps estimate around 80%~~

500 ~~of harvested are in the under farm size 100–200 ha, while Lowder's dataset gives zero. In this case, our downscaled maps are still similar to Lowder's dataset since both indicate large farms are the major farm size in the country even though it was not reflected in the correlations. These differences may be attributed to Lowder's dataset being developed for the year 2000, which is ten years earlier than our focus. Farm sizes may change during ten years in some countries. Besides reporting time, these differences may also be attributed to how different datasets harmonize farm size. The farm size classes collected from the local~~

505 ~~agriculture census usually need to be harmonized into a classification system. Different datasets may have their own choice during this process. This may lead to some differences shown in the comparison, especially when the major farm sizes are similar but not the same.~~

(a) (b)
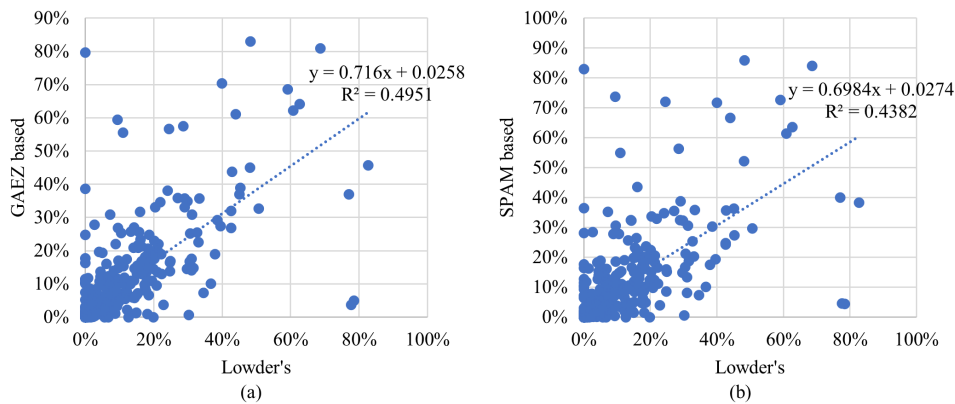
7 and Fig. A3 show the large similarities of farm size structures of 28 countries for 2010, while of the remaining 16 countries,

510 farm size structures of Brazil, Czechia, Ethiopia, Germany, Greece, Poland, and Portugal show good correspondence for 2000. The latter also implies these estimates are similar to the dataset by Lowder et al. (2016).

Not all countries' farm size structure corresponds well between the datasets. Farm size structure according to our datasets for Albania, for example, lies in between the FAOSTAT data for 2000 and 2010, and our datasets farm size structures of Costa Rica, Lithuania, and Mexico also deviate slightly from the FAOSTAT structure data. One explanation for such differences

515 could come from how different datasets harmonize collected data into a farm size classification system. For example, if only farm sizes >100 ha are reported, areas could be classified into farm sizes 100–200 ha or be redistributed to farm sizes 100–200 ha, 200–500 ha, and so on. However, the farm size structure of our datasets is inherited from the dataset by Ricciardi et al. (2018a, b), which in turn was based on highly similar local agricultural census and household surveys which FAOSTAT likewise drew from.

520 While decent overall correspondence between our datasets and either FAOSTAT 2000 or 2010 data might be sufficient grounds to validate our estimates on farm size structure, and particularly correspondence to 2010 being the reference year for our datasets, it should be noted that farm size structures of several countries changed significantly between 2000 and 2010, e.g. Bulgaria and Germany, a period of just 10 years. The FAO themselves also indicate that the robustness of their 2010 estimates is fragile, in part due to significant usage of interpolation (FAO, 2022). Moreover, for 5 of the 44 analyzed countries (i.e.,

525 Burkina Faso, Colombia, Peru, and Russian Federation), it remains unclear what causes these differences.

Comparing our datasets~~Figure 7. Correlations on the percentage of harvested area operated by each farm size between Lowder's dataset (Lowder et al., 2016) and GAEZ based downscaled map (a) and SPAM based downscaled map (b) for 37 countries and 11 farm sizes.~~

~~Compared~~ with the dataset by Mehrabi et al. (2020), Fig. 8 shows that the ~~same pattern~~patterns of ~~the~~ spatial distributions of

530 dominant farm ~~size could be observed in~~sizes are similar across both datasets. For the ~~Mehrabi's~~farm-size- and crop-specific dataset ~~(Fig. 8(a)), the GAEZ downscaled map (Fig. 8(b)), and SPAM~~ based ~~downscaled map (Fig. 8(c)). Overall, for GAEZ based downscaled~~on the GAEZv4 crop map, 54.~~2~~% of grid cells' dominant farm sizes ~~are similar~~correspond to ~~that~~those in ~~Mehrabi's~~the dataset~~, 27.5~~ by Mehrabi et al. (2020), while 28% are larger, and 18.~~3~~% are smaller~~; for SPAM based downscaled map, 52.8%~~. For the SPAM2010 based counterpart, 53% of grid cells' dominant farm sizes are similar~~, 26.0~~ to the dataset by

535 Mehrabi et al. (2020), while 26% are larger, and 21.~~2~~% are smaller ~~([S7 for details). These differences may be partly explained by the above~~. Here, similar means the farm size in our datasets is the same or next to the farm size in the dataset by Mehrabi et al. (2020). [S9] provides a more detailed analysis of this comparison~~ with Lowder's dataset since Mehrabi's dataset has~~. As shown in Fig. 7, there are still differences between our datasets and the dataset by Lowder et al. (2016) (FAOSTAT structure data of 2000). These differences can also be seen in the comparison with the dataset by Mehrabi et al. (2020) since the dataset

540 by Mehrabi et al. (2020) keeps the same country level farm size distribution as ~~Lowder's dataset. Some differences could also be attributed to the comparison of dominant farm size: the dominant farm size in Mehrabi's dataset may be the second dominant farm size in our downscaled map. The~~ the dataset by Lowder et al. (2016). Note, that the comparison of dominant farm size may magnify the differences in farm size structure between our datasets and the dataset by Mehrabi et al. (2020) since the dominant farm size in the dataset by Mehrabi et al. (2020) may be the second-dominant farm size in our datasets.
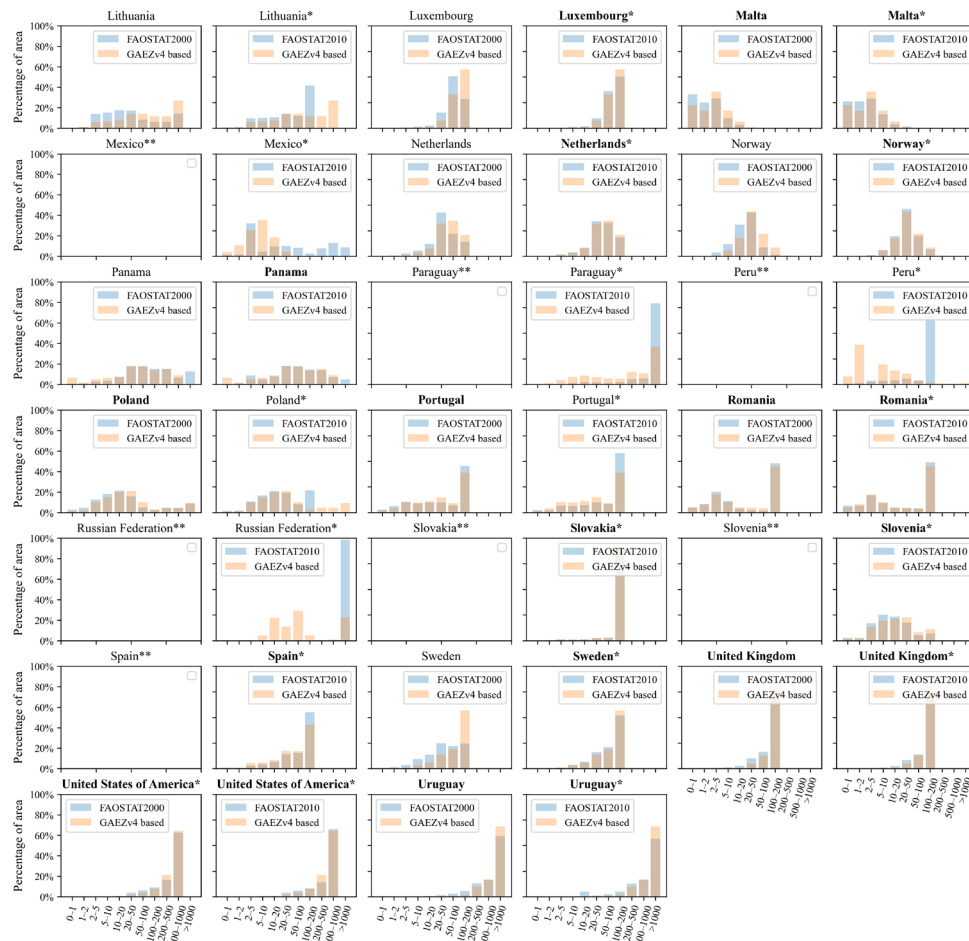
(Continued)

Figure 7. ~~difference in estimating the overall farm sizes. Since Mehrabi's~~Comparison of the percentage of total crop area operated by each farm size (non-crop-specific farm size structure) between FAOSTAT structural data for the year 2000 and 2010 (FAO, 2022) and our farm-size- and crop-specific dataset based on the GAEZv4 crop map. Bold font country titles indicate that farm size structures in FAOSTAT are similar to our **dataset** ~~only include dominant farm size, it~~. Note that for the year 2000, farm size structure from FAOSTAT structural data **is** ~~not clear that how the difference would be estimating the overall farm sizes~~the same with Lowder et al. (2016) except for one country [S5]. Only the countries covered by our dataset and FAOSTAT are shown. The alternative version based on SPAM2010 crop map is given in Fig. A3. * FAOSTAT provides (part of) the structural data by

Formatted: Caption

interpolating other reported data, not directly from countries' official reports. ** FAOSTAT provides no farm size structural data of the year 2000 or 2010 for comparison.
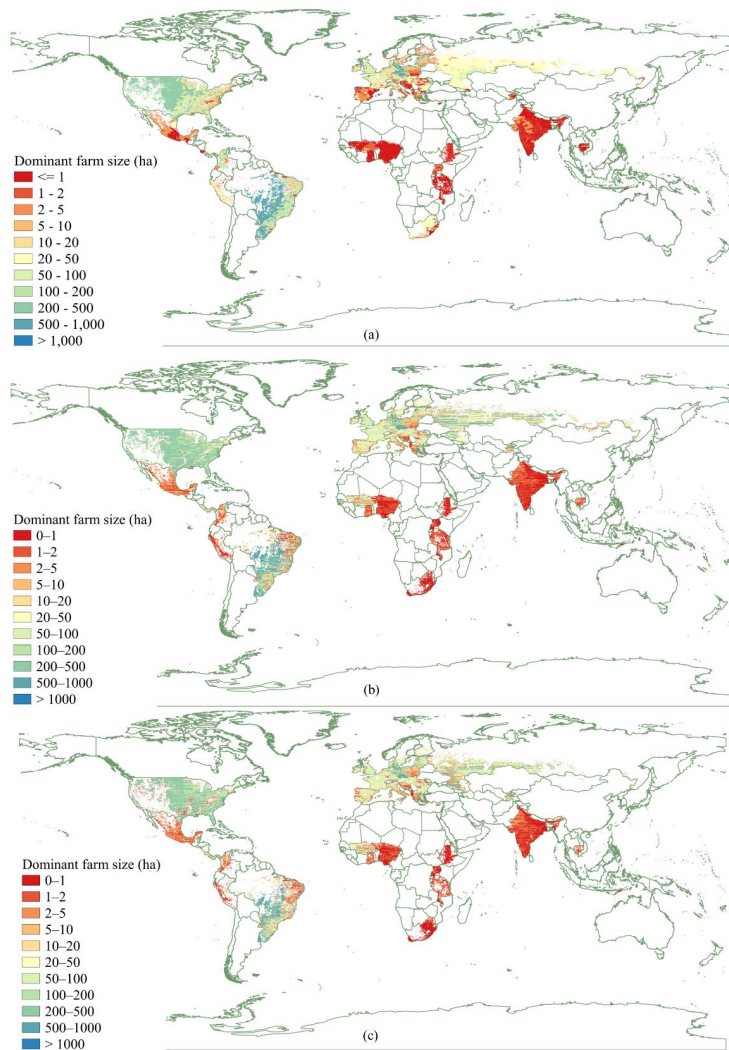
560

Figure 8. Dominant farm size according to ~~Mehrabi's dataset (Mehrabi et al., 2020) (a), GAEZ based downscaled map (b) and SPAM based downscaled map (c). We only show the cells from Mehrabi's~~the **dataset** ~~where our downscaled maps have estimations~~by

4 Discussion

4.1 Potential explanations for irrigation and farm size

Our datasets confirm findings by previous studies that smaller farms have a higher relative irrigation share compared to larger farms. This seems to be the case because relatively many of the small farms are located in severe water scarce regions, which would require them to irrigate more and more often to grow their crops (Fig. 9). However, it remains unclear whether small farms adapt to water scarcity via irrigation or that irrigation practices of small farms increase water scarcity (Grafton et al., 2018). Another explanation relates to the farm size structures between countries. Asian countries are home to the majority of small farms, and previous studies have shown that, on average, the relative share of irrigated area on Asian small farms is indeed much higher than in other countries, regardless regional water scarcity levels (Ricciardi et al., 2020).

(a)   (b)
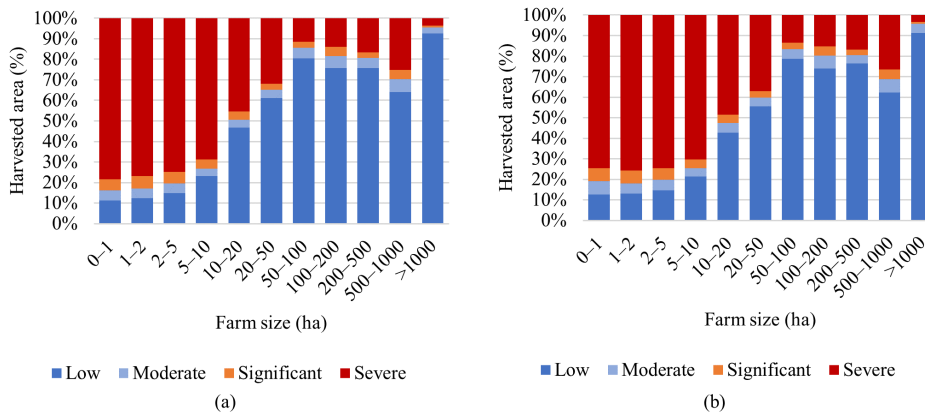
Figure 9. Blue water scarcity levels (WSL) within each farm size according to our farm-size- and crop-specific harvested area dataset based on the GAEZv4 crop map (a) and the SPAM2010 crop map (b) under four blue water scarcity levels (WSL) by Mekonnen and Hoekstra (2016). Low blue WSL indicates blue water consumption does not exceed blue water availability; moderate WSL indicates blue water consumption is 100–150% of blue water availability; significant WSL indicates blue water consumption is 150–200% of blue water availability; and severe WSL indicates blue water consumption is larger than 200% of blue water availability.

The irrigation of >1000 ha farm size shown by our datasets is relatively low, which could be explained by the regional climate and crop characteristics. Sugarcane in São Paulo, Brazil, is one of the main contributors to the significant and severe water scarce area of >1000 ha farm size. In these regions, water scarcity is not present all year round. The level of water scarcity is low from January to June, which is the tillering phase for sugarcane. Sugarcane is usually harvested during the dry season,

33

585 desirably, during which moisture in sugarcane is relatively low and the sugar is highly concentrated (Kavats et al., 2020). This may help to explain why the large farms in this area are rainfed even though under a certain level of water scarcity.

4.2 Uncertainties

We ~~explicitly consider the~~ hypothesized that uncertainties in crop maps ~~by developing~~ might propagate to and influence uncertainties in our gridded datasets. Therefore, we developed two ~~separately downscaled~~gridded datasets based on two
590 different crop maps ~~based on two crop maps, GAEZ v4~~, i.e., GAEZv4 and SPAM2010. From the results and validation~~s~~, we observed some differences in ~~the~~ crop distribution ~~between the two crop maps,~~. especially at the grid cell level. ~~This reflects the~~ These differences reflect uncertainties in farmland location~~. It affects~~ and affected the spatial validations on both farm-size~~-~~specific oil palm and ~~the~~ dominant farm size ~~distribution. However,~~ distributions. At the same time, these uncertainties at ~~the~~ grid cell level have a limited impact on country level ~~results~~statistics and ~~validations which~~validation, as can be seen
595 ~~from~~in Fig. 3, Fig. 4~~5~~, Fig. A2~~6~~, and Fig. A3~~A2~~.
~~Uncertainty in~~Differences—and therefore uncertainties—related to farming systems are more pronounced between the two crop maps~~is more pronounced for farming systems. From~~, also at country level. Fig. 6 and [~~S5~~ we could see the SPAM~~S8]~~ show that our SPAM2010 based ~~downscaled map has a~~dataset yields lower irrigation ratio~~s~~ than ~~GAEZ~~that based ~~downscaled map.on~~ GAEZv4. This is ~~because~~likely the consequence of SPAM2010 defin~~es~~ing irrigation ~~according to~~as the actually
600 irrigated area ~~and GAEZ v4~~, whereas GAEZv4 defines irrigation by the area that is equipped with fully irrigation facilities. ~~The lower irrigation ratio in SPAM2010 does not affect the conclusions and validations drawn~~Despite these differences, however, findings of the overall relative irrigation share being higher of smaller farms and higher absolute irrigation of larger farms under elevated levels of water scarcity are supported by our datasets based on both crop maps.
The uncertainties in the crop maps also affect how we downscaled the dataset by Ricciardi et al. (2018a, b), the core source of
605 our datasets. It occurred that crops could be found in the dataset by Ricciardi et al. (2018a, b) for a given administrative unit but not in the crop maps, or vice versa. The consequence of these inconsistencies was that 23.3% and 21.6% of the crop area in the dataset by Ricciardi et al. (2018a, b) could not be downscaled, respectively because the GAEZv4 or the SPAM2010 crop map indicated no crops were grown in those locations. Vice versa, 17.8% and 12.4% of the harvested area in the GAEZv4 and SPAM2010 crop maps, respectively, could not directly be assigned a farm size due to absent records in the dataset by Ricciardi
610 et al. (2018a, b). Although these are substantial percentages of crop areas, our validation did not detect any peculiarities in outcomes attributable to these inconsistencies. Developing more accurate crop maps should reduce a substantial bit of the abovementioned uncertainties in the future.
Beside uncertainties propagated from ~~the GAEZ based map; for example, the finding of overall higher irrigation of smaller farms is robust under this uncertainty, and so is the observation on higher irrigation of larger farms under the elevated level of~~
615 ~~water scarcity.~~
~~Some~~input data, new uncertainties are introduced ~~by~~through our pre-processing ~~and constraints relaxation during the solving processes. When~~procedures. In estimating crop-specific farm size structures using ~~Ricciardi's datasets, around 12% of them~~the

dataset by Ricciardi et al. (2018a, b), ~12% of our final estimates were based on crop production instead of crop area. According to Ricciardi et al. (2018a, b), the introduced uncertainties are limited when using crop production. ~~In terms of uncertainties~~In addition, the year of the source data of Ricciardi et al. (2018a, b) ranges from 2001 to 2015 with median year of 2013, the transient nature of farm sizes, particularly in developing countries, may not be captured when it is used for the year of 2010.

The way we defined and apply constraints during the optimization process also introduced ~~by constraints relaxation,~~new uncertainties. Solving for ~~GAEZ (SPAM)~~the GAEZv4 and the SPAM2010 based ~~map~~datasets, we ~~solved~~performed 7381 ~~(and 6017)~~ optimizations~~. GAEZ v4 and SPAM2010 based downscaling solved different~~, respectively. Differences in total number of optimizations ~~because of the different~~can be explained by differences in cropland extent ~~which affect the~~underlying both crop maps. Of their total number of ~~grid cells to be allocated. Among all the~~optimizations, 4378 ~~(and 3671)~~ needed to be relaxed using ~~an~~ elastic factor ~~of~~ 0.125 or smaller (Eq. ~~(5));~~(5)), for the respective crop maps, while 239 ~~(and 203)~~ needed to be further relaxed by removing some of the minimum area constraints (Eq. (6) – (9)). ~~Only the~~The latter relaxation of minimum area ~~constraint will introduce additional~~constraints introduced inconsistencies with the ~~datasets used. This means the constraints relaxation introduce additional~~source dominant field size distribution, which further adds uncertainties ~~among~~to our datasets. This affected ~3% of ~~the~~our total calculations.

In ~~addition, we might allocate~~the optimization process, it further occurred that crop area needed to be allocated to a farm size that ~~is~~was not included in ~~Ricciardi's~~the dataset~~.~~ by Ricciardi et al. (2018a, b). This ~~only~~happened ~~when~~in cases where both the crop and part of the eleven farm sizes ~~a~~were included in ~~Ricciardi's~~the dataset ~~but~~by Ricciardi et al. (2018a, b), yet meeting the minimum area constraints ~~requires~~required introducing an additional farm size for the crop at hand. In ~~this case,~~such cases, we still ensured the 10% maximum relative difference with ~~Ricciardi's~~the dataset ~~is still ensured for the available~~by Ricciardi et al. (2018a, b) to ensure the overall farm size~~. Only~~ structures. This uncertainty was introduced for ~0.1% ~~(and 5.0%)~~% of ~~allocated~~harvested area ~~is in this case~~ for ~~GAEZ (SPAM)~~the GAEZv4 and SPAM2010 based ~~downscaled map~~farm-size- and crop-specific datasets, respectively.

~~More uncertainties in the downscaled maps may come from used datasets. Since Ricciardi's dataset was not developed for 2010, farm size may change a lot in some developing countries. This put some uncertainties in our results since we relied on it to estimate farm size structure. The uncertainties in the crop map affect how we downscaled Ricciardi's dataset. Some crops can be found in Ricciardi's dataset for an administrative unit but not in crop map, or vice versa. This means that, on the one hand, 23.3% (21.6%) of the crop area in Ricciardi's dataset was not downscaled because the GAEZ v4 (SPAM2010) crop map indicates no crop. On the other hand, 17.8% (12.4%) of the harvested area in the GAEZ v4 (SPAM2010) crop map was not allocated a farm size because Ricciardi's dataset has no relevant records. These uncertainties may have affected the allocated area in the downscaled maps, but according to validations, they are not high enough to make the downscaled maps lose the utilities. Highly accurate crop maps will reduce this part of uncertainties in the future.~~

~~Despite~~ Finally, despite the uncertainties at the grid cell level, the used datasets and ~~the downscaled maps~~our datasets were found to be more reliable at the country level. For example, the two crop maps were developed by downscaling the agriculture

census at the (sub)national level. Collected agriculture census and social-ecological factors considered during downscaling may lead to some differences at the grid cell level in the two crop maps, while they were all adjusted to the country level data from FAOSTAT ~~(FAO, 2019a).~~. The dominant field size distribution is also uncertain at the grid cell level which was estimated

655 by spatial interpolating of training samples. The uncertainty will decrease when the focus is on the regional level (Lesiv et al., 2019). Validations also show well consistencies with country level observations. ~~Therefore, future uses of our downscaled map are more confident at the country level than grid cell level.~~ Using GAEZv4 based ~~map~~ and SPAM2010 based ~~map~~datasets at the same time helps to reduce uncertainties at the grid cell level.

4.~~2~~3 Limitations

660 With the ambition to map ~~global~~ simultaneously farm-size- and crop-specific harvested area, we were ~~only~~ able to cover 56 countries ~~due to data availability, though this reflected~~based on state-of-the-art recent datasets (e.g. Ricciardi et al. (2018a, b), Lesiv et al. (2019), and Kim et al. (2021)). Although these countries reflect about half of the global cropland. ~~Farm~~, the remaining countries could not be included due to lacking data availability. Particularly farm-size-specific data is scarce ~~and~~or not publicly available ~~in some countries. The datasets we used, like Ricciardi et al. (2018b) and Lesiv et al. (2019), are the~~

665 ~~currently best-available datasets on farm or field sizes (Kim et al., 2021). Data~~ for most of the excluded countries, but across-the-board data availability is the main obstacle ~~to~~in creating a dataset with global ~~map. The development of~~coverage. Approaches based on deep learning and remote sensing, similar to what Descals et al. (2020) did to obtain their oil palm dataset with which we validated some of our findings, may ~~help to map~~prove promising alternatives to mapping the global farm-size- and crop-specific harvested ~~are in another way, like the farm size specific oil palm in Descals et al. (2020). The~~. However, the

670 lack of farm size training samples and the enormous computational requirements are ~~the main~~still challenges for ~~deep learning and remote sensing.~~such approaches (Descals et al., 2020).

Our estimations are based on planted crop and harvested area, ~~which~~ that is ~~static for~~ representative of the year 2010. Farmers' choice of crop will change along with climate, market demands, and ~~so on. Current downscaled maps could only~~ many other factors. While our gridded datasets provide a robust baseline ~~for the distributions of small and large farms. It remains~~

675 ~~challenging~~, it would be insightful to describe ~~the~~developments over time. However, capturing dynamics of harvested area under changing ~~environment.~~

~~The~~ conditions and environments, particularly dynamic in developing countries (Giller et al., 2021), requires even more additional data. Still, our datasets may be updated in the future ~~updates~~for additional years, since many of ~~our downscaled maps rely on~~ the ~~updates of our used~~underlying datasets. ~~Fortunately~~, including GAEZ ~~v4, SPAM2010~~, SPAM, and the

680 cropland extent map ~~have regular update plans according to their document.~~by Latham et al. (2014) and Lu et al. (2020) are planned to be regularly updated. The dominant field size distribution ~~was also~~by Lesiv et al. (2019) has already been updated since ~~the~~its first publication and ~~may have~~announced more updates in the future. ~~Ricciardi's dataset may~~Ricciardi et al. (2018a, b) did not ~~have updated~~share plans to update their dataset (yet), but it could be ~~updated~~done using ~~the~~particularly data from the World Programme for the Census of Agriculture (FAO, ~~2020b~~2015) and EUROSTAT (~EUROSTAT, 2021). ~~Any~~We

685 developed our model and code such that any updates and extensions of ~~Ricciardi's dataset from other data sources~~ in the future are ~~compliable with current model and code~~relatively easily incorporated.

4.~~3~~4 Suggestions on developing farm-size- and crop-specific production dataset

Crop production of small farm~~s~~ is one of the main concerns of the Target 2.3 (double the agricultural productivity and the incomes of small-scale food producers) of SDG 2 (Zero hunger~~). Developing~~) (UNSD, 2022). It would therefore be a major
690 achievement if we could develop farm-size-specific ~~maps on~~agricultural production ~~may be one of the applications of our dataset that directly benefits from the additional dimensionality achieved.~~dataset in support of this Target. However, compared to harvested areas, an empirical farm-size-specific dataset on production or yield is even more scarce. ~~The data on production or yield of farm sizes is available for a limited number of countries, but those countries are not always the most vulnerable in terms of food insecurity. Thus, such datasets would require estimating the~~Thus, developing a farm-size- and crop-specific
695 production dataset requires additional modeling and our datasets could readily be used as input for such development. ~~Developing a farm-size- and crop-specific~~ production ~~or yield based on additional models.~~
~~Current~~dataset requires unpacking the various factors that impact yield and are known or expected to correlate with farm size as recent studies show that the relationship between farm size and crop production ~~or yield~~ is indirect and complex ~~(,~~ cf. Muyanga and Jayne (2019) and Iizumi et al. (2021~~))~~). ~~Many factors contribute to this relationship, including but not limited to~~
700 ~~crop types, fertilizer input,~~). Some factors could be unpacked directly for farm sizes with our datasets. For example, one could overlap our datasets with the soil and climate~~, and~~ datasets to estimate soil and climate production conditions~~. The farm size itself does not directly affect yield, but~~ for each farm size ~~often correlates with factors that affect yield. So, estimating crop yield for different farm sizes requires first unpacking the factors that directly impact yield and correlate with farm sizes. For environmental~~. Other factors ~~like soil conditions and climate, this could be achieved by overlapping our dataset with the soil~~
705 ~~and climate database. Agricultural~~could be unpacked indirectly via agricultural production system, e.g. agricultural management and input factors~~, like fertilizer input, could be inferred from the agricultural production system data~~. Specifying agricultural management and input factors according to farming systems could help to first evaluate crop yield for different farming systems, and then allocate the yield back to farm sizes according to ~~their proportion in each farming system. Such an approach would rely on the assumption that agricultural management practices of different farming systems do not depend on~~
710 ~~farm size. Reliable estimations of yield for different farming systems could be either derived from SPAM2010 and GAEZ v4 or based on crop modeling~~farm size structure in each farming system. With unpacked factors, one could estimate the farm-size- and crop-specific production with our harvested area as input using crop models as well as GAEZv4 and SPAM2010.

5 Code and data availability

The code, source data, and ~~the simultaneously~~resulting farm-size- and crop-specific harvested area~~, including the GAEZ based~~
715 ~~downscaled map and SPAM based downscaled map,~~ datasets are ~~open-access, free, and~~freely available via a Creative

37

Commons Attribution 4.0 International license at https://doi.org/10.5281/zenodo.~~5747616~~6976249 (Su et al., 2022). The ~~downscaled maps~~resulting datasets are available in *.csv ~~files~~and *.nc (netCDF) for each crop and farming system. ~~Each *.csv~~ ~~file provides the grid cell index, administrative unit index,~~For each crop ~~name, ,~~ farming system, and farm size, we provide gridded harvested area~~, and x and y coordinates~~ in the ~~projection~~coordinate Systems of ~~WGS84~~EPSG:4326 - WGS 84. Gridded summaries over crops and farming systems are also available.

6 Conclusions

This study presents ~~a~~ 5-arcmin gridded simultaneously farm~~-~~size- and crop-specific dataset~~s~~ of harvested area for 56 countries. ~~We downscaled the best-available datasets, Ricciardi et al. (2018b) which collected direct reports of farm size and crop area, by using the latest~~ The datasets ~~on~~are based on various state-of-the-art and recent datasets on farm-size- and/or crop-specific land use, cropland extent, and dominant field size distribution. ~~We explicitly addressed the uncertainty in crop maps by using two crop maps separately during downscaling. The downscaled maps are well-consistent with observations on farm size specific oil palm cultivation from satellite images and farm size specific irrigation from household surveys. Our downscaled maps~~The resulting datasets show ~~the planted crops~~strong consistency along multiple variables validated against multiple empirical ~~and irrigation differ among farm sizes which support previous findings. We observed uncertainties in the maps produced at the grid cell level but found country level conclusions to be robust to grid cell level uncertainties, including the uncertainties from crop maps.~~

~~Intended future updates will increase the spatial~~ published sources. While our high-resolution dataset fills a part of the data gap, lacking data availability is still hampering the development of dynamic datasets with full global coverage. ~~Our simultaneously farm size~~Nevertheless, we are confident that our current datasets will prove to be a useful tool for improving our understanding of differences between small- and ~~crop-specific dataset will facilitate studies to explicitly incorporate farm size into global agriculture, water resources, and~~ large-scale farms, e.g. in terms of climate change ~~studies~~adaptation and mitigation strategies, water consumption patterns, and contribution to (local) food security and SDG 2.

Appendices
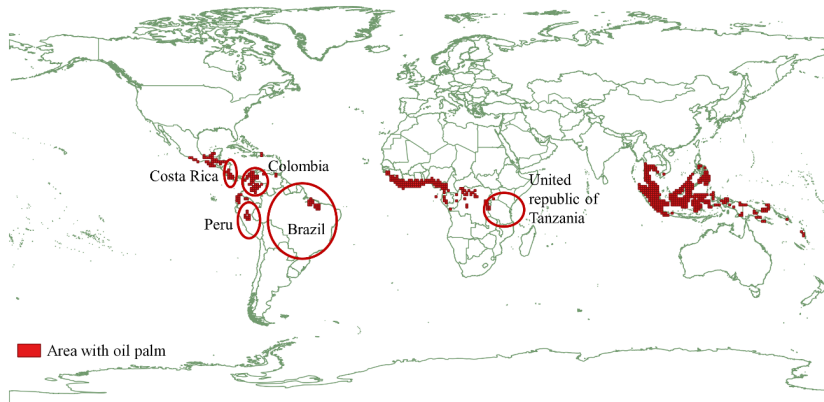
740

Figure A1. The global distribution of oil palm~~s~~ according to Descals et al. (2020~~) and the~~). The five countries ~~to validate~~for which we validated our ~~downscaled maps~~datasets are circled in red.

(a)

(b)
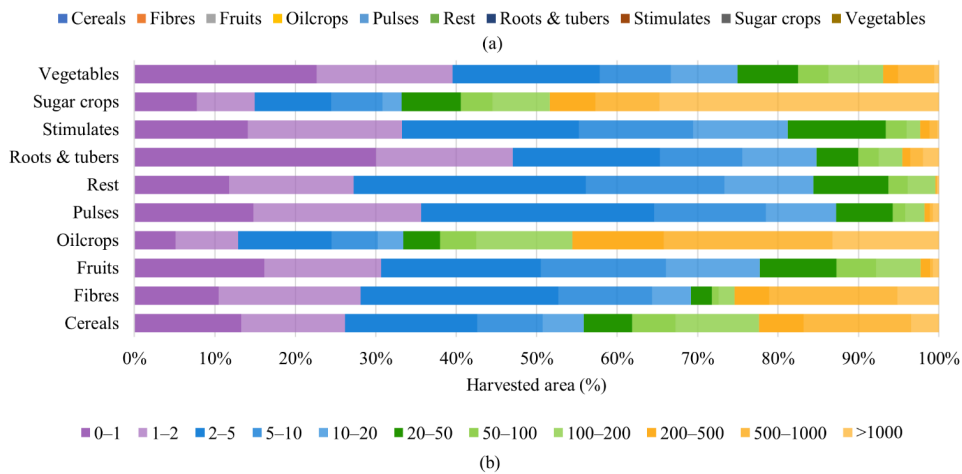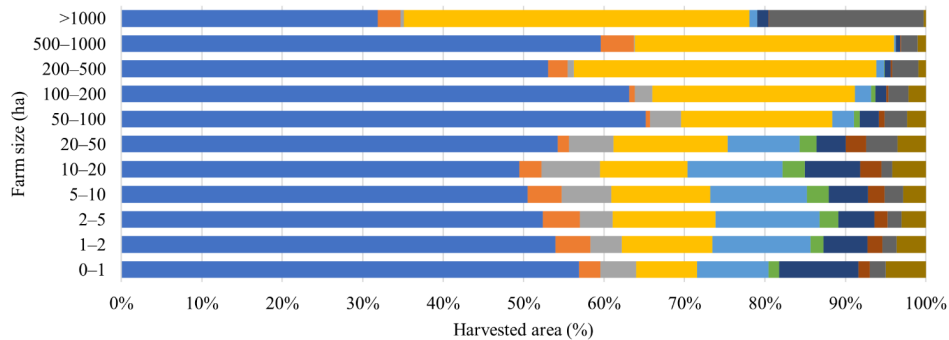
Figure A2. Harvested area of crop groups within each farm size (a) and harvested area of crop groups by farm size (b) according to ~~SPAM-based downscaled map~~our farm-size- and crop-specific harvested area dataset based on the SPAM2010 crop map. The alternative version based on GAEZv4 crop map is given in Fig. 3.
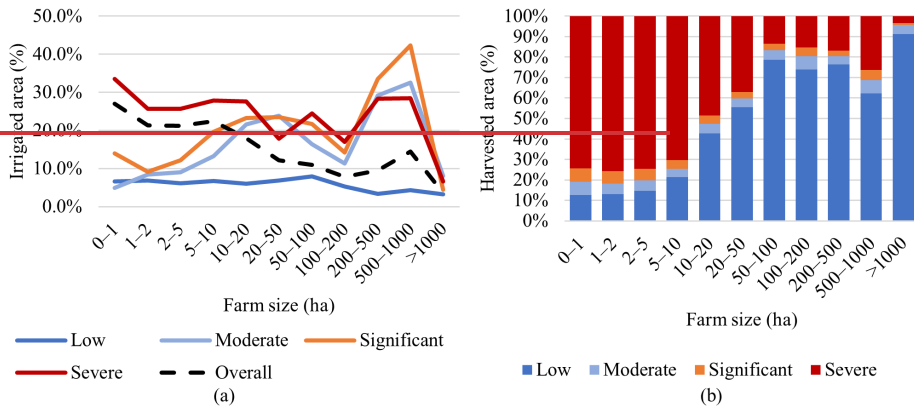
Figure A3. The percentage of the irrigated area by farm size under each water scarcity level (a) and levels of water scarcity within each farm size (b) according to SPAM based downscaled map.

750

41

The figure shows a grid of small bar charts (histograms) comparing FAOSTAT2000, FAOSTAT2010, and SPAM2010 based data across percentage of area for multiple countries. Countries shown include: Albania, Albania*, Austria, Austria*, Belgium, Belgium*, Brazil, Brazil, Bulgaria, Bulgaria*, Burkina Faso**, Burkina Faso*, Cambodia**, Cambodia*, Colombia*, Colombia*, Costa Rica**, Costa Rica*, Croatia, Croatia*, Cyprus*, Cyprus*, Czechia, Czechia*, Denmark, Denmark*, Estonia, Estonia*, Ethiopia, Ethiopia**, Finland, Finland*, France, France*, Germany, Germany*, Greece, Greece*, Hungary**, Hungary*, India, India*, Ireland, Ireland*, Italy, Italy*, Latvia, Latvia*.

The y-axis for each chart shows "Percentage of area" (0% to 100%). The x-axis categories are: 0-1, 1-2, 2-5, 5-10, 10-20, 20-50, 50-100, 100-200, 200-500, 500-1000, >1000.
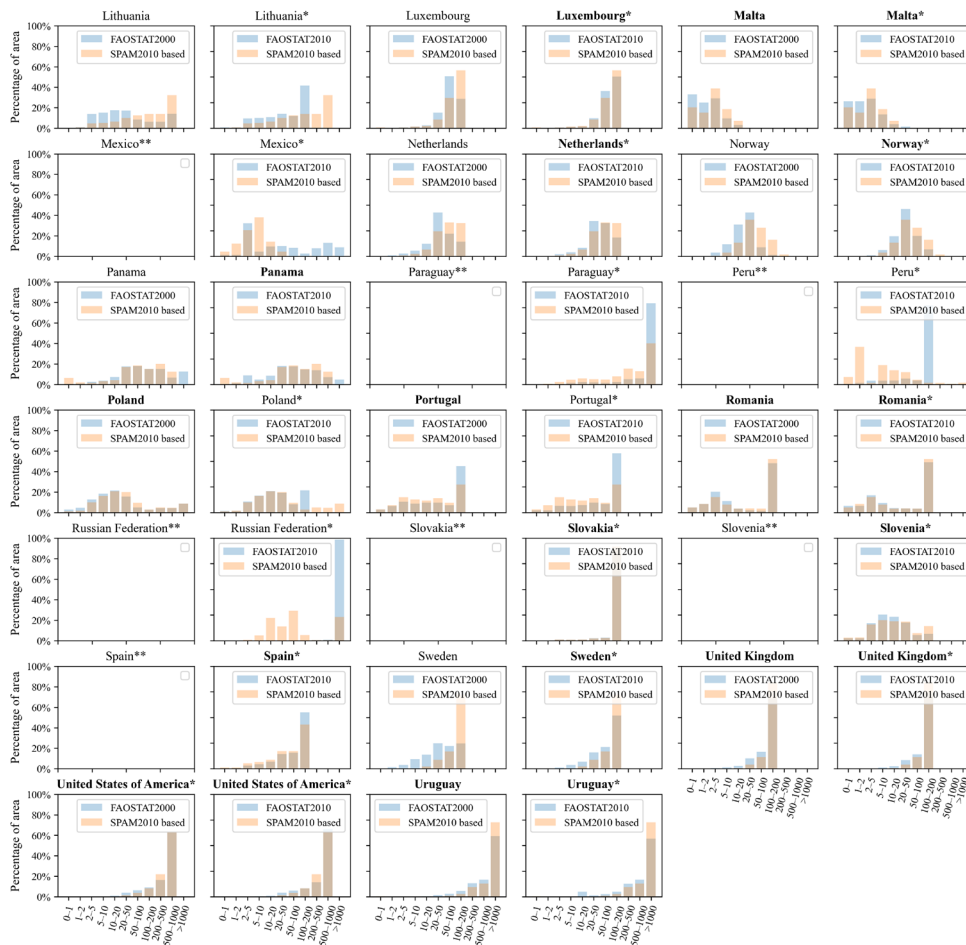
Figure A3. Comparison of the percentage of total crop area operated by each farm size (non-crop-specific farm size structure) between FAOSTAT structural data for the year 2000 and 2010 (FAO, 2022) and our farm-size- and crop-specific dataset based on the SPAM2010 crop map. Bold font country titles indicate that farm size structures in FAOSTAT are similar to our dataset. Note that for the year 2000, farm size structure from FAOSTAT structural data is the same with Lowder et al. (2016) except for one country [S5]. Only the countries covered by our dataset and FAOSTAT are shown. The alternative version based on GAEZv4 crop map is given in Fig. 7. * FAOSTAT provides (part of) the structural data by interpolating other reported data, not directly from countries' official reports. ** FAOSTAT provides no farm size structural data of the year 2000 or 2010 for comparison.

43

Author contribution

765 The concept of this work originated from HS ~~during the discussion~~in collaboration with MSK and RJH. The study was ~~then~~ designed and conducted by HS under the supervision of BM and DLG with feedback from MSK and RJH. HS wrote the draft of this manuscript. HS, BM, DLG, MSK, and RJH participated in the analysis of results and revision and editing of the manuscript.

Competing interests

770 The authors declare that they have no conflict of interest.

Acknowledgements

775 ## References

Bosc, P.-M., Berdegué, J., Goïta, M., van der Ploeg, J. D., Sekine, K., and Zhang, L.: Investing in smallholder agriculture for food security, HLPE, 2013.

Descals, A., Wich, S., Meijaard, E., Gaveau, D., Peedell, S., and Szantoi, Z.: High-resolution global map of smallholder and industrial closed-canopy oil palm plantations, Earth System Science Data, 2020.

780 EUROSTAT: EUROSTAT, Agriculture, forestry and fisheries, Africulture, Farm structure, 2021.

FAO: World programme for the census of agriculture 2020, Volume 1, Programme, concepts and definitions, FAO UN, Rome, Italy, 2015.

FAO: Small family farms data portrait. Basic information document. Methodology and data description, FAO UN, Rome, Italy, 2017.

~~FAO: FAOSTAT, FAO UN, Rome, Italy, 2019a.~~

FAO: Methodology for computing and monitoring the Sustainable Development Goal indicators 2.3.1 and 2.3.2, FAO UN, Rome, Italy,

785 2019~~b~~.

FAO: RuLIS Codebook, Rural Livelihoods Information System, FAO UN, Rome, Italy, 2020~~a~~.

~~FAO: World Programme for the Census of Agriculture, FAO UN, Rome, Italy, 2020b.~~

FAO: RuLIS - Rural Livelihoods Information System, FAO UN, Rome, Italy, 2021.

FAO: FAOSTAT, FAO UN, Rome, Italy, 2022.

790 FAO and IIASA: Global Agro Ecological Zones version 4 (GAEZ v4), FAO UN, Rome, Italy, 2021.

Fischer, G., Nachtergaele, F. O., van Velthuizen, H., Chiozza, F., Francheschini, G., Henry, M., Muchoney, D., and Tramberend, S.: Global Agro-ecological Zones (GAEZ v4)-Model Documentation, FAO Rome, Italy, 2021.

Fritz, S., See, L., McCallum, I., You, L., Bun, A., Moltchanova, E., Duerauer, M., Albrecht, F., Schill, C., Perger, C., Havlik, P., Mosnier, A., Thornton, P., Wood-Sichra, U., Herrero, M., Becker-Reshef, I., Justice, C., Hansen, M., Gong, P., Abdel Aziz, S., Cipriani, A., Cumani,

795 R., Cecchi, G., Conchedda, G., Ferreira, S., Gomez, A., Haffani, M., Kayitakire, F., Malanding, J., Mueller, R., Newby, T., Nonguierma, A., Olusegun, A., Ortner, S., Rajak, D. R., Rocha, J., Schepaschenko, D., Schepaschenko, M., Terekhov, A., Tiangwa, A., Vancutsem, C., Vintrou, E., Wenbin, W., van der Velde, M., Dunwoody, A., Kraxner, F., and Obersteiner, M.: Mapping global cropland and field size, Glob Chang Biol, 21, 1980-1992, 10.1111/gcb.12838, 2015.

Giller, K. E., Delaune, T., Silva, J. V., Descheemaeker, K., van de Ven, G., Schut, A. G., van Wijk, M., Hammond, J., Hochman, Z., and

800 Taulya, G.: The future of farming: Who will produce our food?, Food Security, 13, 1073-1099, 2021.

Gollin, D.: Farm size and productivity: Lessons from recent literature, 1-35, 2019.

Grafton, R. Q., Williams, J., Perry, C. J., Molle, F., Ringler, C., Steduto, P., Udall, B., Wheeler, S. A., Wang, Y., Garrick, D., and Allen, R. G.: The paradox of irrigation efficiency, Science, 361, 748-750, 2018.

Gurobi Optimization, L.: Gurobi Optimizer Reference Manual, 2021.

805 Herrero, M., Thornton, P. K., Power, B., Bogard, J. R., Remans, R., Fritz, S., Gerber, J. S., Nelson, G., See, L., Waha, K., Watson, R. A., West, P. C., Samberg, L. H., van de Steeg, J., Stephenson, E., van Wijk, M., and Havlík, P.: Farming and the geography of nutrient production for human use: a transdisciplinary analysis, The Lancet Planetary Health, 1, e33-e42, 10.1016/s2542-5196(17)30007-4, 2017.

Hoekstra, A. Y., Mekonnen, M. M., Chapagain, A. K., Mathews, R. E., and Richter, B. D.: Global monthly water scarcity: blue water footprints versus blue water availability, PLoS One, 7, e32688, 10.1371/journal.pone.0032688, 2012.

810 Iizumi, T., Hosokawa, N., and Wagai, R.: Soil carbon-food synergy: sizable contributions of small-scale farmers, CABI Agriculture and Bioscience, 2, 43, 10.1186/s43170-021-00063-6, 2021.

Kavats, O., Khramov, D., Sergieieva, K., and Vasyliev, V.: Monitoring of sugarcane harvest in Brazil based on optical and SAR data, Remote Sensing, 12, 1-26, 10.3390/rs12244080, 2020.

Khalil, C. A., Conforti, P., Ergin, I., and Gennari, P.: Defining small scale food producers to monitor target 2.3 of the 2030 Agenda for 815 Sustainable Development, FAO, Rome, 2017.

Kim, K.-H., Doi, Y., Ramankutty, N., and Iizumi, T.: A review of global gridded cropping system data products, Environmental Research Letters, 16, 10.1088/1748-9326/ac20f4, 2021.

Latham, J., Cumani, R., Rosati, I., and Bloise, M.: Global land cover share (GLC-SHARE) database beta-release version 1.0-2014, FAO, Rome, Italy, 2014.

820 Lesiv, M., Laso Bayas, J. C., See, L., Duerauer, M., Dahlia, D., Durando, N., Hazarika, R., Kumar Sahariah, P., Vakolyuk, M., Blyshchyk, V., Bilous, A., Perez-Hoyos, A., Gengler, S., Prestele, R., Bilous, S., Akhtar, I. U. H., Singha, K., Choudhury, S. B., Chetri, T., Malek, Z., 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, Glob Chang Biol, 25, 174-186, 10.1111/gcb.14492, 2019.

825 Lowder, S. K., Sánchez, M. V., and Bertini, R.: Which farms feed the world and has farmland become more concentrated?, World Development, 142, 10.1016/j.worlddev.2021.105455, 2021.

Lowder, S. K., Skoet, J., and Raney, T.: The Number, Size, and Distribution of Farms, Smallholder Farms, and Family Farms Worldwide, World Development, 87, 16-29, 10.1016/j.worlddev.2015.10.041, 2016.

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 830 global synergy cropland map, Earth System Science Data, 12, 1913-1928, 10.5194/essd-12-1913-2020, 2020.

Mehrabi, Z., McDowell, M. J., Ricciardi, V., Levers, C., Martinez, J. D., Mehrabi, N., Wittman, H., Ramankutty, N., and Jarvis, A.: The global divide in data-driven farming, Nature Sustainability, 4, 154-160, 10.1038/s41893-020-00631-0, 2020.

Mekonnen, M. M. and Hoekstra, A. Y.: Four billion people facing severe water scarcity, Science Advances, 2, 10.1126/sciadv.1500323, 2016.

835 Meyfroidt, P.: Mapping farm size globally: benchmarking the smallholders debate, Environmental Research Letters, 12, 10.1088/1748-9326/aa5ef6, 2017.

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, n/a-n/a, 10.1029/2007gb002947, 2008.

Muyanga, M. and Jayne, T. S.: Revisiting the Farm Size-Productivity Relationship Based on a Relatively Wide Range of Farm Sizes: 840 Evidence from Kenya, 101, 1140-1163, https://doi.org/10.1093/ajae/aaz003, 2019.

Noack, F., Larsen, A., Kamp, J., and Levers, C.: A bird's eye view of farm size and biodiversity: The ecological legacy of the iron curtain, American Journal of Agricultural Economics, 10.1111/ajae.12274, 2021.

Ray, D. K., Mueller, N. D., West, P. C., and Foley, J. A.: Yield trends are insufficient to double global crop production by 2050, PloS one, 8, e66428, 2013.

845 Ren, C., Liu, S., van Grinsven, H., Reis, S., Jin, S., Liu, H., and Gu, B.: The impact of farm size on agricultural sustainability, Journal of Cleaner Production, 220, 357-367, 10.1016/j.jclepro.2019.02.151, 2019.

Ricciardi, V., Mehrabi, Z., Wittman, H., James, D., and Ramankutty, N.: Higher yields and more biodiversity on smaller farms, Nature Sustainability, 4, 651-657, 10.1038/s41893-021-00699-2, 2021.

Ricciardi, V., Ramankutty, N., Mehrabi, Z., Jarvis, L., and Chookolingo, B.: How much of the world's food do smallholders produce?, Global 850 Food Security, 17, 64-72, 2018a.

Ricciardi, V., Ramankutty, N., Mehrabi, Z., Jarvis, L., and Chookolingo, B.: An open-access dataset of crop production by farm size from agricultural censuses and surveys, Data Brief, 19, 1970-1988, 10.1016/j.dib.2018.06.057, 2018ab.

Ricciardi, V., Ramankutty, N., Mehrabi, Z., Jarvis, L., and Chookolingo, B.: How much of the world's food do smallholders produce?, Global Food Security, 17, 64-72, 2018b.

855 Ricciardi, V., Wane, A., Sidhu, B. S., Godde, C., Solomon, D., McCullough, E., Diekmann, F., Porciello, J., Jain, M., and Randall, N.: A scoping review of research funding for small-scale farmers in water scarce regions, Nature Sustainability, 3, 836-844, 2020.

Riesgo, L., Louhichi, K., Gomez y Paloma, S., Hazell, P., Ricker-Gilbert, J., Wiggins, S., Sahn, D. E., and Mishra, A. K.: Food and nutrition security and role of smallholder farms: challenges and opportunities, Institute for Prospective Technological Studies; Information for Meeting Africa's Agricultural Transformation and Food Security Goals (IMAAFS),

860   Rudra, A.: Farm size and yield per acre, Economic Political Weekly, 1041-1044, 1968.

Samberg, L. H., Gerber, J. S., Ramankutty, N., Herrero, M., and West, P. C.: Subnational distribution of average farm size and smallholder contributions to global food production, Environmental Research Letters, 11, 10.1088/1748-9326/11/12/124010, 2016.

Savastano, S. and Scandizzo, P.: Farm Size and Productivity: A "Direct-Inverse-Direct" Relationship, 2017.

Su, H., Willaarts, B., Luna Gonzalez, D., S. Krol, M., and J. Hogeboom, R.: Gridded 5-arcmin, simultaneously farm-size-specific and crop

865   -specific harvested area for 56 countries, https://doi.org/10.5281/zenodo.5747616976249, 2022.

UNSD: SDG Indicators, United Nations Statistics Division, https://unstats.un.org/sdgs/indicators/indicators-list/, 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 System Science Data, 12, 3545-3572, 10.5194/essd-12-3545-2020, 2020.

870

Formatted: EndNote Bibliography