EYES IN THE SKY, BOOTS ON THE GROUND: ASSESSING SATELLITE- AND GROUND-BASED APPROACHES TO CROP YIELD MEASUREMENT AND ANALYSIS

DAVID B. LOBELL, GEORGE AZZARI, MARSHALL BURKE, SYDNEY GOURLAY, ZHENONG JIN, TALIP KILIC, AND SIOBHAN MURRAY

Understanding the determinants of agricultural productivity requires accurate measurement of crop output and yield. In smallholder production systems across low- and middle-income countries, crop yields have traditionally been assessed based on farmer-reported production and land areas in household/farm surveys, occasionally by objective crop cuts for a sub-section of a farmer's plot, and rarely using full-plot harvests. In parallel, satellite data continue to improve in terms of spatial, temporal, and spectral resolution needed to discern performance on smallholder plots. This study evaluates ground- and satellite-based approaches to estimating crop yields and yield responsiveness to inputs, using data on maize from Eastern Uganda. Using unique, simultaneous ground data on yields based on farmer reporting, sub-plot crop cutting, and full-plot harvests across hundreds of smallholder plots, we document large discrepancies among the ground-based measures, particularly among yields based on farmer-reporting versus sub-plot or full-plot crop cutting. Compared to yield measures based on either farmer-reporting or sub-plot crop cutting, satellite-based yield measures explain as much or more variation in yields based on (gold-standard) full-plot crop cuts. Further, estimates of the association between maize yield and various production factors (e.g., fertilizer, soil quality) are similar across crop cut- and satellite-based yield measures, with the use of the latter at times leading to more significant results due to larger sample sizes. Overall, the results suggest a substantial role for satellite-based yield estimation in measuring and understanding agricultural productivity in the developing world.

Key words: Agricultural productivity, crop yield estimation, crop cutting, maize, remote sensing, Uganda.

JEL codes: C83, Q12.

Improving the productivity of smallholder farmers is widely considered to be one of the most effective avenues for reducing their poverty and food insecurity (Byerlee et al. 2007). With agriculture contributing up to 69% of rural household income in Africa (Davis et al.

David B. Lobell is a professor in the Department of Earth System Science, and the Center on Food Security and the Environment (FSE), Stanford University. George Azzari is chief technology officer at Atlas AI and Marshall Burke is an assistant professor, both in the Department of Earth System Science and the FSE, Stanford University. Sydney Gourlay is a survey specialist at Living Standards Measurement Study (LSMS), Development Data Group (DECDG). Zhenong Jin is an assistant professor in the Department of Bioproducts and Biosystems Engineering, University of Minnesota. Talip Kilic is a senior economist and Siobhan Murray is a technical specialist, both at LSMS, DECDG, the World Bank. The lead principal investigators were Lobell and Kilic on the Stanford University and the World Bank front, respectively. Field data collection for MAPS II (2016) was financed by the World Bank Innovations in Big Data Analytics Program, the World Bank Trust Fund for Statistical Capacity Building—Innovations in Development Data Window, and the CGIAR Standing Panel on Impact Assessment. Terra Bella (now Skysat) provided free high-resolution satellite imagery for the MAPS remote sensing tasking area for research purposes. MAPS I and MAPS II were both implemented using the World Bank Survey Solutions Computer-Assisted Personal Interviewing (CAPI) platform. The research team would like to thank the dedicated management and field staff of the Uganda Bureau of Statistics regarding fieldwork implementation; Mr. Wilbert Drazi Vundru for Survey Solutions programming, fieldwork supervision and survey data quality control; and Ms. Madeline Lisaius for help with image processing. The authors thank the Global Innovation Fund and USAID/BFS for additional funding. Correspondence may be sent to: dlobell@stanford.edu.

Amer. J. Agr. Econ. 00(0): 1–18; doi: 10.1093/ajae/aaz051

2017), and given high rates of expected poverty reduction associated with agricultural growth (Dorosh and Thurlow 2018), such productivity improvements remain a longstanding goal in many African countries. Similarly, at the international level, doubling the productivity and incomes of smallholders have been identified as a key target within the United Nation's Sustainable Development Goal (SDG) 2 of Ending Hunger.

Accurate measurements of crop production, cultivated area, and yield are at the heart of official agricultural statistics and are key to monitoring progress towards national and international development goals, including SDG 2. Further, the survey data underlying these outcomes are frequently used by agricultural economists to investigate a vast array of policy-relevant research topics, including (*a*) the scale-productivity relationship (Larson et al. 2014; Julien et al. 2019); (*b*) agricultural productivity impacts of fertilizer use (Harou et al. 2017), soil quality (Berazneva et al. 2018), land misallocation (Restuccia and Santaeulalia-Llopis 2017), and sustainable land management practices (Arslan et al. 2015); (*c*) farm- and household-level impacts of exposure to extreme weather events (Wineman et al. 2017; McCarthy et al. 2018); (*d*) the extent and cost of gender differences in agricultural productivity (O'Sullivan et al. 2014; Kilic et al. 2015); (*e*) the relationships between agricultural and welfare outcomes at the household- and/or individual-level (Carletto, Corral, and Guelfi 2017; Darko et al. 2018); and (*f*) the comparative effects of agricultural versus non-agricultural growth on poverty reduction (Dorosh and Thurlow 2018; Ivanic and Martin 2018).

The most common way to assess outcomes related to the productivity of smallholder farmers, including land productivity (e.g., crop yields), is by using information collected through in-person interviews for household and farm surveys. For example, the household surveys supported by the World Bank Living Standards Measurement Study—Integrated Surveys on Agriculture (LSMS-ISA) initiative measure plot areas with handheld GPS units and solicit farmer-reported information on crop production and input use, among other topics, at the plot level. These data, together with the multi-topic information solicited by these surveys, have informed a burgeoning field of development research on Africa over the last decade.

Compared to the body of methodological research that has shown severe systematic biases in farmer-reported plot area measures (Carletto et al. 2017) and that have underlined the increasing use of GPS-based plot area measurement in national household surveys, there is a dearth of evidence on the accuracy of farmer-reported crop production. It is, however, known that the process of soliciting farmer-reported production information is mediated by complexities that include (*a*) potential recall bias, (*b*) tendency to round off numbers, (*c*) the use of non-standard measurement units, (*d*) various conditions and states of crop harvest; and (*e*) partial/early crop harvests, among others (Carletto, Jolliffe, and Banerjee 2015). In fact, the emerging body of evidence from various smallholder production systems across Africa has revealed the systematic measurement errors in self-reported crop production (Gourlay, Kilic, and Lobell 2017; Desiere and Jolliffe 2018; Abay et al. 2019) and their non-negligible implications for questions at the heart of agricultural economics, including the scale-productivity relationship. These findings further highlight the critical need to improve the accuracy of methods used to measure land productivity.

A less common but also well-established approach to measure crop yields is by physically harvesting a sub-section of a farmer's plot, also known as crop cutting (Fermont and Benson 2011). Crop cutting provides a more objective way to measure grain production for a part of the plots, but heterogeneity within a plot can lead to sensitivities of crop cut yields to the precise location and size of the crop cut sub-plot vis-à-vis the entire plot (Fermont and Benson 2011). An alternative is to harvest the entire plot, which avoids most of the problems of the prior methods and is therefore frequently considered the "gold standard" yield measurement (Casley and Kumar 1988; Fermont and Benson 2011). However, full plot harvests require a substantial amount of labor and coordination with farmer harvest schedules, which makes them costly and difficult to scale.

Given the limitations of existing approaches, recent work has explored the ability of satellite data to track crop yields. Burke and Lobell (2017) showed that 1 m resolution data from Terra Bella's Skysat sensors (now owned by Planet Labs) were useful for mapping maize yields for farms in western Kenya. This usefulness was measured both by correlation of satellite-based yield estimates with traditional ground-based yield measures, as well as by the ability of satellite-based yields to detect positive yield associations with fertilizer and hybrid seed inputs. This latter aspect was considered especially important since (*a*) ground-based yield measures are inevitably imperfect themselves, and (*b*) detecting response to inputs or some other aspect of farm management is a common motivation for collecting plot-level yield data in the first place. However, objective ground-based measures of productivity were unavailable in Burke and Lobell (2017), which limited the ability to understand the relative extent of measurement error in ground- versus satellite-based measures.

Here, using unique gold-standard data from full-plot harvests across hundreds of smallholder fields, we assess the ability of satellite-based approaches to measure plot-level maize yields on African smallholder farms and to understand how yields respond to productivity-enhancing factors such as soil quality. The analysis uses data from Eastern Uganda from the 2016 round of MAPS: Methodological Experiment on Measuring Maize Productivity, Soil Fertility and Variety, a survey experiment implemented during the first rainy season of 2016 (June–October) in 45 enumeration areas within a 400 square kilometer area spanning the Iganga and Mayuge districts of Eastern Uganda, the leading maize-producing region of the country.

The analysis extends the work presented in Burke and Lobell (2017) in at least three substantial ways. First, the Ugandan maize systems are considerably more subsistence-focused and heterogeneous than the Kenyan counterparts in Burke and Lobell (2017), with generally smaller plot sizes, lower input use, greater prevalence of under-canopy intercrops such as beans and groundnuts, and frequent occurrence of over-canopy intercrops such as cassava and bananas. Thus, Uganda represents a different and, in many ways, more challenging environment in which to test satellite-based crop yield measurement approaches.

Second, whereas Burke and Lobell (2017) relied on farmer self-reported data on maize production, this paper uses objective measures based on survey field team harvests of maize grain for $64m^2$ subplots within each plot ("crop cuts"), as well as whole plot harvests for a random half of our sample ("full plot harvests"). Thus, we can compare different ground-based measures with each other, and with the satellite data.

Third, the study uses data from the Copernicus program's Sentinel-2A satellite, which has coarser spatial resolution but more spectral bands than the Skysat sensor used in Burke and Lobell (2017). Furthermore, whereas Skysat data are currently only available for a small fraction of the Earth's surface each day, Sentinel-2A and its recently launched sister satellite Sentinel-2B each capture imagery every ten days for the entire land surface of the Earth, with an effective five-day repeat for the Sentinel-2 duo since June 2017. These imagery are quickly made available to the public at no cost. For these reasons, Sentinel-2 represents an attractive option for estimating yields over large regions.[1]

All plot-level measures of maize yield, including farmer-reported self-reported production per hectare (SR), sub-plot crop cut production per hectare (CC), full plot crop production per hectare (FP), and variants of remotely sensed production per hectare (RS) rely on GPS-based plot areas. All such measures are also compared to each other using standard statistical approaches, and are used to study the sensitivity of the associations between maize yield and various production factors measured through a combination of a household survey and extensive soil sampling. Overall, we find that SR yields exhibited significant positive bias when compared to CC or FP yields, with an average yield in SR more than double the other two. Although CC yields agreed well with FP in terms of the overall yield distribution, the correlation between CC and FP yields across fields was relatively low, with CC able to capture roughly one-quarter of the variability in FP yields. The RS yields exhibited significant correlations with the ground measures, in some cases exceeding CC yields in the ability to capture variation in FP yields. Moreover, RS yields exhibit correlations with different production factors (e.g., fertilizer, soil quality) that are very similar to those for CC and FP yields, further indicating that RS yields provide a meaningful measure of land productivity.

[1] Burke and Lobell (2017) focused on field campaigns in 2014 and 2015, before Sentinel-2 was operational.

The paper is organized as follows. The next section describes the data, while the following section presents the comparisons among ground-based yield measures, as well as between ground- and satellite-based yield measures, and the results from the estimations of maize yield regressions for each yield variant of interest. The last section discusses these results and summarizes the main conclusions.

Data

MAPS: Methodological Experiment on Measuring Maize Productivity, Soil Fertility and Variety is a two-round household panel survey that was conducted in Eastern Uganda to test the relative accuracy of subjective approaches to data collection vis-à-vis objective survey methods for maize yield measurement, soil fertility assessment, and maize variety identification. Both survey rounds were implemented by the Uganda Bureau of Statistics, with technical and financial assistance provided by an inter-agency partnership that was led by the World Bank Living Standards Measurement Study (LSMS).

Sampling Design and Fieldwork

Analysis in this paper focused on Round II of MAPS in 2016, building on sampling from earlier MAPS I in 2015. In Round 1, a sample of 75 enumeration areas (EAs) were selected in Eastern Uganda, the top maize-producing region of the country, using the 2014 Population and Household Census (PHC) EA frame. We focus on 45 EAs distributed across a 400 square kilometer remote sensing tasking area spanning the Iganga and Mayuge districts (figure 1). Fieldwork was conducted from June to October 2016, and field teams attempted to track and re-interview 540 households that had been interviewed in Round 1 within the tasking area.

Overall, 489 of the 540 households were successfully re-interviewed.[2] As in MAPS I, one maize plot was selected from each household for crop cutting and variety identification components.

MAPS II implemented full-plot crop cutting for a random sub-sample of plots, and increased the area for sub-plot crop cutting (from 4x4m to 8x8m) on each plot. These decisions were anchored in the concerns around intra-plot variability of maize yields. Given the enhancements in the scope of crop cutting data in MAPS II and the interest in the validation of satellite-based approaches to yield estimation, we rely solely on the MAPS II data on 463 households/plots for which sub-plot crop cutting data are available. The only exception, as explained below, is the plot-level data on soil fertility, which is sourced from MAPS I. Table 1 provides a breakdown of 463 plots in accordance with pure stand versus (type of) intercropped cultivation status.

Three visits were made to each household during MAPS II. During the (first) post-planting visit, enumerators solicited information on (*a*) demographic and socio-economic attributes of household members; (*b*) household dwelling characteristics and ownership of durable assets and agricultural implements; and (*c*) area, cultivation pattern, management, pre-harvest labor and seed inputs for all maize plots that were cultivated during the reference rainy season.[3] Following the completion of the household post-planting interview, each enumerator visited the maize plot that was selected in accordance with the protocol detailed in the previous section. At that time, plot boundaries were mapped with a handheld GPS device and crop-cut sub-plots set up for later harvesting and weighing. The crop cut sub-plot location was chosen at random, in accordance with the protocol detailed by Gourlay, Kilic, and Lobell (2017) and in line with international best practices.

During the (second) crop cutting visit, the enumerator harvested the crop cut sub-plots to obtain objectively measured harvest

[2] In total, 34 out of 51 households that we did not interview in MAPS II were due to the fact that they were not cultivating maize in the first season of 2016. The remaining 17 households can be broken down as follows: 5 households could not be tracked or were outside of the tracking area defined as the Iganga and Mayuge districts (5); 4 households had suffered total crop loss prior to post-planting interview; 7 households had already harvested their maize by the post-planting interview; and 1 household refused. Gourlay, Kilic, and Lobell (2017) report that attrition bias is not a concern.

[3] A parcel is conceptualized as a continuous piece of land under a common tenure system, while a plot is defined as a continuous piece of land on which a unique crop or a mixture of crops is grown under a uniform, consistent crop management system, not split by a path of more than one meter in width, and with boundaries defined in accordance with the crops grown and the operator. Therefore, a parcel can be made up of one or more plots. This distinction is key since for the purposes of within-farm analysis of agricultural productivity, the ideal is to capture within-parcel, plot area measurements linked with plot-level measurement of agricultural production.

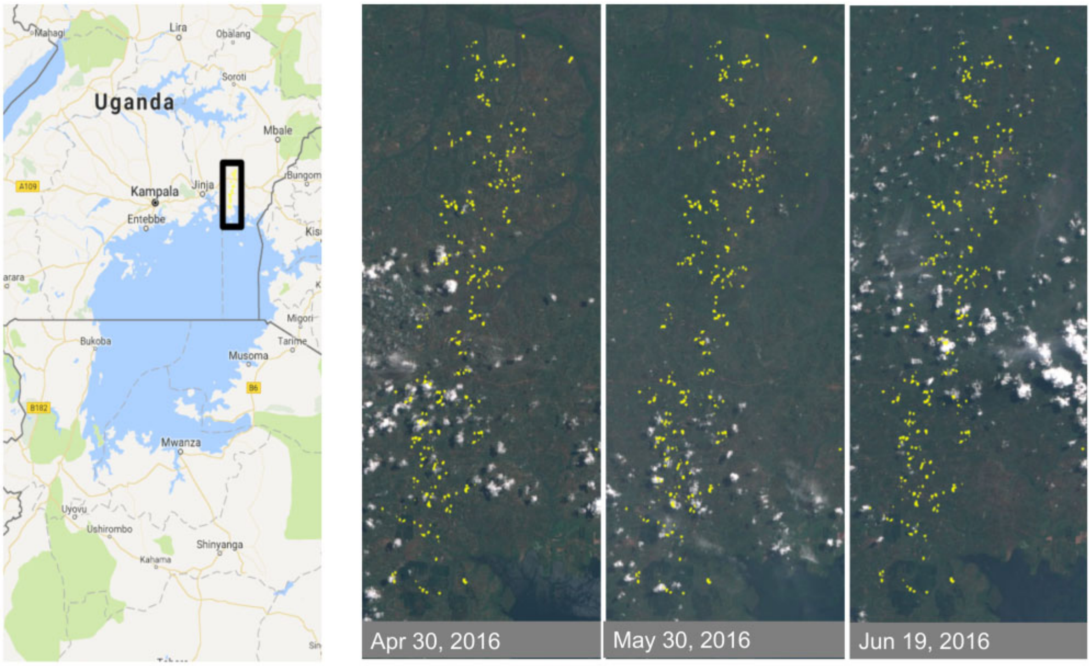

Figure 1 Study region in Eastern Uganda

Note: Three images show Sentinel-2 images and dates used in the study. Polygons indicate outlines of plots where surveys/crop cuts were performed.

Table 1. Distribution of MAPS II Plots by Cultivation Status

Pure stand	Intercropped			
	Maize-Legume	Maize-Cassava	Maize-Legume-Cassava	Maize-Other
124	119	161	52	7

quantities, as detailed in the subsequent section. Finally, during the (third) post-harvest visit, farmer-reported information on total plot-specific maize production, non-labor inputs and harvest labor inputs was solicited for all maize plots that were cultivated during the reference season. The post-harvest visit was scheduled within a two-month period following the completion of each household's harvest.

Key Measurement Domains and Methods

Plot area measurement. After walking the perimeter of a given plot with the plot manager to identify the boundaries, the enumerators re-paced the perimeter and measured the area with a Garmin eTrex 30 handheld GPS device. The area was recorded on the questionnaire in square meters, and the raw GPS track outline was stored. The competing yield measures in our study are all anchored in GPS-based plot area measurement. In MAPS II, the median plot size was 0.11 hectare (ha; roughly one-quarter of an acre), with 46% below 0.10 ha and 17% below 0.05 ha.

Soil fertility assessment. The soil quality index, based on lab analyses of soil samples obtained from the sampled plot locations, is used in our analysis to gauge the possibility of recovering the expected coefficients in production function estimations that use satellite-based yields as dependent variables. Gourlay, Kilic, and Lobell (2017) provide details on the collection of soil samples at each plot location in MAPS I. Briefly, four soil samples were collected at random locations within each plot and were subjected to spectral soil analysis. The resulting data were used to construct a composite soil quality index (SQI), following Mukherjee and Lal (2014). Given the data limitations, the constructed index focuses on nutrient storage capacity but ignores the other two components of soil quality identified by Mukherjee and Lal (2014) related to root development and water storage.[4]

Ground-based maize yield measurement. We construct both farmer self-reported (SR) and crop-cut based yield estimates. For the farmer estimates, plot managers were asked to report their estimate of maize harvest at the parcel-plot-level during the post-harvest visit, as described in Gourlay, Kilic, and Lobell (2017). Each plot manager was allowed to report production in non-standard measurement units, and the dry grain-equivalent harvest quantities in kilograms were calculated by using a conversion factor database developed by UBOS.[5]

To complement SR estimates, we also obtained two crop-cut based measures of plot-level yields. Crop cutting has been recognized as the gold standard for yield measurement since the 1950s by the Food and Agriculture Organization of the United Nations (FAO). Gourlay, Kilic, and Lobell (2017) review the potential concerns regarding yield measurement concerning crop cutting and detail the way in which the MAPS approach to crop cutting and its hands-on supervision overcame them.

In this study, one 8x8m sub-plot (divided into four 4x4m quadrants) was laid on each plot. Each subplot was cordoned off until harvest and was supervised by the EA-specific crop cut monitor between the post-planting and the crop cutting visits. Each plot manager was asked not to harvest any crop from the sub-plots until the crop cutting visit, and not to manage the sub-plot any differently than the rest of the plot. These messages, first communicated by the enumerator, were intended to be enforced by the local crop cut monitors.[6] The shelled maize harvests tied to each of the four adjacent 4x4m quadrants were weighed in the field and then reweighed at a central location in Kampala under strict supervision following additional drying. At the time of the final weighing, the moisture content of each sample was captured to standardize all crop cut sample weights used for our analyses at 12% moisture. The MAPS II sub-plot crop cutting based plot-level maize production estimates are computed by multiplying the crop cut sub-plot production across the $64m^2$ area covered by the 8x8m sub-plot by the ratio of the entire GPS-based plot area in square meters up to $64m^2$.

Furthermore, half of the target household population within each of the pure stand and intercropped domain was selected at random for a full-plot (FP) crop cut. This rare approach to crop production measurement entails the harvesting of the entire plot area, shelling the resulting harvest, weighing it in the field, and capturing its moisture level. This operation was conducted by the enumerators with help from the EA-specific crop cut monitor and the crop cut assistant(s) recruited from within the households. On the MAPS II plots selected for full-plot harvest, the harvest of the designated 8x8m subplot was weighed separately from the full-plot harvest to allow for comparative yield analysis. The full-plot harvests were only weighed in the EAs as their transport to and additional drying and reweighing at a central location was deemed logistically infeasible. Moisture readings taken from the maize grain harvested from the full plot harvests were used to standardize the production quantity to 12% moisture. A total of 211 plots had full-plot harvests. Gourlay, Kilic, and Lobell (2017) detail the approach to full plot harvests. Although farmers were not told the final weight of their harvest, it is likely that the process of harvesting and bagging the maize improved their self-report production values compared to plots without full plot harvests. Therefore, the analyses that use self-reported maize production per hectare rely only on 252 plots without a full plot harvest.

Ground-based SR and FP yields were derived by dividing the reported or measured mass of maize production by the area corresponding to the GPS-based plot area, or $64m^2$, in the case of the 8x8m crop cut sub-plot.

Satellite-based yield measurement. Images from Sentinel-2A, processed to top-of-

[4] The PCA-based soil quality index was constructed for the full MAPS 1 sample, and therefore analyzes the correlation of soil properties and crop cutting yields on a larger sample than MAPS 2.

[5] Refer to Gourlay, Kilic, and Lobell (2017) for more information regarding the conversion factors used in expressing farmer-reported production information in kilogram-equivalent terms.

[6] The lack of statistically significant differences between average CC and FP yields is a finding in support of the assumption that the crop cut sub-plot areas were not managed differently with respect to the rest of the plot. Following the first visit to the sampled households, the supervision of the crop cut sub-plots were conducted on a weekly basis by the local crop cut monitors, who were tasked with visiting the sampled households and sub-plot locations to ensure that the farmers were clear regarding our request for consistency in management practices on the crop cut sub-plot vis-à-vis the rest of the plot. During the fieldwork, the field teams submitted weekly progress reports, none of which referred to any suspected instances of differential management of crop cut sub-plot areas.

Table 2. Spectral Vegetation Indices (VIs) Employed in This Study

Name	Equation	Equation using Sentinel-2 bands	Reference
NDVI (Normalized Difference Vegetation Index)	$(R_{NIR} - R_{RED}) / (R_{NIR} + R_{RED})$	(B8 – B4) / (B8 + B4)	(Rouse et al. 1973)
GCVI (Green Chlorophyll Vegetation Index)	$(R_{NIR} / R_{GREEN}) - 1$	(B8/B3) - 1	(Gitelson et al. 2003)
MTCI (MERIS Terrestrial Chlorophyll Index)	$(R_{NIR} — R_{705}) / (R_{705} — R_{RED})$	(B8-B5) / (B5 – B4)	(Dash and Curran 2004)
NDVI705 (Red-Edge $NDVI_{705}$)	$(R_{NIR} - R_{705}) / (R_{NIR} + R_{705})$	(B8 – B5) / (B8 + B5)	(Gitelson et al. 2003)
NDVI740 (Red-Edge $NDVI_{740}$)	$(R_{NIR} - R_{740}) / (R_{NIR} + R_{740})$	(B8 – B6) / (B8 + B6)	(Gitelson et al. 2003)

Note: R refers to reflectance, and B refers to the corresponding sentinel-2 band number used to compute the VI.

atmosphere reflectance (Level -1 C), were accessed within the Google Earth Engine platform. Sentinel-2A is a polar orbiting satellite carrying a Multi-Spectral Instrument (MSI), which acquires images at ~10:30 a.m. local time for each location on the Earth's land surface roughly every ten days. The MSI measures radiation reflected from the Earth's surface in 13 separate wavelength intervals called "bands", with a spatial resolution of 10m for the visible and near-infrared bands, and 20m to 60m for other bands. For this study, three relatively cloud-free images were available during the growing season, on April 30, May 30, and June 19, 2016. Sentinel-2B, which is identical to Sentinel-2A but staggered by five days, was launched in 2017 and so is not included in this study.

Clouds and shadows were masked from the Sentinel images using a random forest classifier trained on points visually selected from images throughout the region. Five vegetation indices (VIs) that are commonly used in the literature were then calculated for each pixel using the equations shown in table 2. The average value of all bands and VIs within each plot polygon were then extracted for each image date for further analysis, averaging across all pixels with at least half of their area overlapping with the plot. In addition, for comparison with the Sentinel-2A images, an image acquired by Planet Lab's Skysat sensor on May 29, 2016 was accessed. Skysat measures radiance in blue, green, red, and near-infrared channels at a 1m resolution. As with the Sentinel-2 data, clouds and shadows were masked using a random forest classifier trained on several images in the region, including those used in Burke and Lobell (2017).

Satellite-based yields were then derived in two ways, following Burke and Lobell (2017). First, "calibrated" remote sensing yields (**RS_cal**) were obtained from a regression model of FP yields on VI values measured on May 30 and June 19, 2016, using only pure stand maize plots that were at least 0.1ha in size. Since FP yields are expensive to obtain and cannot be considered as part of large-scale operations, an alternative version of the calibrated remote sensing yield was obtained (**RS_cal_cc**), which used CC, rather than FP yields to calibrate the model. These models can be specified as follows:

$$(1)\quad \textbf{RS\_cal Model} : FP\ Yield_i = \alpha_{FP,i} + \beta_{VI,FP,1} * VI_{May\ 30,i} + \beta_{VI,FP,2} * VI_{June\ 19,i} + \varepsilon_{FP,i}$$

$$(2)\quad \textbf{RS\_cal\_cc Model} : CC\ Yield_i = \alpha_{CC,i} + \beta_{VI,CC,1} * VI_{May\ 30,i} + \beta_{VI,CC,2} * VI_{June\ 19,i} + \varepsilon_{CC,i}$$

where i denotes plot, and α and ε denote regression-specific constant and error term, respectively. The calibration was done using only purestand plots since ground-based objective yield estimates were not available for non-maize crops on intercropped plots. The restriction in terms of plot area was driven by smaller plots having larger problems with geolocation accuracies and mixed pixels in Sentinel-2. All VIs shown in table 2 were tested and are discussed below, with the preferred model using the MERIS Terrestrial Chlorophyll Index (MTCI). Yields were estimated as a linear function of VI as shown in equations (1)–(2). Quadratic models were also considered but gave poorer out-of-sample performance.

The second satellite-based approach was to estimate "uncalibrated" yields (**RS_scym**) by using the scalable crop yield mapper (SCYM) approach (Lobell et al. 2015). In this approach, a crop model and local daily weather data were used to simulate crop growth and yield for various realistic combinations of on-farm management, such as sow date, seeding density, and fertilizer rate. The simulated values of total canopy nitrogen on the dates with available images were then translated into MTCI using published relationships (Schlemmer et al. 2013),

$$(3) \quad MTCI = 3.05 + 0.789 * canopyN$$

where *canopyN* is the simulated amount of total nitrogen in aboveground biomass after subtracting the nitrogen in the grain (which is invisible to the sensor). As in the calibrated approach, the yields are then regressed on MTCI, except in the case of SCYM the regression uses simulated yield and MTCI rather than actual values. In this way, SCYM avoids reliance on any ground data for calibration, which is why it is referred to as an "uncalibrated" approach.

Both types of satellite-based yield estimates were tested in two complementary ways. First, the yields were compared directly with the ground-based estimates across both purestand and intercropped plots. However, given that ground-based estimates are subject to (different types of) measurement error and neglect a potentially substantial amount of production from non-maize crops, the direct comparisons between the two yield measures is not a straightforward test of the satellite-based yields. That is, some of the discrepancy will also be due to errors in the ground-based estimates, or discrepancies in the types of outputs that are measured. As a second form of evaluation, we performed regressions of yield on different production factors for both ground-based and satellite-based yields and compared the resulting coefficients. Specifically, we regressed yields on key plot characteristics, including log of plot area, log of distance to household (km), presence of cover crops, log of seed planted (kg), use of inorganic fertilizer, log of household labor days and hired labor days, number of hired laborers, soil quality index (SQI), and household attributes, including wealth index, agricultural asset index, dependency ratio, household size, head of household age, gender, and years of education, and whether the manager was the survey respondent. For regressions including intercropped plots, two additional variables were included: a binary variable indicating the presence of an intercrop, and a variable indicating the log of the intercrop seed rate (i.e., the ratio of quantity of seed planted to quantity of seed that the farmer estimates would have been planted if the plot was pure stand).

Although we include a rich set of controls in our regressions, it is possible that omitted variables may be affecting yields, and therefore the estimated coefficients should not be interpreted as causal. Instead, the primary goal of this analysis is to use independently measured variables—many of which (such as fertilizer or soil quality) are known to affect productivity in a wide array of cropping systems—to further evaluate satellite-based yield measures. This is especially helpful in cases where the ground-based yield measures are thought to be error-prone, or when output is measured only for one crop on an intercropped field.

Results

Given the unique co-occurrence of three different ground-based yield measures in this study, we begin by comparing these measures to each other. We then describe the comparison of satellite and ground measures of maize yield for purestand maize fields, where the comparison is most straightforward because maize harvest alone defines the productivity of the plot. Comparisons are then presented for intercropped fields where ground-based measures provide only a partial measure of crop output. Finally, we present results of regressing the various yield measures on different production factors, both with and without including intercropped fields.

Comparison of Ground-Based Yield Measures

The distributions of yields from the three ground-based approaches are displayed in figure 2a and summarized in table 3. Both

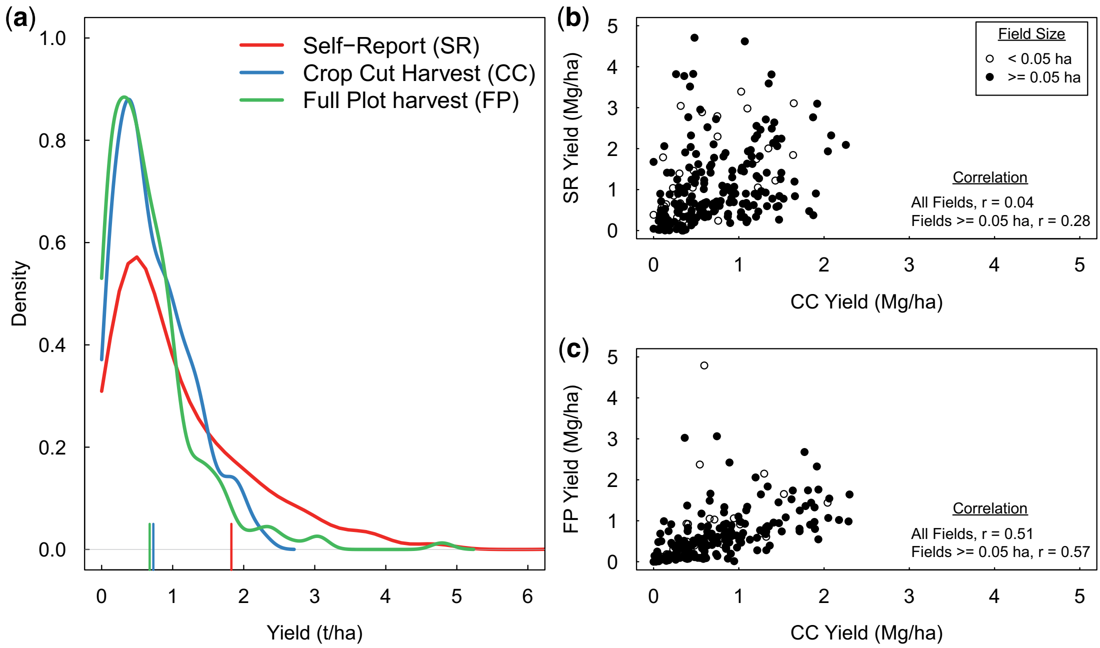

Figure 2. Yield distributions for ground-based measures

Note: (*a*) Vertical bars at bottom indicate the mean yield for each measurement approach. (*b*) Scatter plot of SR and CC yields for all plots, and, separately, for plots above 0.05ha in size (black points). (*c*) Scatter plot of FP and CC yields.

objective, harvest-based approaches show very similar distributions, with a mean CC yield of 0.73 metric tons per hectare (t/ha) and a mean FP yield of 0.68 t/ha. These differences were not statistically significant ($p > 0.2$). In contrast, the farmer self-reported (SR) yields contained many more high yielding values, including 11 (out of 252 total) plots with SR yield greater than 5 t/ha. The highest SR yields tended to occur on very small plots, with 8 of these 11 being on plots smaller than 0.05 ha. The average SR yield of 1.83 t/ha was significantly higher, and indeed more than double, that for CC and FP yields.

Given that SR, CC, and FP yields are competing ground-based measures, a useful question is how well correlated they are across different plots. Correlation between CC and FP yields was significant ($p < 0.01$) but only 0.51 overall (figure 2c). If one views full-plot crop cutting as the "gold standard" of ground-based measures, this indicates that 8x8m crop cuts capture only roughly one-quarter of the variability in actual plot yields. These discrepancies reflect the substantial intra-plot heterogeneity of yields in these systems. The 64 m^2 area of the crop cuts, despite requiring a costly and ambitious effort, are roughly just 6% of the median plot size (0.11 ha or 1,100 m^2) or 4% of the average plot size. The effect of this heterogeneity appears to be greater in intercropped plots, as the correlation between CC and FP yields is higher on pure stand maize plots ($r = 0.70$).

The more subjective SR yields show almost no correspondence ($r = 0.04$) with the crop cutting-based measures (figure 2b). Because correlations may be heavily influenced by errors on especially small fields, figure 2b also reports correlations that are based on the exclusion of plots with areas below 0.05ha. Despite the increase in the correlation coefficient to 0.28, still less than 10% of the variation in CC yields is captured by SR yields.

Comparison of Ground- and Satellite-Based Yield Measures on Pure Stand Plots

We begin the evaluation of satellite VIs by presenting the performance of the calibrated models (in terms of adjusted R^2) using different sources of ground-based yields for calibration, as well as different types of VIs (figure 3). Satellite-based yields were estimated for all plots that did not contain clouds on either May 30 or June 19 (397 out of 463 total plots).

Table 3. Summary Statistics of the Different Ground-Based Yield Measures

	All		Pure stand		Intercropped	
Yields (in kg/ha)	mean	median	mean	Median	mean	median
Self-Reported (SR)	1,826	784	1,878	1,039	1,805	685
Sub-Plot Crop Cutting (CC)	728	595	827	725	692	571
Full Plot Crop Cutting (FP)	676	511	842	740	623	472
	Different means?	Different Distributions?	Different means?	Different Distributions?	Different means?	Different Distributions?
SR vs. CC	***	***	***	***	**	***
CC vs. FP	–	–	–	–	–	–

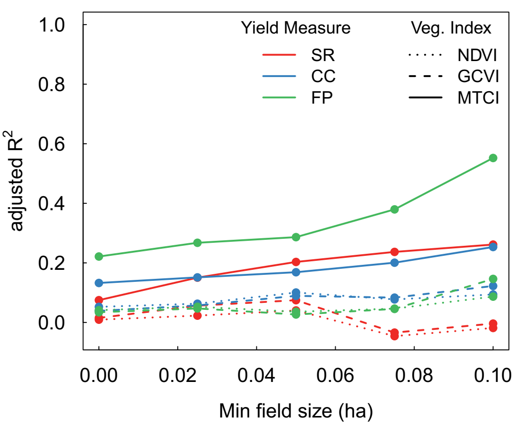

Figure 3. Adjusted R^2 of regressions of yields vs. VI, by VI type and type of ground-based yield measure

Note: Models were run for successive subsets of data by excluding plots below indicated plot size. Results for some VIs in table 2 are not displayed for clarity, but consistently performed worse than GCVI and MTCI.

Four important features are evident in figure 3: (*a*) Adjusted R^2 values were generally higher between VIs and FP yields than between VIs and CC or SR yields, which is consistent with the notion that full-plot crop cutting provides a better measure of plot-level productivity. (*b*) Adjusted R^2 tended to improve when excluding the smallest plot sizes, consistent with the results in Burke and Lobell (2017). A likely explanation for this is the increased importance of georeferencing errors and mixed pixels on the smallest of plots. For example, a 0.05 ha plot covers an area of just five 10x10m Sentinel-2 pixels, and most of these pixels are likely to span the edge of the plot and contain some contribution from neighboring plots. (*c*) The MTCI consistently outperformed the other VIs on both image dates. The MTCI was designed to be sensitive to canopy chlorophyll concentration (Dash and Curran 2004), which is likely a good proxy for yield in the low nutrient setting of Uganda. Perhaps more importantly, MTCI is much less sensitive to atmospheric conditions than other VIs such as NDVI or GCVI (Curran and Dash 2005) because it uses the difference in reflectance between two nearby bands that will be similarly affected by atmospheric scattering. In both images, significant amounts of haze are evident above many of the plot sites in both the raw reflectance and NDVI or GCVI images. However, the MTCI images exhibit much lower sensitivity to haze (see online supplementary appendix figure A1). (*d*) Finally, a substantial fraction of FP yield variability is captured by VIs, with the MTCI-based model capturing 55% of yield variability on plots of at least 0.10ha. Notably, this value is greater than the amount of FP yield variability captured by CC yields on these plots (adjusted $R^2 = 47\%$), indicating that satellite measures are better correlated with full plot harvests than the crop cuts on those same fields. Performance using only May 30 or June 19 was similar but slightly worse than the model using both dates (37% and 49% of yield variation explained for each date, respectively), as shown in online supplementary appendix figure A2.

One potential concern with the calibrated models is that they are unduly influenced by sowing date differences between fields. For example, if rains came early, such that fields planted early in the season had higher yields, but also more mature plants at an earlier stage, the correlation between crop yields

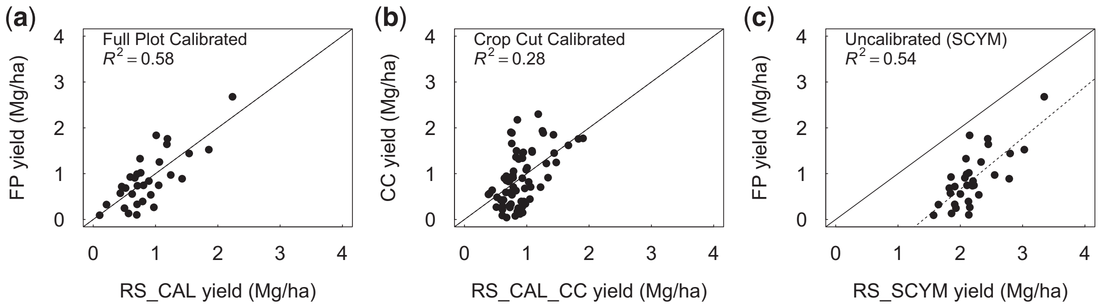

Figure 4. Plot yield comparisons

Note: Comparison of (*a*) full plot yields vs. predictions from a remote sensing model calibrated to full plot yields, (*b*) crop cut yields vs. predictions from a remote sensing model calibrated to crop cut yields, and (*c*) full plot yields vs. "uncalibrated" remote sensing yield estimates, which are based on calibration to crop model simulations. All panels show results for pure stand maize plots at least 0.1 ha in size, which are the subset of plots used to calibrate the models in (*a*) and (*b*).

and the satellite measures would appear positive but could simply arise from plants being at different stages in the season. For our dataset, farmers reported the month of sowing and whether they sowed in the first or second half of the month. These reported sowing dates exhibited a weak negative correlation with yields, with r = −0.22 for all purestand maize fields and r = −0.44 for purestand fields larger than 0.1ha. Moreover, the agreement between VI and yields were not unduly influenced by omitting particular sowing dates from model testing, as shown in online supplementary appendix figure A3. Specifically, removing fields with different sow dates had a negligible impact on the correlation between satellite and full-plot yields, with the exception of one influential field sown in February, which achieved a very high yield and whose removal reduced the adjusted R^2 by roughly 15 percentage points. Nonetheless, even after removing this field the model still explained a highly significant 43% of yield variation in the remaining fields.

At first glance, the results discussed above imply that measuring FP yields will result in a superior calibrated model, given that the adjusted R^2 for the FP model is more than twice that for the CC model when focusing on the performance of pure stand plots larger than 0.10 ha. Individual field predictions are shown in figure 4a and b, along with the calibration statistics. Interestingly, though, the coefficients of the two regressions were very similar, with the model calibrated to CC yields having a slightly lower range of predicted yields. As a result, this model did nearly as well predicting FP yields (R^2 = 0.54) as the model calibrated to FP yields.

This finding suggests that although CC yields are noisier measures of plot-level productivity compared to FP yields, this noise is mostly random and does not significantly bias the estimated coefficients in a model to predict yields from satellite data. Thus, one can expect models calibrated using CC yields (which are much more feasible and common than FP yields) to have lower R^2 but similar out of sample accuracy for predicting true plot productivity as models calibrated with FP yields.

The "uncalibrated" estimates, obtained from a regression of simulated yields versus simulated MTCI on these same dates, resulted in a nearly identical R^2 to models calibrated with FP yields (R^2 = 0.54, figure 4c). The uncalibrated estimates did exhibit significant bias, with a tendency to overestimate yields by roughly 1 ton/ha, because none of the simulated yields were as low as the lowest of the observed FP yields. Nonetheless, the high correlation between uncalibrated estimates and true FP yields indicates that ground calibration is not a prerequisite for capturing a large fraction of spatial yield variability with satellite data.

The "calibrated" and "uncalibrated" models can be viewed as two extremes of using available ground data, with the calibrated model using all purestand maize fields with cloud-free imagery, and the uncalibrated model using only model simulations. In practice, an important question is how much accuracy is retained as the size of the calibration dataset, and the associated costs of field work, is reduced. To explore this further, we randomly selected a subset of fields larger than 0.1ha to train a calibrated model, and

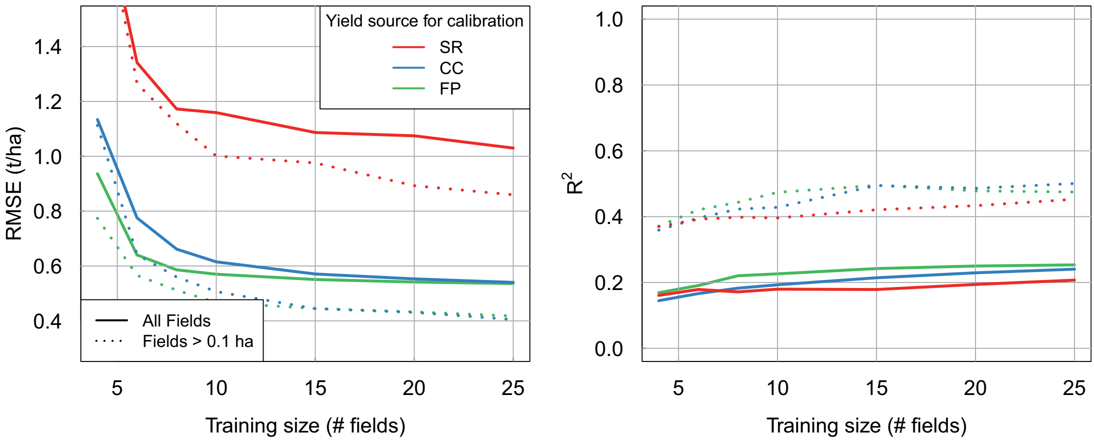

Figure 5. Sample size training effects

Note: The effect of training sample size on the out-of-sample root mean square errors (rmse, left) and squared correlation (R^2, right) for predicting FP yields, using models trained on SR, CC, or FP yields. Dashed lines indicate results for purestand maize fields larger than 0.1ha, while solid lines show results for all purestand fields.

then tested the model on FP yields from fields not used in the calibration. This was repeated for different sizes of the training subset, with the average performance plotted as a function of training sample size (figure 5). We alternatively evaluated predictions on fields above 0.1ha as well as all purestand fields. Results indicate that although the out-of-sample performance continues to improve for additional samples, training on 10 fields does nearly as well as training on 25. Also evident in figure 5, and consistent with the discussion above, is the fact that training on CC results in nearly identical performance as the FP model when tested on FP yields. Training on SR yields, in contrast, results in large root mean square errors because of the substantial bias associated with SR yields (figure 5).

The superior performance of MTCI is noteworthy, especially given that several of the most recent satellite sensors, which possess higher spatial resolution than Sentinel-2, lack the red edge bands needed to calculate MTCI. In this study, we fortuitously had access to a relatively cloud-free image acquired by Terra Bella's Skysat sensor on May 29, one day before a Sentinel-2 image. Skysat was used in Burke and Lobell (2017), and in the context of smallholder mapping has the particularly attractive feature of 1m spatial resolution. Particularly for the small plot sizes in Uganda, we anticipated that the 1m resolution would offer substantial benefits compared to the 10m resolution of Sentinel-2's main bands, and the 20m resolution of Sentinel-2's red edge bands. Surprisingly, we found that Sentinel-2 and Skysat performed very similarly when using GCVI for both, even though many plots contained only a few Sentinel-2 pixels (online supplementary appendix figure A4). The large boost in performance when using MTCI with Sentinel-2 therefore more than outweighed any loss in accuracy from using coarser resolution. This result may be specific to the particular atmospheric conditions, time of growing season, and characteristics of the study site, and therefore we caution against overweighting the benefits of spectral versus spatial resolution. Nonetheless, it is an informative comparison made possible by having two images so close in time over a study site with large amounts of quality ground-based data.

Comparison of Ground- and Satellite-Based Yield Measures on All Maize Plots

Of interest in agricultural regions such as Uganda, where maize is typically intercropped with other species, is how well satellite measures can capture the performance of mixed-crop plots. Of course, ground-based yield measures are readily beset by challenges from intercropping (Carletto, Jolliffe, and Banerjee 2015). In crop cutting applications, pure stand plots are typically prioritized due to (*a*) differences in harvest calendars of crops on intercropped plots (e.g., maize versus root/tuber crops, such as cassava as in our study, whose harvests may span an extended period; take place on a needs basis; and cut across agricultural seasons); (*b*) the

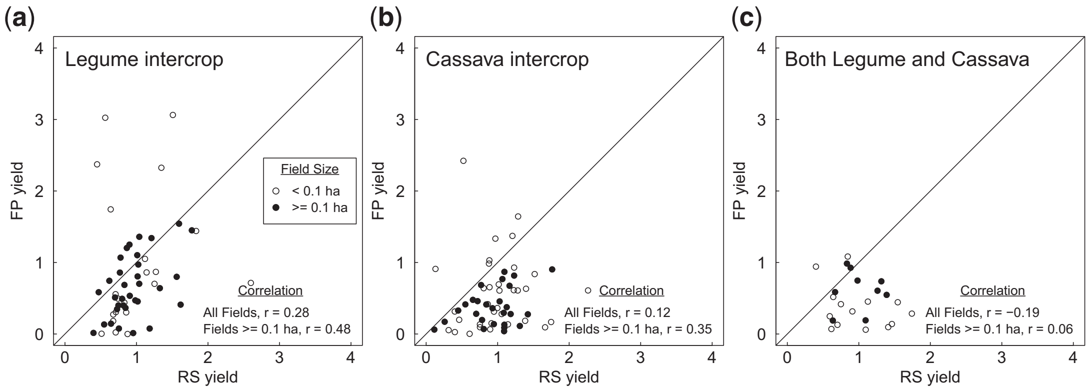

Figure 6. Comparison of calibrated remote sensing yields vs. full plot harvests for different types of intercropped plots

Note: (*a*) maize intercropped with only legumes (beans, groundnuts), (*b*) maize intercropped with only cassava, and (*c*) maize intercropped with both legumes and cassava. All panels show remote sensing yields based on calibration to FP yields in purestand maize plots at least 0.1ha in size (model shown in figure 4a). RS yields tend to be higher than FP yields in intercropped fields since the latter do not account for production from the other crops.

difficulty of conducting multiple crop harvests on intercropped plots and accommodating crop-specific post-harvest processing and drying needs prior to weighing. And in the analysis of surveys soliciting farmer-reported information on crop production, the yields for each crop on an intercropped plot is computed by dividing the production of each crop by either (*a*) the entire plot area, (*b*) the plot area multiplied by the farmer-reported share of the plot area cultivated with the crop, or (*c*) the plot area multiplied by the ratio between the farmer-reported seed use under intercropping and hypothetical seed use under pure stand cultivation.

In our study, the ground-based measures of yield (SR, CC, and FP) were obtained only for maize, irrespective of the pure stand versus intercropped cultivation status. The secondary crop harvests were not considered in our crop cutting operation primarily due to the above referenced reasons that typically lead to the prioritization of pure stand plots in crop cutting applications. In turn, we compared the satellite-based yield measures to FP for different types of plots, grouped based on the presence and type of intercropping (figure 6). The performance on plots intercropped with legumes (beans or groundnuts) was significantly lower than on pure stand plots, with roughly 20% of yield variability captured for plots at least 0.10ha in size (figure 6a). Maize yield estimates were even worse on plots intercropped with cassava (figure 6b) or both legumes and cassava (figure 6c), with less than 10% of the maize yield variability captured by the satellite estimates. The relatively better performance for legume intercrops presumably reflects the fact that both beans and groundnuts grow close to the ground, below the maize crop, whereas cassava intercrops often include very mature cassava plants that exceed the maize crop in height.

The worse performance for satellite- based maize yields on intercropped compared to pure stand plots makes sense, since non-maize crops can be a large contributor to the light reflected from the canopy and measured by satellite sensors, especially in the case of intercrops such as cassava that overhang maize plants. However, in these situations it is doubtful that the yield of maize is the best measure of land productivity. In the absence of other ground-based measures of productivity, we turn instead to assessing the sensitivity of the relationships between yield and factors of production to the choice of the ground- versus satellite-based yield variant.

Assessment of Inter-Relationships between Maize Yields and Factors of Production

Pure stand plot-level maize yield regressions resulted in similar coefficients for models using CC, FP, and satellite-based yields (table 4). The coefficients for the three factors of production of interest—plot area, soil quality index, and incidence of inorganic fertilizer use—are visualized in figure 7a. As also noted by Gourlay, Kilic, and Lobell (2017), the regression using SR yields resulted in a much stronger negative coefficient for plot area than the objective ground-

Table 4. Regression Coefficients for Pure Stand Plots Using Different Yield Measures

	Dependent Variable/Maize Yield Type					
	Self-report (1)	Crop-cut (2)	Full plot (3)	RS_cal_fp (4)	RS_cal, cc (5)	RS_scym (6)
Log Plot Area (GPS, ha)	−1.94*** (0.42)	−0.08 (0.07)	−0.23 (0.14)	−0.15** (0.06)	−0.10** (0.04)	−0.11** (0.05)
Log Plot Distance from Dwelling (GPS, km)	0.10 (0.33)	−0.04 (0.06)	−0.19 (0.12)	−0.11** (0.05)	−0.0024	−0.05 (0.04)
Cover Crops Present Prior to Planting[a]	−0.35 (0.99)	0.01 (0.20)	0.26 (0.48)	−0.08 (0.15)	−0.08 (0.11)	−0.04 (0.13)
Log Maize Seed Planting Rate (Kg/Ha)	1.19** (0.48)	0.09 (0.08)	0.18 (0.14)	0.14** (0.07)	0.10** (0.05)	0.10* (0.05)
Inorganic Fertilizer Application[a]	0.56 (1.14)	0.35** (0.17)	0.98*** (0.28)	0.34** (0.13)	0.26*** (0.09)	0.33*** (0.11)
Log Household Labor Days	0.56* (0.30)	0.05 (0.06)	−0.01 (0.10)	0.05 (0.04)	0.04 (0.03)	0.05 (0.03)
Log Hired Labor Days	0.27 (0.42)	−0.01 (0.06)	−0.03 (0.10)	−0.11** (0.05)	−0.07** (0.03)	−0.09** (0.04)
No Hired Labor[a]	0.13 (0.96)	−0.24 (0.16)	0.09 (0.26)	−0.06 (0.12)	−0.02 (0.09)	−0.05 (0.10)
Soil Quality Index	1.36 (2.64)	1.11** (0.45)	1.84** (0.82)	1.44*** (0.36)	1.02*** (0.25)	1.14*** (0.30)
Wealth Index	0.46 (0.39)	0.09 (0.07)	−0.05 (0.12)	−0.0045	−0.06 (0.04)	−0.06 (0.04)
Agricultural Asset Index	0.43 (0.32)	−0.01 (0.06)	0.09 (0.10)	0.08* (0.04)	0.05 (0.03)	0.06 (0.03)
Dependency Ratio	−0.16 (0.35)	0.01 (0.06)	0.01 (0.10)	−0.01 (0.05)	−0.02 (0.03)	−0.02 (0.04)
Household Size	−0.04 (0.11)	0.01 (0.02)	0.02 (0.04)	0.01 (0.02)	0.003 (0.01)	0.003 (0.01)
Manager = Respondent[a]	0.07 (0.83)	0.03 (0.16)	−0.05 (0.38)	0.02 (0.13)	0.04 (0.09)	0.07 (0.11)
Received Crop-Production Related Extension Services[a]	−0.08 (0.69)	−0.16 (0.12)	0.26 (0.19)	0.06 (0.09)	0.06 (0.06)	0.09 (0.08)
Female[a]	−0.20 (0.73)	−0.09 (0.13)	−0.04 (0.25)	−0.02	−0.16** (0.07)	−0.18** (0.09)
Age (Years)	−0.03 (0.02)	−0.004 (0.004)	0.003 (0.01)	−0.001 (0.003)	−0.002 (0.002)	−0.001 (0.003)
Years of Education	−0.09 (0.07)	−0.01 (0.01)	0.03 (0.02)	0.01 (0.01)	0.002 (0.01)	−0.003 (0.01)
Constant	−4.35 (3.25)	−0.12 (0.60)	−2.01* (1.08)	−0.76 (0.47)	−0.21 (0.33)	0.95** (0.39)
Observations	73	124	51	105	105	105
R^2	0.4	0.19	0.47	0.4	0.39	0.37
Adjusted R^2	0.19	0.05	0.17	0.27	0.27	0.24
Residual Std. Error	2.25 (df = 54)	0.54 (df = 105)	0.55 (df = 32)	0.37 (df = 86)	0.25 (df = 86)	0.31 (df = 86)
F Statistic	1.96** (df = 18; 54)	1.33 (df = 18; 105)	1.55 (df = 18; 32)	3.17*** (df = 18; 86)	3.11*** (df = 18; 86)	2.78*** (df = 18; 86)

a*Note*: denotes a dummy variable. Asterisks ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively. Standard errors appear in parentheses.

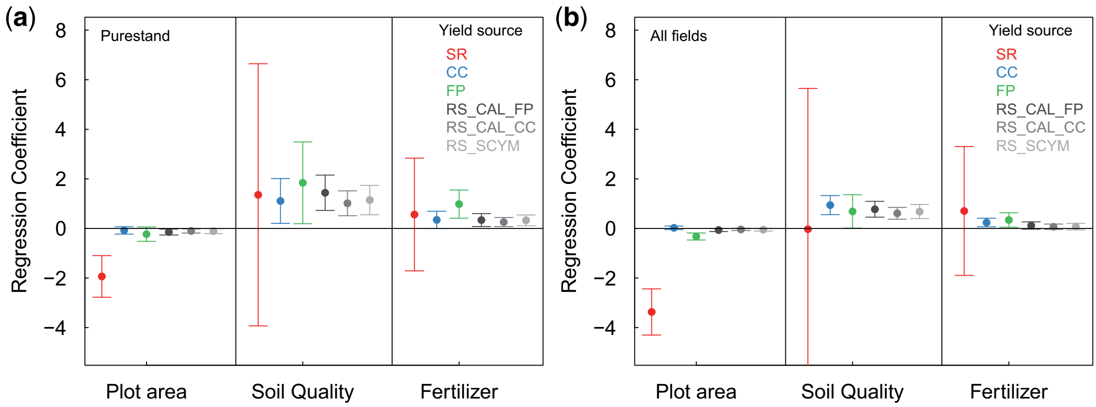

Figure 7. Summary of regression coefficients for three relevant factors using six different models corresponding to six yield measures. Error bars show +/- two standard deviations of the mean estimate.

based measures, indicating that the conventional wisdom of an inverse-relationship between farm size and productivity may be an artifact of measurement error. While the relationship between soil quality and any one of CC, FP, and satellite-based yields was positive and statistically significant at least at the 5% level, the coefficient associated with soil quality failed to be statistically significant in the regression using SR yields. In line with the results of the CC and FP yield regressions, the relationship between fertilizer use and any one of the calibrated or uncalibrated satellite-based yields was positive and statistically significant at the 1% level.

The regressions for all plots, including both pure stand and intercropped plots, show qualitatively similar coefficients, as depicted in figure 7b and online supplementary appendix table A1. The satellite-based regressions still find a significant positive association with soil quality, whereas the coefficients on fertilizer remain positive but become statistically insignificant. A possible explanation for this result is that cassava biomass, which influences the satellite-based yield estimates on intercropped plots, is similar to maize in its responsiveness to soil quality, but less responsive to inorganic fertilizer. In comparison to regressions using FP yields, those using either CC or satellite-based yields generally had smaller confidence intervals for coefficient values, which reflects the fact that full plot harvests were only performed on 211 plots, whereas sub-plot crop cutting was done for all 463, and satellite estimates were available on 397.

Discussion and Conclusions

Despite the importance of agriculture for rural livelihoods, poverty alleviation, and food security across the developing world, household and farm surveys collecting micro data on agriculture exhibit substantial cross-country heterogeneity in terms of access policies, use of international best practice survey methods and dissemination standards, and data quality (Carletto, Jolliffe, and Banerjee 2015). Given the rapid advances in the availability of 10-meter or sub-10-meter spatial resolution satellite imagery, the demand is increasing for understanding how these advances can be leveraged to measure and understand agricultural outcomes with greater accuracy and higher spatial resolution.

Although there is a concerted push to showcase the value of geospatial applications for monitoring and evaluation efforts in the agriculture sector, and for tracking the progress towards the SDGs, multi-disciplinary research efforts aimed at assessing the accuracy and feasibility of the proposed applications, particularly in smallholder production systems, are scant. If validated, satellite-based remote sensing, combined with georeferenced household and farm survey data that could serve as "ground truth", could dramatically enhance not only our ability to fill the data gaps, but also our understanding of the linkages between development and human welfare. The field of agricultural economics, too, has a stake in these developments, given the wide range of research applications in low- and low–middle income contexts that continue to rely on household and farm survey data, and the emerging evidence on systematic measurement errors in farmer-reported crop production estimates that may have a bearing on fundamental relationships in smallholder production systems (Gourlay, Kilic, and Lobell 2017; Desiere and Jolliffe 2018; Abay et al. 2019; Wossen et al. 2019).

Taking advantage of a unique range of ground-based plot-level maize yield measures based on farmer-reporting, sub-plot crop cutting and full-plot harvests that were collected as part of a methodological survey experiment that was conducted in Eastern Uganda, our study showcases the accuracy and empirical utility of satellite-based approaches to plot-level maize yield estimation in smallholder production systems with a median plot size of approximately one-tenth of a hectare.

The satellite-based yield estimates include those that are (*a*) anchored in a calibration model that relates maize yields from full-plot harvests to MTCI values on multiple dates on a subset of pure stand maize plots that were at least 0.1 ha in size; (*b*) based on the same calibration model that uses sub-plot crop cut, as opposed to full-plot, yield; and (*c*) based solely on crop model simulations, without reliance on any ground-based yield measure. While (*a*) and (*b*) are identified as "calibrated" variants of remotely-sensed maize yields, (*c*) is framed as the "uncalibrated" counterpart.

The accuracy of the satellite-based maize yield estimates is found to be very encouraging. The availability of over 200 full plot harvests, which is very rare because of their cost, is a unique situation with which to test satellite estimates, and we find that both calibrated and uncalibrated approaches capture roughly half of the variance in full plot harvests when restricting the analysis to where both ground and satellite approaches are measuring the same output (pure stand plots), and where the satellite pixels corresponding to the plot are less likely to be contaminated by neighboring plots (plots > 0.10 hectare). The uncalibrated approach exhibits, however, a strong tendency to overestimate yields, but adequately captures spatial variation in yield. In fact, the satellite-based estimates explained slightly more variance in full plot harvests than sub-plot crop cuts performed within the plots.

In addition, satellite-based estimates can faithfully reproduce the associations between yield and key production factors such as soil quality and fertilizer use, even when including plots of all sizes and those that are intercropped. The significance levels of the coefficients informed by the satellite-based measures are often even higher than those underlined by the full plot harvests. The cross-sectional nature of our data limits the ability to interpret regression coefficients as the causal effect of a factor on yields. Nonetheless, the fact that factors expected to affect yields (i.e., soil quality, fertilizers) are associated as strongly with satellite-based yield measures as with ground-based yield measures indicates that the errors in both yield measures are of similar magnitude. This finding emphasizes that an imperfect correlation between satellite measures and full plot harvests reflects errors in ground-based estimates as well as those in satellite-based estimates. Moreover, the regression results suggest that even if satellite-based measures are less accurate than full plot harvests, the greater sample size can compensate for any loss in accuracy.

Also noteworthy is the fact that satellite-based models calibrated to CC yields perform similarly to those calibrated to FP yields, in terms of both agreement with FP yields and estimation of yield response to soil quality and fertilizer. These results indicate that although CC yields are imperfect approximations of actual yields, the errors do not substantially bias remote sensing calibrations. Thus, sub-plot crop cutting appears to be a suitable replacement for full-plot harvests when the latter are not possible. We also found that even using just 10 fields of either FP or CC yields for calibration results in accuracies approaching that of the full model. In addition, we show that crop model simulations can be used as a replacement for ground-based measures if the potential bias in estimated yields is recognized and acceptable. The bias may also be reduced in the future, although that is beyond the scope of the current paper.

Overall, our findings suggest that remote sensing approaches to measuring crop yields, particularly when calibrated based on crop cutting operations on the ground, can offer more accurate and precise measurements compared to farmer reporting. At the plot-level, the future models can be trained with sub-plot crop cutting on a subsample of plots identified in a household/farm survey, and subsequently, used to estimate crop yields on the remaining plots that are not subject to crop cutting as part of the same survey.

Our results corroborate and extend those in Burke and Lobell (2017), despite differences in the study region and the sensors used. Burke and Lobell reported higher R^2 between satellite estimates and self-reported yields on purestand maize fields (~0.4 vs. ~0.2 in this study), which could reflect the fact that farmers in the commercial fields of western Kenya have more accurate estimates of their yields than the more subsistence farmers of Eastern Uganda. Unlike in Burke and Lobell (2017), this study had the benefit of objective ground-based measures, including 8x8m crop cuts and full plot harvests, which revealed the low accuracy of self-reports in this region. Similar to Burke and Lobell (2017), this study found that the correlation between yields and different production factors were very similar whether using satellite-based yields or the preferred ground-based yield measures, though in this study a wider range of factors, including objective soil measurements, was considered.

Even though our study emphasized measuring plot-level yields, many applications, such as forecasting regional food supply or assessing local conditions for insurance payouts, require accuracy at more aggregate scales. Our results suggest that the integration of georeferenced micro survey data on agriculture, such as that from the LSMS-ISA, with the expanding, publicly-available high-resolution satellite imagery, will provide a tool to generate the landscape-scale data needed for these aggregate estimates. Not only could these outputs be used in national and international monitoring efforts, they should be expected to create an unparalleled scope for research on entire landscapes of agricultural plots. Collectively, these measurement tools will allow more rapid feedback on the effectiveness of different efforts to raise productivity, which in turn can enable more effective agricultural and development policy.

Supplementary Material

Supplementary materials are available at *American Journal of Agricultural Economics* online.

References

Abay, K.A., G.T. Abate, C.B. Barrett, and T. Bernard. 2019. Correlated Non-Classical Measurement Errors, "Second Best" Policy Inference and the Inverse Size-Productivity Relationship in Agriculture. *Journal of Development Economics* 139: 171–84.

Arslan, A., N. McCarthy, L. Lipper, S. Asfaw, A. Cattaneo, and M. Kokwe. 2015. Climate Smart Agriculture? Assessing the Adaptation Implications in Zambia. *Journal of Agricultural Economics* 66 (3): 753–80.

Berazneva, J., L. McBride, M. Sheahan, and D. Güereña. 2018. Empirical Assessment of Subjective and Objective Soil Fertility Metrics in East Africa: Implications for Researchers and Policy Makers. *World Development* 105: 367–82.

Burke, M., and D.B. Lobell. 2017. Satellite-Based Assessment of Yield Variation and Its Determinants in Smallholder African Systems. *Proceedings of the National Academy of Sciences* 114 (9): 2189–94.

Byerlee, D., A. De Janvry, E. Sadoulet, R. Townsend, and I. Klytchnikova. 2007. *World Development Report, 2008: Agriculture for Development*. Washington DC: World Bank.

Carletto, C., P. Corral, and A. Guelfi. 2017. Agricultural Commercialization and Nutrition Revisited: Empirical Evidence from Three African Countries. *Food Policy* 67: 106–18.

Carletto, C., S. Gourlay, S. Murray, and A. Zezza. 2017. Cheaper, Faster, and More than Good Enough: Is GPS the New Gold Standard in Land Area Measurement. *Survey Research Methods* 11 (3): 235–65.

Carletto, C., D. Jolliffe, and R. Banerjee. 2015. From Tragedy to Renaissance: Improving Agricultural Data for Better Policies. *Journal of Development Studies* 51 (2): 133–48.

Casley, D., and K. Kumar. 1988. *The Collection, Analysis and Use of Monitoring and Evaluation Data*. Washington DC: The World Bank.

Curran, P.J., and J. Dash. 2005. Algorithm Theoretical Basis Document ATBD 2.22 Chlorophyll Index. Southampton-UK: University of Southampton.

Darko, F.A., A. Palacios-Lopez, T. Kilic, and J. Ricker-Gilbert. 2018. Micro-Level Welfare Impacts of Agricultural Productivity: Evidence from Rural Malawi. *Journal of Development Studies* 54 (5): 915–32.

Dash, J., and P.J. Curran. 2004. The MERIS Terrestrial Chlorophyll Index. *International Journal of Remote Sensing* 25 (23): 5403–13.

Davis, B., S. Di Giuseppe, and A. Zezza. 2017. Are African Households (Not) Leaving Agriculture? Patterns of Households' Income Sources in Rural Sub-Saharan Africa. *Food Policy* 67: 153–74.

Desiere, S., and D. Jolliffe. 2018. Land Productivity and Plot Size: Is Measurement Error Driving the Inverse Relationship? *Journal of Development Economics* 130: 84–98.

Dorosh, P., and J. Thurlow, 2018. Beyond agriculture versus non-agriculture: decomposing sectoral growth–poverty linkages in five African countries. *World Development* 109: 440–451.

Fermont, A., and T. Benson. 2011. Estimating Yield of Food Crops Grown by Smallholder Farmers: A Review in the Uganda Context. Washington, DC:IFPRI Discussion Paper 01097 1–57.

Gitelson, A.A., Y. Gritz, and M.N. Merzlyak. 2003. Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. *Journal of Plant Physiology* 160: 271–282.

Gourlay, S., T. Kilic, and D. Lobell. 2017. Could the Debate Be over? Errors in Farmer-Reported Production and Their Implications for Inverse Scale–Productivity Relationship in Uganda. doi: 10.1596/1813-9450-8192.

Harou, A.P., Y. Liu, C.B. Barrett, and L. You. 2017. Variable Returns to Fertiliser Use and the Geography of Poverty: Experimental and Simulation Evidence from Malawi. *Journal of African Economies* 26 (3): 342–71.

Ivanic, M., and W. Martin. 2018. Sectoral Productivity Growth and Poverty Reduction: National and Global Impacts. *World Development* 109 (C): 429–39.

Julien, J.C., B.E. Bravo-Ureta, and N.E. Rada. 2019. Assessing Farm Performance by Size in Malawi, Tanzania, and Uganda. *Food Policy* 84: 153–64.

Kilic, T., A. Palacios-Lopez, and M. Goldstein. 2015. Caught in a Productivity Trap: A Distributional Perspective on Gender Differences in Malawian Agriculture. *World Development* 70: 416–63.

Larson, D.F., K. Otsuka, T. Matsumoto, and T. Kilic. 2014. Should African Rural Development Strategies Depend on Smallholder Farms? An Exploration of the Inverse-Productivity Hypothesis. *Agricultural Economics* 45 (3): 355–67.

Lobell, D.B., D. Thau, C. Seifert, E. Engle, and B. Little. 2015. A Scalable Satellite-Based Crop Yield Mapper. *Remote Sensing of Environment* 164: 324–33.

McCarthy, N., T. Kilic, A. De la Fuente, and J. Brubaker. 2018. Shelter from the Storm? Household-Level Impacts of, and Responses to, the 2015 Floods in Malawi. *Economics of Disasters and Climate Change* 2 (3): 237–58.

Mukherjee, A., and R. Lal. 2014. Comparison of Soil Quality Index Using Three Methods. *PLoS One* 9 (8): e105981.

O'Sullivan, M., A. Rao, R. Banerjee, K. Gulati, and M. Vinez. 2014. *Levelling the Field: Improving Opportunities for Women Farmers in Africa*. Washington DC: World Bank Group.

Restuccia, D., and R. Santaeulalia-Llopis. 2017. Land Misallocation and Productivity. NBER Working Paper No. 23128.

Rouse, J.W., R.H. Haas, J.A. Schell, and D.W. Deering. 1973. Monitoring Vegetation Systems in the Great Okains with ERTS. In *Proceedings of the Third Earth Resources Technology Satellite-1 Symposium*, Washington, DC, USA, eds. Freden, S.C., Mercanti, E.P. Washington, DC: NASA.

Schlemmer, M., A. Gitelson, J. Schepers, R. Ferguson, Y. Peng, J. Shanahan, and D. Rundquist. 2013. Remote Estimation of Nitrogen and Chlorophyll Contents in Maize at Leaf and Canopy Levels. *International Journal of Applied Earth Observation and Geoinformation* 25: 47–54.

Wineman, A., N. Mason, J. Ochieng, and L. Kirimi. 2017. Weather Extremes and Household Welfare in Rural Kenya. *Food Security* 9 (2): 281–300.

Wossen, T., T. Abdoulaye, A. Alene, P. Nguimkeu, S. Feleke, I.Y. Rabbi, M.G. Haile, and V. Manyong. 2019. Estimating the Productivity Impacts of Technology Adoption in the Presence of Misclassification. *American Journal of Agricultural Economics* 101 (1): 1–16.