UNDERSTANDING THE REQUIREMENTS FOR SURVEYS TO SUPPORT SATELLITE-BASED CROP TYPE MAPPING: EVIDENCE FROM SUB-SAHARAN AFRICA

George Azzari[a]
Shruti Jain[a]
Graham Jeffries[a]
Talip Kilic[a]
Siobhan Murray[a]

50x2030 WORKING PAPER SERIES

Abstract

With the surge in publicly available high-resolution satellite imagery, satellite-based monitoring of smallholder agricultural outcomes is gaining momentum. This paper provides recommendations on how large-scale household surveys should be conducted to generate data needed to train models for satellite-based crop type mapping in smallholder farming systems. The analysis focuses on maize cultivation in Malawi and Ethiopia, and leverages rich, georeferenced plot-level data from national household surveys that were conducted in 2018-2020 and that are integrated with Sentinel-2 satellite imagery and complementary geospatial data. 26,250 *in silico* experiments are simulated within a machine learning framework to identify the approach to survey data collection that yields optimal data for training remote sensing models. The best model is then applied to map seasonal maize cultivation from 2016 to 2019 at 10-meter resolution in both countries. The analysis reveals that smallholder plots with maize cultivation can be identified with up to 75 percent accuracy. However, the predictive accuracy varies with the approach to georeferencing plot locations and the number of observations in the training data. Collecting full plot boundaries or complete plot corner points provide the best quality of information for model training. Classification performance peaks with slightly less than 60 percent of the training data and seemingly-small erosion in accuracy under less preferable approaches to georeferencing plots results in total area under maize cultivation to be overestimated by 0.16 to 0.47 million hectares (8 to 24 percent) in Malawi.

JEL Codes: C52, C55, C83, Q12.
Keywords: Agriculture, Maize, Crop Type Mapping, Sentinel-2, Earth Observation, Remote Sensing, Household Surveys, Data Integration, Malawi, Ethiopia.

---

[a] The authors are listed alphabetically. George Azzari, Chief Technology Officer, Atlas AI P.B.C., george@atlasai.us. Shruti Jaina, Geospatial Data Scientist, Atlas AI P.B.C, shruti@atlasai.us. Graham Jeffries, Senior Technical Portfolio Manager, Atlas AI P.B.C, graham@atlasai.us. Talip Kilic, Corresponding author. Senior Economist, Development Data Group, World Bank, tkilic@worldbank.org. Siobhan Murray, Technical Specialist, Development Data Group, World Bank, smurray@worldbank.org.
The authors would like to thank Calogero Carletto, Keith Garrett, and the members of the Earth Observation Task Force of the United Nations Global Working Group for Big Data for their comments on the earlier version of the paper. This is a publication of 50x2030 Initiative to Close the Agricultural Data Gap, a multi-partner program that seeks to bridge the global agricultural data gap by transforming data systems in 50 countries in Africa, Asia, the Middle East and Latin America by 2030. For more information on the Initiative, please visit https://www.50x2030.org/.

1. Introduction

Agriculture is an integral part of livelihoods in Sub-Saharan Africa, where it can contribute up to 69 percent of household income in rural areas (Davis et al. 2017). As such, improving the productivity of smallholder farmers has been a long-standing goal in many African countries that aim to eliminate poverty and food insecurity.

To monitor progress towards national and international development goals related to agricultural productivity, countries need accurate, crop-specific measures of area under cultivation, production and yields – not only at the national-level but with sufficient within-country disaggregation that can guide targeting and evaluation of policies and programs promoting not only agricultural and rural development but also resilience against disasters and extreme weather events.

With the commencement of the European Space Agency's Sentinel-2 mission in 2015 and the subsequent surge in the public availability of high-resolution satellite imagery, research has shown the feasibility of satellite-based monitoring of agricultural outcomes in smallholder farming systems (Becker-Reshef et al., 2020; Burke and Lobell, 2017; Jin et al., 2017, 2019; Lambert et al., 2018; Lobell et al., 2019, 2020). Latest advances in satellite imagery and remote sensing techniques have the potential to provide timely insights into conditions on the ground and can fill gaps in agricultural monitoring and statistics (Nakalembe, 2020).

Satellite-based approaches to mapping agricultural outcomes, such as crop-specific estimates of cultivated areas and yields, require data for training and validating the underlying remote sensing models. The quality and spatial resolution of satellite-based estimates is directly impacted by the data used for model training and validation (Lobell et al., 2019, 2020). Recent earth observation research that has focused on low-income countries has relied largely on two sources of training and validation data: (i) manually-labeled optical imagery (DeFourny et al., 2019; Xiong et al., 2017; Wei et al., 2020)[1], and (ii) ground data collection, including as part of household and farm surveys (Hegarty-Craver et al., 2020; Jin et al., 2017, 2019; Kerner et al., 2020; Lambert et al., 2018; Richard et al., 2017)[2]. Our paper is related to earth observation applications that rely on

[1] In order to map cropland across the African continent, researchers have manually labeled land cover types for sample points using high-resolution optical imagery as a reference. For example, Xiong et al (2017) created a training dataset covering Africa by labeling homogenous crop/non-crop areas with a minimum area of 90 x 90 meters (0.81 ha). Several satellite-based cropland datasets have been created for the African continent (using manually-labeled satellite imagery - see Wei et al. (2020)'s comparison of the available datasets), and at sub-regional scales (DeFourny et al. 2019).

[2] In order to train and validate remote sensing models for mapping cropland and areas cultivated with maize and non-maize crops, including cassava and beans, at sub-national areas in Tanzania and Kenya, researchers either have used existing survey data that were collected by non-governmental organizations and government agencies to (Jin et al., 2017, 2019; Kerner et al., 2020) or have conducted new survey data collection to meet model training needs (Richard et al., 2017). Similarly, Lambert et al. (2018) rely on sub-national in situ data collected in Mali for satellite-based crop type mapping and yield estimation at the village-level in Konigue commune. Relatedly, Hegarty-Craver et al. (2020)

georeferenced survey data to meet model training and validation needs and that in turn attempt to generate agricultural insights in low-income, smallholder contexts and with greater levels of accuracy, spatial resolution, frequency and timeliness vis-a-vis surveys alone.

On surveys, research has revealed the need to use improved methods for household and farm survey data collection for enhancing our understanding of the agricultural sector, particularly in smallholder systems that stand to benefit the most from the resulting data (Abay et al. 2019; Carletto et al., 2015, 2017; Desiree and Jolliffe, 2018; Kilic et al., 2020). Gourlay et al. (2019) demonstrated that the inverse scale-productivity relationship in agriculture (i.e. the hypothesis that smallholders are more productive than their larger counterparts) may be a statistical artifact, driven by systematic measurement errors in farmer-reported survey data on crop production. The authors showcased these errors by comparing objectively-measured crop cutting and farmer-reported maize yields on the same sample of plots in Eastern Uganda.[3] Follow-up research has demonstrated that survey methods for measuring crop yields directly affect the utility of surveys for earth observation applications, and have provided unambiguous support for the use of objective survey methods to generate the required training and validation data for remote sensing models that integrate survey and satellite data to derive high-resolution estimates of crop yields (Lobell et al. 2019, 2020 in Uganda and Mali, respectively).

Despite the expanding knowledge base regarding the use of earth observation techniques in low-income countries that are primarily characterized by smallholder farming, research studies have largely remained sub-national in scope and have exhibited heterogeneity in terms of the content of the ground data used in support of comparable analytical objectives pursued in different settings. Lack of methodological research to identify the required volume of and approach to ground data collection for training and validating remote sensing models is arguably one of the hurdles against the scale-up of satellite-based estimation of agricultural outcomes across countries and expansive geographies. Identifying ground data requirements for key earth observation applications in low-income countries, including high-resolution crop type mapping and crop yield estimation, would be important not only for assessing the utility of existing georeferenced household survey data for earth observation research but also informing the design of future large-scale household and farm surveys that can provide the required training and validation data for downstream earth observation efforts.[4]

Against this background, this paper addresses several operational and inter-related research questions in the context of high-resolution maize area mapping in Malawi and Ethiopia: 1) what

mapped four crop types (maize, beans, bananas, cassava) in Senegal using training data derived from high-resolution UAV imagery which they collected in the field.

[3] These findings are corroborated by Abay et al. (2019) and Desiere and Jolliffe (2018), in Ethiopia and Uganda, respectively.

[4] Future large-scale surveys that can be a source of training and validation data for earth observation efforts include surveys that are supported by the 50x2030 Initiative.

is the minimum volume of household survey data that is required to reach an acceptable level accuracy of a crop classification algorithm? and 2) how does the approach to georeferencing plot locations as part of household surveys impact the accuracy of the same crop classification algorithm? Furthermore, we demonstrate how our crop classification accuracy is affected based on 1) the type of satellite data used (optical only, radar only or both) - given the considerable differences in the complexity and costs of imagery processing across the various options, and 2) whether plots under specific area thresholds are excluded from the training data - given the potential concerns around the mismatch between the relatively small scale of farming in Malawi and Ethiopia and the Sentinel-2 imagery used in our analysis.

The analysis leverages three national multi-topic household surveys that have been implemented by each country's national statistical agency over the period of 2018-2020 with financial support from the World Bank Living Standards Measurement Study - Integrated Surveys on Agriculture (LSMS-ISA) initiative. The surveys include detailed, plot-level data on crop farming and georeferenced plot locations. Each dataset offers a representative snapshot of the smallholder production system in each country for a given reference season. Linking georeferenced plot-level survey data to publicly available Sentinel-2 imagery and other ancillary geospatial data for the reference agricultural season, we conduct a rich array of sensitivity analyses to assess how crop type prediction accuracy changes when trained on different subsets of plot observations in the survey data - not only in terms of the plot observation count but also the approach to georeferencing plot location. Each data subset was designed to simulate a specific ground data collection scenario. We simulate conditions where, for example, only a certain amount or quality of data is available to train a model, and then compare the out-of-sample prediction accuracies across the scenarios. The results of 26,250 *in silico* experiments shed light on the ground data needs that should be met for household surveys to plan a more enabling role in satellite-based crop type mapping. After identifying the best available model, we apply it to map areas cultivated with maize across Malawi and Ethiopia at 10-meter spatial resolution.

There are six headline findings that emerge from our analysis. First, a simple machine learning workflow can classify pixels with maize cultivation with up to 75 percent accuracy - though the predictive accuracy varies with the survey data collection method and the number of observations available for model training. Second, georeferencing plot locations by collecting either full plot boundaries or complete plot corner points provide the best quality of information for model training. Third, classification performance almost always peaks before or around 4,000 plots - corresponding to slightly less than 60 percent of the training data. Fourth, the seemingly-small erosion in predictive accuracy under less preferable approaches to georeferencing plot locations results in total area under maize cultivation to be overestimated by 0.16 to 0.47 million hectares (8 to 24 percent) in Malawi vis-a-vis the results from the best performing model. Fifth, to avoid overfitting models, it is preferable not to exclude from the training dataset any plot observations based on a minimum area threshold. Finally, in the case of maize area mapping in Malawi, optical

features alone can provide sufficient signal to maximize prediction quality and the potential benefits offered by synthetic aperture radar, i.e. satellite data unaffected by cloud cover, can be offset by additional noise introduced with its use.

The paper is organized as follows. Section 2 describes the survey and earth observation data. Section 3 presents the empirical methodology. Section 4 discusses the results and section 5 concludes.

2. Data

2.1. Survey data

We use nationally representative, multi-topic household survey data collected in Malawi and Ethiopia by the respective national statistical office over the period of 2018-2020 with support from the World Bank Living Standards Measurement Study-Integrated Surveys on Agriculture (LSMS-ISA) initiative. The key variables that drive each survey's sampling design is household consumption expenditures and poverty. However, the surveys do provide large samples of agricultural households and extensive data on their agricultural activities. Maize is the primary crop grown in Malawi, while in Ethiopia, small grains are more prevalent, but maize still plays an important role as a staple crop. More details regarding the survey data are provided below.

2.1.1. Malawi

The survey data in Malawi stem from the Integrated Household Panel Survey (IHPS) 2019 and the Fifth Integrated Household Survey (IHS5) 2019/20. The surveys were implemented concurrently by the Malawi National Statistical Office.

IHPS 2019 is the fourth follow-up to a national sample of households and individuals that had been interviewed for the first time in 2010, and later in 2013 and 2016. At baseline, the IHPS was designed to be representative at the national-level and separately for rural and urban domains.[5] Starting in 2013, the IHPS attempted to track all household members that were interviewed in the last survey round and that were projected to be at least 12 years of age and were known to be residing in mainland Malawi during the follow-up survey round.[6] Once a split-off individual was located, the new household that he/she may have joined vis-a-vis the prior survey round was brought into the IHPS sample. Based on these protocols, the dynamically expanding IHPS sample included 3,181 households in 2019, which can be traced back to 1,491 original households that

[5] For more information on the IHPS rounds, please consult the publicly-available basic information document for each survey - IHPS 2010: https://bit.ly/ihps2010; IHPS 2013: https://bit.ly/ihps2013; IHPS 2016: https://bit.ly/ihps2016; and IHPS 2019: http://bit.ly/ihps2019.

[6] The individuals that were residing in prisons, police compounds or army barracks were excluded from tracking.

had been interviewed in 2010. The IHPS 2019 fieldwork was conducted from April to December 2019, and the households that were determined to have owned and/or cultivated land during the 2018/19 rainy season were attempted to be visited twice, once in the post-planting period and once in the post-harvest period, following the same set of fieldwork protocols that had been used in the prior IHPS rounds.

The IHS5 2019/20 is the second source of survey data in Malawi. Unlike the IHPS 2019, the IHS5 is a cross-sectional survey that is designed to be representative at the national-, urban/rural-, regional- and district-levels. The IHS5 sample includes a total of 11,434 households, distributed across 717 EAs throughout Malawi.[7] The fieldwork was implemented from April 2019 to April 2020, and each sampled household was visited once. The households that were determined to have owned and/or cultivated any land reported information on the last completed rainy season, which could have been 2017/18 or 2018/19 depending on the interview date.

The IHPS 2019 and the IHS5 2019/20 used identical, extensive agricultural questionnaires that elicited information at the parcel-, parcel-plot-, and parcel-plot-crop-level, depending on the topic. Of particular importance to our research is that the surveys identified each crop cultivated on each plot, and in the process, determined whether a given plot was monocropped or intercropped. Further, each cultivated plot that was determined to be within 2 hours of travel (irrespective of the mode of transport) was attempted to be visited with the farmer. The plot area was captured with a Garmin eTrex 30 handheld global positioning system (GPS) unit, and the plot location was georeferenced in two ways: (i) the enumerator captured the GPS coordinates for the corner point at which the plot area measurement commenced and manually inputted the GPS coordinate into the computer-assisted personal interviewing (CAPI) application (i.e. the corner point method), and (ii) the enumerator also captured the perimeter of the plot during the plot area measurement exercise and stored the resulting geospatial polygon on the GPS unit following a naming convention that facilitates the linking of the polygon to the plot record in the household survey data (i.e. the full boundary method).

We refined the initial dataset to isolate the best quality data for the analysis. Plot records were retained only if they possessed both a corner point and a full plot boundary and had a crop type record for the reference rainy agricultural season. Furthermore, if the location information (either corner point, or plot boundary, or both) was duplicated across two or more plots, we dropped all duplicated records, except in cases where one, and only one, of the duplicated records had a high degree of confidence assigned to their location data quality - in these cases, the record with the high degree of confidence was kept and the remaining records were dropped. Lastly, only records with a high degree of confidence in the location data quality (both for the corner point and the plot boundary), as indicated by a metric provided by the GPS unit, were retained. We treated plots that were cultivated with any maize as "maize plots", and otherwise labeled them as "non-maize."

[7] See the IHS5 2019/20 basic information document for more information: http://bit.ly/ihs201920.

Tables 1 and 2 show the IHPS 2019 and the IHS5 2019/20 rainy season plot observations, broken down by georeferenced information availability and by maize cultivation status, respectively. The final analysis sample includes 1,470 IHPS 2019 plots and 3,506 IHS5 plots that are specific to 2018/19 rainy season, and 1,926 IHS5 plots that are specific to 2017/18 season. The total number of agricultural households that are associated with these observations is 1,470 in the IHPS and 5,432 in the IHS5.

Table 1: IHPS 2019 and IHS5 2019/20 rainy season plots by georeferenced information availability

Plot category	**IHPS 2019**		**IHS5 2019/20**	
	Obs	%	Obs	%
Plots with no geolocation information	334	6.2	1,105	6.4
Plots with a corner point, but no polygon boundary	1,365	25.4	4,871	28.4
Plots with a corner point and a polygon boundary, but dropped from analysis	874	16.3	2,139	12.5
Plots with a corner point and a polygon boundary, used for analysis	2,792	52.0	9,059	52.7
Total # of Plots	**5,365**	**100.0**	**17,174**	**100.0**
Total # of Associated Households	**2,335**		**8,770**	

Table 2: IHPS 2019 and IHS5 2019/20 rainy season plots by maize cultivation status, conditional on being used for analysis

	IHPS 2019		**IHS5 2019/20**			
Season	2018/19		2017/18		2018/19	
Crop type	Obs	%	Obs	%	Obs	%
Maize	2,033	72.8	2,330	71.4	4,222	72.9
Non-maize	759	27.2	935	28.6	1,572	27.1
Total # of Plots	**2,792**	**100.0**	**3,265**	**100.0**	**5,794**	**100.0**
Total # of Associated Households	**1,470**		**1,926**		**3,506**	

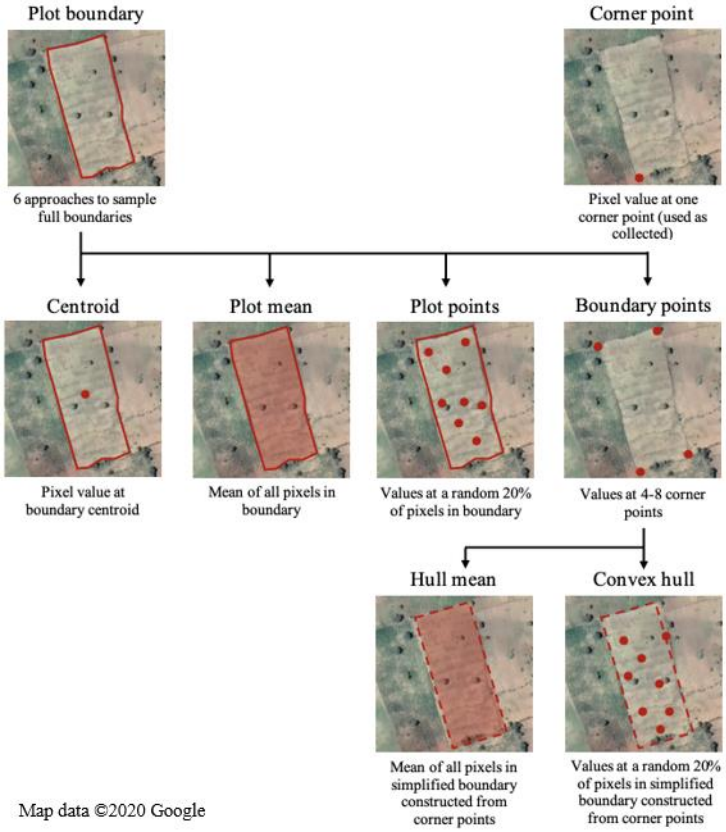
To begin investigating how the approach to georeferencing plot locations would affect the accuracy of remote sensing models that combine survey and satellite data for high-resolution crop type mapping, we used the full plot boundaries to first derive several additional sets of coordinates that could have been generated with alternative plot geolocation methods and that include:

1. The coordinates of one plot corner that is recorded by the enumerator, i.e. “corner point.”
2. The coordinates of the plot centroid that is derived from the full boundary, i.e. “centroid.”
3. The coordinates of 4 to 8 plot corner points that are derived from the boundary, based on the complexity of the plot shape (geometric simplification) and that are in turn used to:

    a. Derive the geospatial predictors for each pixel corresponding to a given corner point and use these pixels and the associated predictors as the training data, i.e. “boundary points.”
    b. Randomly select 20% of the pixels within the convex hull that is formed by the corner points; derive the geospatial predictors of interest for each sampled pixel; and use these pixels and the associated predictors as the training data, i.e. “convex hull.”
    c. Derive the geospatial predictors for all pixels within the convex hull and aggregate the information to the plot-level by taking the average, for each predictor, across all pixels, i.e. “hull mean.”
4. The full plot boundary that is in turn used to:
    a. Randomly select 20% of the pixels from a 10m grid within the plot; derive the geospatial predictors of interest for each sampled pixel; and use these pixels and the associated predictors as the training data, i.e. “plot points.”
    b. Derive the geospatial predictors for all pixels from a 10m grid within the plot and aggregate the information to the plot-level by taking the average, for each predictor, across all pixels, i.e. “plot mean.”

This listing of alternative approaches to georeferencing plot locations on the ground is also indicative of the increasing operational complexity as we move from 1 to 4. Finally, Figure 1 provides a visual overview of these methods and how they impact the computation of geospatial predictors that are specific to each plot.

Figure 1: Plot geolocation methods and approaches for combining plot geometries with pixel-level data



2.1.2 Ethiopia

The survey data in Ethiopia originate from the Ethiopia Socioeconomic Survey (ESS) 2018/19, which was implemented by the Central Statistical Agency as the new baseline for the national longitudinal household survey program. The ESS 2018/19 has been designed to be representative at the national-, urban/rural- and regional-levels, and the sample includes a total of 7,527 households, distributed across 565 EAs throughout Ethiopia. The rural ESS sample includes 3,792 households that originated from 316 EAs that were subsampled from the sample of EAs that were visited by the Annual Agricultural Sample Survey 2018. In each rural EA, the ESS households that cultivated any land during the 2018 (*meher*) agricultural season were visited twice by the resident enumerator, once in the post-planting period and once in the post-harvest period. Similar to the IHS5 and the IHPS, the ESS 2018/19 also used extensive agricultural questionnaires that elicited information at the parcel-, parcel-plot-, and parcel-plot-crop-level, depending on the topic. Each cultivated crop was identified on each plot, and the data are indicative of whether a given plot was monocropped or intercropped.[8] Finally, the ESS CAPI application that leveraged the GPS functionality of the Android tablets enabled each enumerator to georeference the corner point for

[8] See the ESS 2018/19 basic information document for more information: http://bit/ly/ess201819.

starting the plot area measurement (which was then conducted with a Garmin eTrex 30 handheld GPS unit).

For our analysis, the plot records were retained only if they possessed corner point information and had a crop type record for the 2018 *meher* season. Furthermore, if the location information was duplicated across two or more plots, we dropped all duplicated records, except in cases where one, and only one, of the duplicated records had a high degree of confidence assigned to their location data quality - in these cases, the one record with the high degree of confidence was kept and the remaining records were dropped. Lastly, a small share of plots was dropped due to low confidence in their location data quality, as captured by the Android tablet. We treated plots that were cultivated with any maize plantings as maize plots, and otherwise labeled them as "non-maize." Tables 3 and 4 show the breakdown of ESS 2018/19 *meher* season plots, by georeferenced information availability and by maize cultivation status, respectively.[9] The final analysis sample includes 11,095 ESS 2018/19 plots, originating from 2,090 households. Since the ESS 2018/19 did not capture full plot boundaries, our analysis focuses primarily on Malawi, with the findings from Ethiopia playing a supporting role.

Table 3: ESS 2018/19 *meher* season plots by georeferenced information availability

Plot category	**ESS 2018/19**	
	Obs	%
Plots with no geolocation information	1,168	8.7
Plots with a corner point, but dropped from analysis	299	2.2
Plots with a corner point, used for analysis	11,905	89.0
Total # of Plots	**13,372**	**100.0**
Total # of Associated Households	**2,199**	

Table 4: ESS 2018/19 *meher* season plots by maize cultivation status, conditional on being used for analysis

Crop type	**ESS 2018/19**	
	Obs	%
Maize	1,867	15.7
Non-maize	10,038	84.3
Total # of Plots	**11,905**	**100.0**
Total # of Associated Households	**2,090**	

[9] The share of plots with no geolocation information is significantly lower in the ESS 2018/19 data mainly due to the reliance on resident enumerators, as opposed to mobile survey teams in the context of the IHPS 2019 and the IHS5 2019/20.

2.2. Earth observation datasets

Mapping crop area among smallholder plots across a large geographic scale requires satellite remote sensing data sources with high spatial resolution and temporal cadence. Building on precedents set in the research literature, we designed an array of satellite-derived metrics that can be used by a statistical model to distinguish between crop cover types. We used two types of satellite imagery in our maize area mapping experiments - optical and synthetic aperture radar (SAR). Each data source captures different crop properties useful for crop type mapping. For example, optical imagery records information that can be used to characterize a crop's phenology, while SAR imagery captures properties of the canopy structure that may signify differences between crops (Robertson et al. 2020). We processed and extracted both optical and SAR data to the survey plot locations for maize area mapping.

2.2.1 Synthetic aperture radar imagery

Sentinel-1 (S1) satellites carry a Synthetic Aperture Radar (SAR) sensor that operates in a part of the microwave region of the electromagnetic spectrum which is insensitive to water vapor and thus unaffected by clouds or haze. Sentinel-1 Interferometric Wide swath mode (IW) provides images with dual polarization (VV and VH) centered on a single frequency. Google Earth Engine provides S1 images at 10m resolution which are corrected for noise. To use this imagery, we applied Local Incidence Angle (LIA) correction, and computed RATIO and DIFF bands (Table 5).

2.2.2 Optical imagery

Sentinel-2 (S2) satellites provide multispectral imagery for 13 spectral bands with a 10 m resolution for red, green, blue, and near infrared bands used to compute common vegetation indices. We used S2 Level-2A imagery hosted in Google Earth Engine (Gorelick et al. 2017) in our analysis. Image pre-processed included masking out pixels containing clouds, cloud shadows, haze, and snow using quality masks provided with the imagery product (see Appendix Table A1 for summary statistics on imagery counts). Once preprocessed, five vegetation indices (VIs) were calculated for all available S2 images (Table 5).

Table 5: Satellites, bands and indices used in the analysis.

Band / Index	Name	Central wavelength / Index formula	Satellite
VV	Vertically polarized backscatter	5.5465763 cm	Sentinel-1
VH	Horizontally polarized backscatter	5.5465763 cm	Sentinel-1
RATIO	Ratio	VV / VH	Sentinel-1
DIFF	Difference	VV – VH	Sentinel-1
RDED4	Red Edge 4	865 nm	Sentinel-2
GCVI	Green Chlorophyll Vegetation Index	(NIR – GREEN)/1	Sentinel-2
NBR1	Normalized Burn Ratio 1	(NIR – SWIR1) / (NIR + SWIR1)	Sentinel-2
NDTI	Normalized Difference Temperature Index	(SWIR1 – SWIR2) / (SWIR1 + SWIR2)	Sentinel-2
NDVI	Normalized Difference Vegetation Index	(NIR – RED) / (NIR + RED)	Sentinel-2
SNDVI	Smoothed Normalized Difference Vegetation Index	(NIR – RED) / (NIR + RED + 0.16)	Sentinel-2

2.2.3 Harmonic regressions for characterizing crop phenology

We used the multi-temporal collection of bands and indices from S1 and S2 to capture changes in crop phenology over time. To identify temporal patterns that characterize crop phenology, a harmonic regression model was fit at a pixel level to the time series of each unique band and index (Deines et al. 2020, Jin et al. 2019). See Equations 1 and 2 for Malawi and Ethiopia, the latter of which includes an additional pair of harmonic terms. The algorithm produces features that capture the seasonality of different crop types and that include harmonic coefficients, seasonal mean, and goodness of fit measures. These features are useful to map crop types because a maize pixel undergoes seasonal changes in greenness that differ from those of other crops (See Figure 2).

Equation 1

$$GCVI_t = \beta_0 + \beta_1 t + \beta_2 cos(2\pi\omega_1 t) + \beta_3 sin(2\pi\omega_1 t) + \beta_4 cos(2\pi\omega_2 t) + \beta_5 sin(2\pi\omega_2 t) + \epsilon$$

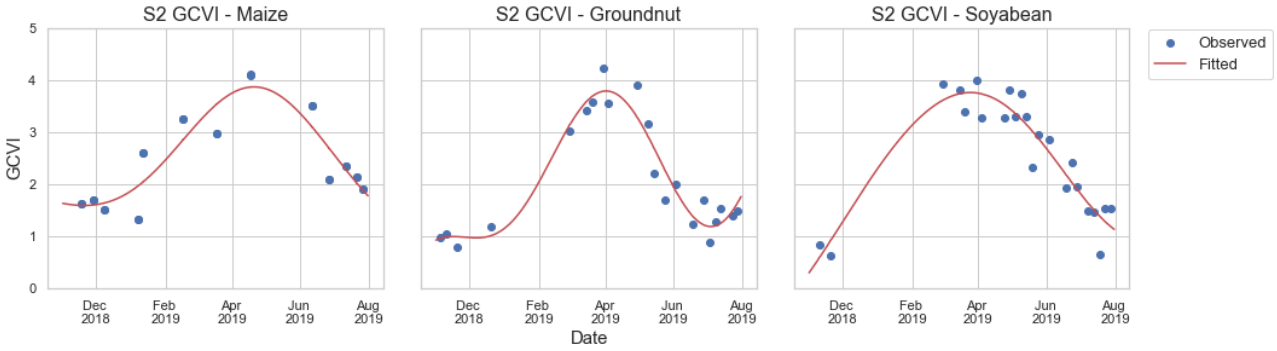
(where $\omega_1 = 1$ and $\omega_2 = 2$)

Equation 2

$$GCVI_t = \beta_0 + \beta_1 t + \beta_2 cos(2\pi\omega_1 t) + \beta_3 sin(2\pi\omega_1 t) + \beta_4 cos(2\pi\omega_2 t) + \beta_5 sin(2\pi\omega_2 t) + \beta_6 cos(2\pi\omega_3 t) + \beta_7 sin(2\pi\omega_3 t) + \epsilon$$

(where $\omega_1 = 1$, $\omega_2 = 2$, and $\omega_3 = 3$)

Figure 2. Examples of the harmonic model smoothing for three different crop types (maize, groundnut, and soybean) using a Sentinel-2 GCVI time series in Malawi. The blue points represent the observed Sentinel-2 GCVI time series at a specific location in Malawi through November 2018 - July 2019. The red line represents the harmonic fitted GCVI time series.



2.2.4 Additional EO data

In addition to multispectral imagery from S2 and SAR imagery from S1, we leveraged data sources that capture landscape and climatological factors correlated with crop type selection. Topography features including elevation, slope, and aspect are commonly incorporated into land cover and land use classifications (Hurskainen et al. 2019). We obtained these three features from the Shuttle Radar Topography Mission (30 m resolution) as proxies for cropland suitability based on the assumption that areas with high slope and elevation are less likely to be suitable for agriculture due to erosion and soil degradation potential. Climate conditions are additional key determinants of crop suitability and therefore can contribute meaningful information in cropland classification models (Konduri 2020). We included weather variables in our models including total precipitation, average temperature, and growing degree days (GDD) during the cropping season. Gridded weather estimates were obtained from the aWhere daily observed weather API (0.1-degree resolution for sub-Saharan African countries, included for Malawi only). Weather data from aWhere was limited to Malawi only due to data licensing constraints. Table 6 shows the additional data used in the pipeline.

Table 6: Additional EO data used in the maize classification pipeline

Feature	Explanation	Data Source	Included in
Elevation	Obtained using GEE's inbuilt *terrain* algorithm that uses an elevation raster to generate slope and aspect bands	Shuttle Radar Topography Mission (30-meter resolution)	Malawi, Ethiopia
Slope			Malawi, Ethiopia
Aspect (direction of slope)			Malawi, Ethiopia
Average temperature	Mean daily temperature during growing season	aWhere daily observed weather API (0.1-degree resolution)	Malawi
GDD	Growing degree days* accumulated during growing season		Malawi
Total precipitation	Total precipitation during growing season		Malawi

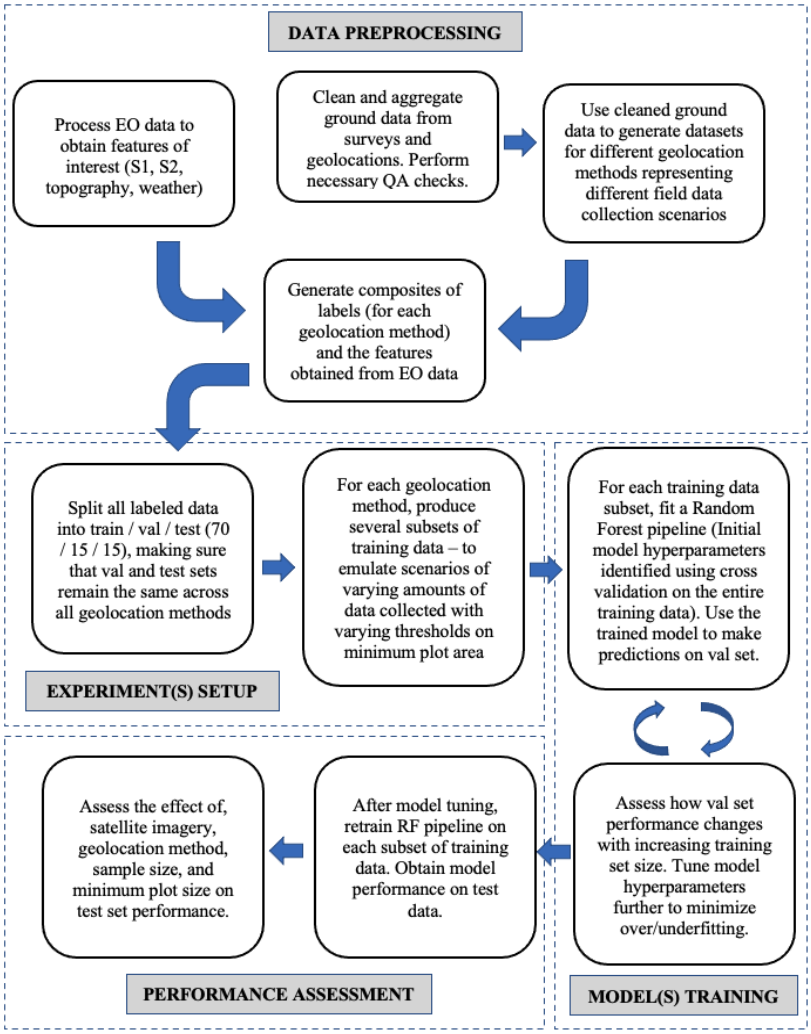
Notes: * A growing degree day is one in which the mean temperature is greater than a base value that must be exceeded for crop growth to occur. For maize, this base value is 10 ℃.

3. Methodology

We developed a methodological framework that is presented in Figure 3 and that is designed to quantify the ability of a machine learning model to identify pixels as maize or non-maize under scenarios with limited training data quantity, various data collection methods, and type of satellite-derived variables used. The overarching approach was to

1. define a common modeling pipeline that trains and evaluates a maize classification model for a given dataset,
2. feed the modeling pipeline with each dataset in a sequence designed to emulate hypothetical scenarios of field data collection (varying the number of observations, the plot geolocation method, and the minimum plot size),
3. vary the type of satellite data used by the modelling pipeline (optical only, radar only, both optical and radar), and
4. compare evaluation metrics across different scenarios. Figure 6 depicts the overall structure of the study.

Figure 3: Workflow



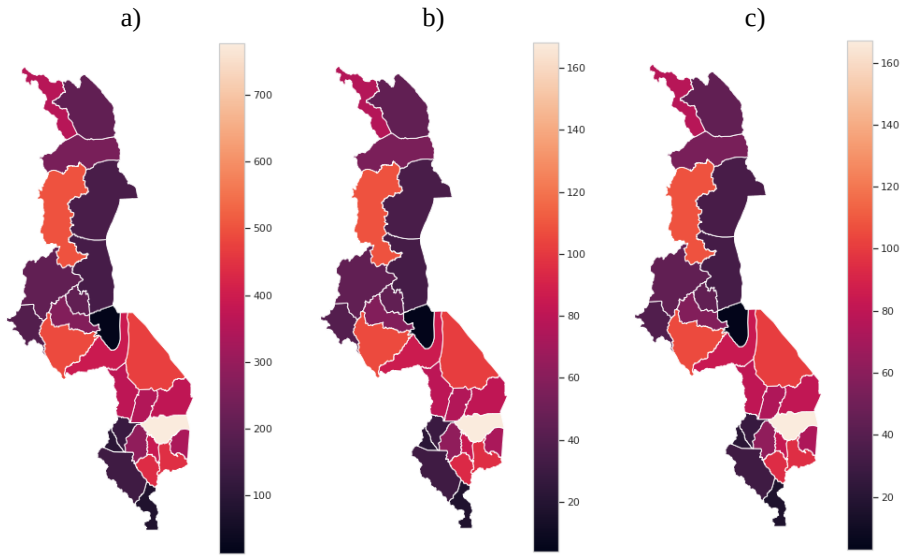
3.1. Maize classification pipeline

A random forest supervised classification model was chosen for the task of pixel-level satellite-based crop type classification. The random forest model was chosen because of its prevalence in related research literature (Jin et al. 2019), due in part to it possessing a good balance between complexity and performance. The maize classification pipeline comprised four stages: 1) feature pre-selection, 2) hyperparameter tuning, 3) model training, and 4) model evaluation. Different portions of the survey data were used for each stage, as explained below.

The complete dataset of surveyed plots in Malawi was divided into subsets for model *training*, *validation*, and *performance testing* (i.e. evaluation). We stratified the dataset by district and crop type (maize and other crops), then divided the records into *train*, *validation*, and *test* subsets (70, 15, and 15 percent of total). Stratifying by geography and crop type ensured that train, validation, and test subsets shared the same balance of crop and non-crop plots. No stratification by year was

applied. Figure 4 below shows the spatial distribution of plots by district across train, validation, and test subsets. The same sampling design was employed in Ethiopia (~13,000 plots)

Figure 4: Distribution of plots across districts in a) train, b) validation, c) and test subsets



Notes: Color indicates number of plots.

Training and validation subsets were used in the maize classification pipeline stages 1 through 3, while the test subset was reserved for model evaluation only.

Feature pre-selection was implemented to prevent model overfitting due to a high number of features (for example, in Malawi: 60 features from S2, 40 from S1, 3 from topography, and 3 from weather). Pre-selection was performed for each dataset passing through the pipeline, rather than the complete dataset, as feature importance may vary with dataset properties (e.g. minimum plot size). Only features with a high Mutual Information score (Equation 3) against the observed dependent variables were kept, such that no two remaining high-ranking features had a correlation of 0.8 or more. See Appendix Table A2 for a listing of all features and selected features.

Equation 3

$$MI(X;Y) = \int \int f(x,y)\, log\, \frac{f(x,y)}{f(x)f(y)}\, dxdy$$

(where $X$and $Y$are two continuous variables with joint p.d.f $f(x,y)$)

A hyperparameter tuning process was designed to minimize overfitting on the training data while maximizing classification performance. A range of values for each of six model parameters were tested in an automated process. Model parameters used in the tuning process included: number of preselected features to use, number of trees in the forest, maximum number of features to consider

when looking for the best split in a tree, maximum tree depth, minimum number of samples required to split an internal node, and minimum number of samples required to be at a leaf node. Model parameters were selected for each dataset by considering feedback from the automated tuning process, in addition to modeler expertise. Models were trained and values for in- and out-of-sample predictions were logged.

Each model was evaluated on its ability to correctly distinguish between maize and non-maize pixels in the testing segment of the dataset (out-of-sample). We calculated two performance metrics: accuracy (Equation 4) and the Matthews' Correlation Coefficient (MCC, Equation 5). Accuracy measures the fraction of correct predictions to total predictions. An accuracy score of 1 represents perfect prediction, and 0 indicates perfectly wrong prediction. MCC improves on the standard accuracy score in cases where the observed prevalence of one prediction class (e.g. not maize) is much larger than other classes. An MCC score of +1 represents a perfect prediction, 0 represents random prediction, and -1 an inverse prediction.

Equation 4

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Equation 5

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

(where TP =True Positives, FP =False Positives, TN =True Negatives, and FN =False Negatives)

3.2. Survey data subsets in accordance with plot area

Burke and Lobell (2017) and Jin et al. (2017) demonstrated that plot size can influence modeled crop yields due to rounding errors. Models trained on observations that exclude very small plots (e.g. < 0.2 ha) commonly perform better because smaller plots can include satellite data pixels that are affected by heterogeneous land use around plot edges. In order to conduct experiments on the effect of a minimum plot size threshold on crop classification accuracy, we created four copies of the stratified and split dataset where training data was filtered to include only plots with areas greater than 0 ha, 0.05 ha, 0.1 ha, 0.15 ha, and 0.2 ha. We retained plots of all sizes in the validation and test subsets to evaluate each model with real-world plot size distributions. The histogram in Figure 5 below shows the distribution of plot areas, and Figure 6 shows the distribution of training plot counts by district when constrained by plot area. Testing the effect of plot area thresholds in Ethiopia was not possible due to the absence of plot boundaries in the survey data.

Figure 5: Distribution of plot areas (ha)

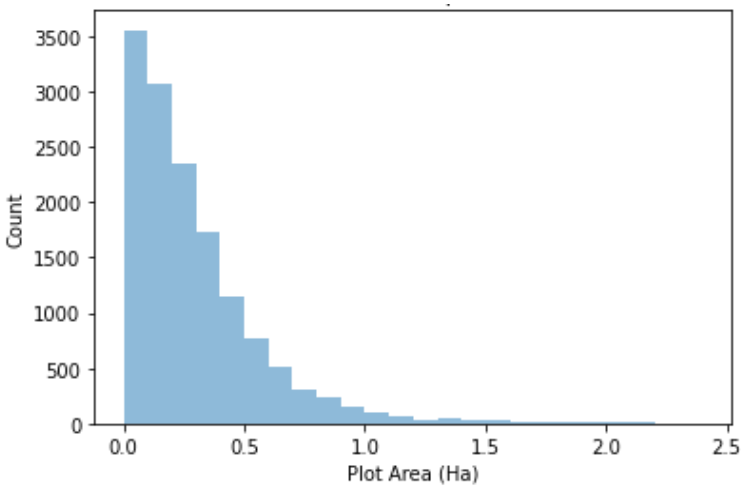
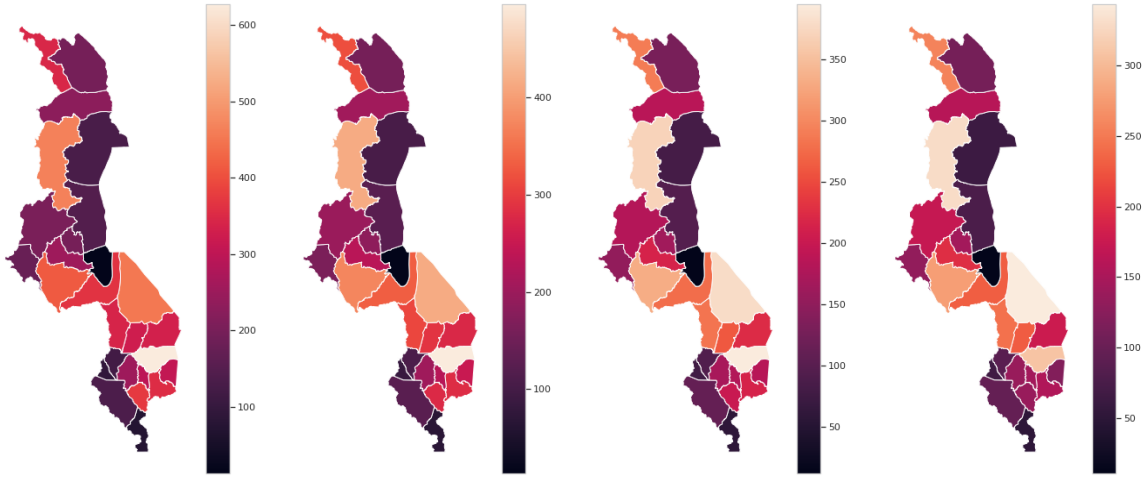


Figure 6: Distribution of plots in train subset across districts if the size of plots were required to be greater than 0.05 ha, 0.1 ha, 0.15 ha, and 0.2 ha (from left to right)



3.3. Modeling data collection scenarios

We applied to the maize classification pipeline a series of datasets designed to emulate data collection scenarios. We described in the "Data" section how, in the case of Malawi, the three types of EO data, seven plot geolocation methods, five minimum plot size thresholds for training data may influence maize classification performance (in Ethiopia there were fewer factors). For each of these 105 scenario datasets, we also included a range of sample size constraints to articulate tradeoffs between data collection effort and classification performance. We defined subsamples of each dataset where the amount of training data was constrained to between 2% and 100% (unconstrained) of the total, iteratively increasing the amount of data available to the modeling pipeline in steps of 2 percentage points. Subsampling of training data was also done in a stratified

manner (by district and class label). Each subsample was passed through the maize classification pipeline and evaluation results were recorded. In total, we tested 26,250 scenarios, comprising:

1. 7 geolocation methods - boundary points, centroid, convex hull, corner, hull mean, plot points, and plot mean
2. 50 data subsets - 2% to 100% subsets of training data, at an increment of 2% points
3. 5 area thresholds - 0, 0.05, 0.1, 0.15, and 0.2 ha
4. 3 feature types - optical only, radar only, both optical and radar
5. 5 replications to capture variability due to random sampling

To compare performance across countries, we applied a similar workflow to the Ethiopia survey dataset. However, due to the limitations of that dataset, we only tested the corner point geolocation method, with no area threshold, and with optical data only.

4. Results and discussion

We conducted sensitivity analyses to understand how survey data properties, especially their number and geolocation method, affect the performance of maize classification predictions. We constructed a maize classification model for each unique combination of the data collection scenarios and compared performance across these scenarios. In this section, we present the drivers of maize classification performance by first focusing on the effects of survey sample size and geolocation method, then plot area thresholds, and finally the source of satellite data used.

4.1. Effect of the approach to georeferencing plot location

Maize classification accuracy scores across geolocation methods typically fell within 2.5 percentage points of each other, with gaps increasing with the number of plots used for model training (Figure 7a). The variation in maize classification performance across geolocation methods was more pronounced in the MCC curves (Figure 7c), with the “hull mean” and “plot mean” methods producing the most significant performance advantage.

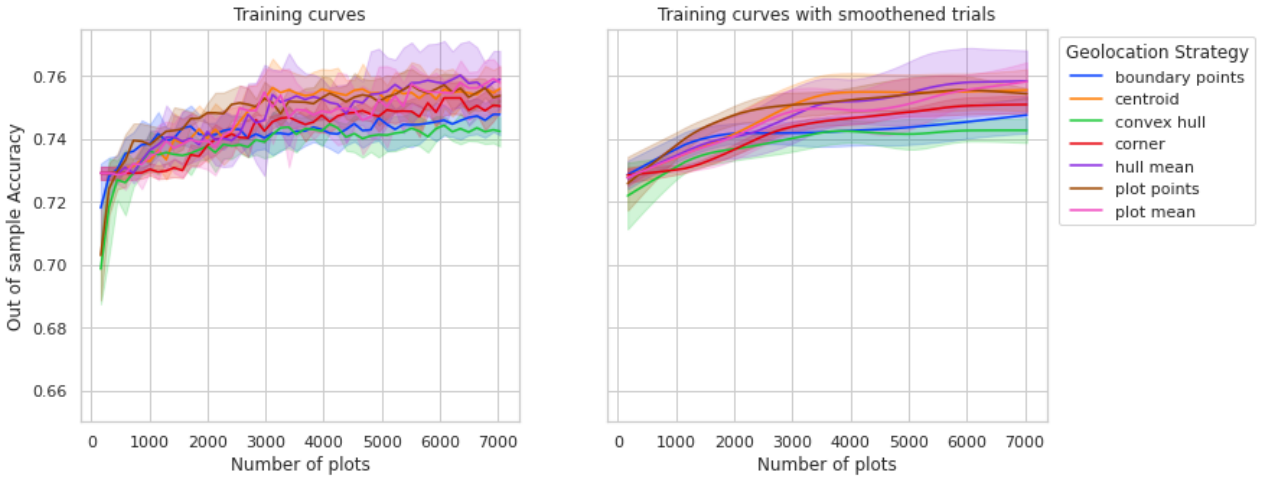
The difference in the accuracy and MCC training curves arises because MCC takes into account all the four confusion matrix categories (true positives, false negatives, true negatives, and false positives), thus providing a more balanced measure of performance. This becomes especially important when we consider geolocation strategies where a number of training points are collected for a single plot, such as “boundary points”, “convex hull”, and “plot points” - they further increase the imbalance between maize and non-maize observations in the training dataset. For this reason, these three geolocation methods show worse MCC performance.

In nearly all cases, the centroid method outperformed corner point method. If only a single GPS point is collected by data collectors, that location should be near the center of the plot. The performance of corner point method was similarly poor in both Malawi and Ethiopia, demonstrated by MCC plots for both countries (Figures 7c and 7d). The plot mean and hull mean methods outperformed all the other methods. Hence, if plot boundaries or multiple corner points are available, the results show that aggregating pixels in the plot (e.g. by taking the mean) is preferable to sampling many pixels from the plot.

Increasing the number of training plots led to an increase in MCC in most cases (Figure 8). When only a small fraction of training observations was available (less than 1,000 plots), geolocation methods where a number of training points can be constructed (such as "boundary points", "convex hull", and "plot points") performed better than others. However, with larger sample sizes (greater than 2,000 plots), "plot mean" and "hull mean" outperformed other methods. The "centroid" geolocation method performed similarly to "plot mean" and "hull mean," except when nearly all of the plots were used for training (around 7,000 plots) in which case "plot mean" and "hull mean" methods outperform the "centroid" method. Lastly, corner points from about 7,000 plots gave roughly the same performance as using the "plot mean" or "hull mean" geolocation strategy on approximately 3,000 plots.

Figure 7: Training curves showing a) test accuracy (Malawi), b) test accuracy (Ethiopia), c) test MCC (Malawi) and d) test MCC (Ethiopia) as a function of training set size for each geolocation strategy in Malawi ((a) and (c)), and for corner point geolocation method in Ethiopia ((b) and (d)). Each training curve is aggregated over 5 trials. The curves shown in the left subplots are aggregated using the trials as is, and the ones on the right are aggregated after first smoothening each trial using a lowess estimator. All figures in the remainder of the results section use the smoothened trials.

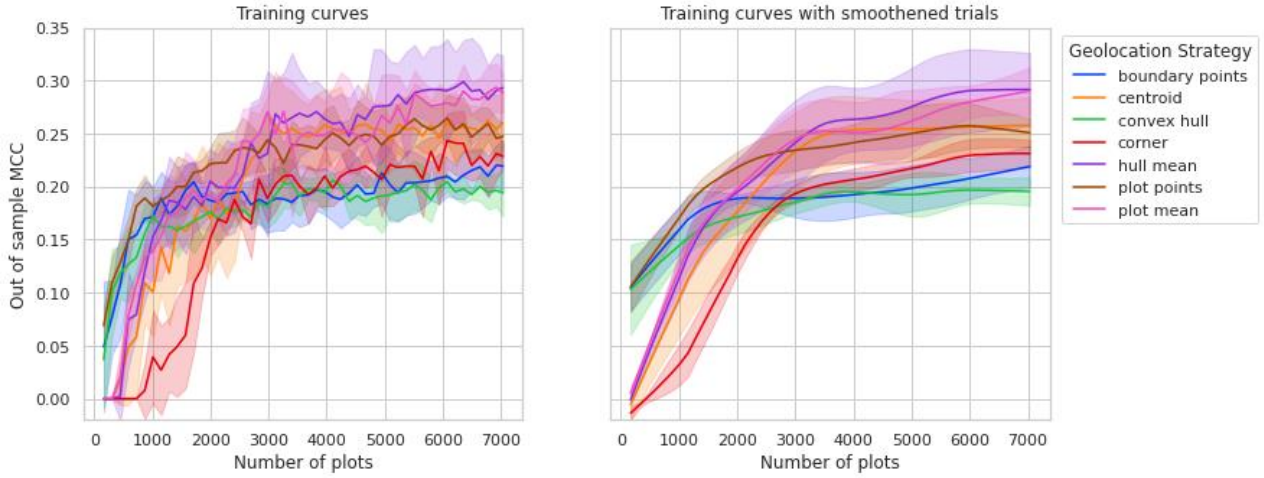
a)



b)

Figure 7 (Continued): Training curves showing a) test accuracy (Malawi), b) test accuracy (Ethiopia), c) test MCC (Malawi) and d) test MCC (Ethiopia) as a function of training set size for each geolocation strategy in Malawi ((a) and (c)), and for corner point geolocation method in Ethiopia ((b) and (d)). Each training curve is aggregated over 5 trials. The curves shown in the left subplots are aggregated using the trials as is, and the ones on the right are aggregated after first smoothening each trial using a lowess estimator. All figures in the remainder of the results section use the smoothened trials.

c)



d)

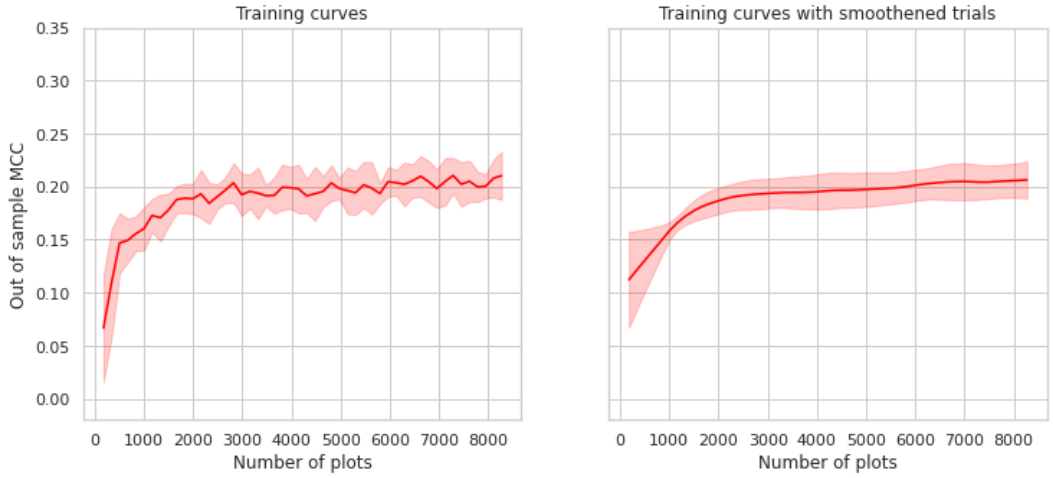
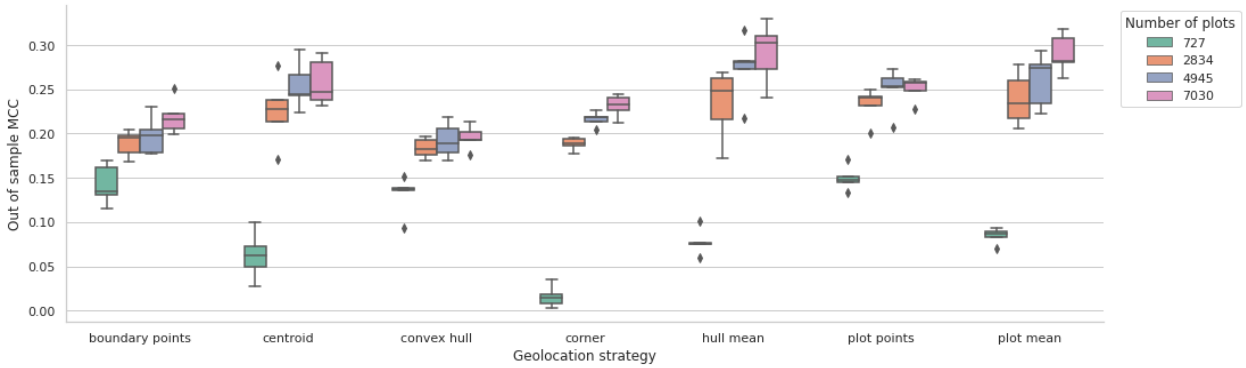


Figure 8: Box plot showing MCC at different training set sizes for each geolocation strategy (Malawi), $n = 727, 2834, 4945, 7030$ correspond to *~10%, 40%, 70%*and *100%*of the training set respectively

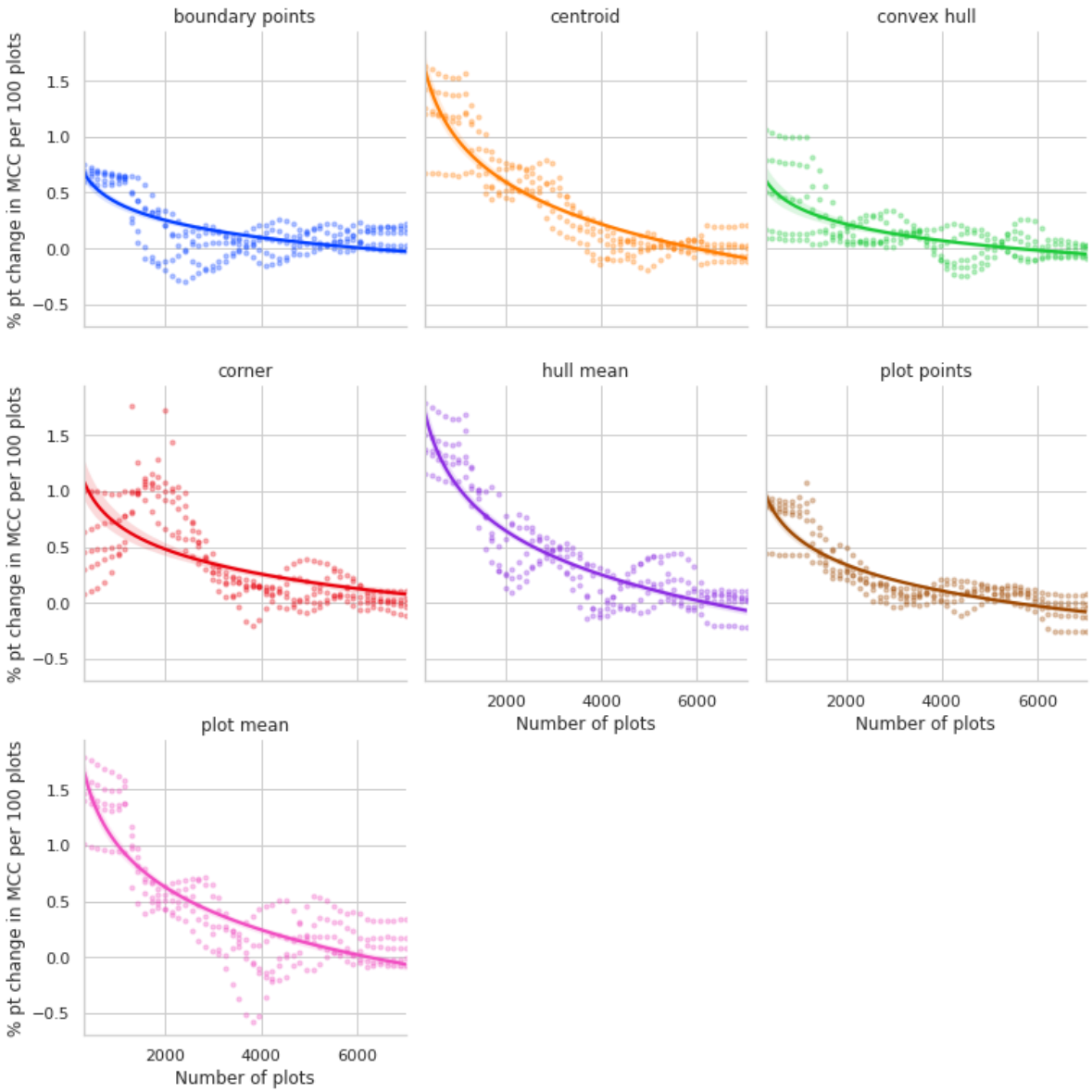


4.2. Effect of sample size

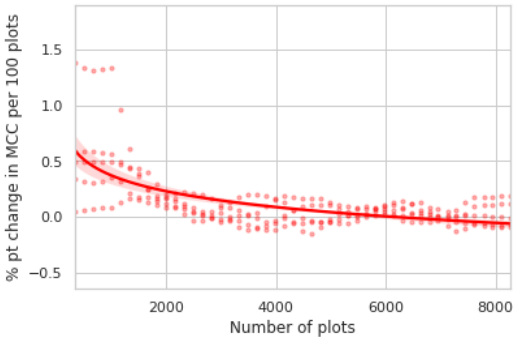
Maize classification performance improves with additional observations, but marginal improvements rapidly diminish after only a small amount of data is available for training (Figure 9). We found that the geolocation method affects not only classification performance, but also how quickly the model improves when provided with incrementally more observations (the "learning rate"). The "centroid," "hull mean," and "plot mean" geolocation methods typically had the highest learning rates when fewer data points were available, in addition to having better overall performance. The "corner point" method showed poor learning rates, especially when more than 3,000 plots are available. We observed that the "corner point" learning rates for a given number of plots were higher in the Malawi case than in Ethiopia.

Figure 9: Trends showing diminishing marginal returns to sample size a) across all geolocation strategies in Malawi and b) for the corner point geolocation strategy in Ethiopia

a)



b)



Returns to sample size slowly vanish depending on geolocation strategy - at around 2,500 plots for “boundary points”, around 4,000 plots for “convex hull”, “hull mean”, “plot points”, and “plot mean,” and around 4,500 plots for “corner” and “centroid.” Peak MCC is calculated as the point where the returns to sample size diminish to <= 0.01 percentage points per 100 plots. The peak MCC obtained at these sample sizes (0.2 - 0.28) also varies by geolocation strategy (Figure 10a). In Ethiopia, the corner point method gave similar values for peak MCC, however the sample size needed to reach that peak was around 3,000 plots which is much lower than that for Malawi, indicating that the model stopped learning sooner in the case of Ethiopia (Figure 10b).

The overall maximum MCC (0.21 - 0.31) varies by geolocation method as well (Figure 11a). The “boundary points” geolocation strategy is able to reach 90% of its maximum at around 2,000 plots, “centroid,” “convex hull,” and “plot points” at around 3,000 plots, and “corner,” “hull mean,” and “plot mean” at around 4,000 plots. Similar behavior was observed for the corner point method in Ethiopia (Figure 11b)

Figure 10: Box plots showing the peak MCC and minimum sample size required to reach the same, a) for each geolocation strategy in Malawi, and b) for corner point method in Ethiopia.

a)

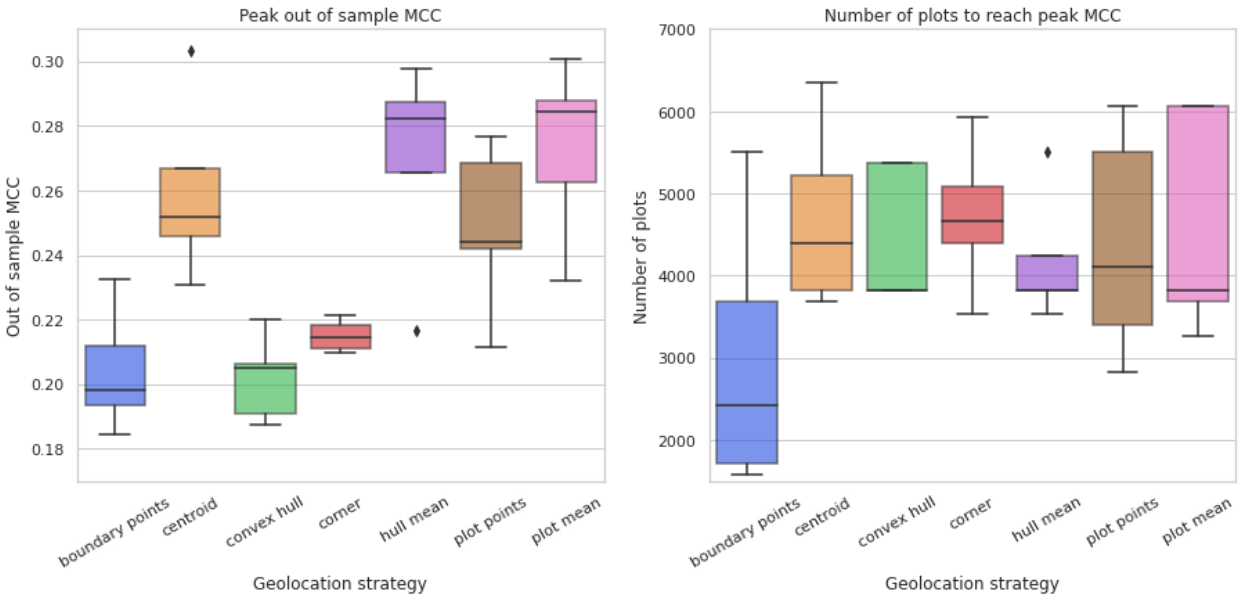


Figure 10 (Continued): Box plots showing the peak MCC and minimum sample size required to reach the same, a) for each geolocation strategy in Malawi, and b) for corner point method in Ethiopia.

b)

Figure 11: Box plots showing the maximum MCC and minimum sample size required to reach ~90% of the same, a) for each geolocation strategy in Malawi, and b) for corner point method in Ethiopia

a)

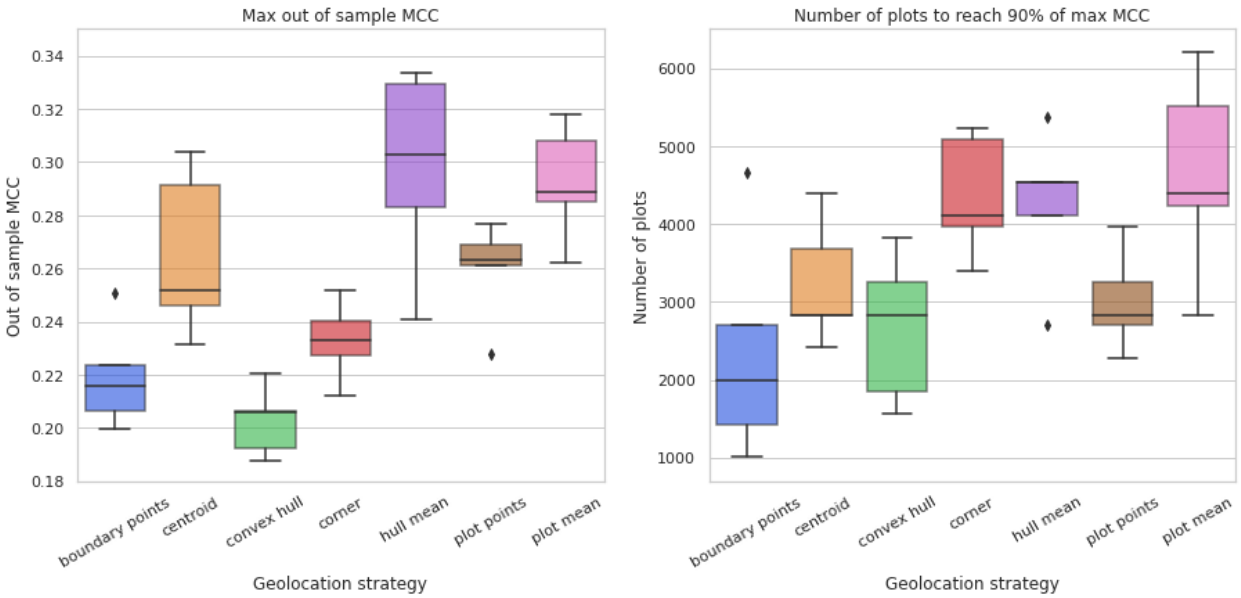
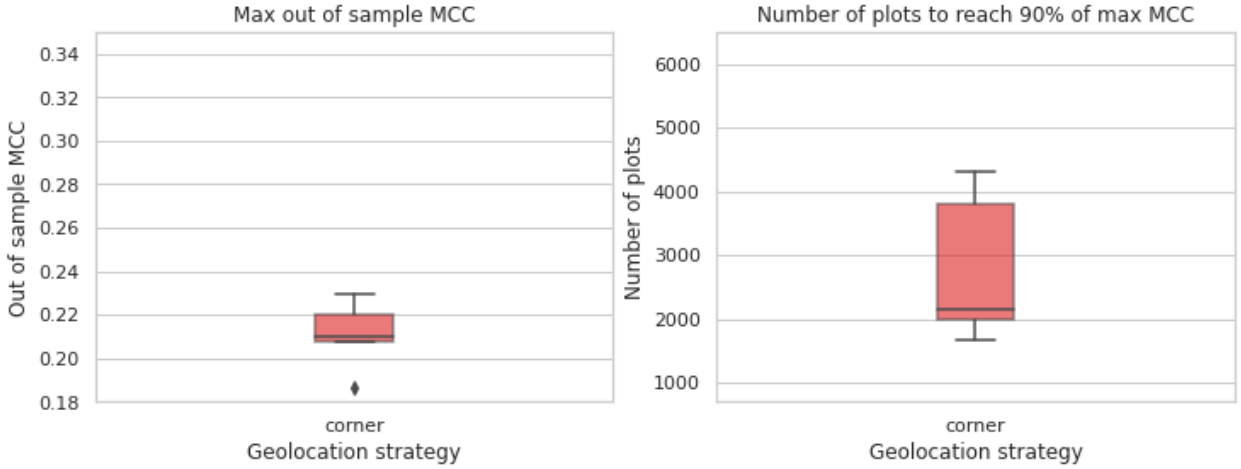


Figure 11 (Continued): Box plots showing the maximum MCC and minimum sample size required to reach ~90% of the same, a) for each geolocation strategy in Malawi, and b) for corner point method in Ethiopia

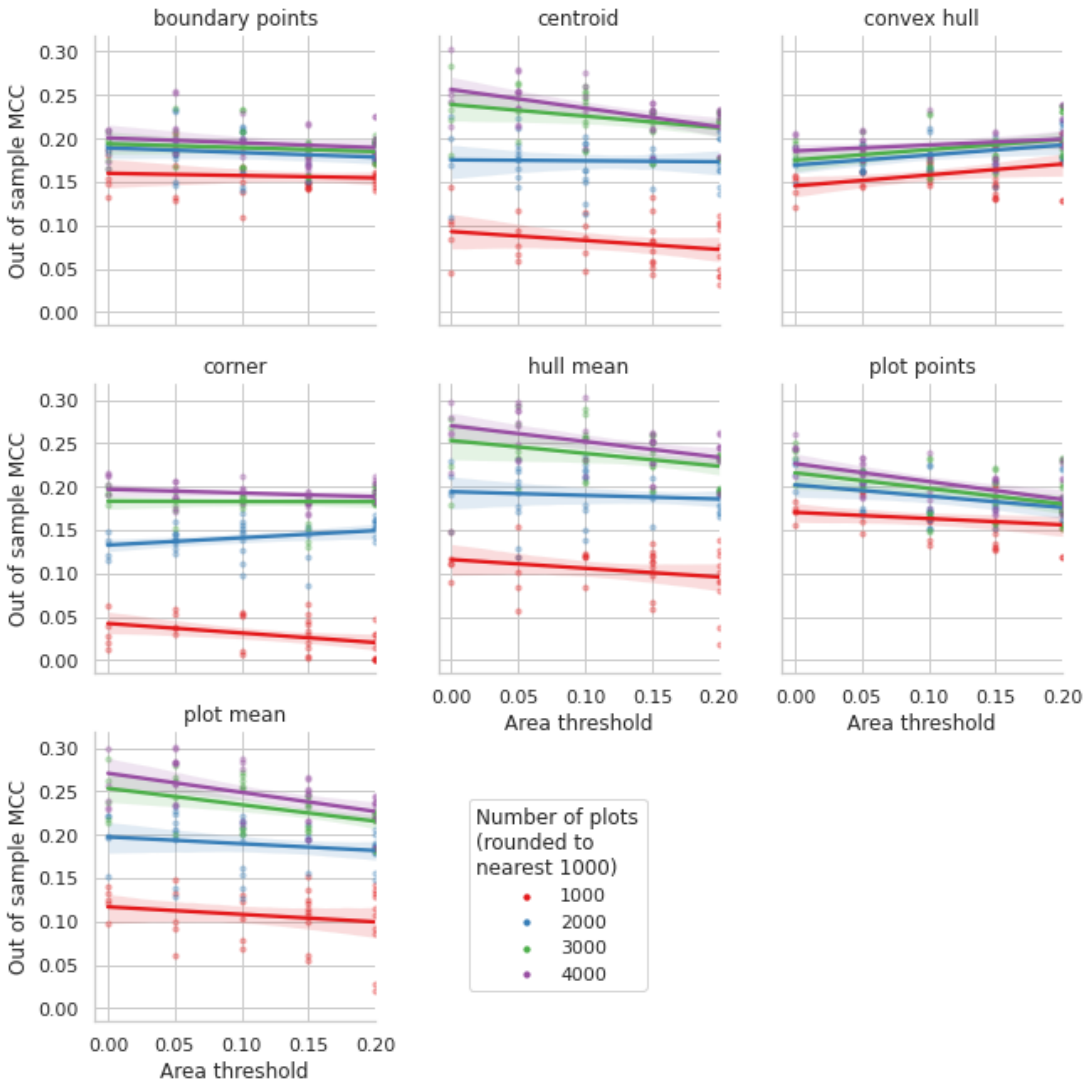
b)



4.3. Effect of minimum plot size

Trends in the relationships between plot size thresholds and MCC suggest that in most cases, limiting training data with plot size criteria decreased MCC scores (Figure 12). The finding is consistent with the intuition that filtering out plots based on a minimum area threshold can significantly change the training data distribution as compared to the validation and/or test data, leading to overfitting. However, in the case of convex hull, filtering out plots by size had a positive effect on MCC. This could be attributed to the fact that in case of smaller plots, the convex hull approximation of a polygon boundary might lead to greater errors, making it preferable to train on bigger plots. We observed that crop type predictions performed better in large plots than in small plots, and that including a plot size threshold exacerbated differences in performance across plot sizes and decreased overall performance.

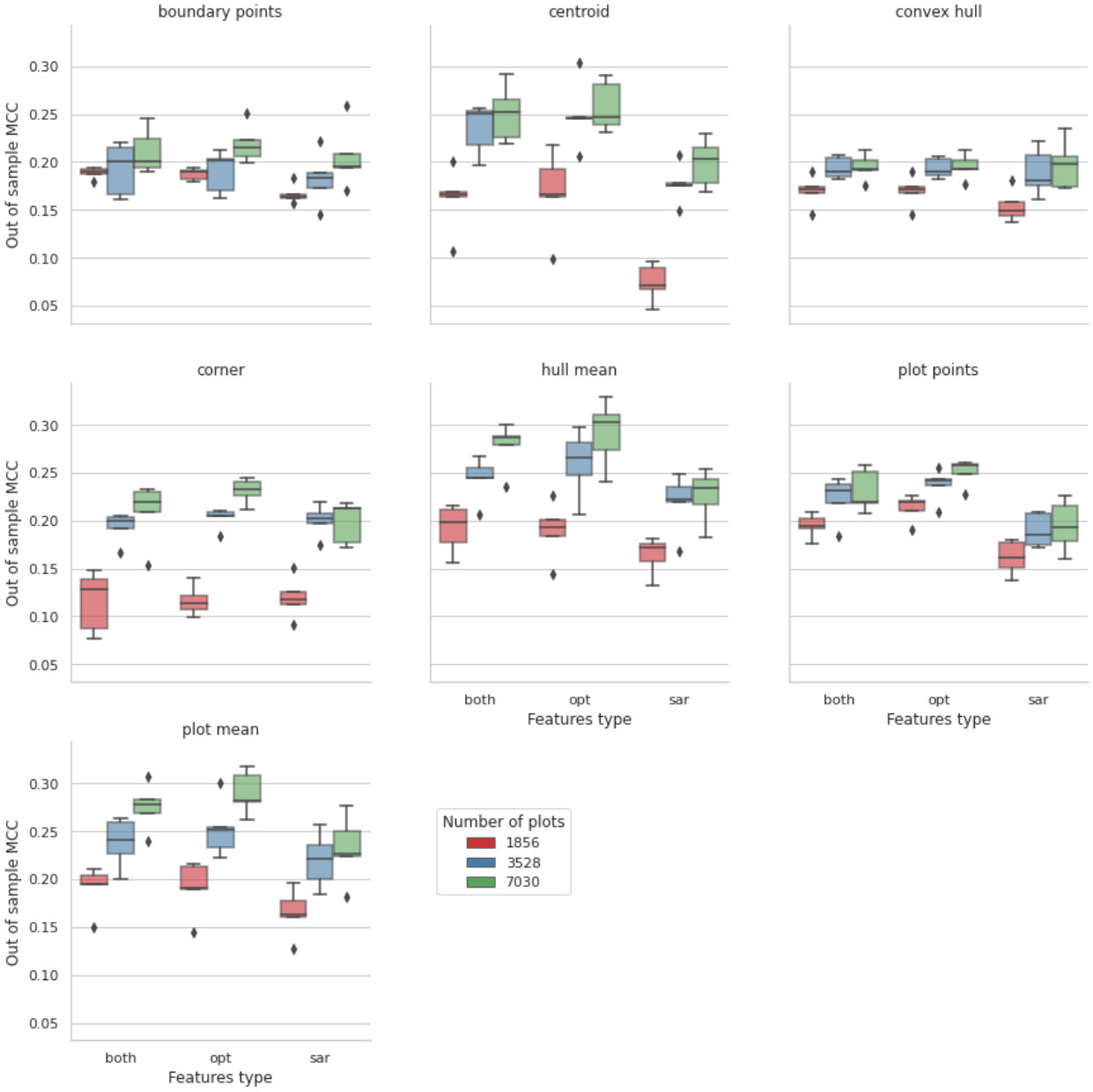
Figure 12: Effects of train plot size thresholds on test MCC (Malawi)



4.4. Effect of satellite data type

We tested the hypothesis that SAR imagery from S1 can be used to detect crop types, alone and in conjunction with optical imagery. However, we found that it was optical features alone that generally produced the most accurate predictions, irrespective of geolocation strategy and sample size. Using both optical and SAR features did not confer MCC gains over a baseline of just using optical features (Figure 13).

Figure 13: Effect of model features on prediction MCC (Malawi)



4.5. Spatial variability of classification performance

We inspected the spatial distribution of maize classification performance to identify significant spatial correlation that may suggest region-specific issues in the model configuration. Figure 14 shows that while there exists a north-south gradient in accuracy (Figure 14a) the same pattern is not detectable for MCC (Figure 14b). This suggests that higher accuracy in the southern part of Malawi may be attributable to a higher concentration of maize production that results in an imbalance in crop types observed in the sample - precisely the motivation for including MCC as an evaluation metric. Figure 15 confirms that regional variations in classification performance are

not strongly correlated with the number of plots. In other words, the model successfully avoids overfitting to regions with a higher density of surveys.

Figure 14: Map of Malawi showing test performance by district, using all the training data from the plot mean sampling strategy, with optical features, and no area threshold (single trial). The performance metric is a) accuracy and b) MCC.

a) b)

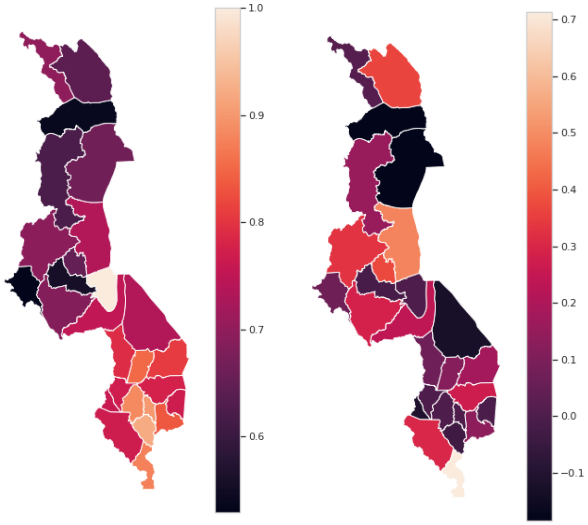
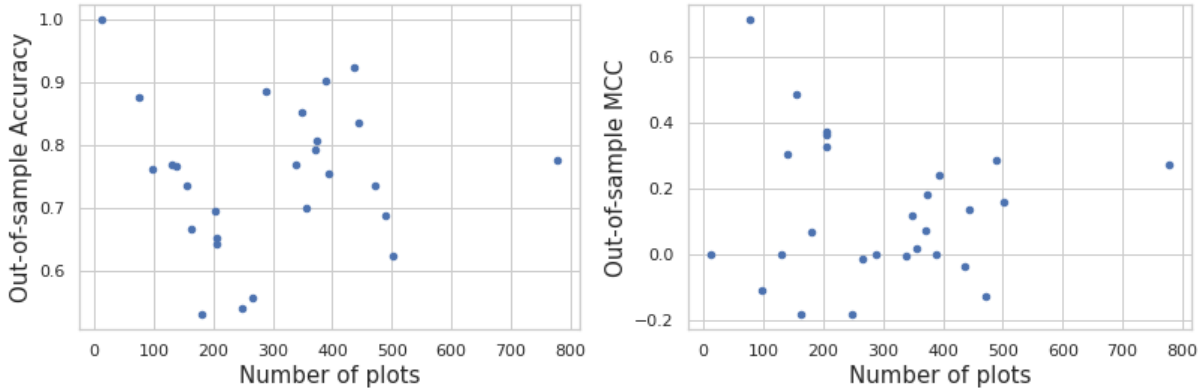


Figure 15: Scatterplots showing the relationship between test performance and the number of training plots by district, using all the training data from the plot mean sampling strategy in Malawi, with optical features, and no area threshold (5 trials). The performance metric is a) accuracy and b) MCC.

a) b)



4.6. Implications of small changes in classification performance

Through the above results, we've demonstrated that certain geolocation methods, area thresholds and satellite features perform better than the others in terms of accuracy and MCC metrics. However, the differences in performances are very small - e.g. Figure 10a shows that the peak MCC for all geolocation strategies varies in the range 0.20 - 0.28. We evaluated how these relatively variations in performance translate into differences in national-level maize area estimations. We trained seven different models, one for each geolocation method, and using the

area threshold and satellite feature set that performed best for each geolocation method (Table 7). Henceforth in this section, each model is referenced by the name of the geolocation method.

Table 7: Testing scenarios used for model training

Geolocation method	**Area threshold**	**Satellite features**	**Out of sample MCC**
Boundary points	0	Optical only	0.21
Centroid	0.05	Optical and SAR	0.24
Convex hull	0.2	Optical only	0.21
Corner	0.05	Optical only	0.23
Hull mean	0	Optical only	0.25
Plot points	0	Optical only	0.24
Plot mean	0.05	Optical only	0.26

Each model was then used to estimate the probability that each 10-meter pixel in Malawi was maize (a 0 to 1 continuous variable) during the 2018/19 rainy season. The pixel-level maize probabilities were converted into a binary classification using a threshold. Pixels with a maize probability above 0.6 were classified as maize, and otherwise are classified as non-maize. Absent objective data on which to empirically calibrate the classification threshold value, we selected a threshold higher than the typical value (0.5) in order to reduce the overclassification of pixels in maize resulting from the overrepresentation of maize plots in our training dataset. Dataset users can select a threshold value that suits their use case.

We used these maizeland maps in conjunction with a cropland mask (showing seasonal cropland coverage) trained on crowdsourced land cover labels (see Appendix) over Malawi to estimate which pixels were cropped with maize in a particular season. Specifically, we first used the cropland mask to remove all pixels in Malawi that were not cropped. We then used each of our trained maize classification models to identify cropped pixels where maize was present. The process resulted in seven different maizeland maps, one for each geolocation method. We calculated the sum of maize pixel areas in the country separately for each geolocation method (Table 8).

Table 8: Malawi maize area as obtained by seven different classification models

Classification model	Out of sample MCC	Total maize area - 2018/19 rainy season (million ha)
Boundary points	0.21	2.27
Centroid	0.24	2.17
Convex hull	0.21	2.46
Corner	0.23	2.15
Hull mean	0.25	1.94
Plot points	0.24	2.41
Plot mean	0.26	1.99
Mean across models	0.23	2.19

We observed in the maize area tabulations that the convex hull and plot points geolocation methods tended to over-classify pixels as maize - consistent with early observations in this paper. Hull mean and plot mean methods showed the most conservative area estimates, possibly because these models take advantage of most information from a single plot while also preventing over-representation of maize in the training data.

Furthermore, Table 9 below shows the relative difference between the plot mean method and each other maize classification method. We chose the "plot mean" method to represent the "best available" method and for each other model we tallied the number of pixels (and their total area) that disagreed with its maize/non-maize prediction. For example, pixels classified as one crop type by the plot mean method, but as the other crop type by centroid method, are considered to be in disagreement.

Table 9: Area misclassified as maize/non-maize under different classification models, as compared to 'Plot mean' which was the best performing model

Classification model	Difference in out of sample MCC	Total area with disagreement (million ha)
Boundary points	-0.05	0.84
Centroid	-0.02	0.48
Convex hull	-0.05	0.69
Corner	-0.03	0.95
Hull mean	-0.01	0.22
Plot points	-0.02	0.55

These results show that while the differences in performance metrics between different modelling scenarios are not very large, small differences can multiply over space leading to substantial differences in maize area estimation. Hence, there is value in achieving small performance gains anchored in better training data.

Finally, after evaluating maize classification performance in Malawi and Ethiopia, we generated 10-meter resolution rasters of area cultivated with maize for both countries over the period of 2016-2019. Table 10 provides an overview of these rasters. After creating the rasters of probability of maize cultivation, we generated binary maizeland masks for each country and season in two steps. We first used our country- and season-specific cropland rasters to remove all pixels that were not cultivated with any crops. Pixels with probability of (any) crop cultivation less than 40 percent were assumed to be non-cultivated. Subsequently, we used our country- and season-specific maizeland rasters to identify which of the cultivated pixels were cultivated with maize. In Malawi, pixels with probability of maize cultivation greater than or equal to 60 percent were assumed to be cultivated with maize. The comparable threshold was 50 percent in Ethiopia.

Table 10: Specifications of predicted maizeland rasters in Malawi and Ethiopia

Country	**Maize classification model specifications**	**Seasons trained on**	**Seasons predicted on**
Malawi	Plot mean geolocation method, 0.05 ha area threshold, Optical features only	2017/18 rainy season, 2018/19 rainy season	2015/16 rainy season 2016/17 rainy season 2017/18 rainy season 2018/19 rainy season
Ethiopia	Corner point geolocation method, No area threshold, Optical features only	2018 meher season	2016 meher season 2017 meher season 2018 meher season 2019 meher season

5. Conclusions

Satellite data sources have tremendous potential for amplifying insights available from household and farm surveys. The research presented here advances our understanding of how to collect optimal plot-level survey data that can train and validate remote sensing models for high-resolution crop type mapping. Specifically, we quantify the interactive effects of (i) plot size, (ii) approach to georeferencing plot locations, and (iii) size of the training dataset on the performance of a machine learning-based maize classification model.

The results show that collecting a complete plot boundary is preferable to competing approaches to georeferencing plot locations in large-scale household surveys and that seemingly-small erosion in maize classification accuracy under less preferable approaches to georeferencing plot locations

consistently results in total area under maize cultivation to be overestimated - in the range of 0.16 to 0.47 million hectares (8 to 24 percent) in Malawi vis-a-vis the results from the best performing model (i.e. plot mean). The analysis reveals that collecting GPS coordinates of the complete set of plot corners, as a second-best strategy, can approximate full plot boundaries and can in turn train models with comparable performance.

Furthermore, when only a few observation plots (less than 1,000 plots) can be visited, full plot boundaries or multiple corner points provide significant gains vis-a-vis plot corner points or plot centroid. With mid-sized samples (3,000 to 4,000 plots), plot centroids can produce similar performance to full plot boundaries. With large sample sizes (around 7,000 plots), plot centroids fall behind full plot boundaries. If only a single GPS point is to be gathered by data collectors, that location should be near the center of the plot rather than at the plot corner. However, georeferencing plot centroids should be understood as a third-best strategy for remote sensing model training purposes. The findings suggest that classification performance almost always peaks before or at around 4,000 plots under the preferred geolocation strategies - corresponding to roughly less than 60 percent of the training data. As such, it is better to collect high quality plot boundaries from 4,000 plots as opposed to corner points from 7,000 plots.

Moreover, we demonstrate that no plot observations should be excluded from model training based on a minimum plot area threshold - another important note for future surveys. Finally, the experiments to quantify the effect of satellite data sources on crop type classification performance suggests that optical features alone can provide sufficient signal to maximize prediction quality. We observed only small differences between models built only with optical features and those using optical and SAR features. In the case of maize area mapping in Malawi, the potential benefits offered by SAR - providing signals unaffected by cloud cover - were offset by additional noise introduced with SAR imagery.

Many outstanding questions remain for future research. Ongoing and future work should focus on improving the predictive power of the maize classification pipeline and gauge the sensitivity of our recommendations for alternative crops and countries. Crop classification accuracies of 0.9 or greater are not unusual in the literature, though small plot sizes in the African context may limit realistically attainable accuracy. Improvements may be gained by removing pixels with few cloud-free observations during the growing season. Experimenting with alternative machine learning approaches and an expanded set of geospatial covariates may increase performance as well. Further work is also needed to distinguish between intercropped and monocropped (“pure stand”) maize plots in order to a) improve classification performance, b) support the creation of intercrop maps and area estimates, and c) lead to continued refinements of downstream research related also to satellite-based crop yield estimation. Related to the latter, future research should similarly identify the minimum-required volume of and approach to survey data collection that would yield optimal data for training and validating remote sensing models for high-resolution crop yield

estimation. Moreover, documenting the accuracy of out-of-season predictions (e.g. using data collected in 2018 to train a model to predict 2019 outcomes) and the extent of decay in model accuracy over time would reveal the required temporal frequency of ground data collection and the relative importance of capturing season-specific conditions. Finally, research on object-based classification and automated detection of plot boundaries using computer vision techniques may additionally help in reducing the data collection requirements for crop area and yield estimation.

References

Abay, K., Abate, G. T., Barrett, C. B., and Bernard, T. (2019). "Correlated non-classical measurement errors, 'second best' policy inference, and the inverse size-productivity relationship in agriculture." *Journal of Development Economics*, 139, pp. 171-184.

Becker-Reshef, I. et al. (2020). "Strengthening agricultural decisions in countries at risk of food insecurity: The GEOGLAM Crop Monitor for Early Warning." *Remote Sensing of Environment*, 237, 11553.

Burke, M., and Lobell, D. B. (2017). "Satellite-based assessment of yield variation and its determinants in smallholder African systems." *PNAS*, 114.9, pp. 2189–2194.

Carletto, C., Gourlay, S., and Winters, P. (2015). "From guesstimates to GPStimates: land area measurement and implications for agricultural analysis." *Journal of African Economies*, 24.5, pp. 593-628.

Carletto, C., Gourlay, S., Murray, S., and Zezza, A. (2017). "Cheaper, faster, and more than good enough: is GPS the new gold standard in land area measurement?" *Survey Research Methods*, 11.3, pp. 235-265.

Davis, B., Di Giuseppe, S., and Zezza, A. (2017). "Are African households (not) leaving agriculture? patterns of households' income sources in rural Sub-Saharan Africa." *Food Policy*, 67, pp. 153– 174

Defourny, P. et al. (2019). "Near real-time agriculture monitoring at national scale at parcel resolution: performance assessment of the Sen2-Agri automated system in various cropping systems around the world." *Remote Sensing of Environment*, 221, pp. 551–568.

Deines, J. M., Patel, R., Liang, S.-Z., Dado, W. and Lobell, D. B. (2020). "A million kernels of truth: insights into scalable satellite maize yield mapping and yield gap analysis from an extensive ground dataset in the US Corn Belt." R*emote Sensing of Environment*, 253, 112174.

Desiere, S. and Jolliffe, D. (2018). "Land productivity and plot size: is measurement error driving the inverse relationship." *Journal of Development Economics*, 130, pp. 84-98.

Gorelick, N. et al. "Google Earth Engine: planetary-scale geospatial analysis for everyone." *Remote Sensing of Environment*, 202, pp. 18-27.

Gourlay, S., Kilic, T. and Lobell, D. B. (2019). "A new spin on an old debate: errors in farmer-reported production and their implications for inverse scale - productivity relationship in Uganda." *Journal of Development Economics*, 141, 102376.

Hegarty-Craver, M. et al. (2020). "Remote crop mapping at scale: using satellite imagery and UAV-acquired data as ground truth." *Remote Sensing*, 12.12, 1984.

Hurskainen, P., Adhikari, H., Siljander, M., Pellikka, P. K. E. and Hemp, A. (2019). "Auxiliary datasets improve accuracy of object-based land use/land cover classification in heterogeneous savanna landscapes." *Remote Sensing of Environment*, 233, 111354.

Jin, Z. et al. (2019). "Smallholder maize area and yield mapping at national scales with Google Earth Engine." *Remote Sensing of Environment*, 228, pp. 115–128.

Jin, Z., Azzari, G., Burke, M., Aston, S. and Lobell, D. B. (2017). "Mapping smallholder yield heterogeneity at multiple scales in Eastern Africa." *Remote Sensing*, 9, 931.

Kerner, H., Nakalembe, C., and Becker-Reshef, I. (2020). "Field-level crop type classification with k nearest neighbors: a baseline for a new Kenya smallholder dataset." Paper presented at the ICLR 2020 Workshop on Computer Vision for Agriculture. Retrieved on March 3, 2021 from https://arxiv.org/abs/2004.03023v1.

Kilic, T., Moylan, H., Ilukor, J., and Pangapanga-Phiri, I. (2020). "Root for the tubers: extended-harvest crop production and productivity measurement in surveys." *Food Policy,* https://doi.org/10.1016/j.foodpol.2021.102033.

Konduri, V. S., Kumar, J., Hargrove, W. W., Hoffman, F. M. and Ganguly, A. R. (2020). "Mapping crops within the growing season across the United States." *Remote Sensing of Environment,* 251, 112048.

Lobell, D. B., Azzari, G., Burke, M, Gourlay, S., Jin, Z., Kilic, Talip, and Murray, S. (2019). "Eyes in the sky, boots on the ground: assessing satellite- and ground-based approaches to crop yield measurement and analysis." *American Journal of Agricultural Economics*, 102, pp. 202–219.

Lobell, D.B., Di Tommaso, S., You, C., Yacoubou Djima, I., Burke, M. and Kilic, T. (2020). "Sight for sorghums: comparisons of satellite-and ground-based sorghum yield estimates in Mali." *Remote Sensing,* 12.1, https://doi.org/10.3390/rs12010100.

Nakalembe, C. (2020). "Urgent and critical need for sub-Saharan African countries to invest in Earth observation-based agricultural early warning and monitoring systems." *Environmental Research Letters*, 15, 121002.

Richard, K. et al. (2017). "Maize cropping systems mapping using rapideye observations in agro-ecological landscapes in Kenya." *Sensors*, 17, 2537.

Robertson, L. D. et al. (2020). "Synthetic Aperture Radar (SAR) image processing for operational space-based agriculture mapping." *International Journal of Remote Sensing,* 41, pp. 7112–7144.

Xiong, J. et al. (2017). "Nominal 30-m cropland extent map of continental africa by integrating pixel-based and object-based algorithms using Sentinel-2 and Landsat-8 data on Google Earth Engine." *Remote Sensing,* 9, 1065.

Wei, Y., Lu, M., Wu, W. and Ru, Y. (2020). "Multiple factors influence the consistency of cropland datasets in Africa." *International Journal of Applied Earth Observation and Geoinformation,* 89, 102087.

Appendix

Appendix Table A1: Summary statistics of pixel-level S2 observation frequency (after pre-processing) within each agricultural season in Malawi

	2016	**2017**	**2018**	**2019**
Mean	22.12	17.79	27.23	27.64
Median	19.04	15.00	23.05	24.05
Variance	216.06	129.74	301.34	324.91
Min	1	1	1	1
Max	178	135	216	227

Appendix Table A2: Example of feature pre-selection for the case of Malawi

GDD*	GCVI_sin2	NDTI_cos1*	NDVI_rmse	RDED4_variance
P_{tot}*	GCVI_t	NDTI_cos2*	NDVI_sin1*	SNDVI_constant
T_{avg}	GCVI_variance*	NDTI_mean	NDVI_sin2*	SNDVI_cos1
aspect*	NBR1_constant*	NDTI_r2*	NDVI_t	SNDVI_cos2
elevation*	NBR1_cos1*	NDTI_rmse*	NDVI_variance	SNDVI_mean
slope*	NBR1_cos2*	NDTI_sin1	RDED4_constant*	SNDVI_r2*
COUNT	NBR1_mean	NDTI_sin2*	RDED4_cos1*	SNDVI_rmse*
GCVI_constant	NBR1_r2*	NDTI_t	RDED4_cos2	SNDVI_sin1
GCVI_cos1	NBR1_rmse*	NDTI_variance*	RDED4_mean	SNDVI_sin2
GCVI_cos2*	NBR1_sin1	NDVI_constant	RDED4_r2*	SNDVI_t
GCVI_mean	NBR1_sin2*	NDVI_cos1*	RDED4_rmse*	SNDVI_variance*
GCVI_r2*	NBR1_t	NDVI_cos2*	RDED4_sin1*	NDVI_sin2
GCVI_rmse*	NBR1_variance	NDVI_mean*	RDED4_sin2*	
GCVI_sin1	NDTI_constant*	NDVI_r2	RDED4_t*	

Notes: * Indicates that the feature was retained after pre-selection process

We constructed a cropland mask layer to capture where annual crops are grown in the region's primary season in a given year. The value of each pixel is a continuous value between 0 and 1 indicating the estimated probability that the land in the pixel was predominantly cropped. Derived from Sentinel-2 imagery, the nominal spatial resolution is 10-meter.

The methods for developing the cropland maps were similar to those employed for crop type mapping, with a few differences. The cropland maps were created by combining various earth observation (EO) datasets with land cover type labels in order to train a random forest model that predicts the probability that a pixel is cropped or not. The EO data sources used to create independent variables were the same as for crop type mapping - Sentinel-2 for multispectral reflectances (10-meter resolution) and Shuttle Radar Topography Mission (30 m resolution) for topography features including elevation, slope, and aspect, and (in the case of Malawi) the aWhere daily observed weather API (0.1 deg resolution for sub-Saharan African countries) for total precipitation, average temperature, and growing degree days (GDD) during the cropping season.

Sentinel-2 imagery (S2) was preprocessed by 1) converting of top-of-atmosphere reflectance values to surface reflectances using a simple linear regression model, and 2) masking out of pixels containing clouds, shadows, haze, snow etc. using Atlas AI's custom anomaly detection model. Once preprocessed, one band and five vegetation indices (VIs) were retained or calculated for all available S2 images (Appendix Table A3). Similar to crop type mapping, multi temporal collection of bands and indices was utilized to capture changes in vegetation phenology over time using harmonic regression models.

Appendix Table A3: Sentinel-2 bands and indices used for land cover mapping

Band / Index	Name	Central wavelength / Index formula	Satellite
RDED4	Red Edge 4	865 nm	Sentinel-2
GCVI	Green Chlorophyll Vegetation Index		Sentinel-2
NBR1	Normalized Burn Ratio 1		Sentinel-2
NDTI	Normalized Difference Temperature Index		Sentinel-2
NDVI	Normalized Difference Vegetation Index		Sentinel-2
SNDVI	Smoothed Normalized Difference Vegetation Index		Sentinel-2

We developed a collection of land cover type observations by manually labelling randomly selected locations within the target geographies. Referring to high resolution basemaps from Google Maps, users were asked to select the land cover type best describing the 10x10-meter pixel

around each random point. Land cover classes included field crop, tree crop or plantation, other vegetation, water, or swamp, building or road, and desert or bare. We assumed that land cover types remained constant over the time period of mapping (2016-2019) and did not collect year-specific land cover records. Limited availability of high-resolution basemaps, and lack of temporal information about them, prevented year-specific data collection. Frequencies of land cover types used for cropland mapping in Malawi and Ethiopia are shown in Appendix Table A4. We collapsed land cover types other than "Field crop" into a single category "other".

Appendix Table A4: Observation counts of land cover classes by country

	Malawi	Ethiopia
Field crop	464	477
Tree crop or plantation	21	66
Other vegetation	711	1,251
Water or swamp	166	24
Building or road	73	59
Desert or bare	71	193
Total	1,506	2,070

The pipeline for cropland classification comprised three stages: 1) feature pre-selection, 2) hyperparameter tuning, and 3) model training. The process for feature pre-selection was the same as described in Section 3.1 - only features with a high Mutual Information score against the observed dependent variables were kept, such that no two remaining high-ranking features had a correlation of 0.8 or more.

Hyperparameter tuning was designed to minimize overfitting on the training data while maximizing classification performance. A range of values for each of five model properties were tested using a 5-fold cross validation approach with folds stratified by district. Stratifying by geography ensured that all five folds shared the same distribution. Model parameters were selected for each dataset by considering feedback from the automated tuning process, in addition to modeler expertise.

The best model was chosen for its ability to correctly distinguish between crop and non-crop pixels in the validation segment of the dataset (out-of-fold). We selected the random forest parameter set that maximized the out-of-fold Matthews Correlation Coefficients (MCC). The best models in Ethiopia and Malawi had MCC scores of 0.52 and 0.44 and accuracies of 0.85 and 0.75, respectively.

The selected models for each country were used to estimate the probability that each pixel in the related region was cropland (0 to 1 continuous variable). The pixel-level maize probabilities were converted into a binary classification using a threshold. Pixels with a maize probability above 0.4 were classified as crop, and otherwise are classified as non-crop.