CropHarvest Datasheet

Gabriel Tseng, Ivan Zvonkov, Catherine Nakalembe, Hannah Kerner

August 2021

1 CropHarvest Datasheet

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

CropHarvest was created to enable global crop type classification and generation of agricultural classification maps, particularly for data-sparse regions and under-represented crops. In particular, it is designed to be a dataset on which models can be pretrained before being finetuned on other tasks of interest. This dataset is also intended to spur new directions of ML research for satellite and geospatial datasets and provide a starting place for AI researchers looking to use satellite data to contribute to global challenges.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

This dataset was created by researchers at NASA Harvest and University of Maryland College Park. It was aggregated from a variety of agricultural datasets.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

This work was funded by the following grants:

- NASA Harvest. Sponsor: NASA Goddard Space Flight Center, Award Number 80NSSC17K0625
- Helmets Labeling Crops. Sponsor: Meridian Institute, Lacuna Fund. Award Number 305334-00001
- Earth Observations for Field Level Agricultural Resource Mapping (EO-FARM): Pilot in Kenya and Mexico in Support of Smallholders. Sponsor: SwissRe Foundation Grant No. 302916-00001
- Estimating Cropped Area and Production in the Feed the Future/Mali Zone of Influence. Sponsor: NASA Goddard Space Flight Center, US Agency

for International Development. Award No. 3302915-00001

- Earth Observation for National Agricultural Monitoring. Sponsor: NASA Goddard Space Flight Center. Award No. 80NSSC20K0264

Any other comments? No

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The dataset consists of a satellite time series from 4 satellite datasets. The data were retrieved using Google Earth Engine and each time series has the following metadata:

- Latitude and longitude location of the time series
- The date of the final timestep in the timeseries
- Whether crops are being grown at that date/location or not
- The collection date of the dataset (i.e., when the labels were collected)
- A key representing the source of the data

In addition, the following metadata is available for some (but not all) the samples:

- The harvesting and planting dates of the field at the latitude and longitude
- A more granular crop type/land use label (e.g. "maize" or "pasture")
- A higher level label determined from the FAO's indicative crop classification

How many instances are there in total (of each type, if appropriate)?

There are 88,145 instances in total. Of these, 28,564 have more granular multi-class crop type/land use labels. The remaining instances only contain binary crop vs. non-crop labels. These labels are drawn from a variety of datasets described in Table 1.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

We filtered the source datasets in the following ways. We excluded samples in which multiple crops were grown during the time series or there was intercropping. In addition, the [6] and [1] datasets contained many more datapoints than the other datasets. We therefore subsampled (at most) 100 labels from each crop type to reduce the geographic imbalance of the combined dataset. Finally, we only included labels collected after 2016 to ensure satellite data could be acquired for each label; at the time of export, Google Earth Engine Sentinel-2 L1C data was available starting from June 2015.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each data instance consists of a crop label and a satellite data time series. We combined 4 different remote sensing datasets to construct a one-year timeseries, with 12 timesteps each representing an aggregated value over 30 days of data. All datasets were retrieved using Google Earth Engine. All data products were upsampled to 10 m/pixel resolution.

- **Multispectral Optical Images** Multispectral information is critical for crop identification: a crop's composition, growth stage, canopy structure, and leaf water content can all affect how it reflects light at different wavelengths [17]. We used Sentinel-2 multispectral observations since this dataset has the highest spatial (10-60 m/px) and temporal (5 day revisit) resolution of current publicly-available satellite datasets. Sentinel-2 has 13 spectral bands including the visible color RGB wavelengths typically used in ML datasets, near-infrared wavelengths which are useful for detecting chlorophyll, and short wave infrared (SWIR) wavelengths which are sensitive to water content [3]. We used the Sentinel-2 Top of Atmosphere Reflectance (Level 1C) available on Google Earth Engine [link]. Optical satellite images contain cloud artifacts that need to be removed. We constructed a cloud-free time series of Sentinel-2 images by using [12] to find the least-cloudy pixel within each 30-day window. We used all bands except B1 (coastal aerosols, used to detect fine particles in the air or contaminants in the water) and B10 (cirrus SWIR, used for cloud detection). In addition, we appended normalized difference vegetation index (NDVI) which is the normalized difference between the near-infrared (B08) and red (B04) bands (NDVI = $\frac{\text{B08}-\text{B04}}{\text{B08}+\text{B04}}$). Because vegetation absorbs red light and reflects near-infrared light, high NDVI values often indicate healthy vegetation [14].

- **Synthetic Aperture Radar (SAR) Data** SAR differs from optical imagery in that instead of passively measuring light reflected from the Earth, SAR sensors beam down radio signals and measure what is reflected back, providing information about about the geometry and water content of the crop. SAR sensors can penetrate cloud cover, making it useful for providing coverage in very cloudy regions and seasons. We used the Sentinel-1 C-band Synthetic Aperture Radar (SAR) Ground Range Detected (GRD) dataset [link]. Sentinel-1 has a 10m resolution with near-global coverage and a highly variable revisit time depending on the region of interest (ranging from several days to several months). For each input, we used either imagery taken during an ascending or descending orbit (depending on which was available for the location). Sentinel-1 emits and receives radio signals at certain polarizations. We used the VV (emit at a vertical and receive at a vertical polarization) and VH (emit at a vertical and receive at a horizontal polarization) bands. We took the median of all available observations within each 30-day window. If no observation was available in the window, we used the temporally closest available observation for that location.

- **Meteorological Data** Crops have distinct spectral-temporal profiles and crop development can be delayed or accelerated by weather conditions [15, 4]. Variations in climate should be considered when classifying crop types, particularly the temperature and precipitation distribution throughout the growing seasons [4]. We used the ERA5 meteorological reanalysis dataset [link], which provides a variety of meteorological data globally at 31 km/px and hourly resolution to capture this information. We used the monthly means product, which gives the monthly average of the reanalysis dataset. We included total precipitation and ground temperature (at 2 m height) from the ERA5 dataset. For each input, we selected the month with the most overlap with each 30 day time period.
- **Topographic Data** The topography of an area can affect its suitability for certain crops [9]. The Shuttle Radar Topography Mission (SRTM) Digital Elevation Model (DEM) [link] provides global elevation at 30 m/px resolution. For each label, we selected the elevation of the latitude and longitude nearest to the label location. We used the surrounding elevations to calculate the slope.

Is there a label or target associated with each instance? If so, please provide a description.

Yes; each instance has a label describing whether the location of the time series contains a crop or not. In addition, a subset of the labels contain more granular crop type/land use labels (e.g., “maize”).

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

To the best of our knowledge, no information is missing from the individual instances.

Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)? If so, please describe how these relationships are made explicit.

The spatial and temporal relationship between the instances are made explicit in the associated latitude and longitude location and export dates.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

Three benchmark tasks are provided, containing in-distribution training and test data sets:

- **Maize vs. rest in Kenya** The goal in this task is to classify samples in Kenya as maize vs. rest using a dataset collected by PlantVillage [7]. This dataset consists of a training set of 1,345 samples (266 maize and 1,079 rest) collected in 2020-2021. We used two evaluation bounding boxes with labels from 2020 which covered approximately $764\ \mathrm{m}^2$ and $1,178\ \mathrm{m}^2$ (total coverage of $1,942\ \mathrm{m}^2$), resulting in 45 polygons total. We used two boxes to increase the diversity of test samples.

- **Coffee vs. rest in Brazil** We used the LEM+ dataset [10] to construct a coffee vs. rest task in Brazil. The training set has 203 samples (21 coffee, 182 rest) collected in 2020-2021. The test set included all polygons from 2020 or 2021 within a $4.2\ \mathrm{km}^2$ bounding box, resulting in 62 polygons total.

- **Crop vs. non-crop in Togo** We used the dataset collected by [8] to construct a binary crop vs. non-crop classification task in Togo. This dataset contains 1,319 samples in the training set 1,319 and 306 samples in the test set.

The motivation behind using these evaluation methods (specifically, generating maps for an area and comparing them to ground truth polygons for Brazil and Kenya, and using a randomly sampled test set in Togo) is because these remote sensing classification models are typically used to create land cover and land use maps; these evaluation methods reflect how these maps would be evaluated, and therefore the real-world utility of models trained using this dataset.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

Some labels were collected by agents in the field using GPS locators. These locators can have an error of several metres.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

To the best of our knowledge, the dataset does not contain confidential data. The labels were collected from existing open source datasets or annotated based on visual interpretation of satellite imagery. All time series were created using publicly available satellite and remote sensing datasets. Data collected in the field only includes the crop type and location and no other personally-identifying information.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.
No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
N/A

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.
N/A

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.
N/A

Any other comments? No.

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
Data collection was done in 3 ways:

- Reported by farmers (e.g., [6])
- Observed from satellite imagery by experts in satellite photo-intepretation
- Collected by observers in the field during the growing season

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?
Data was collected using the following procedures:

- **Photointerpretation** Experts in photointerpretation (e.g., [8]) and volunteers (e.g., [13]) used satellite imagery to label pixels. Validation was done by having multiple labellers label the same points or by comparing to expert annotations, but this was not done for all datasets.

- **Self-reporting** Farmers reported which crops they were going to grow on a plot of land [6].

- **Ground truth collection** Observers visited fields to observe which crop was being grown, and recorded location using OpenData Kit (ODK) on GPS enabled devices. In some cases fields were randomly sampled and others were opportunistically selected. The labels in [11] were validated by experts using remote sensing data.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

Some large datasets (e.g., [6]) were subsampled to prevent geographic over-representation in the dataset. In this case, the subsampling was random within each class.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Cropland data were collected by NASA Harvest team members through on-screen labeling in Google Earth and Collect Earth Online. Field-based crop-type labels from Mali were collected by field agents working with Lutheran World Relief. Agents were equipped with an Android tablet with the ODK application and data collection form. Each agent was paid daily in accordance with their organization's procedures. Other constituent datasets (Table 1) describe their respective data collection practices. Collection practices include crowd-sourcing [13], in-situ data collection [10], and more.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The labels were collected between 2017-2021. Photointerpreted labels used historical satellite data to create the labels (e.g., to observe changes in vegetation cover over a period time). All satellite data was exported in 2021.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

Not to the best of our knowledge.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

N/A

Were the individuals in question notified about the data collection?

If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Any other comments?

No.

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

The following processing was done on the satellite data:

- **Sentinel-2 top of atmosphere reflectance (L1C)**: We constructed a cloud-free representation of S2 L1C data by using [12] to find the least-cloudy pixel within the 30-day time period. We removed the B1 (coastal aerosols, used to detect fine particles in the air or contaminants in the water) and B10 (cirrus SWIR, used for cloud detection) bands.In addition, we included normalized difference vegetation index (NDVI) ($NDVI = \frac{B08-B04}{B08+B04}$), a commonly used index for crop type mapping.
- **Sentinel-1 C-band synthetic aperture radar ground range detected**: We took either the ascending or descending imagery (depending on which was available for the location) and took the VV and VH bands. We took a median of all available pixels within the 30-day time period, and if no pixel was available we took the (temporally) closest available pixel at that location.
- **ERA5 Monthly Means**: No processing
- **SRTM DEM**: In addition to the elevation product provided by the Digital Elevation Model (DEM), we calculated the slope at each location based on the surrounding elevations.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

No. The data was processed prior to export using Google Earth Engine. However, the raw data is obtainable from Google Earth Engine using code included in the public repository.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes; the Google Earth Engine code used to process and export the satellite imagery is available on GitHub (https://github.com/nasaharvest/cropharvest).

Any other comments?

No.

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No; benchmark models have been trained using the dataset.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

All papers which link the dataset (which we are aware of) will be linked in the GitHub repository.

What (other) tasks could the dataset be used for?

This dataset is intended to be used to train machine learning models which can then be used to generate cropland and crop-type maps for regions. It may also be useful as a pre-training for other land use mapping and remote sensing tasks (e.g. yield estimation).

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

Not to the best of our knowledge.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Not to the best of our knowledge.

Any other comments? No.

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

No.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?

The dataset is available on Zenodo (https://zenodo.org/record/5106542). A python package to interact with the dataset is available on GitHub (https://github.com/nasaharvest/cropharvest). The dataset has the following DOI: 10.5281/zenodo.5106542.

When will the dataset be distributed?

The dataset is currently available on Zenodo (https://zenodo.org/record/5106542) and a python package to interact with the dataset (including downloading and extracting files from Zenodo) is available on GitHub (https://github.com/nasaharvest/cropharvest).)

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset will be distributed under a CC BY-SA-4.0 license.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No, to the best of our knowledge. The licenses of all constituent datasets are recorded in Table 1.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No, to the best of our knowledge.

Any other comments?

No.

Maintenance

Who will be supporting/hosting/maintaining the dataset?

Gabriel Tseng and Ivan Zvonkov will be maintaining the dataset, through NASA Harvest. The dataset will be supported and hosted by the NASA Harvest program.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Gabriel Tseng is reachable at gabrieltseng95@gmail.com. Ivan Zvonkov is reachable at izvonkov@umd.edu. Users can also contact Hannah Kerner (NASA Harvest AI/ML Lead) at hkerner@umd.edu.

Is there an erratum? If so, please provide a link or other access point.

No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

Yes. Updates will be communicated through GitHub, specifically through new releases of the CropHarvest python package.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

This dataset does not relate to people.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Yes; datasets are automatically versioned through both Zenodo and Github, with older versions remaining accessible.

If others want to extend / augment / build on / contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

Yes; this mechanism is described on the GitHub repository (https://github.com/nasaharvest/cropharvest). Contributions will be reviewed by the dataset owners/maintainers (Gabriel Tseng and Ivan Zvonkov).

These contributions will be distributed via update communication (through new releases of the CropHarvest python package).

Any other comments?

No.

References

[1] Agriculture and Agri-Food Canada. Annual crop inventory ground truth data. https://open.canada.ca/data/en/dataset/503a3113-e435-49f4-850c-d70056788632, 2021.

[2] Christophe Bocquet. Dalberg data insights uganda crop classification. https://doi.org/10.34911/RDNT.EII04X, 2019.

[3] Yaping Cai, Kaiyu Guan, Jian Peng, Shaowen Wang, Christopher Seifert, Brian Wardlow, and Zhan Li. A high-performance and in-season classification system of field-level crop types using time-series landsat data and a machine learning approach. *Remote sensing of environment*, 210:35–47, 2018.

[4] Saskia Foerster, Klaus Kaden, Michael Foerster, and Sibylle Itzerott. Crop type mapping using spectral–temporal profiles and phenological information. *Computers and Electronics in Agriculture*, 89:30–40, 2012.

[5] Great African Food Company. Great african food company tanzania ground reference crop type dataset. https://doi.org/10.34911/RDNT.5VX40R, 2019.

[6] IGN. Registre parcellaire graphique (rpg). https://geoservices.ign.fr/, 2019.

[7] Annalyse Kehs, Peter McCloskey, John Chelal, Derek Morr, Stellah Amakove, Bismark Plimo, John Mayieka, Gladys Ntango, Kelvin Nyongesa, Lawrence Pamba, Melodine Jeptoo, James Mugo, Mercyline Tsuma, Winnie Onyango, and David Hughes. From village to globe: A dynamic real-time map of african fields through plantvillage. *bioRxiv*, 2019.

Table 1: A list of datasets combined, including their source, country of focus and license. The "Crop type" column describes whether the dataset contained more granular land use labels (e.g. describing a specific crop being grown) - datasets without these labels all had binary "crop" or "non-crop" labels.

Area of Focus	Labels Used	Source	License
Ethiopia	830	NASA Harvest	CC BY-4.0
Sudan	422		
Rwanda	3,600		
Mali	142		
Brazil	36		
Kenya	2,704		
Togo	1,582	[8]	
France (Ile de France)	6,184	[6]	etalab Open License
France (Réunion)	2,776		
France (Martinique)	2,421		
Canada	9,088	[1]	Open Government License
Zimbabwe	49	FEWS NET	CC BY-SA-4.0
Mali	148	Harvest Partner	CC BY-4.0
Kenya	319	[7]	CC BY-SA-4.0
Brazil	800	[10]	CC BY-4.0
Global	35,866	[13]	CC BY-3.0
Uzbekistan, Tajikistan	5,302	[11]	CC BY-4.0
Uganda	233	[2]	CC BY-4.0
Tanzania	392	[5]	CC BY-4.0
Global	14,976	[16]	LP DAAC[1]

[1] All LP DAAC current data and products acquired through the LP DAAC have no restrictions on reuse, sale, or redistribution.

[8] Hannah Kerner, Gabriel Tseng, Inbal Becker-Reshef, Catherine Nakalembe, Brian Barker, Blake Munshell, Madhava Paliyam, and Mehdi Hosseini. Rapid response crop maps in data sparse regions. In *ACM SIGKDD Conference on Data Mining and Knowledge Discovery Workshops*, 2020.

[9] T.J. Mary Silpa and P.T. Nowshaja. Land capability classification of ollukara block panchayat using gis. *Procedia Technology*, 24:303–308, 2016. International Conference on Emerging Trends in Engineering, Science and Technology (ICETEST - 2015).

[10] Lucas Volochen Oldoni, Ieda Del'Arco Sanches, Michelle Cristina A. Picoli, Renan Moreira Covre, and José Guilherme Fronza. Lem+ dataset: For agricultural remote sensing applications. *Data in Brief*, 33:106553, 2020.

[11] Ruben Remelgado, Sherzod Zaitov, Shavkat Kenjabaev, Galina Stulina, Murod Sultanov, Mirzakhayot Ibrakhimov, Mustakim Akhmedov, Victor Dukhovny, and Christopher Conrad. A crop type dataset for consistent land cover classification in central asia. *Scientific Data*, 7(1):250, Jul 2020.

[12] Michael Schmitt, Lloyd Hughes, Chunping Qiu, and Xiao Zhu. Aggregating cloud-free sentinel-2 images with google earth engine. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-2/W7:145–152, 09 2019.

[13] Linda See. A global reference database of crowdsourced cropland data collected using the Geo-Wiki platform, 2017.

[14] John Weier and David Herring. Measuring vegetation (ndvi and evi). https://www.earthobservatory.nasa.gov/features/MeasuringVegetation, 2000.

[15] Wiebke Weymann, Ulf Böttcher, Klaus Sieling, and Henning Kage. Effects of weather conditions during different growth phases on yield formation of winter oilseed rape. *Field Crops Research*, 173:41–48, 2015.

[16] Jun Xiong, Prasad Thenkabail, J Tilton, Murali Krishna Gumma, Pardhasaradhi Teluguntla, Russell Congalton, Kamini Yadav, J Dungan, Adam Oliphant, J Poehnelt, C Smith, and R Massey. Nasa making earth system data records for use in research environments (measures) global food security-support analysis data (gfsad) cropland extent 2015 africa 30 m v001, 2017.

[17] Zhiwei Yi, Li Jia, and Qiting Chen. Crop classification using multi-temporal sentinel-2 data in the Shiyang river basin of China. *Remote Sens.*, 12(24):1–21, dec 2020.