Causal Inference with Predicted Outcomes: A satellite-based impact assessment of 'Direct Seed Marketing' in Ethiopia

Johanne Pelletier, Mira Korb, Solomon Alemu,
Manex Bule Yonis, Travis J. Lybbert, Matthieu Stigler

August 2024

Abstract

Recent advances in earth observation and machine learning have opened new frontiers in impact evaluation that appear well-suited for agricultural settings. We apply these promising methods in the context of Ethiopia's Direct Seed Marketing (DSM) program, which rolled out after 2011 and aims to enhance farmer access to improved seed varieties. Our satellite-based impact assessment focuses on maize productivity as a summary outcome. Satellite-based yield predictions enable a high-resolution, landscape-level analysis of DSM impacts using a difference-in-difference identification strategy, but yield prediction errors introduce new sources of potential bias in subsequent causal inference. We test for this prediction error bias and compare our DSM impact estimates to those that use farmer-reported and crop cut yield measures. We find evidence of small positive but insignificant effects of the DSM on maize yield, explore how errors in predicted yields introduce bias in causal estimation, and discuss implications for the selection of prediction models.

JEL Codes: O13, Q12, C81, C52

Keywords: Earth observation, Impact evaluation, Agriculture, Maize, Ethiopia, Crop yield, Direct Seed Marketing

Highlights:

- We use satellite-based maize yield predictions to evaluate the impacts of Ethiopia's Direct Seed Marketing (DSM) program.
- Using predicted outcomes for causal inference can introduce an additional source of non-classical measurement error that leads to biased results, underestimating impact and overestimating precision.
- Difference-in-difference estimates of the DSM impact on maize yield are positive but are smaller and less precise than comparable survey-based estimates, which may be driven by yield prediction errors.
- When predicting outcomes for causal inference, the best performing prediction models are not necessarily the best for causal inference, raising a potentially important tradeoff for satellite-based impact evaluation.

1 Introduction

Satellite remote sensing has dramatically expanded in recent decades, providing Earth Observation (EO) data at higher spatial, temporal, and spectral resolution than ever before. This unprecedented availability of satellite data, much of which is freely available to users (Gorelick et al., 2017), offers great promise for measurement and empirical analysis across many fields of study. In development economics, this satellite-powered revolution has unleashed exciting new possibilities for impact evaluation of large-scale interventions that can elucidate key spatial and temporal impact dynamics and heterogeneity, thereby generating new insights into the design of programs and policies and into the mechanisms that drive impact. The growing interest in these methods is evident not only among researchers (e.g., Burke et al., 2021; Jain, 2020; Donaldson and Storeygard, 2016; Rolf et al., 2021), but also within the wider development community[1] as evident in the recent emergence of guidelines for satellite-based monitoring and impact evaluation (Serrat Capdevila, Aleix; Herrmann, Stefanie Maria, 2018; Pelletier et al., 2023; Independent Office of Evaluation of IFAD, 2023; Space, 2023).

[1]For example, see the GeoField 2023 Convening on *Leveraging Earth Observation for Impact Evaluations of Climate-Sensitive Agriculture*, link

Given the persistent importance of agriculture among the world's poor and its critical spatial features, enthusiasm is understandably high for satellite-based impact evaluation applications in international agricultural development. In this setting, there are several notable advantages to satellite data when combined with machine learning (ML). They allow researchers to 'gather' data from unsampled locations or periods, with the potential to significantly reduce data collection costs. By obtaining data from unsampled locations – effectively characterizing entire landscapes – researchers can in principle analyze a full range of heterogeneous effects and spatial spillovers. Collecting data from unsampled periods further enables the acquisition of pre-intervention benchmark data or the ability to track the dissemination of innovations over time (e.g., BenYishay et al., 2024; Salazar et al., 2021; Al Rafi, 2023; Ferguson and Govaerts, 2024; Deines et al., 2019).

Despite the justifiable excitement for novel uses of EO data in applied economics, this new research territory also raises new questions about the applicability of remote sensing for causal impact evaluation. For example, using EO-based predictions as outcome variables potentially introduces new sources of measurement error, which may have important implications for subsequent inference. ML methods typically achieve higher predictive accuracy over classical methods by reducing variance at the expense of introducing some bias. While this bias-variance trade-off might be optimal from a prediction perspective, such a narrow prediction-optimal criteria may be distinctly sub-optimal and even misleading from a causal impact evaluation perspective. We aim to contribute to the emerging literature of satellite-based impact evaluation by building greater appreciation for this potential tension between what is optimal for prediction and what is optimal for causal inference. We illustrate this possible tension in a specific application of satellite-based impact evaluation and propose an error test to guide model selection and explore bias correction methods.

We pursue these methodological objectives by using satellite data as the basis for estimating the causal impacts of a national roll-out of an agricultural intervention in Ethiopia. The Direct Seed Marketing (DSM) program, which was piloted by the Government of Ethiopia in 2011 and then scaled up in the subsequent years, aimed to strengthen the engagement of the private sector in the seed system and thereby enhance Ethiopian farmers' access to improved seeds. Since maize is one of the most important staple crops for food security in Ethiopia (Abate et al., 2015; Van Dijk et al., 2020) and a previous study estimated that DSM increased maize yields by (a remarkable) 26 percent

(Mekonnen et al., 2021), we take maize yield as our primary outcome of interest and as a summary of the many ways the DSM program could have plausibly impacted agricultural practice and production as it scaled up. In contrast to Mekonnen et al. (2021), who use conventional farmer-reported survey-based measures of maize yield, we use ground-truth yield data from crop cuts to train an ML model to predict maize yield and use predicted yield for administrative units in all maize growing areas of Ethiopia as our outcome of interest.

Specifically, we use 30-m Landsat data to map maize, which we combine with plot-level maize crop cut data and 250-m MODIS vegetation index to predict maize yield from 2010 to 2020 for Ethiopia using ML-methods. We then aggregate pixel-level predicted maize yield to averages at the district- (*woreda*) level (the administrative unit of DSM implementation) and the smaller ward- (*kebele*) level. Next, we deploy these average predicted maize yields as our DSM outcome variable using a staggered difference-in-differences (DiD) estimator (Callaway and Sant'Anna, 2021). Compared to the 26 percent yield increase of Mekonnen et al. (2021), our results suggest that DSM had much more modest (3 percent or less) and statistically insignificant impacts on maize yield. To investigate this empirical discrepancy, we explore the role that yield prediction error may play in attenuating the true DSM effect on maize yield using crop cut data and a conceptual framework of DiD and prediction errors. Using a large set of ML models ranked according to conventional measures of predictive performance, we illustrate the tension between the best model for prediction and best (least biased) model for causal inference. We find that in satellite-based impact evaluation, what is optimal for predicting outcomes may not necessarily be optimal for causal inference based on analysis of these predicted outcomes.

This paper makes two contributions. First, we contribute satellite-based impact evidence to the emerging literature that uses EO-ML methods to evaluate agricultural interventions. To date, much of this literature has focused on remotely tracking the adoption and landscape-level diffusion of new agricultural practices or innovations. Recent studies along these lines assess the adoption of irrigation on crop productivity in Mali (BenYishay et al., 2024) and in the Dominican Republic (Salazar et al., 2021), of stress-tolerant rice varieties resilience to flood in Bangladesh (Al Rafi, 2023), of sustainable agriculture practices on crop residue burning and health outcomes in Mexico (Ferguson and Govaerts, 2024), of conservation tillage on yields (Deines et al., 2019; Cambron et al., 2024), and of customized soil nutrient manage-

ment advice on crop yields (Cole et al., 2020). The evidence we provide, while more cautionary than celebratory, uses similar methods to evaluate a seed marketing innovation. We generate several different estimates of DSM impacts in order to compare the results with directly measured (crop cut) yield and conventional self-reported yield, shedding light on the methodological decisions that affect such satellite-based impact evidence.

Second, and more substantively, we make a methodological contribution by addressing how new sources of (potentially non-classical) measurement error that emerge from prediction models used to convert satellite data into outcome variables affect subsequent causal estimation of impact. In their enthusiasm for EO-enabled measures, researchers may fail to appreciate that these EO-powered outcome variables are also affected by measurement error. While measurement error in EO-based outcome variables can arise from a host of sources (e.g., choice of satellite data, pre-processing of these data, field data to train and validate prediction models for both classification and regression), we highlight how the choice of a predictive model used to combine these elements and predict outcomes may result in prediction errors that bias subsequent causal inference. Formally, we show that a placebo test, i.e. running the causal design using the prediction errors as outcome variable, can be used to diagnose potential bias in the causal inference and to correct for it. If the resulting "prediction error placebo test" does not reject the null hypothesis of absence of DiD effects, using ML predictions should not introduce bias in the causal analysis. Furthermore, drawing on very recent developments in the field of prediction-powered inference (Angelopoulos et al., 2023a), we show how the "prediction error placebo test" can be used to correct for bias.

In the next section, we provide relevant background details for Ethiopia as our study context and for DSM in particular. In Section 3, we provide a conceptual framework for how prediction errors can become a source of causal inference bias and describe emerging attempts to remedy similar sources of bias from recent literature. Section 4 describes the data sources we use for the roll-out of the DSM and as training data for maize area and yield. Section 5 describes our yield prediction methods, causal inference and identification strategy, and DiD prediction error tests and corrections. We conclude with reflections on the contributions and limitations of our work and what it means for satellite-based impact evaluation.

2 Background

The agricultural sector is central to Ethiopia's economy, employing more than 75 percent of the labour force and accounting for nearly 40 percent of the GDP.[2] In the past 40 years, maize has become a dominant crop in this sector with rapid increases in both area and yield (Abate et al., 2015). Favored for its wide adaptability to different growing conditions, maize is cultivated in pure monocrop stands and intercropped plots. While agricultural productivity in Ethiopia has increased markedly, especially for cereals, yields are still well below their potential; a recent productivity study suggests that average maize yields in Ethiopia are only 30 percent of their potential due to low (albeit increasing) adoption rates of fertilizer and improved seeds (World Bank, 2022).

Among the many factors that historically discouraged Ethiopian farmers' adoption of profitable and improved inputs in agriculture, rigidity and dysfunction in its centralized seed system has loomed large. Until recently, Ethiopia's seed sector was centralized into state-controlled public organizations that were responsible for the entirety of the seed system, from breeding to distribution, which entailed complex coordination of seed production, processing, transportation to district agricultural offices, then to development agents and cooperatives, and finally to farmers. This state-run seed system was inefficient, introducing more delays and dysfunction than innovation and investment in quality. Even progressive farmers willing to pay a premium for improved maize hybrid seeds did not generally have much choice in this system. Delayed delivery of seeds meant delayed cultivation, which combined with low quality seed, often of the wrong variety, translated into disappointingly low yields at harvest. For decades, experts advocated for innovation in government policy to facilitate access and promote the adoption of improved seed varieties or cultivars as an essential part of improving productivity and sustainability of agricultural intensification by smallholder farmers. These recommendations were based on evidence that the development of resilient and inclusive national seed systems is one of the most promising strategies to increase crop yield (Jain et al., 2023), thus reducing the existing yield gap (Aramburu-Merlos et al., 2024), and to adapt to climate change via the adoption of climate-resilient crop varieties (Acevedo et al., 2020; Westengen et al., 2023).

[2] https://www.usaid.gov/ethiopia/agriculture-and-food-security

In response to these concerns and recommendations, the DSM program was introduced in 2011 as a maize seed pilot by the Government of Ethiopia and its partner, Integrated Seed Sector Development, in two districts of the Amhara region. Success of this early pilot and subsequent expansion to other districts and regions led to the scale up of DSM after 2013. By 2020, DSM was operational in 320 districts and had expanded from maize alone to more than 10 crops (see Figure 1). The DSM program, which is administered at the district level, aims to improve efficiency in the seed system and thereby supply an expanded menu of high-quality and affordable improved seeds to farmers in a timely manner (Benson et al., 2014). In contrast to the previous seed system, DSM includes the following salient characteristics: 1) it allows both public and private certified seed producers; 2) it shortens the seed distribution chain by allowing seed producers to market directly to farmers; 3) it creates a competitive environment between seed producers which brings the costs down; and 4) it improves seed traceability and accountability of seed producers to farmers (for an assessment of the DSM pilot performance relative to these objectives, see Mekonnen et al. (2019)).

Mekonnen et al. (2021) leverage a panel household survey to conduct a quantitative evaluation of DSM's impact on different outcomes for farming households, including on-farm productivity of smallholder farmers. Using a DiD approach, they find evidence that the DSM program improved eight measures of seed system performance (e.g., seed availability, pricing, quality, timeliness, easy of purchase, etc.). Based on farmer-reported yields, the authors estimate that DSM led to an economically and statistically significant increase in maize yields of 26 percent with statistically insignificant yield effects for other crops. Like this study and as explained in detail below, we also adopt a DiD strategy and focus on maize yield as a summary outcome based on the assumption that DSM efficiency gains and improvements ultimately manifest as increased on-farm productivity, even if this ultimate impact is transmitted through a variety of on-farm adjustments to the seed system improvements provided by DSM. In contrast to Mekonnen et al. (2021), we use crop cut yield data to train a yield prediction model and use predicted maize yield as our (summary) outcome of interest, which enables us to extend this impact evaluation to all maize growing areas of Ethiopia rather than just those included in the household survey.

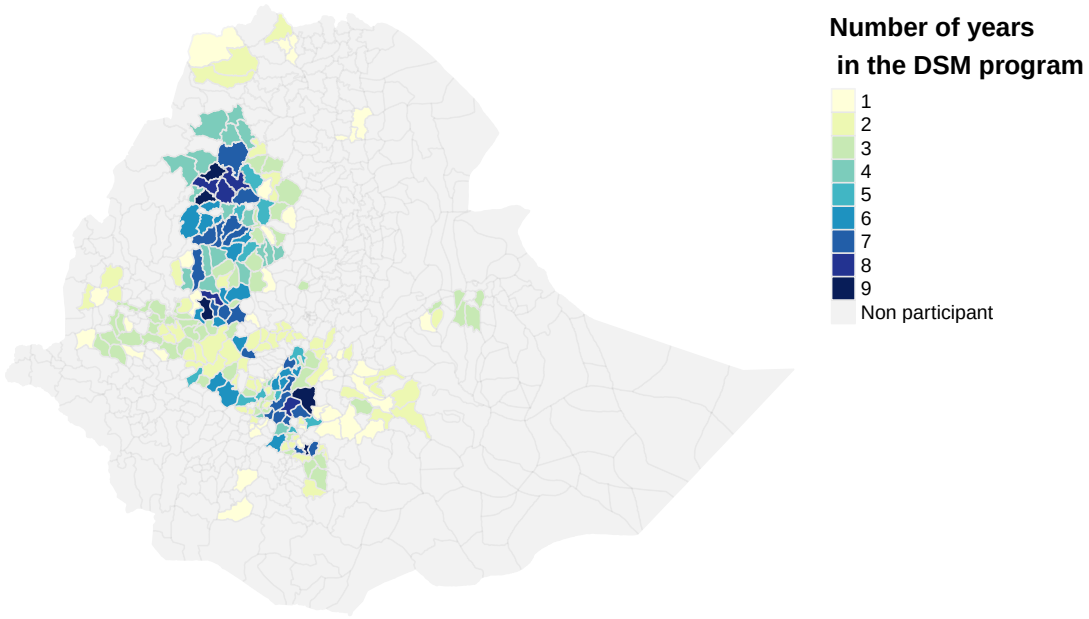

Figure 1: Map of DSM roll-out for maize, showing the number of years that a district was part of the program.

3 Conceptualizing prediction error as source of causal bias

In this section, we discuss the implications of using a predicted continuous outcome variable, $\hat{y}$, instead of the true outcome variable, y, for causal inference. With a classical measurement error structure $\hat{y} = y + u$, measurement errors in a continuous y variable do not bias coefficients and only increase their variance – unlike errors in a discrete y or errors in x – and therefore have received relatively less attention in the literature (Hausman, 2001). However, as we show below, errors in remotely-sensed ML predictions tend to have a non-classical error structure, which warrants further investigation. To start with, we assume a fairly general prediction error structure:

$$\hat{y}_{it} = \gamma + \lambda y_{it} + \delta x_{it} + u_{it} \tag{1}$$

where $\hat{y}$ is the ML predicted value, y the true value, and x is a potential covariate affecting measurement error. Setting $\gamma = \delta = 0$, $\lambda = 1$ and $\text{Cov}(y, u) = 0$ leads to the so-called "classical" measurement error model, $\hat{y} = y + u$, under which using $\hat{y}$ does not lead to bias. Denoting by $\widehat{\text{DiD}}(z)$ the DiD estimator using outcome variable z and denoting the prediction error by $e \equiv \hat{y} - y$, the linearity of both the DiD and the error structure implies that $\widehat{\text{DiD}}(\hat{y})$ can be algebraically decomposed as:[3]

$$\widehat{\text{DiD}}(\hat{y}) = \lambda\widehat{\text{DiD}}(y) + \delta\widehat{\text{DiD}}(x) + \widehat{\text{DiD}}(u) \tag{2}$$

$$= \widehat{\text{DiD}}(y) + \widehat{\text{DiD}}(e) \tag{3}$$

From this algebraic property, the bias and variance of $\beta^{DiD}(\hat{y})$, the DiD coefficient using $\hat{y}$ as outcome variable, are:

$$\text{E}\left[\beta^{DiD}(\hat{y})|X\right] = \lambda\beta^{DiD} + \delta\beta^{DiD}(x) + \beta^{DiD}(u) \tag{4}$$

$$V\left[\beta^{DiD}(\hat{y})|X\right] = \lambda^2 V\left(\beta^{DiD}\right) + \delta^2 V\left(\beta^{DiD}(x)\right) + V\left(\beta^{DiD}(u)\right) \tag{5}$$

[3] This is a simple numerical property of the OLS estimator that holds algebraically. Under standard assumptions guaranteeing the consistency of $\widehat{\text{DiD}}(y) \to \text{DiD}(y)$, equation (2) and (3) will hold at the estimand level. For the proof, see Supplementary Material (SM) section D.

Equation (2) indicates that in order to obtain an unbiased estimation with $\widehat{\text{DiD}}(\hat{y})$, the ML prediction of $\hat{y}$ should ensure that λ (the slope of the predicted versus true values) is equal to 1 and that a difference-in-difference on both the covariates, x, and the residuals u, should be zero (or alternatively, that the three terms cancel out). Equation (3) shows that the three conditions above can be tested using the $\widehat{\text{DiD}}$ (e) instead. Equation (3) is particularly interesting as it does not rely on a specific measurement error structure, unlike Equation (2) which is specific to the assumptions laid out in Equation (1). Equation (3) is thus very general, and applies to any identification strategy that relies on a linear-in-y estimator: the same applies to an RCT, IV or matching estimator.[4] In the analysis below, we use the Callaway and Sant'Anna (2021) estimator, which is a (weighted) average of DiD coefficients and is thus linear in y. Therefore, the same identity holds for the CS-DiD design, that is CS-DiD$(\hat{y})$ = CS-DiD(y) + CS-DiD(e).

Taken together, Equations (2) and (3) suggest two simple tests to gauge the potential bias of a satellite-based impact evaluation. The first test involves evaluating how/whether λ differs from 1, which can be a first indication of bias.[5] The second involves what we call here a *prediction error placebo test*, i.e. making sure that evaluating the causal design of interest using e as outcome variable instead of $\hat{y}$ does not appear significant.

Turning to the variance of the estimator, equation (5) shows that the second part of the common wisdom, i.e. that measurement error in y only increases variance, may not hold with non-classical measurement error. The variance is now affected by λ^2 and thus might be smaller than the variance of the estimator without measurement error when $\lambda < 1$.

Unfortunately, usual ML procedures do not seek to minimize the inference-specific bias $\widehat{\text{DiD}}(e)$ or the λ value, but rather to minimize the root mean squared error (RMSE) associated with $e = \hat{y} - y$. Whereas an RMSE of 0 will necessarily lead to a $\widehat{\text{DiD}}(e)$ of 0, there is no guarantee that reducing RMSE leads to an automatic reduction in $\widehat{\text{DiD}}(e)$. Remembering that the RMSE loss corresponds to the squared bias and variance, a reduction in

[4]Equation (4) covers also the case when the treatment variable is continuous as discussed in Proctor et al. (2023), noting that the second term in their equation 6 $\text{E}[\hat{\beta}] = \lambda\beta + \sigma_{xu}/\sigma_x^2$ corresponds to $\beta_{u\sim x}$, the coefficient of a regression of the measurement errors on the variable x of interest.

[5]As Equation (2) shows, however, such test is not sufficient as $\lambda \neq 1$ is only one part of the total bias. Furthermore, Proctor et al. (2023) find that it accounts for only a small share of the bias they observe in several simulations.

RMSE could amount to reducing variance at the expense of increasing (ML) bias, thus potentially increasing the causal bias. Unfortunately, the focus on RMSE in the ML paradigm means that simple statistics like λ are rarely reported, obscuring the potential bias in using satellite-based predictions for impact assessment. There is growing evidence, however, that satellite-based ML predictions tend to have a λ value below 1. Proctor et al. (2023) use ML to predict six outcomes commonly used in economics and obtain λ values between 0.47 and 0.9. Out of four studies seeking to predict wealth based on satellite data, we find that all of them obtain $\lambda < 1$, with Ratledge et al. (2022) reporting a value of 0.69, appendix data in Yeh et al. (2020) showing values between 0.19 and 0.73, and visual interpretation of the scatterplots in Jean et al. (2016) and Chi et al. (2022) indicative of $\lambda < 1$. Similarly, we find that a majority of crop yield predictions surveyed below have $\lambda < 1$. This evidence suggests that an impact assessment using these satellite-based predictions is likely to be biased downwards with possibly smaller standard errors, falsely indicating small or null results with confidence.

Despite the extensive research on machine learning for causal inference in econometrics, only a handful of papers directly address the EO-ML setting where ML is used to train a model over a small ground-truth sample and predict over a distinct and much larger sample.[6] The studies addressing prediction error in EO-ML settings can be broadly classified into three groups depending on how they address the equation $\hat{\beta}(\hat{y}) = \hat{\beta}(y) + \hat{\beta}(e)$. In the first, which we call the traditional *measurement error correction* approach, the researcher models ex-post the relationship between $\hat{y}$ and y to correct for prediction errors in the causal parameter of interest $\beta(\hat{y})$. Proctor et al. (2023) use multiple imputation error techniques based on $\hat{y}$ and y to obtain "debiased" $\hat{y}$ predictions. Wang et al. (2020b) likewise model the relationship between $\hat{y}$ and y to adjust the inference step. Alix-García and Millimet (2023) use a binary choice model that accounts for misclassification by plugging-in external assessments of classification errors from deforestation maps.

A second approach, which we call *inference-targeted prediction*, applies

[6]Most of the current research in econometrics focuses on the case where ML is used to obtain the best conditional prediction $\hat{x}|z$ of x while observing x for all individuals (for example Belloni et al., 2014 focus on IV and use ML to predict treatment D using many instruments z). On the other hand, the EO-ML field seeks to obtain the best prediction $\hat{y}$ of y (or $\hat{x}$ of x) over a larger sample where one does not observe the true y (or x). Future research ought to investigate how the debiasing approaches of Chernozhukov et al. (2018, 2022) can be applied to the EO-ML context.

to the case where the analyst is producing her own ML prediction $\hat{y}$ and can therefore proceed ex-ante to an upstream adjustment of the ML predictions tailored to the causal problem at hand. Ratledge et al. (2022) use a convolutional neural network (CNN) to predict livelihood in Uganda in order to analyze the effect of electrification on livelihood. They augment the traditional loss function used in a CNN with a *causal* term penalizing the bias in e, showing that their new causal-penalty parameter effectively leads to a λ closer to 1 and therefore a potentially lower DiD bias. Gordon et al. (2023) likewise adjust their loss function to incorporate an adversarial debiasing term.

Finally, a last set of recent papers goes beyond the comparison between $\hat{y}$ and y as done in the *measurement error correction* approach by leveraging the fact that the analyst can compare $\beta(\hat{y})$ to $\beta(y)$ on the ground-truth dataset to conduct a downstream adjustment of $\beta(\hat{y})$. Angelopoulos et al. (2023a) show how a large class of estimators can be corrected by combining $\hat{\beta}(\hat{y}_{out})$ (the estimator obtained using the out-of-sample $\hat{y}_{out}$) with what they call the reducer, i.e. an estimate of the in-sample (or ground truth) estimation error $\hat{\beta}(\hat{y}_{in}) - \hat{\beta}(y_{in})$. They develop a method to derive confidence intervals for their estimator, $\beta^{PPI} \equiv \hat{\beta}(\hat{y}_{out}) - \left(\hat{\theta}(\hat{y}_{in}) - \hat{\theta}(y_{in})\right)$ with theoretical guarantees on their coverage. Interestingly, applied to the DiD case, their estimator amounts to $\widehat{\text{DiD}}^{PPI} = \widehat{\text{DiD}}(\hat{y}_{out}) + \widehat{\text{DiD}}(\hat{e}_{in})$, which naturally arises from (3). Angelopoulos et al. (2023b) and Zrnic and Candès (2023) provide further discussions and extensions of the PPI estimator.

4 Data

4.1 Administrative data on DSM roll-out

Ethiopia has four levels of sub-national administrative units: regions, zones, districts (*woredas*; n=691) and wards (*kebeles*; n=15,670). We obtained administrative district-level data on improved seed varieties supplied and sold by the DSM program from 2011 to 2020 from the Ethiopia Agricultural Transformation Agency (ATA) and the Ministry of Agriculture. Maize and wheat were the primary crop types for which seeds were supplied, both over time and in terms of amount of seed supplied, though seeds for ten crops were provided by the program by the end of the roll-out period. However, there was substantial heterogeneity in the diversity in type of seed provided

across regions participating in DSM. For example, in the Amhara region, only improved maize seeds were provided over the course of the program.

4.2 ESS crop location and crop cut data

For this study, we obtain restricted-access geo-coordinates for plots linked to the surveys conducted during the Ethiopia Socioeconomic Survey (ESS). We use two waves of this survey, ESS 3 and ESS 4, a long-term panel household survey data collection project conducted by the ESS and the World Bank Living Standards Measurement Study-Integrated Surveys on Agriculture (LSMS-ISA) team. The survey collects detailed household-level data on agriculture, such as yields and input use. In the third wave of the panel, ESS 3, 1,255 households were re-interviewed households from previous waves. In the fourth wave, ESS 4, a new panel was created that included a total of 6,770 households interviewed for agriculture and household characteristics. Most of the data collected through this effort is shared publicly and supported by extensive documentation including the surveys and the variables collected (Central Statistical Agency of Ethiopia, 2016, 2019).

The ESS 3 covers the 2015 main cropping season while ESS 4 covers the 2018 main cropping season. For ESS 3 and ESS 4, we use georeferenced points from two different surveys: 1) the household agricultural survey, which provides the plot GPS location per crop type for each household, and 2) the crop cut survey, which provides the yield estimate obtained from crop cutting at harvest for one subplot (or quadrat), that is a sample of one farm plot (not of the full plot area). The crop cut survey was collected for a subset of farm plots. For both surveys, the GPS waypoint was collected at only one corner of the plot. We performed a series of verification steps to assess field data quality and filter out mis-located point coordinates, as we described further in Supplementary Materials (SM) C.2.

4.3 TAMASA crop location and crop cut data

We augment the ESS data with geolocated maize plots and crop cuts data from the Taking Maize Agronomy to Scale in Africa (TAMASA) project.[7] This six-year project (2014-2020) sought to improve productivity and profitability for small-scale maize farmers in Ethiopia, Tanzania and Nigeria. For

[7] https://www.cimmyt.org/projects/taking-maize-agronomy-to-scale-in-africa-tamasa/

Ethiopia, we obtained georeferenced crop cuts data for 2015, 2017, 2018, and 2019. The sampling design from crop cutting vary between years, but most often used 3 subplots from the same maize farm plot. The GPS waypoint was collected at the center of the first quadrat, located at the center of the farm plot. Two other subplots were collected in opposite direction, on a diagonal from the center subplot. The maize yield for one GPS location was estimated as the mean of the subplots.

4.4 Satellite-based maize mapping

To focus our analysis on areas where maize is grown, we map maize crop area for the main growing (*Meher*) season over the country from 2010 to 2020. For each year from 2010-2020, we create an annual maize mask that identifies maize pixels using predictors constructed from the 30-m Landsat surface reflectance data. The higher resolution of Landsat was more appropriate than MODIS for identifying maize area, given that the average size of a smallholder farmer's plot is around or below 1 ha (Headey et al., 2014). We select images from June to the end of December for each year, to cover the main cropping season as identified in the ESS 3 and 4, as well as in previous work about maize and/or cropping season onset (Dunning et al., 2016; Edao et al., 2018; Alhamshry et al., 2020; Guo et al., 2023).

We evaluated the image count at each pixel location to see how many cloud-free images (or pixels) per season were available to use time series analysis for crop mapping (e.g. harmonic modelling), but the image count was too low (2 to 3 images per season, per year). We therefore create annual Landsat median mosaics with cloud/shade-free pixels. For each Landsat median image, we calculate four vegetation indices at the pixel-level including the Normalized Difference Vegetation Index (NDVI), the Enhanced Vegetation Index (EVI), Green Chlorophyll Vegetation Index (GCVI), and the Land Surface Water Index (LSWI).

To map maize, in addition to time-varying Landsat bands and indices, we use time-invariant predictors that are relevant to maize suitability, including topography, climate, and soil characteristics. We use a random forest algorithm for classification, and use hyperparameter tuning to set the optimal model specification. The best results for the maize/non-maize classification were obtained when using the TAMASA datasets for the "maize" category and the ESS 3 and 4 household survey for the "non-maize crop" category as reference data, with a final sample size of 13,501 points for training and

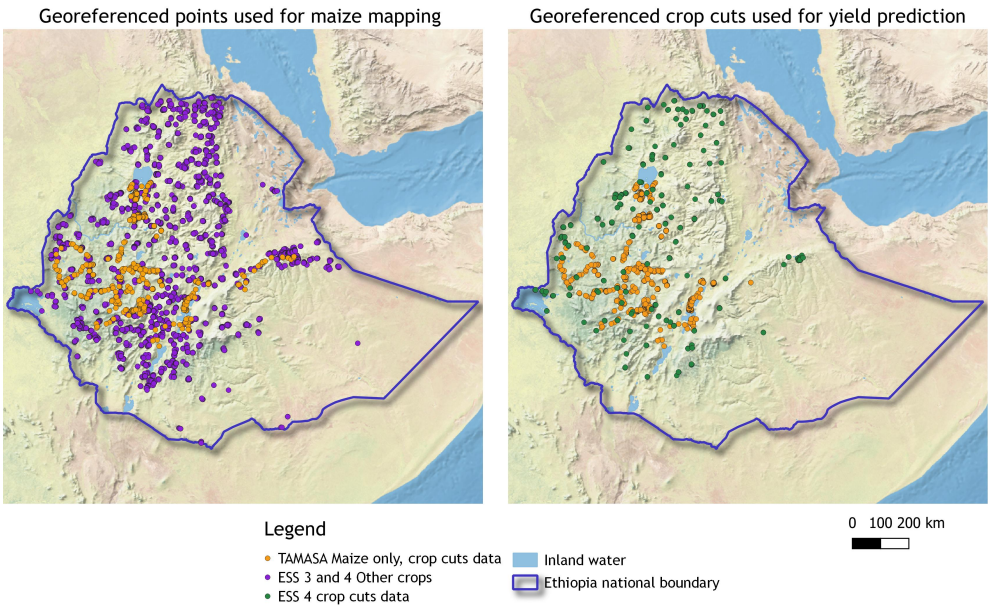

Figure 2: Distribution of the field data used in final models, for maize mapping on the left and maize yield prediction in the right.

4,903 points for validation (Figure 2). We provide accuracy metrics obtained from different combination of the tested datasets in Table SM.2. We then use Landsat-derived cropland masks to exclude non-cropland pixels (Potapov et al., 2022). These cropland extent maps have a good accuracy, [8] but are not available annually. We therefore use the 2011 cropland mask for the 2010-2014 period, the 2015 cropland mask for 2015-2018, and the 2019 cropland mask for 2019-2020.

The final model had 97.8% overall accuracy, with a 78.2% maize producer accuracy (omission error of 21.8%) and 97.2% maize user accuracy (commission error of 2.8%). The choice of a model with a very low commission error of 2.8% despite a larger omission error of 21.8% was made with the final purpose of the model in mind. Admittedly, this model would not be accurate if its goal was to quantify planted area. In fact, comparison with FAOSTAT indicates that the model significantly underestimates maize planted area to

[8] Table 3 from Potapov et al. (2022) report 97.2 (0.6)% overall accuracy for stable cropland for the Africa region.

approximately a fifth of the national total (Figure SM.1). However, since the final goal of the maize map is to select the pixels over which to predict yields, the cost of falsely predicting maize pixel (and thereby predicting maize yield on non-maize pixels) is certainly much higher than the cost of omitting maize pixels. We therefore adopt a conservative approach using a model with a very low commission error, giving us confidence that what we consider maize has a high probability to be maize. At the inference level, this means that the causal effect we are estimating is not strictly over the whole "population" of maize pixels in Ethiopia, but rather over the pixels that have a high probability to be maize pixels. These annual maize/non-maize maps served as maize mask for the annual maize yield predictions discussed in the next section.

5 Methods

5.1 Satellite-based maize yield predictions

We use ML methods to predict maize yield over the pixels that were predicted as planted with maize according to the maize map described above. Our empirical analysis is at administrative units much larger than pixels, so we subsequently aggregate these pixel-level predictions of maize yields to the district- and ward-level. Here we outline the specific data and methodology used to construct these maize yield predictions.

We run a large collection of ML yield prediction models, based on nine algorithms including linear regression models such as OLS, LASSO, Ridge, elastic net to more non-parametric models such as random forest, gradient boosting, a classification and regression tree (CART), a bagged CART and support vector machines (SVM) (see Table SM.1), available in R's package tidymodels (Kuhn and Wickham, 2020).

As predictors in our ML model, we use the Aqua and Terra MODIS Vegetation Indices 16-Day Global 250m.[9] Combined together, these products provide a 8-day cloud-free time series of vegetation indices, with 23 images for the main cropping season each year.[10] The phenological signal carried by

[9]MYD13Q1.061 Aqua Vegetation Indices 16-Day Global 250m and MOD13Q1.061 Terra Vegetation Indices 16-Day Global 250m

[10]Since only 2-3 Landsat images were available per cropping season, it did not contain sufficient information about the crop growth status to predict yields. The temporal reso-

MODIS vegetation indices time series helps predict crop yield for the cropping season. However, the time series can still exhibit missing observations and potential outlier values. To create a regularly-spaced and filtered MODIS vegetation index time-series, we apply pixel-level temporal interpolation and the Savitzky and Golay (1964) smoothing method. Then, we compare four different versions of the data set: 1) unfiltered NDVI time series; 2) gap-filled and filtered NDVI time series; 3) unfiltered EVI time series, and 4) gap-filled and filtered EVI.

For each cropping season, we aggregate the monthly rainfall from CHIRPS data set (Funk et al., 2015), as well as the monthly maximum daytime and minimum nighttime land surface temperature from MODIS.[11] These three predictors are each defined with 6 variables, one per month for the six-month period, for each year. We include the same time invariant variables for elevation and soil conditions as those used for maize mapping (see Section C.3).

To train and validate the yield model, we use ground truth subplot crop cutting data from two sources, the TAMASA project and ESS 4. These datasets covered four different years: 2015, 2017, 2018 and 2019. We extracted the value of each predictors at the location of the maize field, matching the specific season for time-varying variables (MODIS NDVI and EVI, Landsat, and climate related variables).

The ML models are trained on the crop cut dataset using initially 75% of observations for training and the remaining 25% for validation. Model hyperparameters are selected based on a grid search. Out-of-sample accuracy measures values are calculated using a 10 fold cross-validation of the training dataset, repeated five times. We therefore obtain 25 hyperparameters combinations for each model (except for OLS), totaling 201 different ML models.

The pixel-level maize yield predictions are then aggregated to district or ward level means, which we use to produce several DiD estimates of the impact of DSM. The pixel-level predictions are aggregated to the ward and district level using the maize mask. Specifically, we aggregate the maize map pixels (30m resolution) to the MODIS pixel level (250m resolution) to obtain the maize *coverage*, defined as the number of Landsat pixels detected as maize within a MODIS pixel. We only predict yields on MODIS pixels with

lution of MODIS provides better information about the maize crop phenology to support yield prediction

[11]Specifically, we used the merged collections of Aqua and Terra Land Surface Temperature and Emissivity 8-Day Global 1km (MYD11A2.061 and MOD11A2.061 products)

a positive maize coverage, and aggregate the MODIS pixels to the district level by taking a weighted mean of predicted yields across all pixels within a district, using the maize coverage as weights.

5.2 Causal identification and inference

We use a DiD identification strategy to evaluate the causal effects of DSM on average maize yield over time. The key identifying assumption in this approach is that maize yield trends in non-DSM districts provide a reliable counterfactual for DSM districts. The validity of this assumptions hinges on the selection and sequencing of districts into the DSM program as it was scaled up. Based on extensive discussions with government administrators charged with implementing the program and farmer focus groups that we conducted as part of this study, as well as secondary data analysis, there are a number of reasons why the DiD identifying assumption is reasonable in this case. However, as is often true with DiD designs, we cannot completely dismiss all potential concerns about this strategy. At no point during the DSM roll-out were districts randomly assigned to be 'treated' by DSM. Among maize growing districts of Ethiopia, however, the initial piloting and early selection of districts into DSM was not based explicitly on suitability for maize production. To illustrate, we use three rounds of ESS data that span early (2013), middle (2015) and late (2018) stages of this roll-out to construct distributions of land suitability for maize production for DSM and non-DSM districts in each of these years. Figure SM.2 shows considerable overlap of the share of land in DSM and non-DSM districts that are highly and (especially) moderately suitable for maize production.[12] This overlap is sensible given that a number of selection criteria for DSM roll-out were exogenous to maize productivity.

We test for parallel trends in average maize yield between DSM and non-DSM districts and find mixed evidence. For some subsets of the data, we fail to reject parallel trends; for others, we reject parallel trends. We hasten to note, however, that conducting these tests with predicted maize yields raises precisely the same prediction error concerns that are our methodological focus in this analysis. With attenuation bias in trend estimates and overestimated precision, it is unclear how much stock we should place in these tests. In sum,

[12]The unsurprising exception is a significant difference between the two groups for land with 0 percent of land highly or moderately suitable for maize and for land with 100 percent of land not suitable for maize.

the identifying assumption of our DiD strategy seems plausible qualitatively, but it is impossible to defend conclusively. Given the nature of the DSM roll-out, however, it is difficult to imagine a better identification strategy despite these lingering potential concerns.

For this DiD analysis, we account for staggered adoption with multiple time periods. We define DSM adoption at two administrative scales: at the district level (n=691; the scale at which the DSM program was administered) and which serve as our main analysis, and at the ward level (n=15,944), which provides greater statistical power and allows for direct comparison with Mekonnen et al. (2021).[13] Recent studies have questioned the suitability of the two-way fixed effects (TWFE) linear regression model, with unit and time fixed effects, for settings such as this one, where groups get treated at different times (Callaway and Sant'Anna, 2021; Goodman-Bacon, 2021; Athey and Imbens, 2022). In the case of staggered adoption, the standard TWFE estimator can be a weighted average of treatment effects with some of the weights negative (Imbens, 2024). The problem of negative weights comes from the use of already-treated groups as comparison group (Baker et al., 2022) for groups that are treated later on.

To circumvent the issue with the TWFE approach, we apply an alternative estimator proposed by Callaway and Sant'Anna (2021) (CS-DiD), which is intended for settings with staggered treatment timing. Their approach estimates group-time average treatment effects (ATT(g,t)), where each ATT(g,t) is the average treatment on treated effect for a group g, at a specific time t. A unit is defined as belonging to the group g if it was treated at time g. Importantly, DiD with staggered adoption assumes that once units are treated at a particular point in time, they remain treated for all periods afterwards. With DSM, there were 45 districts that entered and exited the program during the 2010 - 2020 panel. These districts were excluded from the analysis.

When estimating the group-time average treatment effects, ATT(g,t), for group g in year t, we use both districts that never adopted DSM, as well as districts that had not yet been enrolled in DSM by year t, as control groups for districts that adopted DSM in year g. This allows us to avoid the issues

[13]DSM adoption at the ward level is defined based on the district that the ward falls in. That is, all wards in the same district follow the same treatment timing path. Note, we cannot be sure that all wards in a district received seeds from DSM. However, given the lack of granular data on DSM adoption at the ward-level, this is a reasonable assumption to make.

of negative weights entering our estimates due to faulty control groups.

Callaway and Sant'Anna (2021) discuss several ways to average the group-time specific ATT(g,t): an overall parameter averaging over all groups and time periods, a group-specific parameter averaging over all years for each group, a calendar-specific parameter averaging over groups for each calendar year or a event-time specific parameter averaging groups for each event-time period. In the following analysis, we focus on the overall average of ATT estimates, which we denote as CS-DiD.

We first estimate CS-DiD impacts of DSM on average predicted maize yields at the district level. To compare these results with those of Mekonnen et al. (2021), we conduct similar CS-DiD estimation with more disaggregated average predicted yields at the ward level. This provides greater statistical power, but also enables a more direct satellite-based replication of Mekonnen et al. (2021). Specifically, we estimate impacts on all maize growing wards as well as for those wards sampled in the household survey by Mekonnen et al. (2021). To do this, we use the centroid of the households surveyed for each enumeration area in their study (provided by their study team) and selected the wards that contained the surveyed households and surrounding agricultural land (n=374). We further test the CS-DiD estimation over only sampled wards in the survey years of Mekonnen et al. (2021) (2012, 2016, 2019).

In addition to comparing our CS-DiD estimates using predicted yields to the DiD estimates using household-reported yields from Mekonnen et al. (2021), we compare the CS-DiD estimates for average predicted maize yield by administrative units with CS-DiD estimates for the actual plot-level crop cut maize yield. As described above, these data were collected by the ESS 4 and the TAMASA project in 2015, 2017, 2018, and 2019.

5.3 Correcting prediction error bias in DiD estimates

In this paper, we combine the *inference-targeted prediction* and *prediction-powered inference* approaches in an attempt to correct for prediction error bias in causal estimates. A drawback of the inference-targeted prediction method of Ratledge et al. (2022) is that customizing the loss function to the causal task at hand is a complicated endeavour that prevents the use of many off-the-shelf highly-optimized algorithms. Instead, we use a large set of popular ML models and assess whether some of the models provide predictions with a lower causal bias $\widehat{\text{DiD}}(e)$. We here depart from the typical

ML workflow where a practitioner would select the best model based on RMSE, refit the model using all ground-truth observations and use that model to generate predictions. Instead of selecting the single model leading to the lowest RMSE, we initially keep all the models and compute the DiD criterion $\widehat{\text{DiD}}(e)$ and $\widehat{\text{DiD}}(\hat{y})$ on each model. We evaluate all models based on both RMSE and causal bias $\widehat{\text{DiD}}(e)$, and ask whether reducing RMSE necessarily leads to a lower causal bias. Finally, we select two combinations per model, the *best ML model* (minimizing RMSE(e)) and *best causal model* (minimizing $\widehat{\text{DiD}}(e)$), which we use to predict yields over the pixels identified as maize-cropped area in Ethiopia and compare the different $\widehat{\text{DiD}}(\hat{y})$.

In a second step, we take a different approach to addressing the causal bias following the *prediction-powered inference* framework (PPI) by Angelopoulos et al. (2023a). Whereas in the first step we were seeking predictions with small causal bias $\widehat{\text{DiD}}(e)$, here we directly debias the estimator $\beta^{DiD}(\hat{y})$ using $\beta^{DiD}(e)$. Angelopoulos et al. (2023a) argue that their PPI estimator combines the strength of the machine-learning predictions (a large dataset $\hat{y}$ presumably leading to a low-variance but possibly biased $\beta^{DiD}(\hat{y})$) and those of the ground-truth dataset (a high-variance but unbiased estimator $\beta^{DiD}(y^{ground})$). In our case, this argument needs to be slightly nuanced as our initial pixel-level predictions $\hat{y}$ are aggregated at the administrative level and thus the sample size of the predicted dataset is thus not much larger than the ground-truth dataset.

6 Results

6.1 Satellite-based maize yield predictions

The annual yield prediction model was trained using the 250-m MODIS vegetation time series together with temperature, rain and terrain and soil variables as described in Section 5.1. In our initial exploration, the filtered EVI vegetation index emerged as the most predictive vegetation index, which we subsequently used to train our nine algorithms. Three algorithms consistently outperformed the others in terms of RMSE: gradient boosting, bagged CART and random forest (RF) (see Figure SM.4). Among them, gradient boosting achieved the lowest RMSE, and was thus retained as the *best-RMSE* model in the subsequent analysis.

The performance of the *best-ML* model evaluated on the independent

validation sample had an RMSE of 1,658 kg/ha and an R^2 of 0.50, while the final model refitted and re-evaluated on the whole cropcut dataset had an RMSE of 1,238 kg/ha and R^2 of 0.75 (Figure 3). While an R^2 value of 0.5 might seem low, this is on par with results reported elsewhere in the literature: Burke and Lobell (2017) obtained a value of 0.39 for maize yields in Kenya, Guo et al. (2023) 0.54 for maize yields in Ethiopia,[14] Jin et al. (2017) 0.3-0.6 for maize in Kenya, Jin et al. (2019) 0.39-0.54 for maize in Kenya and Uganda, and Jain et al. (2016) 0.33 for wheat in India. Even in the USA, where larger and more uniform fields, as well as high-quality ground-truth data should make prediction more accurate, Deines et al. (2021) obtain a R^2 of 0.45. Statistics on λ were unfortunately rarely available in these studies. However, it appears that $\lambda < 1$ in almost all cases. For instance, the seminal paper of Burke and Lobell (2017) likely obtained a λ value of 0.39, Jin et al. (2017) likely obtained λ values between 0.25 and 0.56.[15]

Figure SM.3 shows the distribution of ten-year average yields over space, highlighting a clustering pattern with relatively higher yields in the Western region.

6.2 DSM impacts on maize yield

We first estimate the effect of DSM at the district level using the best-RMSE yield predictions. Table 1 shows the CS-DiD coefficient using different subsets of the districts, using either all district-years available (Column 1), only the years for which we had training data (Column 2) or only the years and districts for which we had training data (Column 3). In each specification, we observe a small and positive coefficient, a result at odds with the findings of Mekonnen et al. (2021) who found a sizeable impact.

Indeed, Table 2 highlights how even our attempts to replicate the analysis of Mekonnen et al. (2021) as closely as possible, by using the smaller spatial units (wards) and restricting to the survey years used in their analysis (Column 3), is unable to recover their findings. While Mekonnen et al.

[14] Guo et al. (2023) report a higher R^2 value of 0.62 using a neural network model, however, they noted the complexity and practical challenges associated with training and scaling up these models, which aligns with our rationale for excluding neural networks from the model we considered.

[15] These numbers were obtained by digitizing figure 2 panel C in Burke and Lobell (2017) and figure 5 in Jin et al. (2017) and should therefore be interpreted with caution.

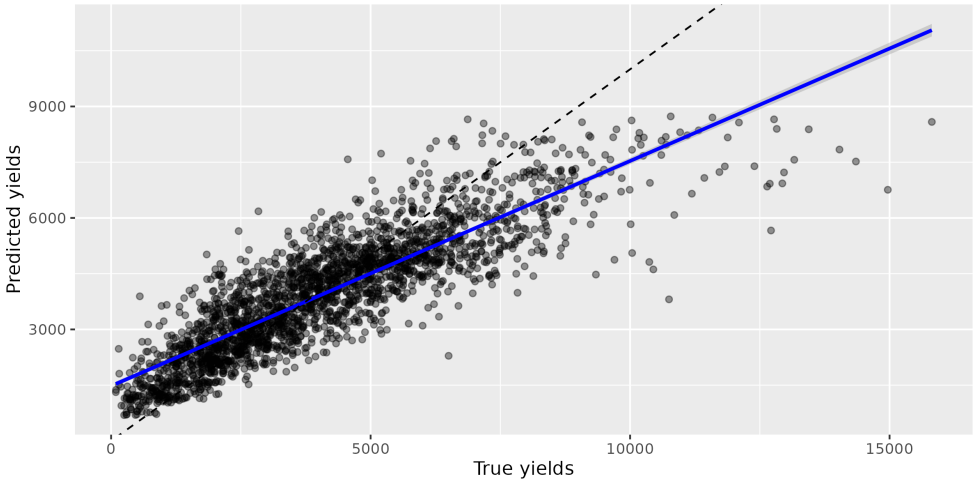

Figure 3: Scatterplots showing the fit between the predicted and the actual maize yield.

(2021) find statistically significant increases in maize yields on the order of greater than 20% using household-reported yields, we find small, statistically insignificant increases or even decreases in predicted maize yields. In addition to prediction errors introduced by model selection, our inability to recover the findings of Mekonnen et al. (2021) could be driven by a variety of other factors, such as low quality input data to the prediction model. We discuss these issues in more detail in SM Section C.

Table 1: CS-DiD estimates based on district-level predicted yields

	Full	Cropcut years	Cropcut years and districts
CS-DiD	70.79	96.46	75.03
	[−36.00; 177.58]	[−49.84; 242.76]	[−138.32; 288.38]
N units	2859	892	418
N years	11	4	4
N groups	9	4	4

* 0 value outside the 95% confidence interval

Table 2: CS-DiD analysis with the sample used in Mekonnen et al. (2021).

	Pred. Yields	Ln(Pred. Yields)	Ln(Pred. Yields), Subset Years
CS-DiD	91.24	0.03	−0.02
	[−67.84; 250.32]	[−0.03; 0.09]	[−0.13; 0.08]
N units	940	940	256
N years	11	11	3
N groups	8	8	7

* 0 value outside the 95% confidence interval

6.3 Causal bias due to prediction error

The result above is obtained using the typical naive prediction-inference workflow where the best RMSE model is used for inference without paying attention to the *causal accuracy of the model* and its potential bias. We revisit here the model and assess its properties in a causal sense.

Using first a model-agnostic criterion, we look at λ, the coefficient of the regression on predicted yields $\hat{y}$ on true yields y (based on the crop cuts data). This coefficient is 0.60 [0.59, 0.62], indicating a relatively high mean-reverting bias. According to Equation (2), this is likely to lead to attenuation bias with confidence, i.e. an underestimation of the true effect and of its variance. To verify this, we compute the value of the overall CS-DiD as well as the individual ATT-DiD coefficients using y, $\hat{y}$, and e from the crop cut data. Results, shown in Table 3, confirm the intuition based on observing $\lambda < 1$: the value of CS-DiD($\hat{y}$) is indeed smaller than CS-DiD(y) by 259 [kg/ha], indicating an attenuation of the true effect. However, the value of CS-DiD(e) indicates that this difference is not significantly different from 0. The variance of CS-DiD($\hat{y}$) is also smaller than that of CS-DiD(y), consistent with the finding that $\lambda < 1$. Turning to the individual group-time ATT-DiD coefficients, we similarly observe the two phenomena of an attenuation of the coefficient and a smaller variance in almost all cases for DiD($\hat{y}$). Comparing the values of DiD($\hat{y}$) to DiD(y) across groups,

While estimates using the true yield y show large and statistically significant variations both across groups and across time, the DiD(e) estimates show a smaller variation. This variation across time is not statistically significant, suggesting that the prediction error can be considered as constant over time.

Table 3: DiD and CS-DiD analysis on crop cut data.

	Yield y	Predicted yield $\hat{y}$	Error $\hat{y} - y$
CS-DiD	976.4***	717.7***	−258.7
	(253.2)	(117.9)	(180.6)
ATT(g=2016,t=2018)	2846.14***	1967.22***	−878.92**
	(565.31)	(353.69)	(334.82)
ATT(g=2017,t=2018)	400.32	491.39	91.07
	(695.74)	(322.75)	(522.01)
ATT(g=2018,t=2017)	61.43	−3.15	−64.58
	(349.61)	(178.71)	(265.24)
ATT(g=2018,t=2018)	559.71	420.52**	−139.19
	(300.93)	(142.81)	(214.83)
ATT(g=2019,t=2017)	914.45*	841.09***	−73.36
	(386.92)	(177.06)	(272.19)
ATT(g=2019,t=2018)	−727.76	−609.82***	117.94
	(405.84)	(145.87)	(362.20)
N units	1248	1248	1248
N years	4	4	4
N groups	4	4	4

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

* 0 value outside the 95% confidence interval. DiD refers to the DiD estimation of the ATT at time t of the group that received treatment at time g. Values with $t < g$ therefore refer to estimates of pre-treatment placebo effects. CS-DiD refers to the overall treatment averaged over all DiD such that $t \geq g$.

The previous results in Table 1 were based on the model selected based on the RMSE criterion. As we argue above, this model is not necessarily the best model from a causal perspective. To investigate this, we revisit the entire collection of machine learning models that were trained, which comprised nine different algorithms resulting in 201 model-parameter specifications (see Table SM.1 for an overview).

The first panel of Figure 4 shows the progression of improvement in RMSE starting from the worst (left) to the best model (right).[16] Whereas the gradient boosting model achieves the lowest RMSE, random forests (RF) and bagged CART model also perform comparatively well, and the fifty best models appear to have a similar RMSE if one considers their confidence interval.[17] The second panel shows the λ coefficients. All models seem to exhibit a strong mean-reverting bias, with λ never above the value of 0.6. Improvements in RMSE do not seem to necessarily increase the λ coefficient toward 1: models with λ closest to 1, mostly from gradient boosting, rank between 75th and 50th in RMSE. Interestingly, some models such as the CART seem to be facing a clear causal versus ML accuracy trade-off, wherein an increase in RMSE seems to lead to a reduction in the mean-reversion bias. The third panel of Figure 4 shows the value of the prediction error placebo test CS-DiD(e), i.e. the CS-DiD coefficient estimated on the errors from the crop cut model.[18] As one could have conjectured from the observation that $\lambda < 1$, almost all these point estimates are negative, ranging between -250 and -750. Confidence intervals are rather wide for many model specifications, and often include zero. Qualitatively, we observe the same results as for the λ coefficients, whereas the best models in a RMSE sense do not necessarily have CS-DiD(e) closest to zero.

To evaluate the CS-DiD coefficient on alternate prediction models, we generated predictions for each class of model, using each time either the model specification that led to the lowest RMSE or to the lowest value of CS-DiD(e). Table 4 shows the CS-DiD coefficients based on each of these model specifications. Comparing within each class of model, we observe that using the predictions from the model with the lowest CS-DiD(e) tends to deliver point estimates that are slightly higher compared to using best-

[16] For readability, Figure 4 shows only the first 175 models since some model specifications performed particularly badly.

[17] The confidence intervals are based on the ten-fold cross-validation repeated five times.

[18] Whereas the first two panels show coefficients estimated on the validation sample, the CS-DiD(e) coefficient was estimated on the full sample after models were refitted.

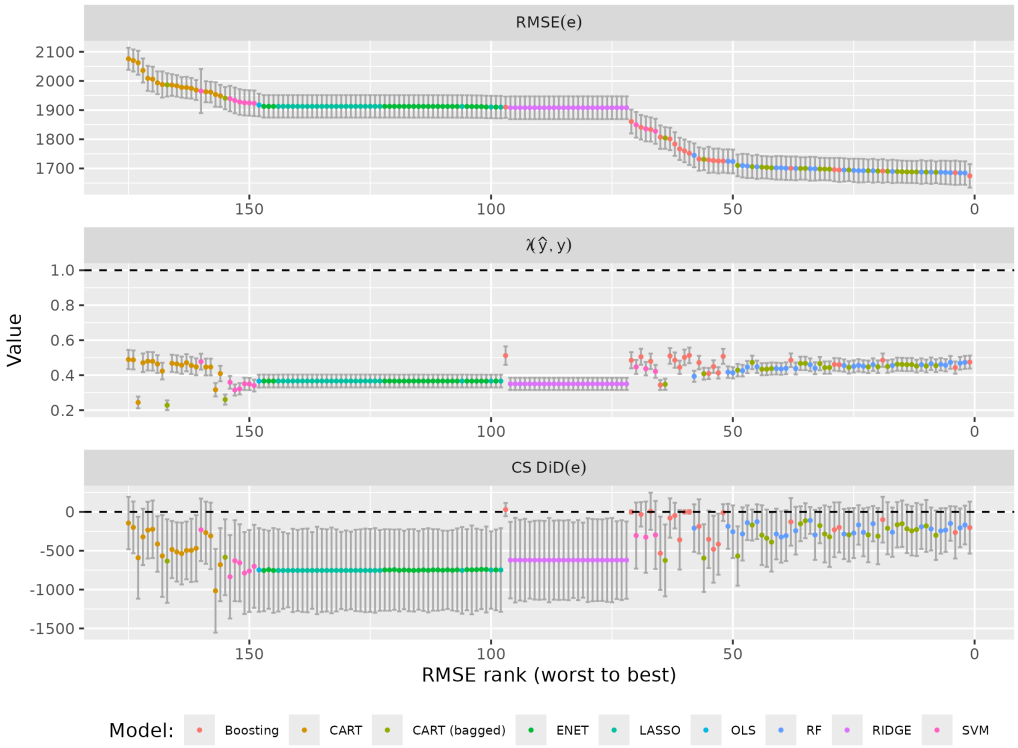

Figure 4: Comparison of RMSE(e), $\lambda(\hat{y}, y)$ and CS-DiD(e) for each model specification.

Table 4: CS-DiD analysis on predictions from each well-performing model.

	Boosting		CART		CART (bagged)		RF	
	Best ML	Best Causal	Best ML	Best Causal	Best ML	Best Causal	Best ML	Best Causal
CS-DiD	70.79	70.25	−22.35	304.91	11.57	59.63	75.85	70.71
	(55.41)	(71.41)	(97.17)	(193.37)	(68.60)	(89.36)	(47.03)	(42.47)
RMSE	1658.28	1809.42	1915.44	2491.50	1661.03	1718.48	1689.35	1703.12
R^2	0.50	0.42	0.36	0.20	0.49	0.46	0.48	0.47
λ	0.44	0.51	0.46	0.50	0.45	0.47	0.44	0.46

RMSE predictions. This is consistent with the phenomenon observed above that most of the predictions had a negative CS-DiD(e).

We turn now to the *prediction-powered inference* method of Angelopoulos et al. (2023b), which consists of correcting the satellite-based estimate with the bias observed in the ground-truth crop cut data. Figure 5 shows the raw and *prediction-powered inference*-corrected estimators for the four model types shown above. As expected, the corrected coefficients tend to be higher than the raw ones, though they have systematically wider confidence intervals. The *prediction-powered inference* correction has a much more limited effect on the estimates that are using the best-causal predictions, which is to be expected given that these predictions were specifically based on their small value of CS-DiD(e). The main insight from Figure 5 is that using a bias-corrected estimator raises the familiar bias-variance trade-off: reducing the bias of the raw estimator comes at the expense of increasing by an important amount the width of the confidence interval.

7 Conclusion

Advances in satellite remote sensing and artificial intelligence provide exciting new opportunities to inform sustainable development efforts by tracking and measuring impacts of agricultural interventions at scale and at low cost. However, remote sensing applications for causal impact evaluation bring new methodological challenges. In particular, satellite-based predictions introduce new sources of measurement error that may bias subsequent causal impact estimation. In this paper, we explore the trade-off between the predictive power of ML models to generate outcome variables and causal bias by estimating the impact of Ethiopia's Direct Seed Marketing program (DSM) on maize yields.

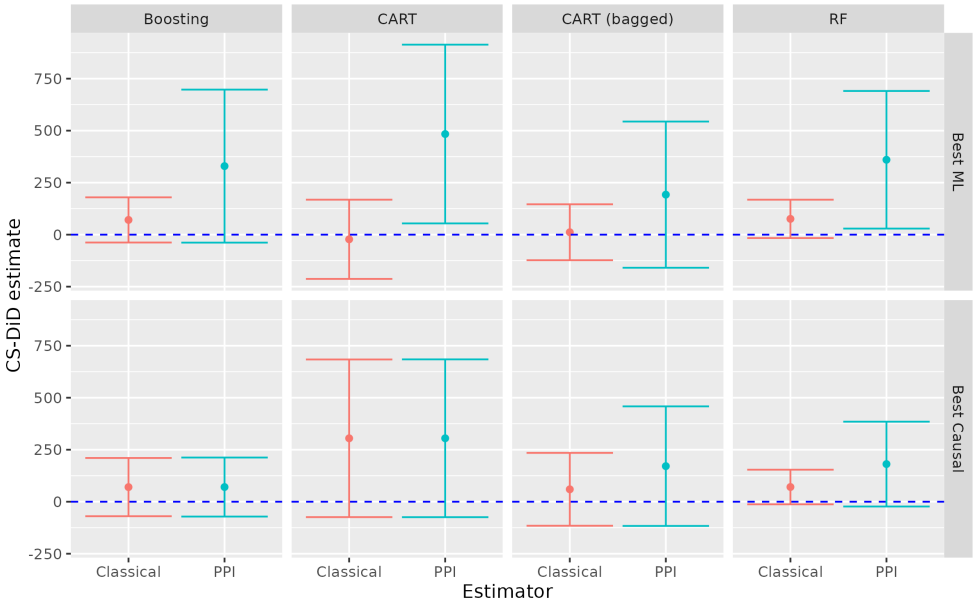

Figure 5: Comparison of naive and *prediction-powered inference* estimates.

Empirically, we contribute to the ongoing expansion of the literature on satellite-based impact evaluation of agricultural interventions, a lively contemporary field of development economics. The DSM program we study aims to overhaul and modernize the seed system in Ethiopia and thereby provide farmers greater choice of and more timely access to improved seeds. We find some positive evidence of DSM on maize yields when measured with satellite-based yield predictions, but this estimated impact is smaller and less precise than the DiD estimates of DSM impact obtained using crop cut or farmer-reported yield as outcome variables.

Methodologically, we investigate the extent to which ML prediction models that minimize RMSE and therefore appear best suited for predicting outcomes may be sub-optimal when using these outcomes for causal impact analysis. Using 30-m Landsat maize maps and 250-m MODIS maize yield predictions from 2010 to 2020, we find that predicted yields from models that minimize RMSE can introduce non-classical measurement error that biases DiD estimates. We provide a theoretical framework that generalizes this finding to other contexts where outcomes predicted from remotely sensed data with ML methods are used for causal inference. This tension underscores

how for researchers interested in causal inference, selection of predictive models should not be based solely on the traditional ML criteria of minimized RMSE. When choosing between ML prediction models, researchers should also consider a causal accuracy criterion by conducting a prediction error placebo test. Inspired by the *prediction-powered inference* (PPI) approach (Angelopoulos et al. (2023a,b); Zrnic and Candès (2023)), we propose a new way to correct for this bias in causal estimates based on the raw ML predictions. By leveraging ground truth data, this PPI bias correction is able to produce correct confidence interval coverage.

The source of the prediction errors that are the methodological focus of this study range from upstream measurement issues in the maize geolocation and yield training data to downstream modelling choices, but the upstream data issues loom large as limitations for this analysis as they do for most similar studies. First, the type of satellite imagery that is available for our research question and historical context lacks the spatial and temporal resolution of smallholder maize farming. More recent interventions will be able to access newer sensors and satellites, and satellite data will also continue to improve and provide additional options for agricultural impact evaluation in such settings (e.g., mapping maize more accurately with time series analysis as in Jin et al., 2019; Wang et al., 2020a). Second, ground-truth field data quality issues can introduce prediction error (Elmes et al., 2020). One such source of error in our case was the geo-referencing protocol used to collect our ground-truth data: a single GPS point taken at a random corner of the surveyed plot, an unfortunate source of geo-location measurement error that can be easily avoided by following recent recommendations (Azzari et al., 2021). It is important to appreciate the out-sized implications of such seemingly trivial data collection decisions for follow-on yield prediction, as described in recent recommendations for crop cutting methods (Lobell et al., 2020; Kosmowski et al., 2021; Tiedeman et al., 2022). The good news is that some of these limitations will fade with adherence to available recommendations for field data sampling and the availability of new and better satellite imagery. But for the foreseeable future researchers will have to grapple with prediction errors and understand their empirical implications.

Our analysis in this paper sheds light on a more general trade-off in applied economics between EO data and survey data as the basis for empirical analysis and impact evaluation specifically. Given their comparative strengths and weaknesses, these approaches will coexist for the foreseeable future. Appreciating the comparative advantage of each will therefore remain

essential to research in many fields. In the context of our study, survey-based methods can provide information that is closer to the decision-making process of farmers, which can reveal the mechanisms at work and can generate a much richer set of outcomes. In contrast, our use of EO data focused only on maize productivity as a summary outcome variable that subsumes a host of on-farm responses and decisions that are (potentially) influenced by the DSM program but invisible to satellites. Advances in sensors and ML methods, combined with richer ground-truth data, may continue to expand the set of outcomes that are possible to predict using EO data. There will, however, always be limits on what researchers are able to 'see' in EO data – and such data limitations will constrain our ability to tell a complete story about how and why an intervention is effective. On the other hand, for understanding *where* an intervention is most (or least) effective, EO data may be more promising than survey data. Spatial, pixel-level data can enable researchers to conduct analysis at different scales or in different time periods in a way that can often be modified or extended at relatively little additional cost. This EO advantage can not only open new lines of inquiry (e.g., related to heterogeneous effects and spillovers), but also open new geographies and contexts that are difficult to study via survey data because they are difficult places to survey (Porteous, 2022). We contribute to a greater appreciation for how and when to leverage these important advantages of EO data for impact evaluation.

While it is easy to get swept up in technical problems and sophisticated solutions, it is also important to appreciate the more mundane key ingredients to the kind of satellite-based impact evaluation contained in this paper. In particular, these empirical methods, as powerful and promising as they can be, require reliable on-the-ground monitoring data that track the scale up of programs or dissemination of products and services. Such monitoring data are rarely, if ever, collected with this use in mind, but they are an essential element of ex-post impact evaluation – and limitations in these monitoring data can directly hamper subsequent evaluation, as illustrated in our case. Since its inception, the DSM program was closely monitored by implementing organizations and partners, which included collecting, recording and archiving information about the seed supplied and sold by variety to every district. Unfortunately, since 2020, the program is managed by the Ministry of Agriculture and the Regional Agriculture Bureaus, which have stopped publishing the district-level seed distribution data. The accounting of the seed distribution by DSM is done at the local level in Ethiopia, but

only provides aggregated numbers of the seed distribution to the Ministry. Our analysis would be more compelling with greater statistical power from additional years, but this is not an option under the present conditions for want of disaggregated monitoring data. Collecting programmatic information, especially during scale-up and roll-out phases, is essential for estimating development impacts of agricultural interventions. That will continue to be the case even as sensors and estimation techniques become ever more sophisticated.

References

Abate, Tsedeke, Bekele Shiferaw, Abebe Menkir et al. (2015) "Factors that transformed maize productivity in Ethiopia," *Food Security*, 7 (5), 965–981, doi:10.1007/s12571-015-0488-z.

Acevedo, Maricelis, Kevin Pixley, Nkulumo Zinyengere et al. (2020) "A scoping review of adoption of climate-resilient crops by small-scale producers in low- and middle-income countries," *Nature Plants*, 6 (10), 1231–1241, doi:10.1038/s41477-020-00783-z.

Al Rafi, Dewan Abdullah (2023) "A Geospatial Impact Evaluation of Stress-Tolerant Rice Varieties in Flood Prone Bangladesh," Master's thesis, The University of Arizona, United States – Arizona, https://www.proquest.com/dissertations-theses/geospatial-impact-evaluation-stress-tolerant-rice/docview/2847323030/se-2?accountid=10267.

Alhamshry, Asmaa, Ayele A. Fenta, Hiroshi Yasuda, Reiji Kimura, and Katsuyuki Shimizu (2020) "Seasonal Rainfall Variability in Ethiopia and Its Long-Term Link to Global Sea Surface Temperatures," *Water*, 12 (1), doi:10.3390/w12010055.

Alix-García, Jennifer and Daniel L. Millimet (2023) "Remotely Incorrect? Accounting for Nonclassical Measurement Error in Satellite Data on Deforestation," *Journal of the Association of Environmental and Resource Economists*, 10 (5), 1335–1367, doi:10.1086/723723.

Angelopoulos, Anastasios N., Stephen Bates, Clara Fannjiang, Michael I.

Jordan, and Tijana Zrnic (2023a) "Prediction-powered inference," *Science*, 382 (6671), 669–674, doi:10.1126/science.adi6000.

Angelopoulos, Anastasios N., John C. Duchi, and Tijana Zrnic (2023b) "PPI++: Efficient Prediction-Powered Inference," doi:10.48550/ARXIV.2311.01453.

Aramburu-Merlos, Fernando, Fatima A. M. Tenorio, Nester Mashingaidze, Alex Sananka, Stephen Aston, Jonathan J. Ojeda, and Patricio Grassini (2024) "Adopting yield-improving practices to meet maize demand in Sub-Saharan Africa without cropland expansion," *Nature Communications*, 15 (1), 4492, doi:10.1038/s41467-024-48859-0.

Athey, Susan and Guido W. Imbens (2022) "Design-based analysis in Difference-In-Differences settings with staggered adoption," *Annals Issue in Honor of Gary Chamberlain*, 226 (1), 62–79, doi:10.1016/j.jeconom.2020.10.012.

Azzari, George, Shruti Jain, Graham Jeffries, Talip Kilic, and Siobhan Murray (2021) "Understanding the Requirements for Surveys to Support Satellite-Based Crop Type Mapping: Evidence from Sub-Saharan Africa," *Remote Sensing*, 13 (23), doi:10.3390/rs13234749.

Baker, Andrew C., David F. Larcker, and Charles C.Y. Wang (2022) "How much should we trust staggered difference-in-differences estimates?" *Journal of Financial Economics*, 144 (2), 370–395, doi:10.1016/j.jfineco.2022.01.004.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen (2014) "Inference on Treatment Effects after Selection among High-Dimensional Controls," *The Review of Economic Studies*, 81 (2 (287)), 608–650, http://www.jstor.org/stable/43551575.

Benson, Todd, David Spielman, and Leulsegged Kasa (2014) *Direct seed marketing program in Ethiopia in 2013: An operational evaluation to guide seed-sector reform*, 1350: Intl Food Policy Res Inst, http://dx.doi.org/10.2139/ssrn.2483989.

BenYishay, Ariel, Rachel Sayers, Kunwar Singh, Seth Goodman, Madeleine Walker, Souleymane Traore, Mascha Rauschenbach, and Martin Noltze

(2024) "Irrigation strengthens climate resilience: Long-term evidence from Mali using satellites and surveys," *PNAS Nexus*, 3 (2), pgae022, doi:10.1093/pnasnexus/pgae022.

Burke, Marshall, Anne Driscoll, David B. Lobell, and Stefano Ermon (2021) "Using satellite imagery to understand and promote sustainable development," *Science*, 371 (6535), doi:10.1126/science.abe8628.

Burke, Marshall and David Lobell (2017) "Satellite-based assessment of yield variation and its determinants in smallholder African systems," *PNAS*, 114:9, 2189–2194, doi:doi:10.1073/pnas.1616919114.

Callaway, Brantly and Pedro H.C. Sant'Anna (2021) "Difference-in-Differences with multiple time periods," *Journal of Econometrics*, 225 (2), 200–230, doi:https://doi.org/10.1016/j.jeconom.2020.12.001, Themed Issue: Treatment Effect 1.

Cambron, Trevor W, Jillian M Deines, Bruno Lopez, Rinkal Patel, Sang-Zi Liang, and David B Lobell (2024) "Further adoption of conservation tillage can increase maize yields in the western US Corn Belt," *Environmental Research Letters*, 19 (5), 054040, doi:10.1088/1748-9326/ad3f32.

Central Statistical Agency of Ethiopia (2016) "Ethiopia Socioeconomic Survey,Wave 3 (ESS3) 2015-2016. Ref: ETH_2015_ESS_v02_M.," https://doi.org/10.48529/ampf-7988.

——— (2019) "Ethiopia Socioeconomic Survey 2018-2019 (ESS4). Ref: ETH_2018_ESS_v03_M.," https://doi.org/10.48529/k739-c548.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins (2018) "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, 21 (1), C1–C68, doi:10.1111/ectj.12097.

Chernozhukov, Victor, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K. Newey, and James M. Robins (2022) "Locally Robust Semiparametric Estimation," *Econometrica*, 90 (4), 1501–1535, doi:10.3982/ecta16294.

Chi, Guanghua, Han Fang, Sourav Chatterjee, and Joshua E. Blumenstock (2022) "Microestimates of wealth for all low- and middle-income

countries," *Proceedings of the National Academy of Sciences*, 119 (3), doi:10.1073/pnas.2113658119.

Cole, Shawn, T. Harigaya, G. Killeen, and Aparna Krishna (2020) "Using satellites and phones to evaluate and promote agricultural technology adoption: Evidence from smallholder farms in India,"Technical report, PxD Precision Development, https://precisiondev.org/wp-content/uploads/2023/07/SoilFertility_Jul21.2023.pdf.

Deines, Jillian M., Rinkal Patel, Sang-Zi Liang, Walter Dado, and David B. Lobell (2021) "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," *Remote Sensing of Environment*, 253, 112174, doi:https://doi.org/10.1016/j.rse.2020.112174.

Deines, Jillian M, Sherrie Wang, and David B Lobell (2019) "Satellites reveal a small positive yield effect from conservation tillage across the US Corn Belt," *Environmental Research Letters*, 14 (12), 124038, doi:10.1088/1748-9326/ab503b.

Donaldson, Dave and Adam Storeygard (2016) "The View from Above: Applications of Satellite Data in Economics," *Journal of Economic Perspectives*, 30 (4), 171–98, doi:10.1257/jep.30.4.171.

Dunning, Caroline M, Emily CL Black, and Richard P Allan (2016) "The onset and cessation of seasonal rainfall over Africa," *Journal of Geophysical Research: Atmospheres*, 121 (19), 11–405, doi:https://doi.org/10.1002/2016JD025428.

Edao, Agere Lupi, Kibebew Kibert, and Girma Mamo (2018) "Analysis of start, end and length of the growing season and number of rainy days in semi-arid central Refit Valley of Oromia State, Ethiopia," *Advances in Crop Science and Technology*, 6 (386), 1–6, doi:10.4172/2329-8863.1000386.

Elmes, Arthur, Hamed Alemohammad, Ryan Avery et al. (2020) "Accounting for Training Data Error in Machine Learning Applied to Earth Observations," *Remote Sensing*, 12 (6), 1034, doi:10.3390/rs12061034.

Fang, Jing and Yongzhong Su (2019) "Effects of Soils and Irrigation Volume on Maize Yield, Irrigation Water Productivity, and Nitrogen Uptake," *Scientific Reports*, 9 (1), 7740, doi:10.1038/s41598-019-41447-z.

Ferguson, Joel and Bram Govaerts (2024) "Economic and Environmental Impacts of Sustainable Agriculture in Practice and at Scale: Evidence from Mexico,"Technical report, AgEcon Search, doi:10.22004/ag.econ.343753.

Funk, Chris, Pete Peterson, Martin Landsfeld et al. (2015) "The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes," *Scientific Data*, 2 (1), 150066, doi:10.1038/sdata.2015.66.

Gitelson, Anatoly A., Yuri Gritz, and Mark 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 (3), 271–282, doi:10.1078/0176-1617-00887.

Goodman-Bacon, Andrew (2021) "Difference-in-differences with variation in treatment timing," *Themed Issue: Treatment Effect 1*, 225 (2), 254–277, doi:10.1016/j.jeconom.2021.03.014.

Gordon, Matthew, Megan Ayers, Eliana Stone, and Luke C Sanford (2023) "Remote Control: Debiasing Remote Sensing Predictions for Causal Inference," in *ICLR 2023 Workshop on Tackling Climate Change with Machine Learning*, `https://www.climatechange.ai/papers/iclr2023/22`.

Gorelick, Noel, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore (2017) "Google Earth Engine: Planetary-scale geospatial analysis for everyone," *Remote Sensing of Environment*, 202, 18–27, doi:https://doi.org/10.1016/j.rse.2017.06.031, ISBN: 0034-4257 Publisher: Elsevier.

Guo, Zhe, Jordan Chamberlin, and Liangzhi You (2023) "Smallholder maize yield estimation using satellite data and machine learning in Ethiopia," *Crop and Environment*, 2 (4), 165–174, doi:10.1016/j.crope.2023.07.002.

Hausman, Jerry (2001) "Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left," *The Journal of*

Economic Perspectives, 15 (4), 57–67, http://www.jstor.org/stable/2696516.

Headey, Derek, Mekdim Dereje, and Alemayehu Seyoum Taffesse (2014) "Land constraints and agricultural intensification in Ethiopia: A village-level analysis of high-potential areas," *Boserup and Beyond: Mounting Land Pressures and Development Strategies in Africa*, 48, 129–141, doi:10.1016/j.foodpol.2014.01.008.

Hengl, Tomislav, Jorge Mendes de Jesus, Gerard BM Heuvelink et al. (2017) "SoilGrids250m: Global gridded soil information based on machine learning," *PLOS One*, 12 (2), e0169748.

Huete, AR, HQ Liu, KV Batchily, and WJDA Van Leeuwen (1997) "A comparison of vegetation indices over a global set of TM images for EOS-MODIS," *Remote sensing of environment*, 59 (3), 440–451, doi:https://doi.org/10.1016/S0034-4257(96)00112-5.

Imbens, Guido W (2024) "Causal Inference in the Social Sciences," *Annual Review of Statistics and Its Application*, 11, 123–152, doi:10.1146/annurev-statistics-033121-114601.

Independent Office of Evaluation of IFAD (2023) "Geospatial tools and applications to support IOE," https://ioe.ifad.org/en/w/geospatial-tools-and-applications-to-support-ioe.

Jain, Meha (2020) "The Benefits and Pitfalls of Using Satellite Data for Causal Inference," *Review of Environmental Economics and Policy*, 14 (1), 157–169, doi:10.1093/reep/rez023.

Jain, Meha, Christopher B Barrett, Divya Solomon, and Kate Ghezzi-Kopel (2023) "Surveying the evidence on sustainable intensification strategies for smallholder agricultural systems," *Annual Review of Environment and Resources*, 48 (1), 347–369, doi:https://doi.org/10.1146/annurev-environ-112320-093911.

Jain, Meha, Amit Srivastava, Balwinder-Singh, Rajiv Joon, Andrew McDonald, Keitasha Royal, Madeline Lisaius, and David Lobell (2016) "Mapping Smallholder Wheat Yields and Sowing Dates Using Micro-Satellite Data," *Remote Sensing*, 8 (10), 860, doi:10.3390/rs8100860.

Jarvis, A., H. I. Reuter, A. Nelson, and E. Guevara (2008) "Hole-filled SRTM for the globe Version 4," http://srtm.csi.cgiar.org, CGIAR-CSI GeoPortal.

Jean, Neal, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon (2016) "Combining satellite imagery and machine learning to predict poverty," *Science*, 353 (6301), 790–794, doi:10.1126/science.aaf7894.

Jin, Zhenong, George Azzari, Marshall Burke, Stephen Aston, and David Lobell (2017) "Mapping Smallholder Yield Heterogeneity at Multiple Scales in Eastern Africa," *Remote Sensing*, 9 (9), 931, doi:10.3390/rs9090931.

Jin, Zhenong, George Azzari, Calum You, Stefania Di Tommaso, Stephen Aston, Marshall Burke, and David B. Lobell (2019) "Smallholder maize area and yield mapping at national scales with Google Earth Engine," *Remote Sensing of Environment*, 228, 115–128, doi:https://doi.org/10.1016/j.rse.2019.04.016.

Kosmowski, Frederic, Jordan Chamberlin, Hailemariam Ayalew, Tesfaye Sida, Kibrom Abay, and Peter Craufurd (2021) "How accurate are yield estimates from crop cuts? Evidence from smallholder maize farms in Ethiopia," *Food Policy*, 102, 102122, doi:10.1016/j.foodpol.2021.102122.

Kuhn, Max and Hadley Wickham (2020) *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.*, https://www.tidymodels.org.

Lobell, David B, George Azzari, Marshall Burke, Sydney Gourlay, Zhenong Jin, Talip Kilic, and Siobhan Murray (2020) "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 (1), 202–219, doi:10.1093/ajae/aaz051.

Mekonnen, Dawit K., Gashaw Abate, Seid Yimam, Rui Benfica, David J. Spielman, and Frank Place (2021) "The Impact of Ethiopia's Direct Seed Marketing Approach on Smallholders' Access to Seeds, Productivity, and Commercialization,"Technical report, IFPRI Discussion Paper 01998, doi:10.22004/AG.ECON.312925.

Mekonnen, Leulsegged Kasa, Nicholas Minot, James Warner, and Gashaw T Abate (2019) *Performance of Direct Seed Marketing pilot program in Ethiopia: Lessons for scaling-up*, 132: Intl Food Policy Res Inst.

Pelletier, Johanne, Casey Maue, Mina Karasalo, Kelsey Jack, and Julio Barrios (2023) "Remote sensing for impact evaluation of agriculture and natural resource management research: Guidelines for use in One CGIAR.,"Technical report, Standing Panel on Impact Assessment, Rome, https://cgspace.cgiar.org/server/api/core/bitstreams/70377fe2-b568-467c-8b40-b458989d074e/content.

Porteous, Obie (2022) "Research deserts and oases: Evidence from 27 thousand economics journal articles on Africa," *Oxford Bulletin of Economics and Statistics*, 84 (6), 1235–1258.

Potapov, Peter, Svetlana Turubanova, Matthew C. Hansen et al. (2022) "Global maps of cropland extent and change show accelerated cropland expansion in the twenty-first century," *Nature Food*, 3 (1), 19–28, doi:10.1038/s43016-021-00429-z.

Proctor, Jonathan, Tamma Carleton, and Sandy Sum (2023) "Parameter Recovery Using Remotely Sensed Variables,"Technical report, National Bureau of Economic Research, doi:10.3386/w30861.

Ratledge, Nathan, Gabe Cadamuro, Brandon de la Cuesta, Matthieu Stigler, and Marshall Burke (2022) "Using machine learning to assess the livelihood impact of electricity access," *Nature*, 611 (7936), 491–495, doi:10.1038/s41586-022-05322-8.

Rolf, Esther, Jonathan Proctor, Tamma Carleton, Ian Bolliger, Vaishaal Shankar, Miyabi Ishihara, Benjamin Recht, and Solomon Hsiang (2021) "A generalizable and accessible approach to machine learning with global satellite imagery," *Nature communications*, 12 (1), 4392, doi:https://doi.org/10.1038/s41467-021-24638-z.

Rouse, John Wilson, Rüdiger H Haas, John A Schell, and Donald W Deering (1974) "Monitoring vegetation systems in the Great Plains with ERTS," *NASA Spec. Publ*, 351 (1), 309.

Salazar, Lina, Ana Claudia Palacios, Michael Selvaraj, and Frank Montenegro (2021) "Using satellite images to measure crop productivity: Long-term impact assessment of a randomized technology adoption program in the Dominican Republic,"Technical report, IDB Working Paper Series, https://hdl.handle.net/10419/252357.

Savitzky, Abraham and Marcel JE Golay (1964) "Smoothing and differentiation of data by simplified least squares procedures.," *Analytical chemistry*, 36 (8), 1627–1639, Publisher: ACS Publications.

Serrat Capdevila, Aleix; Herrmann, Stefanie Maria (2018) "Mainstreaming the use of remote sensing data and applications in operational contexts.,"Technical report, World Bank Group, Washington, D.C. http://documents.worldbank.org/curated/en/154221517427945919/Mainstreaming-the-use-of-remote-sensing-data-and-applications-in-operational-co

Space, Caribou (2023) "Earth Observation for Monitoring and Evaluation,"Technical report, Global Development Assistance and European Space Agency, https://gda.esa.int/wp-content/uploads/2023/05/Caribou-Space-GDA-EO-for-ME-Public.pdf.

Tiedeman, Kate, Jordan Chamberlin, Frédéric Kosmowski, Hailemariam Ayalew, Tesfaye Sida, and Robert J. Hijmans (2022) "Field Data Collection Methods Strongly Affect Satellite-Based Crop Yield Estimation," *Remote Sensing*, 14 (9), doi:10.3390/rs14091995.

Van Dijk, Michiel, Tomas Morley, Marloes van Loon, Pytrik Reidsma, Kindie Tesfaye, and Martin K van Ittersum (2020) "Reducing the maize yield gap in Ethiopia: Decomposition and policy simulation," *Agricultural Systems*, 183, 102828, doi:https://doi.org/10.1016/j.agsy.2020.102828.

Wang, Sherrie, Stefania Di Tommaso, Jillian M. Deines, and David B. Lobell (2020a) "Mapping twenty years of corn and soybean across the US Midwest using the Landsat archive," *Scientific Data*, 7 (1), 307, doi:10.1038/s41597-020-00646-4.

Wang, Siruo, Tyler H. McCormick, and Jeffrey T. Leek (2020b) "Methods for correcting inference based on outcomes predicted by machine learning," *Proceedings of the National Academy of Sciences*, 117 (48), 30266–30275, doi:10.1073/pnas.2001238117.

Westengen, Ola T, Sarah Paule Dalle, and Teshome Hunduma Mulesa (2023) "Navigating toward resilient and inclusive seed systems," *Proceedings of the national academy of sciences*, 120 (14), e2218777120, doi:https://doi.org/10.1073/pnas.2218777120.

World Bank (2022) "Ethiopia Rural Income Diagnostics Study: Leveraging the Transformation in the Agri-Food System and Global Trade to Expand Rural Incomes,"Technical report, World Bank, https://openknowledge.worldbank.org/entities/publication/e9a029b2-9c57-5be4-9abd-996f55e7f513.

Xiao, Xiangming, Stephen Boles, Steve Frolking, William Salas, B Moore Iii, C Li, L He, and R Zhao (2002) "Observation of flooding and rice transplanting of paddy rice fields at the site to landscape scales in China using VEGETATION sensor data," *International Journal of Remote Sensing*, 23 (15), 3009–3022, doi:https://doi.org/10.1080/01431160110107734.

Yeh, Christopher, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke (2020) "Using publicly available satellite imagery and deep learning to understand economic well-being in Africa," *Nature Communications*, 11 (1), doi:10.1038/s41467-020-16185-w.

Zrnic, Tijana and Emmanuel J. Candès (2023) "Cross-Prediction-Powered Inference," doi:10.48550/ARXIV.2309.16598.

A Supplementary figures and tables

We compared our estimated maize area and estimated yield at national scale to the values provided by FAOSTAT (Figure SM.1). It should be noted, however, that the maize/non-maize prediction model does consider intercropped fields as non-maize given that the ground-truth TAMASA dataset only includes pure maize (monocrop maize) fields. On the other hand, the FAOSTAT statistics likely include maize from intercropped fields, possibly explaining part of the discrepancy observed. We expect that the model also somewhat underestimates pure maize area as well.

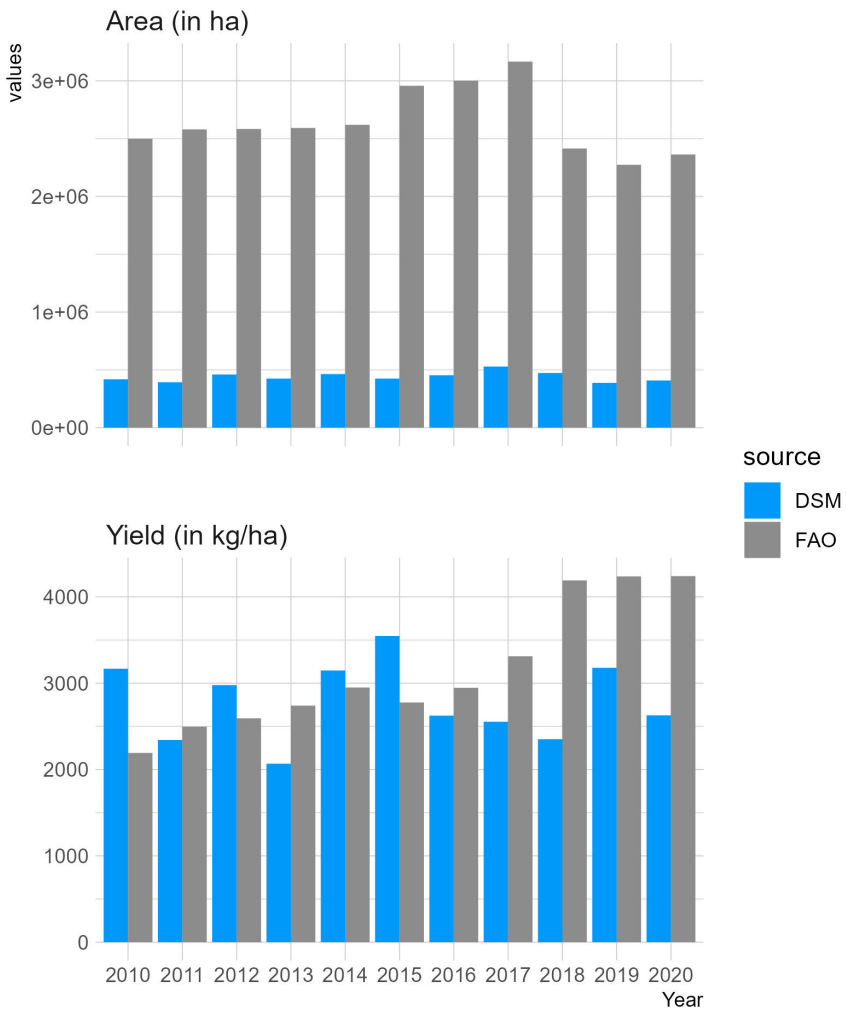

Figure SM.1: Comparison between the annual maize area and yield quantified with this study and the national figure from FAOSTAT (Food and Agriculture Organization food and crop statistics database)

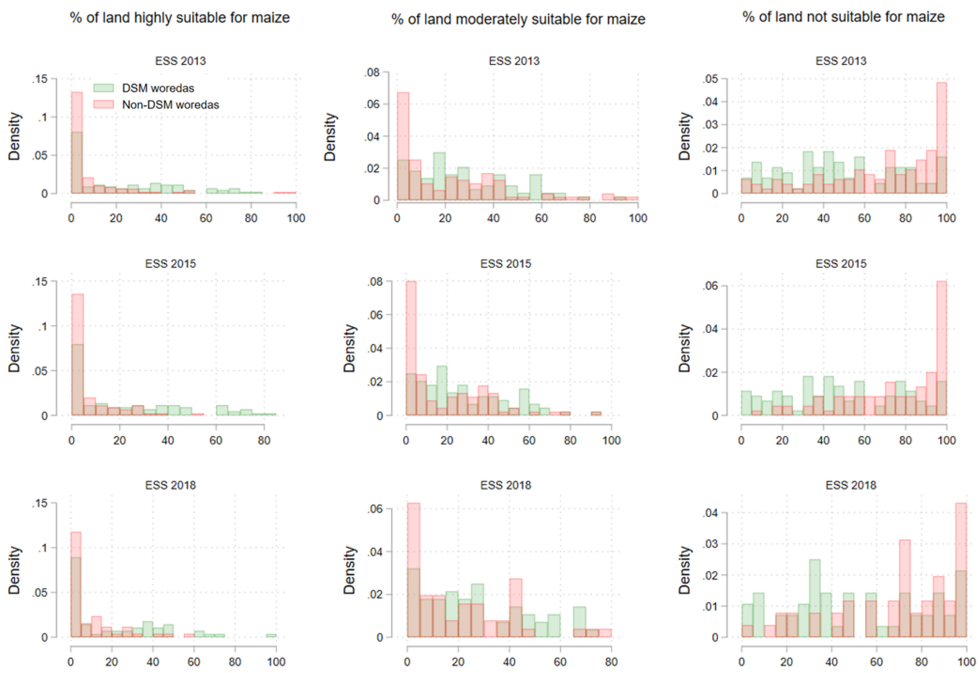

Figure SM.2: Share of land at district-level in different maize suitability categories for DSM and non-DSM districts in 2013, 2015 and 2018.

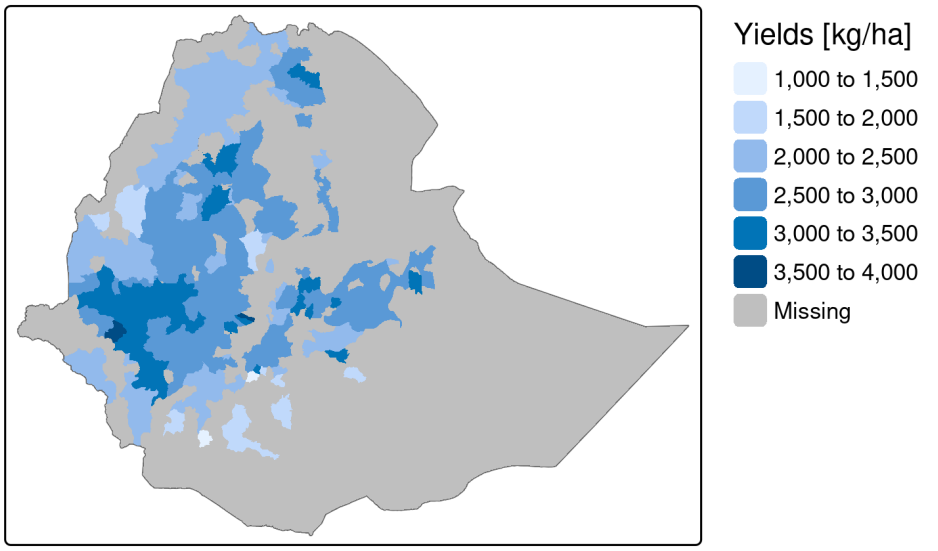

Figure SM.3: District-level statistics of predicted average maize yield in Ethiopia averaged over 2010 to 2020.

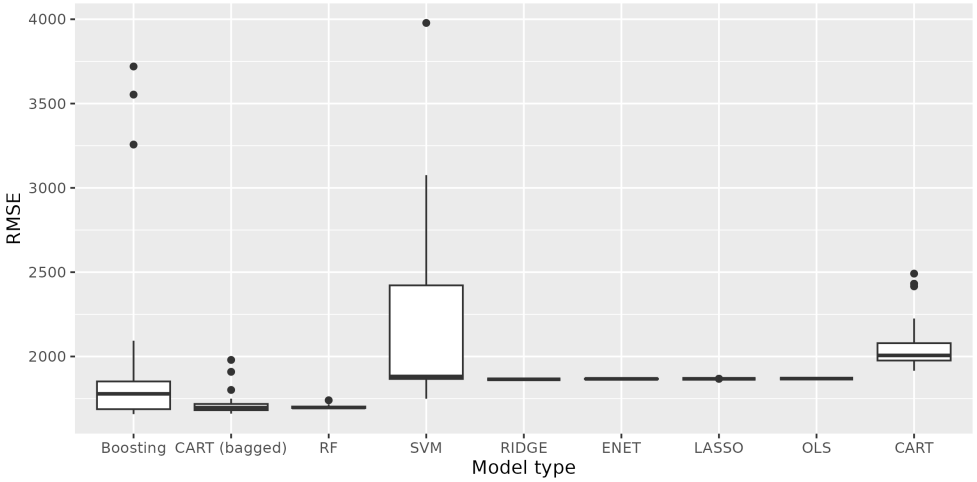

Figure SM.4: RMSE of the different models evaluated

Table SM.1: ML model specifications.

Model Name	R Package	Hyperparameter Name	Hyperparameter Value
CART	rpart	cost_complexity	tune
		tree_depth	
		min_n	tune
CART_bagged	rpart	cost_complexity	tune
		tree_depth	
		min_n	tune
		class_cost	
RF	ranger	mtry	tune
		trees	1000
		min_n	tune
boosting	xgboost	mtry	
		trees	tune
		min_n	tune
		tree_depth	tune
		learn_rate	tune
		loss_reduction	tune
		sample_size	tune
		stop_iter	
LASSO	glmnet	penalty	tune
		mixture	1
RIDGE	glmnet	penalty	tune
		mixture	0
ENET	glmnet	penalty	tune
		mixture	tune
OLS	lm	penalty	
		mixture	

Column *Hyperparameter Name* indicates the hyperparameters of each model, while column *Hyperparameter Value* indicates whether the parameter was 1) left to its default value (no mention), 2) tuned, 3) set to a given number.

B Heterogeneous Treatment Effects

In addition to aggregating the Callaway and Sant'Anna (2021) group-time average treatment effects into an "overall" estimate of the effect of DSM adoption on predicted yields, we also examined heterogeneity in treatment effects across time and treatment adoption groups. For our heterogeneity analysis, we selected the boosting ML algorithm which performed well according to both the ML and causal criteria, with the lowest RMSE and the lowest value of CS-DiD(e). This is the model used to generate the results in section 6.2.

To estimate how the effect of DSM adoption on yields might change over time, we take a weighted average of the group-time average treatment effects at different lengths of exposure to DSM. An event study with unbalanced panel shows some evidence of parallel trends before DSM, with yield increases after DSM was adopted (Figure SM.5). We can notice large confidence intervals around these point estimates.

To estimate whether the effect of DSM differs depending on when DSM was adopted (i.e. heterogeneity by treatment adoption group), we take a weighted average of the group-time average treatment effects across all years, by treatment adoption group. Figure SM.6 shows that districts adopting DSM in 2015 and 2016 experienced yield decreases on average during their involvement in the program, while later-adopting groups experienced yield increases on average. Further investigation into whether the execution of the DSM program varied across time is needed to understand why groups that adopted DSM in different years experienced such starkly contrasting impacts on yields.

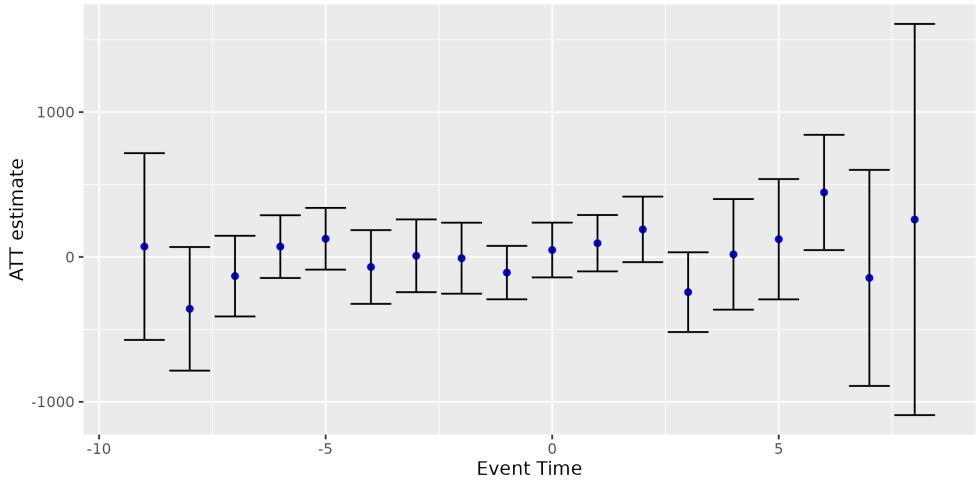

Figure SM.5: Event study with unbalanced panel which shows the average treatment effects by the length of exposure (x-axis) to the DSM program. The length of exposure equal to 0 provides the average effect of participating in the treatment across groups in the time period when they first participate in the treatment (or instantaneous treatment effect). Length of exposure equal to -1 corresponds to the time period before groups first participate in the treatment, and length of exposure equal to 1 corresponds to the first time period after initial exposure to the treatment.

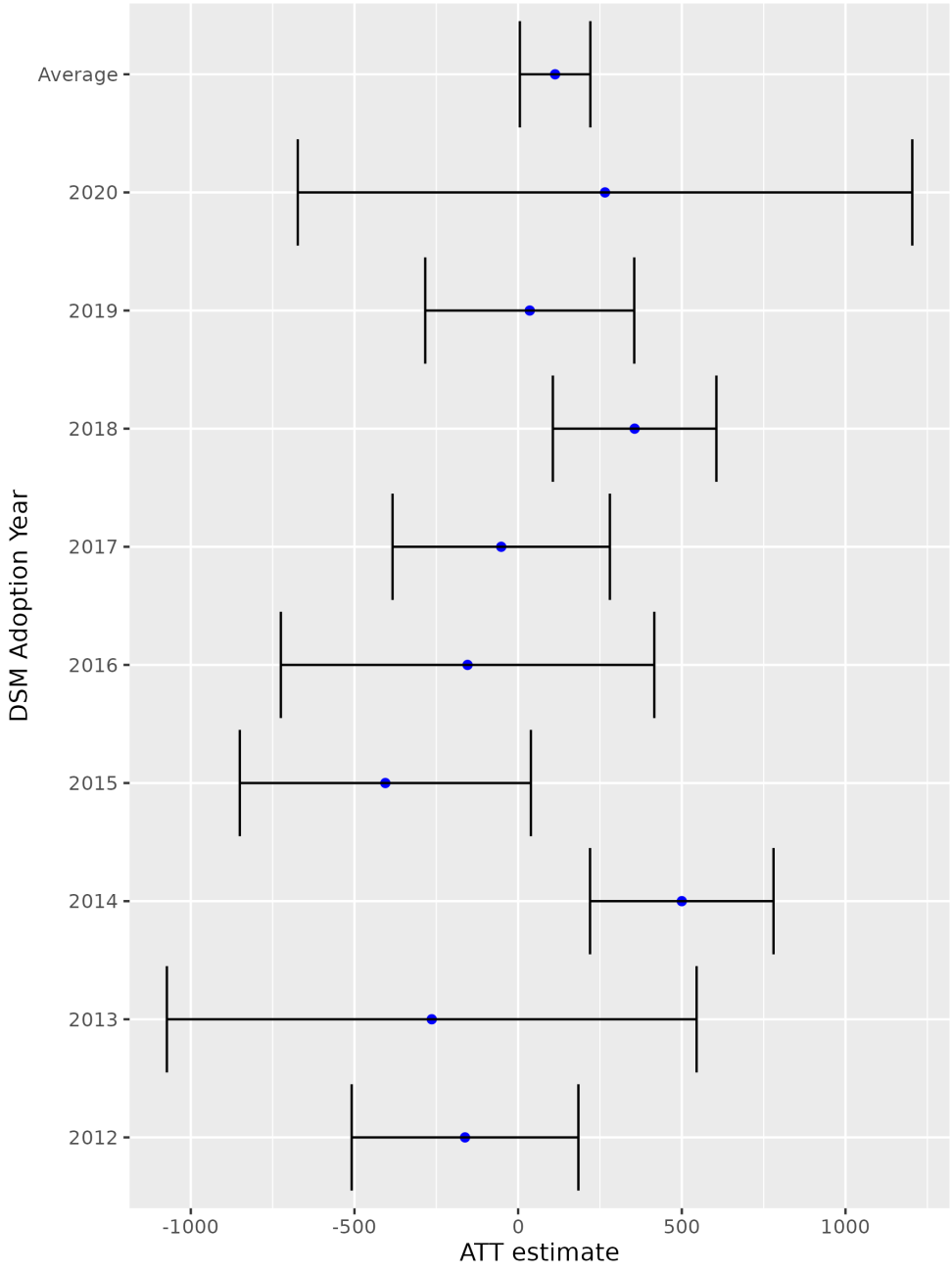

Figure SM.6: Group-specific effect of DSM on yields: group-time average treatment effects are aggregated across all years within a group.

C Supplementary information on field and satellite data processing

We provide additional information on the data used for this study, the data processing, as well as on the remote sensing analysis for mapping maize.

C.1 Administrative data about the DSM roll-out

The administrative data on the DSM program roll-out included the region, zone, and district (woreda; administrative level 3), and the amount seed supplied and sold for maize by the DSM program as a spreadsheet. This data was merged to the spatial district-level administrative data file, which included 691 districts (woredas) and was used as the standard for this analysis because it most closely matched the DSM administrative data. Since the names of administrative units are not standardized amongst data sources, a substantial effort was dedicated to ensure the match between the names and spatial location. We had to assess any name changes and re-districting over time so that it could be properly linked to the original name in our spatial administrative data file, and verify that the spatial coverage on the map match the administrative level identified by the DSM program. To do this over time, especially for the most recent years, we compared the spatial area (polygons) of recent 2021 administrative data file (modified was by OCHA, obtained from CSA (Central Statistics Agency) and the Regional Bureau of Finance and Economic Development (BoFED)) with our standard administrative boundaries in QGIS.

The DSM program roll-out grew over time from 2 district in 2011 to 320 districts in 2020. Not all of these districts received maize however. The number of participating districts for maize increased over time to cover a larger portion of the country (Figure SM.7).

To link our evaluation to the study produced by IFPRI with panel household surveys, we obtained the village centroids of their survey sample, in which they had interviewed 20 maize cultivating households in each village. For matching the survey and the area cultivated by these surveyed households, we used the ward administrative level (administrative level 4, equivalent to ward; 15,670 records) which was obtained from the Ethiopia Geoportal (on Africa Geoportal) and covered most of the maize producing areas in the country, but not the whole country area. The village centroids

Figure SM.7: DSM roll-out from 2011 to 2020 with DSM participant districts shown in green and the non-participant district shown in light brown color. (Data source: ATA for the DSM status and from Africa Open Data)

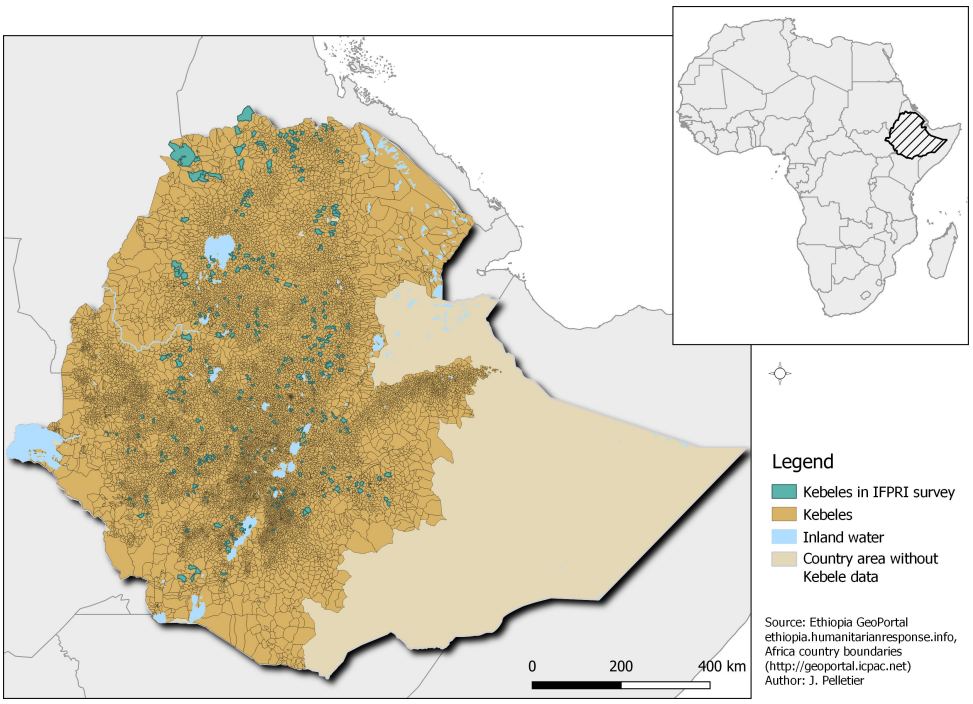

Figure SM.8: Map of the ward administrative level (Fourth administrative level (brown color); equivalent to ward), along with the wards sampled (turquoise color) during the IFPRI study that evaluated DSM impact, using self-reported yields through household survey method.

were used to select the wards where the IFPRI study took place (Figure SM.8). For each ward, the belonging to a district was done by location in QGIS. Since the program was implemented at the district level, the district belonging determined the DSM participation status for the ward.

C.2 Field data management

We undertook several data management steps to reduce potential error in our reference data. We described the steps we took for the different data sets we used in the following sections.

C.2.1 ESS 3 and 4 data sets

The Ethiopia Socioeconomic Survey provides a large data set with crop type information. The ESS 3 contained 22,792 complete point geocoordinates for one corner of the farm plot and crop type information associated with that plot. Maize constituted about 14 percent of the total sample of ESS3, with 2,639 observations for maize as the primary crop, with 2,000 of these planted as monocrop and 639 as intercrop with 2 or more crops, as well as 556 observations of maize as a secondary crop (so intercropped). The ESS 4 had 12,207 complete point geocoordinates for one corner of the farm plot and crop type data. Maize made up about 16 percent of the total sample, with 1,578 observations for maize as primary crop. Out of these 1,267 observations were cultivated as monocrop and 311 were intercropped. Another 323 observations were intercropped, with maize planted as a secondary crop. In order to assess and improve field data quality, we first tried to eliminate erroneous points for the maize. We extracted the value to points of the Copernicus Global land cover maps of 2015 for the ESS 3 and 2018 for ESS 4 to identify any points outside cropland. We found that 897 and 586 "maize" points fell outside cropland (in other land covers including shrubs, herbaceous vegetation, open forest, and urban) for the ESS 3 and ESS 4, respectively. Since the land cover maps also contain error and that it was important for our purpose to maintain as many maize points as possible, we exported these points to keyhole markup language (kml) format and loaded them in Google Earth Pro (GE). We then used historical imagery available in GE to validate the land cover at each of these points. We used the time slider to identify the relevant dates before and after the year of data collection for the points. The timing and quality of the very high imagery available in GE for this verification did not allow to identify the crop type most of the time; most of the imagery were from the dry season or when the crops were not well-developed. We were only able to identify if the point fell into cropland or not, the closest date of the image available, and a qualitative level of certainty for this identification (4 classes 25, 50, 75, and 100 percent certainty). Any

points that did not have an image for the season of interest was ranked lower in terms of certainty. Some points were unequivocally erroneous, which may arise when/if the enumerator did not visit the field as required; they fell on roads or in villages. The points that fell outside cropland or for which the land cover was too uncertain were removed for this analysis. It was also clear from this verification process that relying on one field corner point necessarily involved substantial noise, since the field corner is always adjacent to other fields or other land uses.

Similarly, for the points representing other crop types (non-maize), we removed the points that fell outside cropland using the value extracted from the year relevant land cover maps, but without further verification. The removed "other crops" points included 9,116 observations for ESS 3 and 4,662 for ESS 4, so a substantial portion of the sample failed a basic quality verification. Even if the size of the sample was large, it became clear from this verification process that the data sets suffered some data quality issue for the purpose of crop mapping.

The first attempts at mapping maize/non-maize with the ESS 3 and 4 household data was unsuccessful. We tried using different data strategy, including: 1) combining all verified (cropland in GE with certainty > 75 percent) maize monocrop and intercrop into one class versus 'Other crops'; 2) same as 1, but removing points that are less than 10 meters apart; 3) verified maize with 100 percent certainty for monocrop and intercrop in one class versus 'Other crops'; 4) same as 3, but removing points less than 10 meters apart; 5) verified maize monocrop only versus 'Other crops' (excluding intercropped maize); 6) three classes with maize monocrop, maize intercropped, and non-maize. None of these tests provided a maize class producer's and user's accuracy greater than 50 percent. It was thus clear from these tests that other field data sets were required to map maize with enough accuracy for this study.

C.2.2 ESS 3 and 4 crop cut data sets

We obtained the ESS 3 and 4 crop cut data sets for pure maize, with the double purpose of improving maize mapping accuracy and for predicting yields. Since the crop cut survey required presence in the field for data collection, we expected less error in the georeferenced coordinates. Yet, the protocol also involved the georeferencing of only one corner per farm plot. We still verified the location with random checks of 30 points for each year

in GE. The crop cut information was collected for selected fields, from a 4m by 4m crop cut or 16m2, and included fresh and dry weight, excluding permanent, tree, and root crops. Upon inspection of the yield distribution of these crop cut data sets, we decided to include only the crop cut data from ESS 4 (n=550) for yield prediction.

C.2.3 TAMASA data sets

The TAMASA data sets consist of only for pure maize and they were used for both maize crop mapping and maize yield predictions. They consisted of three datasets for 2015 covering different areas of the country with a total of 717 georeferenced maize plots, 553 for 2017, 469 for 2018, and 230 for 2019.

The plot georeferencing was again done with point data, instead of the complete plot boundary. The experimental design involved three 4m by 4m quadrats placed on a diagonal to capture within-field heterogeneity, except for one dataset for the year 2015 (69 observations) which included only two quadrats. The first quadrat was installed at the center of the field, and the two other quadrats were placed on opposite side, at the same distance to the middle quadrat. Some datasets had the GPS coordinates taken from the center of the field (middle quadrat) only, while other datasets had GPS coordinates for each quadrat. The plot locations for each dataset were exported to .kml and randomly check to verify the quality of the geolocation for the period matching the field data collection in Google Earth Pro. These maize plot locations were used as reference data for creating annual maize maps from 2010 to 2020 over Ethiopia.

The crop cuts data was collected during the Meher season (main cropping season) for a large subset of the georeferenced maize plots, including 659 plots for 2015, 542 for 2017, 464 for 2018, and 230 for 2019. Two 2015 data sets provided the yield estimate directly, while other provided the variables for yield calculation (the weight of grain relative the weight of cobs, as well as the moisture content ratio calculated on a dried subsample). We calculated the yield for each quadrat and them computed the mean yield of the quadrats for field in kilograms per hectare (kg/ha). The yield information for these crop cuts data sets was used as reference data for predicting maize yield, with remotely sensed predictors.

C.3 Maize crop mapping

For mapping maize cropland, we used Landsat surface reflectance Tier 1 collections, including data from Landsat 5 TM, Landsat 7 ETM+, Landsat 8 OLI. We limited the use of Landsat 7 ETM+ collection up until Landsat 8 imagery became available (March 18th, 2013), to reduce potential artefacts caused by the Scan Line Corrector (SLC) failure missing data. The pre-processing of each collection was done separately for the images covering Ethiopia. It included the application of scaling factors for radiometric calibration, cloud and cloud shade masking, and the renaming of the bands from the Landsat 5 and 7 collection to match the Landsat 8 collection. The collections were merged together and included 7,349 images.

We compared two methods to create annual seasonal mosaic: a median composite, which consist in calculating the median of all the cloud/shade free pixels at a location and a greenest pixel composite, which consist in selecting the pixel with the highest NDVI value. After evaluation, we chose the median mosaic because the greenest pixel composite resulted in shade artefacts in mountainous agricultural areas which characterize are large part of Ethiopia.

The annual maize/non-maize maps created annually from 2010 to 2020, were based on the Landsat median mosaic. For each pre-processed Landsat image, we calculated four indices, the NDVI (Rouse et al., 1974), the GCVI (Huete et al., 1997), the GCVI (Gitelson et al., 2003) and the LSWI (Xiao et al., 2002). We provide the equation for these indices below:

$$NDVI = \frac{NIR - RED}{NIR + RED} \tag{6}$$

$$EVI = 2.5 \times \frac{NIR - RED}{NIR + 6 \times RED - 7.5 \times BLUE + 1} \tag{7}$$

$$GCVI = \frac{NIR}{GREEN} - 1 \tag{8}$$

$$LSWI = \frac{NIR - SWIR1}{NIR + SWIR1} \tag{9}$$

In addition to Landsat imagery, we included different time-invariant covariates that are relevant to maize suitability. These covariates included the STRM 90m elevation dataset (Jarvis et al., 2008) and elevation-derived slope and aspect. It incorporated climate related variables, including the

mean, maximum and minimum *Meher* seasonal rainfall for the period 2000 to 2020 from the CHIRPS pentad dataset (Funk et al., 2015), the *Meher* seasonal mean and maximum day time land surface temperature as well as the minimum nighttime temperature from the MODIS/Terra Land Surface Temperature/Emissivity 8-Day (MOD11A2) Version 6.1. We added soil characteristics important for maize cultivation (Fang and Su, 2019), including depth to bedrock, as well as soil pH, percent sand content and soil organic carbon at two depths, 0 to 20 cm and 20 to 50 cm (Hengl et al., 2017).

C.3.1 Maize map model selection and accuracy assessment

The spatio-temporal maize/non-maize crop classification was supported by field data collected over different years for the main cropping season. For each field data set, we extracted the values of the year-specific as well as time-invariant covariates. Figure SM.9 shows the maize/non-maize classification for year 2018, with close up comparison with higher resolution imagery from Google Earth Pro.

For the classification model, we first partitioned the field data using an 80/20 ratio for training and validation. The validation data was used uniquely for the purpose of evaluating the model's performance. To address possible spatial correlation that may inflate the classification accuracy, we enforced a minimum distance by removing training points that were less than 100 meters from validation data points.

We compared different data strategies to map maize annually, by testing different combinations of input datasets and comparing the maize classification accuracy (Table SM.2). While, the "Other crops" (or non-maize) class and the overall accuracy were consistently high, the most importance metrics for our purpose were the accuracy of the Maize class, including the maize producer's accuracy (1-Omission Error) and the maize user's accuracy (1-Commission Error).

We were interested to strike a good balance with high value from both the producer's accuracy and the user's accuracy for maize, yet, it was more important to minimize commission error (non-maize classified as maize) for the purpose of this analysis. The best model was obtained by combining the TAMASA datasets (for maize only) and the ESS 3 and 4 agricultural household surveys for the non-maize crop class. We can observe from the district-level statistics, that the pure maize area remains more or less stable over time (Figure SM.10), with few districts showing high areas of pure maize.

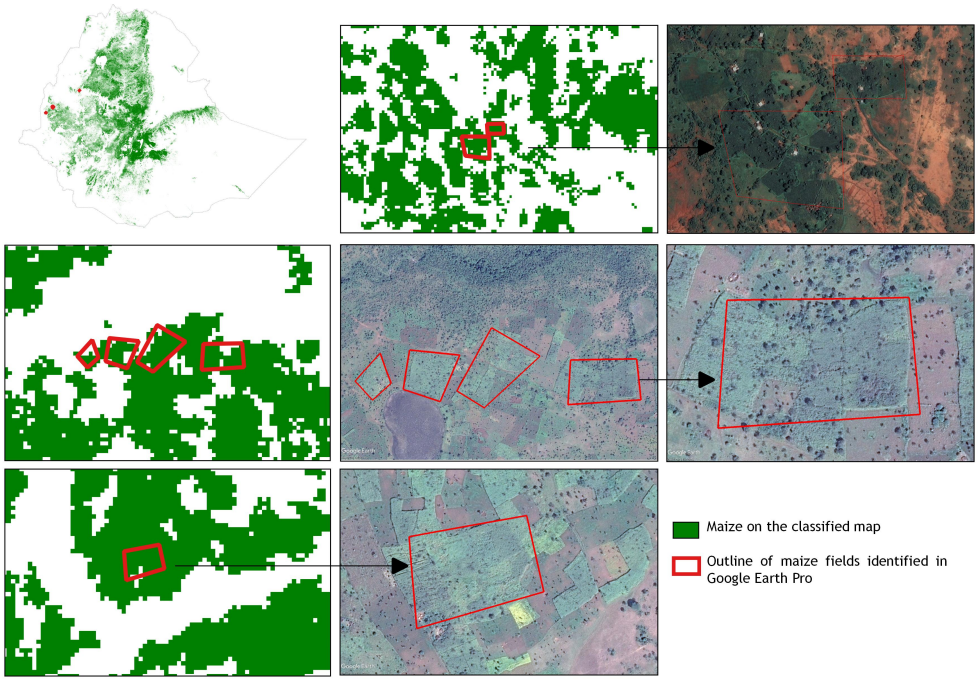

Figure SM.9: Maize and non-maize map of Ethiopia from 2018, with close up areas from Google Earth Pro for the same year.

Table SM.2: Accuracy Assessment of different field data strategy for maize mapping in Ethiopia. The values are expressed in the range of 0 to 1, but can also be expressed as a percentage. The final model selected is shown in bold font. The four last rows are the same model when splitting the validation data per year including 2015, 2017, 2018, and 2019. Note that the year 2017 and 2019 do not have non-maize data, only maize from the TAMASA datasets.

			User accuracy		**Producer accuracy**	
Strategy	**Validation sample size**	**Overall accuracy**	**Non-maize**	**Maize**	**Non-maize**	**Maize**
ESS 3 and 4 Maize crop cut data + ESS 3 and 4 survey non-maize data	4221	0.941	0.941	1	1	0.02
ESS 3 and 4 survey for maize and non-maize data	4543	0.887	0.893	0.507	0.992	0.068
ESS 3 and 4 survey for maize and non-maize data + TAMASA	4656	0.874	0.884	0.694	0.981	0.252
ESS 3 and 4 Maize crop cut data + ESS 3 and 4 survey non-maize data + TAMASA	4385	0.925	0.927	0.845	0.996	0.229
ESS 3 and 4 Maize crop cut data + ESS 3 and 4 survey for maize and non-maize data	4830	0.84	0.846	0.511	0.989	0.062
ESS 3 and 4 Maize crop cut data + ESS 3 and 4 survey for maize and non-maize data + TAMASA	5103	0.838	0.978	0.354	0.839	0.828
ESS 3 and 4 survey non-maize data + TAMASA	**4273**	**0.978**	**0.978**	**0.972**	**0.998**	**0.782**
ESS 3 and 4 survey non-maize data + TAMASA, validation for 2015	2674	0.963	0.965	0.898	0.996	0.465
ESS 3 and 4 survey non-maize data + TAMASA, validation for 2017 (maize only)	123	0.959	NA	1	NA	0.959
ESS 3 and 4 survey non-maize data + TAMASA, validation for 2018	1567	0.99	0.992	0.961	0.998	0.86
ESS 3 and 4 survey non-maize data + TAMASA, validation for 2019 (maize only)	49	0.98	NA	1	NA	0.98

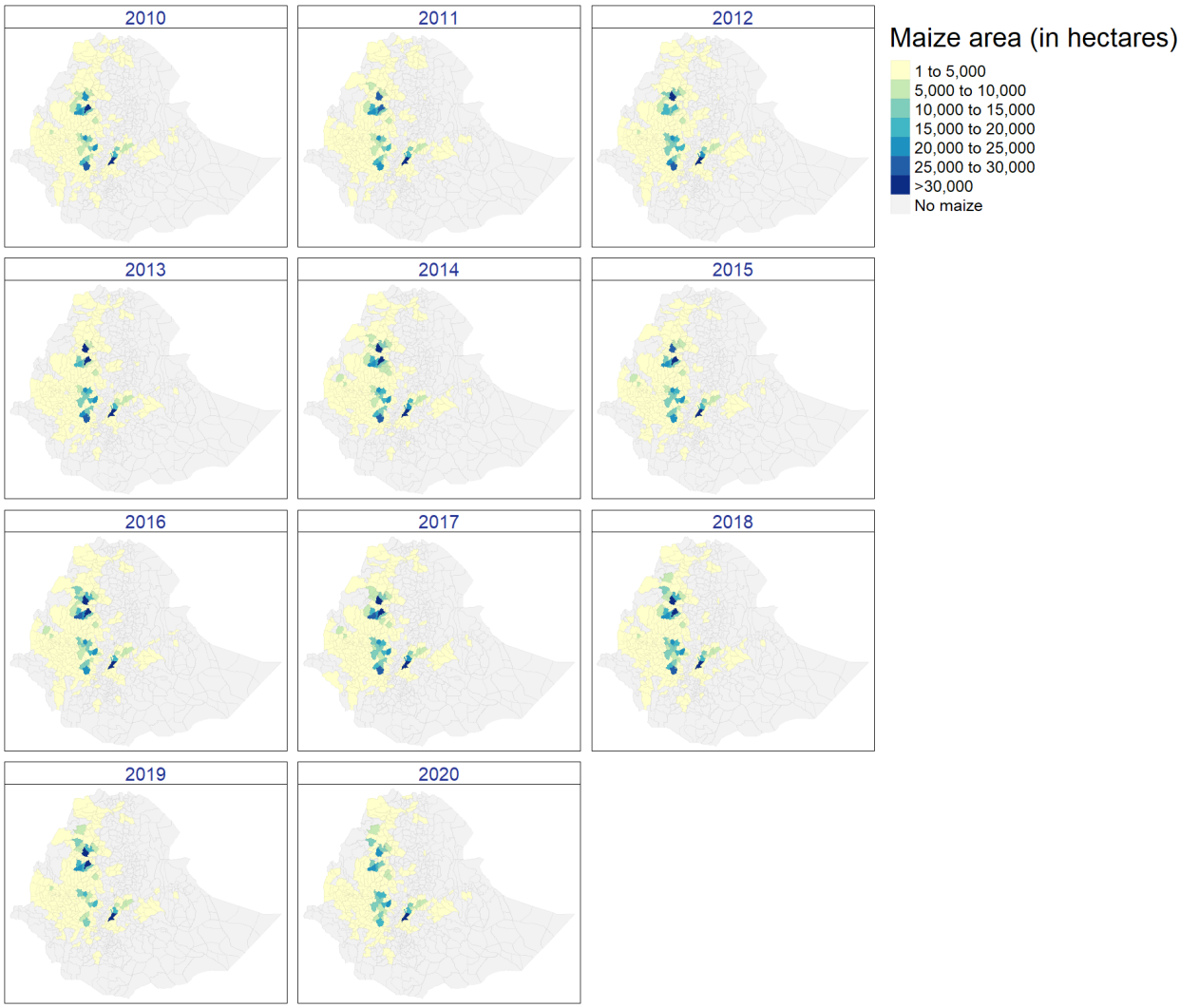

Figure SM.10: District-level statistics about pure maize area (in hectares) in Ethiopia from 2010 to 2020.

We also performed a model validation per year, by doing the data partitioning between training and validation for each year separately, and then merging the training data together for building the model. We obtained four validation samples, one for each year, that we used to check for the model consistency between years. The results were consistent for 2017, 2018, and 2019, but the maize producer's accuracy as lower for the year 2015, which also had the largest validation sample size. This may thus underestimate the maize area for that year. We could not validate the other years because of the lack of field data.

We believe that the confusion in the maize class may stem from at least two main potential sources of error. First, the georeferenced field corner point collected during the ESS 3 and 4 is adjacent to other fields or land uses, with some of them containing maize crop and providing a mixed signal at the 30-m Landsat pixel resolution. The ESS 3 and 4 household survey and crop cut data sets have deficiencies in the geo-positioning of field (with field corner point) which is not ideal for the purpose of crop type mapping (Azzari et al., 2021). Second, the maize training data from TAMASA was for pure maize alone, as a monocrop, while we know that maize is often intercropped in Ethiopia, that is, cultivated jointly and simultaneously with two or more crops interspersed on the same field. It is thus possible that part of the non-maize on the ground that is classified as maize and vice versa by our model come from the combination of these sources of error.

Another limitation is related the difficulty to identify the crop type for smallholder farms on very high-resolution imagery. We do not control the timing of the image that are publicly available (in Google Earth) , most of which are for the dry season. We have thus to rely on already collected field data. This means that we cannot generate our own set of random points per class, that we could use to verify independently our maize/non-maize map results. Even if we counted on field data from four different years to support our spatio-temporal predictions, we were not able to verify the quality of our maize/non-maize classification for the years for which we do not have field data.

D Mathematical derivations

To understand Equation (2), it is best to start from the OLS representation of the DiD coefficient, remembering that the DiD is equivalent to a two-way

fixed effects estimator when there is a single common intervention time:

$$\widehat{\mathrm{DiD}}(y) = \frac{\mathrm{Cov}(\tilde{y}, \tilde{D})}{\mathrm{Var}(\tilde{D})} = \frac{\mathrm{Cov}(y, \tilde{D})}{\mathrm{Var}(\tilde{D})}$$

where the tilde $\tilde{x}$ corresponds to the two-way within transformation, $\tilde{x}_{it} \equiv x_{it} + \bar{\bar{x}}_{i\cdot} + \bar{\bar{x}}_{\cdot t} + \bar{\bar{\bar{x}}}_{\cdot\cdot}$, and the second equality comes from the properties of the Frish-Waugh theorem.

Looking now at:

$$\begin{aligned}
\widehat{\mathrm{DiD}}(\hat{y}) &= \frac{\mathrm{Cov}(\hat{y}, \tilde{D})}{\mathrm{Var}(\tilde{D})} \\
&= \frac{\mathrm{Cov}(\gamma + \lambda y_{it} + \delta x_{it} + u_{it}, \tilde{D})}{\mathrm{Var}(\tilde{D})} \\
&= \lambda \frac{\mathrm{Cov}(y_{it}, \tilde{D})}{\mathrm{Var}(\tilde{D})} + \delta \frac{\mathrm{Cov}(x_{it}, \tilde{D})}{\mathrm{Var}(\tilde{D})} + \frac{\mathrm{Cov}(u_{it}, \tilde{D})}{\mathrm{Var}(\tilde{D})} \\
&= \lambda \widehat{\mathrm{DiD}}(y) + \delta \widehat{\mathrm{DiD}}(x) + \widehat{\mathrm{DiD}}(u)
\end{aligned}$$

Second, looking at the observed prediction error $e \equiv \hat{y} - y$, one has:

$$e \equiv \hat{y} - y = \gamma + (\lambda - 1) y_{it} + \delta x_{it} + u_{it}$$

Then

$$\begin{aligned}
\widehat{\mathrm{DiD}}(e) &= \frac{\mathrm{Cov}(e, \tilde{D})}{\mathrm{Var}(\tilde{D})} \\
&= \frac{\mathrm{Cov}(\gamma + (\lambda - 1) y_{it} + \delta x_{it} + u_{it}, \tilde{D})}{\mathrm{Var}(\tilde{D})} \\
&= \lambda \frac{\mathrm{Cov}(y_{it}, \tilde{D})}{\mathrm{Var}(\tilde{D})} - \frac{\mathrm{Cov}(y_{it}, \tilde{D})}{\mathrm{Var}(\tilde{D})} + \delta \frac{\mathrm{Cov}(x_{it}, \tilde{D})}{\mathrm{Var}(\tilde{D})} + \frac{\mathrm{Cov}(u_{it}, \tilde{D})}{\mathrm{Var}(\tilde{D})} \\
&= \lambda \widehat{\mathrm{DiD}}(y) - \widehat{\mathrm{DiD}}(y) + \delta \widehat{\mathrm{DiD}}(x) + \widehat{\mathrm{DiD}}(u) \\
&= \widehat{\mathrm{DiD}}(\hat{y}) - \widehat{\mathrm{DiD}}(y)
\end{aligned}$$