https://doi.org/10.5194/essd-2021-125
Preprint. Discussion started: 25 May 2021

Open Access Earth System Science Data Discussions

Harmonized in situ JECAM datasets for agricultural land use mapping and monitoring in tropical countries

Audrey Jolivot[1,2], Valentine Lebourgeois[1,2], Mael Ameline[1,2], Valérie Andriamanga[1,3], Beatriz Bellón[1,2], Mathieu Castets[1,2], Arthur Crespin-Boucaud[1,2], Pierre Defourny[4], Santiana Diaz[1,2], Mohamadou Dieye[5], Stéphane Dupuy[1,2], Rodrigo Ferraz[6], Raffaele Gaetano[1,2], Marie Gely[1,2], Camille Jahel[1,2], Bertin Kabore[7], Camille Lelong[1,2], Guerric le Maire[8,9], Louise Leroux[10,11,12], Danny Lo Seen[1,2], Martha Muthoni[13], Babacar Ndao[12,14], Terry Newby[15], Cecília Lira Melo de Oliveira Santos[16], Eloise Rasoamalala[1,3], Margareth Simoes[6], Ibrahima Thiaw[11], Alice Timmermans[4], Annelise Tran[1,2], Agnès Bégué[1,2]

[1]CIRAD, UMR TETIS, F-34398 Montpellier, France.
[2]TETIS, Univ Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France.
[3]Centre National de la Recherche Appliquée au Développement Rural (FOFIFA), Antsirabe, Madagascar.
[4]Université Catholique de Louvain (UCLouvain), Louvain-la-Neuve, Belgique.
[5]Institut Sénégalais de Recherches Agricoles (ISRA), Dakar, Senegal.
[6]Brazilian Agricultural Research Corporation (EMBRAPA), Rio de Janeiro, Brazil.
[7]Independent consultant, Ouagadougou, Burkina Faso
[8]CIRAD, UMR Eco&Sols, F-34398 Montpellier, France.
[9]Eco&Sols, Univ Montpellier, CIRAD, INRAE, Institut Agro, IRD, Montpellier, France.
[10]CIRAD, UPR AIDA, Dakar, Senegal.
[11]AIDA, Univ Montpellier, CIRAD, Montpellier, France.
[12]Centre de Suivi Ecologique (CSE), Dakar, Senegal.
[13]Independent consultant, Nairobi, Kenya.
[14]Université Cheikh Anta DIOP de Dakar (UCAD), Dakar, Senegal
[15]Agricultural Research Council (ARC), Pretoria, South Africa.
[16]Interdisciplinary Center on Energy Planning, NIPE, University of Campinas, UNICAMP, Campinas, Sao Paulo 13083-896, Brazil.

Correspondence to: Audrey Jolivot (audrey.jolivot@cirad.fr)

Abstract. The availability of crop type reference datasets for satellite image classification is very limited for complex agricultural systems as observed in developing and emerging countries. Indeed, agricultural land use is very dynamic, agricultural census are often poorly georeferenced, and crop types are difficult to photo-interpret directly from satellite imagery. In this paper, we present nine datasets collected in a standardized manner between 2013 and 2020 in seven tropical and subtropical countries within the framework of the international JECAM (Joint Experiment for Crop Assessment and Monitoring) initiative. These quality-controlled datasets are distinguished by in situ data collected at field scale by local experts, with precise geographic coordinates, and following a common protocol. Altogether, the datasets completed 27 074 polygons (20 257 crop and 6 817 non-crop) documented by detailed keywords. These datasets can be used to produce and validate agricultural land use maps in the tropics, but also, to assess the performances and the robustness of classification methods of cropland and crop types/practices in a large range of tropical farming systems. The dataset is available at https://doi.org/10.18167/DVN1/P7OLAP.

https://doi.org/10.5194/essd-2021-125
Preprint. Discussion started: 25 May 2021

Open Access Earth System Science Data Discussions

1. Introduction

Land use and land cover (LULC), and their changes, are key information to study and monitor carbon and water cycles, threats to biodiversity, but also to set up land use planning and public policies. In particular, accurate mapping of cropland and associated cropping practices is of primary importance for food security, agricultural and environmental monitoring as well as land management. However, cropland and crop type mapping using Earth observation data is still challenging as it requires large sets of training and validation data, and as the land use (field limits and content) generally changes annually, even seasonally. If large data sets on cropping practices are available in the Global North, it is not the case in most of the developing and emerging countries. In these countries, cropland and crop types can be particularly difficult to map (Waldner et al., 2015) because the fields are often small to medium size (Fritz et al., 2015), the crops are easily confused with natural vegetation and fallows, and cropping systems are typically highly variable in time and space. Each farming system has its own specificities in terms of crop type and composition, field size, cropping calendar, irrigated/rainfed mode and other practices (Bégué et al., 2018). It is thus necessary to adapt the classification approaches (satellite data and algorithms as well as training and validation in situ data) to the large variability of the farming systems in the world (Dixon et al., 2001), and thus to have access to appropriate training data.

The arrival of Sentinel-1 and 2 satellite image time series, the emergence of new classification algorithms in the domain of machine learning and artificial intelligence, and an easy access to pre-processed images and image processing tools on web-platforms, have democratized image processing, and opened-up new avenues for LULC mapping over large areas. Following this trend, large benchmark datasets acquired using annotation tools of satellite images all over the world have multiplied to train algorithms and validate remote sensing-derived products (Long et al., 2020). However, these datasets have a broad LULC nomenclature, and agricultural land use is often reduced to a single class due to difficulties in discriminating cropping practices from satellite images. The main data sources currently available for agricultural land use mapping in the Southern countries are listed below.

At a global and continental scale, initiatives that freely distribute land cover reference datasets exist (see review by Tsendbazar et al. (2015)) such as GOFC-GOLD (Global Observation for Forest and Land Cover Dynamics; http://www.gofcgold.wur.nl/sites/gofcgold_refdataportal.php) that regroups existing global datasets prior 2015, or the LULC reference dataset (150,000 samples ; 300 m-1 km resolution) (Fritz et al., 2017) and the cropland dataset (36 000 points ; 300 m resolution) (Laso Bayas et al., 2017b) both collected through crowdsourcing campaigns using the Geo-Wiki tool (photo-interpretation of very high spatial resolution satellite images). Lately, the LandCoverNet dataset has been released for the African continent (Alemohammad et al., 2020) with 130 million of labelled 20 m pixels of 1 980 image chips (256 x 256 pixels), spanning 66 tiles of Sentinel-2 acquired in 2018. These data are used to validate global (Hoskins et al., 2016) or

https://doi.org/10.5194/essd-2021-125
Preprint. Discussion started: 25 May 2021

national cropland maps (Laso Bayas et al., 2017a) as the nomenclature used for labelling the classes does not specify the crop type.

At a national scale, ground campaigns, such as those carried out as part of the Sen2Agri project in South Africa and Mali, collected data on the main crop types (Defourny et al., 2019). But these data are generally not available to validate global maps or train new classification algorithms, as they are often the responsibility of national sovereignty.

At a local scale, datasets on crop types have been acquired, and are still acquired, across multiple world regions within the context of the JECAM (Joint Experiment for Crop Assessment and Monitoring; Available online: http://www.jecam.org/; accessed on 10 February 2020) international network. The JECAM initiative was first developed under the GEO (Group of Earth Observations) umbrella and then became the research and development component of GEOGLAM (GEO Global Agricultural Monitoring), to enable the global agricultural monitoring community to carry out cross-sites experiments and compare results based on disparate sources of data, using various methods, over a variety of local or regional cropping systems. Data are acquired following a given protocol and nomenclature (see Defourny et al. (2014)). The experiment has been operating since 2013, and some in situ datasets produced at field scale have been used in different benchmarking mapping studies (Waldner et al., 2016; Inglada et al., 2015).

The aim of this data paper is to share with the community, harmonized in situ agricultural land use datasets mostly acquired within the JECAM initiative, at local scale, and focusing on emerging/developing countries on the tropics. These datasets include data collected on nine sites in seven countries of the tropical belt (Figure 1) with various farming systems (Figure 2; Table 1). The acquisition protocol has been adapted from Defourny et al. (2014) to take in account the characteristics of tropical agriculture (e.g. small field size, accessibility). Information on crop type and cropping practices was collected locally, at the field level, with a detailed nomenclature. The acquisition period is between 2013 and 2020, and the number of monitoring years per site is between 1 and 7.

https://doi.org/10.5194/essd-2021-125
Preprint. Discussion started: 25 May 2021

Open Access Earth System Science Data Discussions

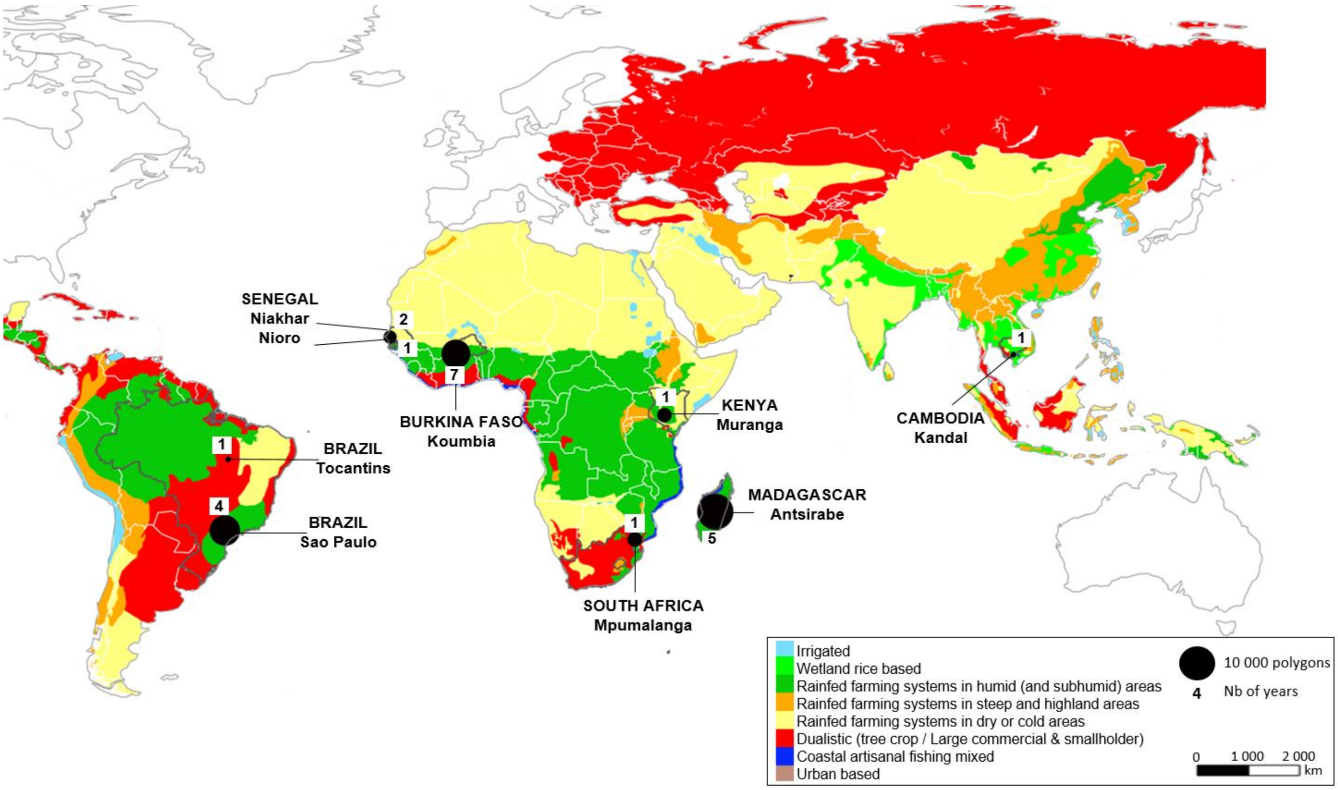
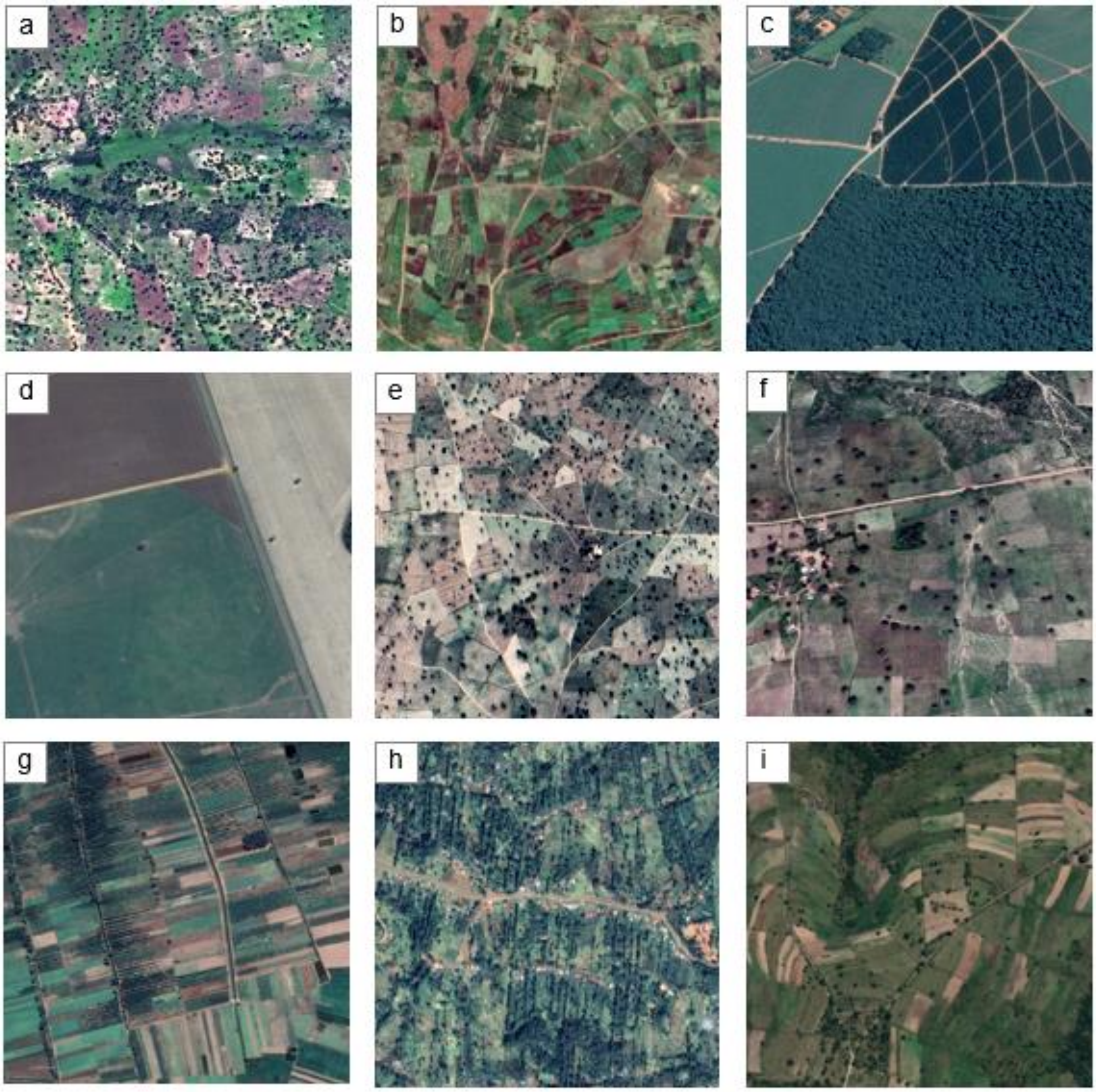


Figure 1. Location map of the study sites, and the associated number of collection years and sampled plots (symbolized by the size of the red circles), displayed on the FAO (broad) farming system map (Dixon et al., 2001) .

https://doi.org/10.5194/essd-2021-125
Preprint. Discussion started: 25 May 2021



95 **Figure 2. A 1 km² sample of land showing the landscape variety across the sampled sites due to the farming system in place** : (a) rainfed cereals in Burkina-Faso; (b) rice systems in Madagascar; (c) agropastoral systems in Brazil-Tocantins; (d) mixed agriculture in Brazil-São Paulo; (e) rainfed groundnut and millet agropastoral systems in Niakhar and (f) in Nioro, Senegal; (g) irrigated rice systems and orchards in Cambodia; (h) agroforestry in Kenya; (i) mixed agriculture in South Africa. *Images* **© Google Earth 2020**

https://doi.org/10.5194/essd-2021-125
Preprint. Discussion started: 25 May 2021

2. Methods

2.1 Study sites

Except for Cambodia, the study sites belong to the JECAM network (http://www.jecam.org/), and cover several hundred squared kilometers each. The table 1 provides a synthesis of the database by site.

The JECAM Burkina Faso study site is a 60 x 60 km² area located around the commune of Koumbia, Tuy province, in the South-West of the country. The climate is tropical. The absence of significant relief and the relatively good conditions in terms of soil and climate favoured the densification of cropped surfaces, which span the majority of the area: arable lands cover more than 60% of the site, the remaining surface being either unsuitable for cultivation (e.g. rocky) or protected areas for nature conservation. The landscape is characterized by an alternation of large cropland areas made up of a patchwork of diversified small cropped fields (about 1 ha) and areas covered by natural vegetation. With the exception of few lowland rice plots, all crops are rainfed, hence cultivated during the rainy season that occurs from May to October (around 1000 mm average annual rainfall). Main crops are more or less equally distributed between cash crops (mainly cotton) and staple crops, with a significant predominance of cereal crops (maize, sorghum, millet, and locally rice) over oleaginous (sesame, groundnuts) and leguminous (peas/cow peas, soybeans).

The JECAM Madagascar study site is a 60 x 60 km² zone located in the Vakinankaratra region, around the Anstirabe city, in the central highlands of the country. It is characterized by mountainous terrain of terraced from 1200 to 1500 m of altitude, rice-growing valleys positioned between grassy hills and rocky outcrops. The climate is subtropical with a rainy season from December to February. The average annual precipitation is 1300 mm. The growing season occurs from October to June. Cultivated crops are diversified, although maize and rice predominate. Fruit production is also present in the area. The mean size of an agricultural field in the area is very small (about 0.05 ha), but contiguous fields of the same crop type occasionally give rise to larger single crop patches. Rice is mainly grown in irrigated areas, but has recently mingled with other rainfed crops on slopes (called tanetys). Other main crops are carrots, potatoes, sweet potatoes, soybeans or cassava.

The JECAM São Paulo site in Brazil is a large area of 90 x 130 km² located in the São Paulo State, close to Botucatu city. It is composed of a relatively smooth relief with slopes mostly <5%. The region is classified as subtropical humid-dry in the winter. Average temperature is 19°C and average annual precipitation is 1400 mm with a rainy season from December to March. The area is diversified and can be divided into four main agricultural sub-regions: (1) in the South-West, annual crops (maize, wheat, soybean) including summer (growth cycle from October to May) and winter crops (June to September) - some of them irrigated with centre pivot systems; (2) in the Centre forest plantations for wood production; (3) in the East pastures, and (4) in the North sugarcane, which has variable planting and harvesting dates: the first sugarcane cycle occurs between September and March, and is grown for around 12–18 months. Sugarcane reaches maximal growth in April, in this region. After the first harvest, the cycle of the ratoon sugarcane starts, with the annual cut between April to December. Natural forests,

https://doi.org/10.5194/essd-2021-125
Preprint. Discussion started: 25 May 2021

mostly along the rivers, and orange orchards are present in these four sub-regions. Field size is generally larger than 10 ha, and can reach more than 200 ha for pastures and forest plantations. A detailed description of this site, including crop and rotation descriptions, is given in de Oliveira Santos et al. (2019).

The JECAM Tocantins site in Brazil is part of the MATOPIBA (Maranhão, Tocantins, Piauí and Bahia) region, a new agricultural frontier in Brazil. It is a 25 x 25 km² site situated in the Municipality of Pedro Afonso and surroundings, in the Cerrado biome. The climate is tropical, with a rainy season from October to Mars. The landscape is composed of a mosaic of large fields (generally around 100 ha), native forest remnants and rangelands, with mild relief, and the annual rainfall is between 1700-1800 mm. The main agricultural systems are soybean single cropping, double cropping of summer soybean from November to February followed by a cereal crop (maize, millet or sorghum) from March to June, some sugarcane, and planted pastures that are increasingly being implemented in the region as part of integrated crop-livestock systems (soybean-corn-planted pasture). Sugarcane crops are irrigated with centre pivot systems.

The Niakhar and the Nioro Senegalese study sites are located in the Senegalese Peanut Basin, in the central western part of the country. The Niakhar site spans the districts of Fatick and Bambey in the Northern part of the Peanut Basin, and the Nioro site is located in the district of Nioro du Rip at the border of the Gambia. Each site covers about 400 km². The climate is Sahelo-Sudanian with one rainy season (400 to 600 mm) that lasts from July to October. The relief is relatively flat. As in many parts of the Sahelian zone, the smallholder farming systems are dominated by tree-based agricultural landscapes, forming what is so-called parklands. The Niakhar site is dominated by Faidherbia albida trees, while the Nioro site is dominated by Cordyla pinnata trees. Livelihoods of rural populations are centered on small-scale rainfed agriculture, with low usage of mineral fertilizer. Pearl millet and groundnut are the main staple crops mainly cultivated in biennial rotation. Other crops are sorghum, cowpea, bissap and maize cultivated during the rainy season.

The JECAM Kenya study site is a 25 x 10 km² area located about 50 km north of Naïrobi, including Kangema and Muranga towns, in the central province of Kenya. It is settled in a very hilly landscape with steep slopes and strong local relief variations in a general toposequence trend following an East-West altitude gradient from 1000 m to 2800 m. Climate is wet tropical, somewhat temperate by the altitude and regularized by two rainy seasons (from Mars to May or June, and from October to November) with 1200 to 2000 mm annual rainfall depending on the altitude. The permanent moisture and good natural drainage of a rich volcanic loam allows for intensive agriculture, mainly based on perennial crops (mostly banana, various fruits, coffee, and tea) associated with dairy farming and rainfed horticultural as well as food crops (eg. French beans, cabbage, maize, cassava). These latter are cropped all year long, except in January and July which are dry months, and without a defined

https://doi.org/10.5194/essd-2021-125
Preprint. Discussion started: 25 May 2021

seasonal calendar (maize, for instance, can have three cycles per year). The mean size of an agricultural field in the area is very small (about 0.08 ha) resulting in a patchwork landscape of heterogeneous fields with a great diversity of structures.

The Cambodia study site corresponds to a 30 km radius buffer area around Wat Pi Chey Saa Kor, Kom Poung Kor village, Kandal Province, where the ecology of fruit bats Pteropus lylei was recently investigated (Choden et al., 2019). The area is characterized by a tropical climate with a rainy season from May to October. The annual rainfall is between 1000 and 1500 mm. Two main rivers, the Mekong and the Bassac, cross the area. In this flat region, rice is the dominant crop, mainly grown in irrigated areas from May to October. Fruit plantations (mango, sapodilla) and natural wetlands are also present. The mean field size is small (around 1 ha). The population lives in villages along the roads composed of small houses with fruit trees backyards.

The JECAM South Africa study area is a 60 x 60 km² site, located in the Mpumalanga province in the North-East of the country, close to the Mozambique border corresponding mostly to a subsistence agriculture area. The climate is subtropical with a rainy season from November to February. The annual rainfall is between 600 and 800 mm. The site is characterized by a bush-clad plain between the Drakensberg Mountains (West) and savannahs (East) with several wildlife reserves (e.g. Kruger Park). The study area is characterized by smallholder's agriculture (generally less than 1 ha), with diversified crops: cereals, groundnuts, potatoes, vegetables, fruit crops. Important timber plantations are present on the west part of the site.

Country, Site name (Area km²)	**Cropping pattern**	**Number of collection years**	**Total number of polygons (percentage of crop polygons in the dataset)**	**Mean polygon size (ha)***	**Percentage of polygons obtained from ground survey**	**Nb. of crop type classes**
Burkina Faso, Koumbia (3 600)	Single cropping	7 (2013 to 2020)	6 264 (79%)	0.60	89%	23
Madagascar, Antsirabe (3 600)	Single / Double cropping	5 (2015 to 2019)	8 351 (87%)	0.35	95%	47
Brazil, São Paulo (11 700)	Multiple cropping	4 (2014 to 2017) **	6 149 (66%)	22	96%	21

https://doi.org/10.5194/essd-2021-125
Preprint. Discussion started: 25 May 2021

Brazil, Tocantins (625)	Double cropping	2 (2015 and 2016)	533 (56%)	150	67%	7
Senegal, Niakhar (400)	Single cropping	2 (2018 and 2019)	1 403 (74%)	0.54	83%	5
Senegal, Nioro (400)	Single cropping	1 (2018)	457 (46%)	1.17	48%	6
Kenya, Muranga (250)	Multiple cropping	1 (2015)	1 647 (77%)	0.14	100%	26
Cambodge, Kandal (2 826)	Double cropping	1 (2014 / 2015)	529 (25%)	*** Small fields	28%	5
South Africa, Mpumalanga (3 600)	Single cropping	1 (2017)	1 741 (59%)	*** Small fields	38%	10

* Areas calculated on cropland polygons
** 16 field campaigns in 4 years
*** The digitized boundaries of the polygons correspond to homogeneous crop areas (collections of adjacent small fields), and not necessarily to single fields.

Table 1. Synthesis of the database by site.

2.2 Data collection

The acquisition protocol is based on the JECAM guidelines (Defourny et al., 2014) with adaptations to consider some characteristics of tropical agriculture (mainly small field size and accessibility). Field surveys were conducted yearly in each study zone, either around the growing peak of the cropping season, for the sites with a main growing season linked to the rainy season such as Burkina Faso, or seasonally, for the sites with multiple cropping (e.g. São Paulo site). Except for Senegal where a stratified sampling plan for field surveys was used (Ndao et al., 2021), the GPS waypoints were gathered following an opportunistic sampling approach (called the “windshield survey”) along the roads or tracks according to their accessibility (that can be difficult during the rainy season, leading to less surveys in secondary roads or tracks in some study areas), while ensuring the best representativity of the existing cropping systems in place (Defourny et al., 2014; Waldner et al., 2019). GPS waypoints were also recorded on different types of non-crop classes (e.g. natural vegetation, settlement areas, water bodies) to allow differentiating crop and non-crop classes. Waypoints were only recorded for homogenous fields/entities of at least 20 x

20 m² (against a minimum of 0.25 ha with a minimum width of 30 m in JECAM guidelines). To facilitate the location of sampling areas and the remote acquisition of waypoints, field operators were equipped with GPS tablets (Trimble - Yuma2 or Handheld - Algiz 10X) providing access to a QGIS project with Very High Spatial Resolution (VHSR) images (orthorectified Pleiades or SPOT 6/7 images ordered just before the surveys, or PlanetScope images). This equipment allowed in situ recording of attributes relative to each waypoint on data entry forms (with automatic filing of IDs or dates and scrollable lists for other attributes to avoid data entry errors). For each waypoint, a set of attributes, corresponding to the cropping practices (crop type, cropping pattern, management techniques) were recorded. An attribute referred to as "Keywords" was also created in order to associate various generic terms (land cover, crop group, crop type, cropping practice, etc.) to each polygon. This attribute has two objectives: (i) facilitating keyword search for the user, (ii) allowing the user to create his own nomenclature (hierarchic or not) with different levels of detail so that the nomenclature can be dedicated to the user's needs. These terms are based on the FAO land use definitions (FAO, 2020) and JECAM hierarchic nomenclature (Defourny et al., 2014), which were adapted to take into account the diversity of the farming systems in the surveyed sites. All these attributes are described in Table 2.

In the specific case of Burkina Faso, Senegal-Niakhar and Brazil-São Paulo sites, the same fields were revisited each year, in order to study crop rotations and fallow practices in the region. For the South Africa site, some points were collected by helicopter using the Producer Independent Crop Estimates System (PICES (Fourie, 2009)) method developed by the National Crop Statistics Consortium. Flights were performed at an average altitude of 500 feet and a low flying speed allowing to record GPS points and to determine land use using a GPS tablet associated with a GIS interface and a recent VHRS image. Only clearly identifiable land covers have been kept in the database.

2.3 Post-processing

Once the waypoints were acquired, the boundaries of each field or non-crop entity were digitized on the VHSR images in the QGIS software, and the class labels (and other attributes, see Table 2) were attached to the polygon database. Additional non-crop polygons were added by CAPI (Computer Assisted Photo Interpretation) of the VHSR images for the built-up areas, water bodies, wetlands, mineral surfaces, and natural forest classes (land covers clearly identifiable on images).

To avoid digitizing errors, this step was performed by the same operator as the one who did the field surveys. Despite this, if there was doubt on the delineation of a given entity (e.g. fuzzy boundaries, high heterogeneity), the given entity was removed from the database. Finally, the topology of each entity was controlled externally.

3. Data Records

This database, which contains 27 197 records, is a geographic layer in Shapefile format. Each record corresponds to a polygon with 16 attributes (Table 2). Because of the dispersion of study sites on the globe, the layer is in a Geographic Coordinates System with Datum WGS84.

https://doi.org/10.5194/essd-2021-125
Preprint. Discussion started: 25 May 2021

Twenty different land cover types and 102 different crop types have been observed. More than ¾ of the observations are agricultural land, the most represented crop types being maize, rice and sugarcane. The distributions of the main land cover and crop types are represented in Figures 3 and 4.

Attribute Name	Data Type	Description / available arguments
Id	Numeric	Unique ID
Country	Text	Country name
SiteName	Text	Site name (generally related to the biggest city around or to the region name)
DataSource	Numeric	Discrimination between land uses acquired from *in situ* surveys or satellite image CAPI (computer assisted photointerpretation) 0: Land use from *in situ* survey 1: Land use from satellite image interpretation 2: Land use from aircraft observation
AcquiDate*	Date	*In situ* survey acquisition date or satellite image acquisition date (when the land use is photointerpreted, see "DataSource" attribute) – Format: yyyy-mm-dd
LandCover	Text	Land cover of the polygon. If value is "Cropland", see CropType 1, 2 and 3 attributes for more information
CropType1	Text	Main crop type of the polygon
CropType2	Text	Secondary crop type of the polygon (in case of intercropping)
CropType3	Text	Tertiary crop type of the polygon (in case of intercropping)
SOS*	Date	Start of season date in the site (if empty, this means that no specific season exists in the study area) – Format: yyyy-mm-dd
EOS*	Date	End of season date in the site (if empty, this means that no specific season exists in the study area) – Format: yyyy-mm-dd
Irrigated	Numeric	Presence/absence of an irrigation system 0: No information available 1: Rainfed 2: Irrigated Empty: For polygons other than cropland

https://doi.org/10.5194/essd-2021-125
Preprint. Discussion started: 25 May 2021

Intercrop	Numeric	Presence/absence of intercropping 0: Single crop 1: Mixed crop or row inter-crop 2: Agroforestry Empty: For polygons other than cropland
Weeding	Numeric	Presence/absence of weeds 0: No information available 1: Presence of weeds Empty: For polygons other than cropland
Area_ha	Numeric	Polygon area in hectares
KeyWords	Text	Set of terms associated to the land use of the polygon (separated by semicolons ";")

* For each field in the Tocantins site, the operator recorded the crop type of the 2 seasons (summer / winter) by observing the crop residues on the field or by interviewing the farmers. Consequently, the acquisition date of those polygons does not always correspond to the actual land cover of the field. The user must refer to the SOS and EOS dates to identify the season corresponding to the crop type recorded.

Table 2. Description of the attributes recorded for each polygon of the database.

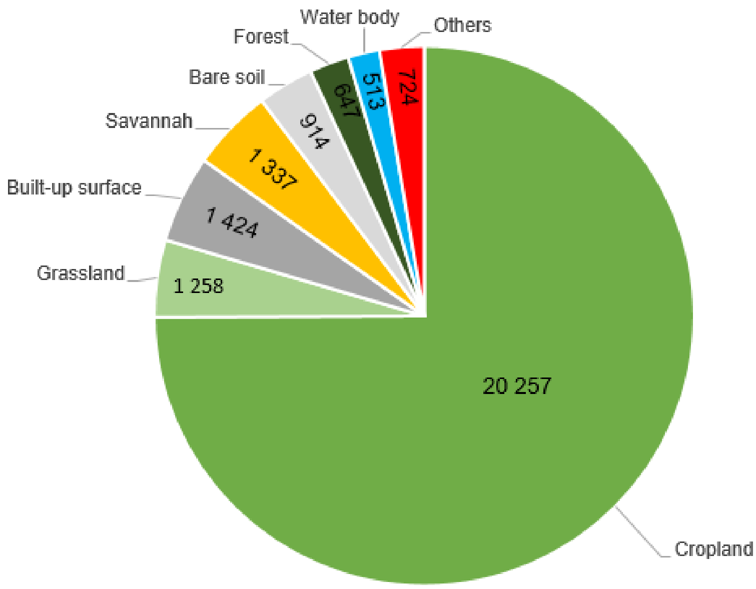


Figure 3. Distribution of the main land cover types (in number of polygons).

https://doi.org/10.5194/essd-2021-125
Preprint. Discussion started: 25 May 2021

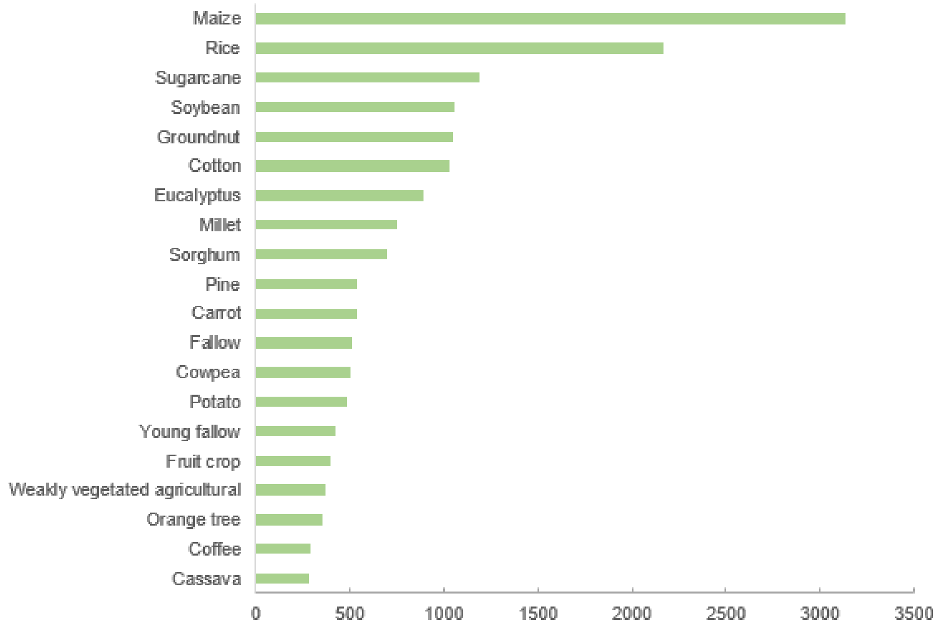


Figure 4. Distribution of the main crop types (in number of polygons).

4. Technical Validation

4.1 Quality Checking

Due to the nature of the dataset (in situ observation), validation is not possible. However, quality control was performed all along the data chain, from the acquisition to the post-processing, to ensure the quality of the datasets and their homogeneity throughout sampled years and locations.

First, the acquisition protocol was described in a technical guide provided to the field teams so that nothing was forgotten during the campaigns. The dropdown list in the data entry form reduced input and post-processing errors. The surveys were carried out by agronomists with geoprocessing skills, accompanied by a national researcher or technician with expertise in the local farming systems.

Second, during the post-processing step, the orthorectification of the VHSR images used to digitize the fields was checked from one year to the next, for multi-year sites, and corrected if necessary by taking homologous points. The fields were then

https://doi.org/10.5194/essd-2021-125
Preprint. Discussion started: 25 May 2021

manually digitized on the VHSR images, and the photos taken in situ were used whenever necessary. In case of doubtful data, these have been discarded and removed from the dataset.

Finally, the fact that the same person performed the whole acquisition and processing chain - from waypoint collection to polygon labelling - minimizes errors and contributes to the overall quality of the datasets.

4.2 Representativeness of data sets

Because of their small size, these sites cannot be considered representative of the entire country in which they are located; however, they are claimed to be representative of an area that encompasses more than the JECAM site. In order to specify the extent of this representative area, we referred to existing zoning maps. We used the two reference maps available for Southern countries: the FEWS-NET livelihood zones map (https://fews.net/fews-data/335) and the FAO farming systems map (http://www.fao.org/farmingsystems/mapstheme_01_en.htm). The livelihood zones are produced at national scale and are available for 38 developing countries. The zones are defined as geographical areas within which people share broadly the same patterns of livelihood (i.e., broadly the same production system, the same income earning opportunities and patterns of trade) (see Grillo and Holt (2009) for more details). Farming system maps are available for the Global South (covering 130 countries). The classes are defined as a population of individual farm systems that have broadly similar resource bases, enterprise patterns, household livelihoods and constraints (Dixon et al., 2001; Auricht et al., 2014).

Although these two maps were not produced for the same purposes, they are derived using similar criteria (agro-climatology, elevation, landscape, dominant pattern of farm activities, etc.) that are closely related to the agricultural land use as recorded in the database. In Table 3, are given the type and extent of the zones where are located our JECAM study sites, for both maps. Unfortunately, the livelihood maps are available only for four of the JECAM countries presented here.

	Livelihood zone (FEWS-NET)		Farming systems (FAO)	
Country	Livelihood type (year of production))	km²	Farming system type (year of production)	km²
KENYA	Central Highlands, High Potential Zone (2011)	19 689	Maize mixed (2014)	615 593
MADAGASCAR	Ankaratra: staple crops, horticulture, milk (2017)	15 675	Rice-tree crop (Maize-mixed) 2014	308 489

https://doi.org/10.5194/essd-2021-125
Preprint. Discussion started: 25 May 2021

Earth System Science Data
Open Access Discussions

SENEGAL	Rainfed groundnut and millet (2015)	10 256	Agro-pastoral millet/sorghum (2014)	1 238 113
SENEGAL	Rainfed groundnut and cereals (2015)	22 087	Agro-pastoral millet/sorghum (2014)	1 238 113
BURKINA FASO	West cotton and cereals (2014)	35 813	Cereal-root crop mixed (2014)	1 931 654
SOUTH AFRICA			Large commercial and smallholder (Maize-mixed or Perenial mixed) (2014)	1 010 746
BRAZIL (SP)			Intensive mixed (2001)	812 259
BRAZIL (TO)			Extensive mixed (Cerrados & Llanos) (2001)	1 744 804
CAMBODGE			Low-land rice (2001)	526 678

Table 3*: Agricultural types and extent of study sites' belonging zones: FEWS-NET livelihood zones* *(source:* https://fews.net/fews-data/335*)* **and FAO farming system zones** *(*http://www.fao.org/farmingsystems/mapstheme_01_en.htm*).*

With a mean size of the zone around 20 000 km² (Table 3), we are pretty confident that our JECAM sites are representative of the livelihood zone they belong to. The datasets presented here can thus be used to train or validate land cover maps of the corresponding zones. The farming system zones are much larger (between 300 000 km² and 2 Mkm²) and include a larger diversity of environmental and farming conditions; in these conditions it is not possible to argue that the JECAM sites are representative of such large areas; thus, the JECAM datasets need to be completed with other datasets belonging to the same

https://doi.org/10.5194/essd-2021-125
Preprint. Discussion started: 25 May 2021

farming system class before to be used for training land cover classification algorithms. However, they can still be used for algorithm / product validation or comparison.

It is also important to mention that other agro-ecological zoning (AEZ) can be used (even if only few are directly related to the agricultural land use) or that each user can produce their own AEZ and use it to delineate the area in which the JECAM dataset can be used to train classification algorithms.

4.3 Validation through study cases

In addition, the in situ JECAM dataset and its derived land use/land cover products have been used in a wide spectrum of studies covering several aspects linked to agricultural monitoring attesting the good quality of the dataset and good spatial representativeness of tropical countries farming systems.

First specific site studies have been conducted to test several methodological aspects. For instance, land use maps combining a supervised object-based approach with multi sources high spatial resolution time series were developed in Madagascar (Lebourgeois et al., 2017) and in Brazil (de Oliveira Santos et al., 2019). The brazilian site (São Paulo) was also included in a broader study presenting an inter-comparison of several cropland mapping methodologies over 5 contrasting JECAM sites (Brasil, Ukraine, Russia, Argentina and China) in terms of growing conditions, characteristics and cropping practices (Waldner et al., 2016). Very recently, following the rapid dissemination of up-to-date artificial intelligence approaches, Gbodjo et al. (2020) and Ienco et al. (2020) proposed to test the potential of deep learning architectures for land cover mapping respectively in Senegal (Niakhar) and Burkina Faso.

Second, in situ data coming from the Burkina Faso site and the Madagascar site have been included as test sites in the Sen2-Agri system. The Sen2-Agri system is an operational processing system that provides several agricultural products from Sentinel-2 and Landsat-8 time series along the cropping season. The two sites have been included in preliminary studies preparing the Sen2-Agri system processing chain (Bontemps et al., 2015; Valero et al., 2016), while the Madagascar site has been considered later in the demonstration phase of the system at local scale (http://www.esa-sen2agri.org/system-demonstration/).

Lastly, the different in situ data and the derived products have been valorized in studies covering different aspects of agricultural monitoring. For instance a semi-automated clustering approach has been proposed for the cropping system mapping over the Tocantin’s region in Brazil (Bellón et al., 2018). Using the land use map derived from the Burkina Faso site and the Senegal site (Niakhar), remote sensing-based statistical crop yield models have been proposed for maize (Leroux et al., 2019) and pearl millet (Leroux et al., 2020b). Based on the land use map derived from the Niakhar and Nioro sites in Senegal, Ndao et al. (2021) proposed an approach to characterize the agricultural landscape heterogeneity in agroforestry

https://doi.org/10.5194/essd-2021-125
Preprint. Discussion started: 25 May 2021

parklands, which was then used to analyse how far the agricultural landscape diversity contribute to the household food security (Leroux et al., 2020a).

5. Data availability

The dataset is ready for use on any GIS software, and can be filtered by region, year or key words. It is distributed with a CC-BY licence. The database, as well as the Kmz file locating the study areas, are available online on the CIRAD DataVerse in Jolivot, Audrey *et al*., 2021, "Harmonized in situ JECAM datasets for agricultural land use mapping and monitoring in tropical countries", doi:10.18167/DVN1/P7OLAP, CIRAD Dataverse, V2 (https://doi.org/10.18167/DVN1/P7OLAP)

6. Conclusion

Accurate mapping of cropland and associated cropping practices in smallholder farming systems of tropical countries is crucial for the improvement of agricultural monitoring systems at local and or global scales. The essential prerequisite to reach such objectives is to have available in situ datasets representative of the diverse agricultural practices in tropical countries. This paper presented an harmonized in situ crop type dataset acquired between 2013 and 2020 over nine sites spread over seven tropical countries. This dataset collected in the framework of the JEAM initiative is unique and very valuable because it is produced at the field scale, based on in situ observation, quality-controlled, and standardized observation for various tropical cropping systems, including small-holder farming systems. These characteristics allow this dataset to be used as a benchmark to assess the performances and the robustness of newly developed classification algorithms for cropland and crop type/practices mapping in diverse and documented agricultural conditions. In addition, this dataset can also be used to validate the cropland class of existing global or national LULC products, in particular those recently produced with Sentinel/Landsat image time series, and some crop type and practices (fallow, double cropping) classes. In the end it should be part of publicly online datasets and algorithm sharing platforms as promoted by the JECAM network and Long et al. (2020) who encourage the sharing of datasets for remote sensing applications, and more broadly to the scientific community, land use planners and agricultural monitoring agencies. This dataset will be further enriched with new ground surveys that are already planned on many on the presented sites.

Author contributions

AJ, AB and VL wrote the manuscript with substantial contributions of the PI's site: VL (Madagascar), RG (Burkina Faso), LL (Senegal), GL (Brazil – Sao Paulo), BB (Brazil – Tocantins), AT (Cambodia), CL (Kenya) and TN and AJ (South Africa)

AJ, VL and RG designed the database.

AJ harmonized and compiled the data.

Ground data collection and pre-processing : BN and MD (Senegal Niakhar), IT (Senegal Nioro), AC, ER, MA, SaD, StD, VA and VL (Madagascar), AB, AJ, CJ, DL, LL, MC, RG and StD (Burkina), GL (Brazil – Sao Paulo), AB and BB (Brazil – Tocantins), CL and MM (Kenya), AJ and MG (Cambodia), AJ and AT (South Africa)

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements

This research was mainly supported by CIRAD (scientists funding), and by various projects (CNES APR TOSCA projects, SIGMA FP7 n°603719; ESA Sen2Agri n° ESRIN 400109979/14/I-AM; CNPq n°454292/2014-7 and n°307560/2016-3; FAPESP-Microsoft Research n°2014/50715-9, the SERENA project funded by the Cirad-INRA metaprogramme GloFoodS). The SPOT and Pleiades images were acquired through the GEOSUD Program EQPX-20 funded by French National Research Agency. The PlanetScope images were acquired through the Planet's Ambassador Program.

Thanks to Embrapa Pesca e Aquicultura (Palmas, TO Brazil), Centre de Suivi Ecologique (Senegal), Agricultural Research Council (South Africa), Pasteur Institut (Cambodia), FOFIFA / DP SPAD (Madagascar), DP ASAP (Burkina), ICRAF (World Agroforestry Center, Kenya), NIPE and FEAGRI (Brazil), and Eder Araujo da Silva (Floragro Apoio, Brazil) for their technical support.

References

Alemohammad, S.H., Ballantyne, A., Bromberg Gaber, Y., Booth, K., Nakanuku-Diggs, L., and Miglarese, A.H.: LandCoverNet: A Global Land Cover Classification Training Dataset: https://radiant-mlhub.s3-us-west-2.amazonaws.com/landcovernet/Documentation.pdf, last access: last access: 7 September 2020.

Auricht, C., Dixon, J., Boffa, J.-M., and Garrity, D.: Farming Systems of Africa, in: Atlas of African agriculture research and development: Revealing agriculture's place in Africa, 14–15, 2014.

Bégué, A., Arvor, D., Bellon, B., Betbeder, J., de Abelleyra, D., P. D. Ferraz, R., Lebourgeois, V., Lelong, C., Simões, M., and R. Verón, S.: Remote Sensing and Cropping Practices: A Review, Remote Sens., 10, 99, https://doi.org/10.3390/rs10010099, 2018.

Bellón, B., Bégué, A., Lo Seen, D., Lebourgeois, V., Evangelista, B. A., Simões, M., and Demonte Ferraz, R. P.: Improved regional-scale Brazilian cropping systems' mapping based on a semi-automatic object-based clustering approach, Int. J. Appl. Earth Obs. Geoinformation, 68, 127–138, https://doi.org/10.1016/j.jag.2018.01.019, 2018.

https://doi.org/10.5194/essd-2021-125
Preprint. Discussion started: 25 May 2021

Bontemps, S., Arias, M., Cara, C., Dedieu, G., Guzzonato, E., Hagolle, O., Inglada, J., Matton, N., Morin, D., Popescu, R., Rabaute, T., Savinaud, M., Sepulcre, G., Valero, S., Ahmad, I., Bégué, A., Wu, B., de Abelleyra, D., Diarra, A., Dupuy, S., French, A., ul Hassan Akhtar, I., Kussul, N., Lebourgeois, V., Le Page, M., Newby, T., Savin, I., Verón, S., Koetz, B., and Defourny, P.: Building a Data Set over 12 Globally Distributed Sites to Support the Development of Agriculture Monitoring Applications with Sentinel-2, Remote Sens., 7, 16062–16090, https://doi.org/10.3390/rs71215815, 2015.

Choden, K., Ravon, S., Epstein, J. H., Hoem, T., Furey, N., Gely, M., Jolivot, A., Hul, V., Neung, C., Tran, A., and Cappelle, J.: Pteropus lylei primarily forages in residential areas in Kandal, Cambodia, Ecol. Evol., 9, 4181–4191, https://doi.org/10.1002/ece3.5046, 2019.

Defourny, P., Jarvis, I., and Blaes, X.: JECAM Guidelines for cropland and crop type definition and field data collection, JECAM, 2014. http://jecam.org/wp-content/uploads/2018/10/JECAM_Guidelines_for_Field_Data_Collection_v1_0.pdf: last access: 17 May 2021

Defourny, P., Bontemps, S., Bellemans, N., Cara, C., Dedieu, G., Guzzonato, E., Hagolle, O., Inglada, J., Nicola, L., Rabaute, T., Savinaud, M., Udroiu, C., Valero, S., Begue, A., Dejoux, J.-F., El Harti, A., Ezzahar, J., Kussul, N., Labbassi, K., Lebourgeois, V., Miao, Z., Newby, T., Nyamugama, A., Salh, N., Shelestov, A., Simonneaux, V., Traore, P. S., Traore, S. S., and Koetz, B.: Near real-time agriculture monitoring at national scale at parcel resolution: Performance assessment of the Sen2-Agri automated system in various cropping systems around the world, Remote Sens. Environ., 221, 551–568, https://doi.org/10.1016/j.rse.2018.11.007, 2019.

Dixon, J., Gulliver, A., and Gibbon, D.: Farming Systems and Poverty: Improving Farmers' Livelihoods in a Changing World, FAO, World Bank, Rome, Washington, 2001.

FAO: Land use, irrigation and agricultural USE, IRRIGATION AND AGRICULTURAL PRACTICES - DEFINITIONS: http://www.fao.org/fileadmin/templates/ess/ess_test_folder/Definitions/Land_Use_Definitions_FAOSTAT.xlsx, last access: 9 September 2020.

Fourie, A.: Better Crop Estimates in South Africa Integrating GIS with other business systems, in: GIS Best Practices - GIS for Agriculture, 9–13, 2009.

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, 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., Velde, M., Dunwoody, A., Kraxner, F., and Obersteiner, M.: Mapping global cropland and field size, Glob. Change Biol., 21, 1980–1992, https://doi.org/10.1111/gcb.12838, 2015.

Fritz, S., See, L., Perger, C., McCallum, I., Schill, C., Schepaschenko, D., Duerauer, M., Karner, M., Dresel, C., Laso-Bayas, J.-C., Lesiv, M., Moorthy, I., Salk, C. F., Danylo, O., Sturn, T., Albrecht, F., You, L., Kraxner, F., and Obersteiner, M.: A global dataset of crowdsourced land cover and land use reference data, Sci. DATA, 4, https://doi.org/10.1038/sdata.2017.75, 2017.

Gbodjo, Y. J. E., Ienco, D., Leroux, L., Interdonato, R., Gaetano, R., and Ndao, B.: Object-Based Multi-Temporal and Multi-Source Land Cover Mapping Leveraging Hierarchical Class Relationships, Remote Sens., 12, 2814, https://doi.org/10.3390/rs12172814, 2020.

Grillo, J. and Holt, J.: Application of the Livelihood Zone Maps and Profiles for Food Security Analysis and Early Warning, https://fews.net/fews-data/335, May 2009.

Hoskins, A. J., Bush, A., Gilmore, J., Harwood, T., Hudson, L. N., Ware, C., Williams, K. J., and Ferrier, S.: Downscaling land-use data to provide global 30″ estimates of five land-use classes, Ecol. Evol., 6, 3040–3055, https://doi.org/10.1002/ece3.2104, 2016.

Ienco, D., Gbodjo, Y. J. E., Gaetano, R., and Interdonato, R.: Weakly Supervised Learning for Land Cover Mapping of Satellite Image Time Series via Attention-Based CNN, IEEE Access, 8, 179547–179560, https://doi.org/10.1109/ACCESS.2020.3024133, 2020.

Inglada, J., Arias, M., Tardy, B., Hagolle, O., Valero, S., Morin, D., Dedieu, G., Sepulcre, G., Bontemps, S., Defourny, P., and Koetz, B.: Assessment of an Operational System for Crop Type Map Production Using High Temporal and Spatial Resolution Satellite Optical Imagery, Remote Sens., 7, 12356–12379, https://doi.org/10.3390/rs70912356, 2015.

Laso Bayas, J., See, L., Perger, C., Justice, C., Nakalembe, C., Dempewolf, J., and Fritz, S.: Validation of Automatically Generated Global and Regional Cropland Data Sets: The Case of Tanzania, Remote Sens., 9, 815, https://doi.org/10.3390/rs9080815, 2017a.

Laso Bayas, J. C., Lesiv, M., Waldner, F., Schucknecht, A., Duerauer, M., See, L., Fritz, S., Fraisl, D., Moorthy, I., McCallum, I., Perger, C., Danylo, O., Defourny, P., Gallego, J., Gilliams, S., Akhtar, I. ul H., Baishya, S. J., Baruah, M., Bungnamei, K., Campos, A., Changkakati, T., Cipriani, A., Das, K., Das, K., Das, I., Davis, K. F., Hazarika, P., Johnson, B. A., Malek, Z., Molinari, M. E., Panging, K., Pawe, C. K., Pérez-Hoyos, A., Sahariah, P. K., Sahariah, D., Saikia, A., Saikia, M., Schlesinger, P., Seidacaru, E., Singha, K., and Wilson, J. W.: A global reference database of crowdsourced cropland data collected using the Geo-Wiki platform, Sci. Data, 4, 170136, https://doi.org/10.1038/sdata.2017.136, 2017b.

Lebourgeois, V., Dupuy, S., Vintrou, É., Ameline, M., Butler, S., and Bégué, A.: A Combined Random Forest and OBIA Classification Scheme for Mapping Smallholder Agriculture at Different Nomenclature Levels Using Multisource Data (Simulated Sentinel-2 Time Series, VHRS and DEM), Remote Sens., 9, 259, https://doi.org/10.3390/rs9030259, 2017.

Leroux, L., Castets, M., Baron, C., Escorihuela, M.-J., Bégué, A., and Lo Seen, D.: Maize yield estimation in West Africa from crop process-induced combinations of multi-domain remote sensing indices, Eur. J. Agron., 108, 11–26, https://doi.org/10.1016/j.eja.2019.04.007, 2019.

Leroux, L., Jahel, C., Ndeye Faye, F., Ndao, B., and Bouquet, E.: In Senegal, agroforestry parklands improves household food accessibility through an improvement of farming system production, 2020a.

Leroux, L., Falconnier, G. N., Diouf, A. A., Ndao, B., Gbodjo, J. E., Tall, L., Balde, A. A., Clermont-Dauphin, C., Bégué, A., Affholder, F., and Roupsard, O.: Using remote sensing to assess the effect of trees on millet yield in complex parklands of Central Senegal, Agric. Syst., 184, 102918, https://doi.org/10.1016/j.agsy.2020.102918, 2020b.

Long, Y., Xia, G., Li, S.,Yang, W., Yang, M.Y., Zhu, X., Zhang, L., Li, D.: DiRS: On Creating Benchmark Datasets for Remote Sensing Image Interpretation: https://www.researchgate.net/publication/342377115_DiRS_On_Creating_Benchmark_Datasets_for_Remote_Sensing_Image_Interpretation, last access: 7 September 2020.

Ndao, B., Leroux, L., Gaetano, R., Diouf, A. A., Soti, V., Mbow, C., Bégué, A., and Sambou, B.: Landscape heterogeneity analysis using geospatial techniques and a priori knowledge in Sahelian agroforestry systems of Senegal, Ecol. Indic, 2021.

de Oliveira Santos, C. L. M., Lamparelli, R. A. C., Dantas Araújo Figueiredo, G. K., Dupuy, S., Boury, J., Luciano, A. C. dos S., Torres, R. da S., and le Maire, G.: Classification of Crops, Pastures, and Tree Plantations along the Season with Multi-Sensor Image Time Series in a Subtropical Agricultural Region, Remote Sens., 11, 334, https://doi.org/10.3390/rs11030334, 2019.

Tsendbazar, N. E., de Bruin, S., and Herold, M.: Assessing global land cover reference datasets for different user communities, ISPRS J. Photogramm. Remote Sens., 103, 93–114, https://doi.org/10.1016/j.isprsjprs.2014.02.008, 2015.

Valero, S., Morin, D., Inglada, J., Sepulcre, G., Arias, M., Hagolle, O., Dedieu, G., Bontemps, S., Defourny, P., and Koetz, B.: Production of a Dynamic Cropland Mask by Processing Remote Sensing Image Series at High Temporal and Spatial Resolutions, Remote Sens., 8, 55, https://doi.org/10.3390/rs8010055, 2016.

https://doi.org/10.5194/essd-2021-125
Preprint. Discussion started: 25 May 2021

Waldner, F., Fritz, S., Di Gregorio, A., and Defourny, P.: Mapping Priorities to Focus Cropland Mapping Activities: Fitness Assessment of Existing Global, Regional and National Cropland Maps, Remote Sens., 7, 7959–7986, https://doi.org/10.3390/rs70607959, 2015.

Waldner, F., De Abelleyra, D., Verón, S. R., Zhang, M., Wu, B., Plotnikov, D., Bartalev, S., Lavreniuk, M., Skakun, S., Kussul, N., Le Maire, G., Dupuy, S., Jarvis, I., and Defourny, P.: Towards a set of agrosystem-specific cropland mapping methods to address the global cropland diversity, Int. J. Remote Sens., 37, 3196–3231, https://doi.org/10.1080/01431161.2016.1194545, 2016.

Waldner, F., Bellemans, N., Hochman, Z., Newby, T., de Abelleyra, D., Verón, S. R., Bartalev, S., Lavreniuk, M., Kussul, N., Maire, G. L., Simoes, M., Skakun, S., and Defourny, P.: Roadside collection of training data for cropland mapping is viable when environmental and management gradients are surveyed, Int. J. Appl. Earth Obs. Geoinformation, 80, 82–93, https://doi.org/10.1016/j.jag.2019.01.002, 2019.