Title: Mapping Field-level Maize Yields in Ethiopian Smallholder Systems Using Sentinel-2 Imagery

Abstract:
Remote sensing offers a low-cost method for estimating yields at large spatio-temporal scales. However, the use of high spatial and temporal resolution remote sensors to map field-level yields in heterogeneous smallholder systems in the developing world is not well understood. Here we examined the ability of Sentinel-2 satellite imagery to map field-level maize yields across smallholder farms in two regions in Oromia district, Ethiopia. We specifically evaluated how effectively different indices, MTCI, GCVI, and NDVI, and different models, linear regression and random forest regression, can be used to map field-level yields. We also examined how generalizable our models were if trained in one region and applied to another region, where no data were used for model calibration. We found that linear regression models that use MTCI led to the highest yield prediction accuracies (R^2 ranging from 0.24 to 0.47), particularly when using only localized data for training the model. These models were not very generalizable, especially when models were applied to regions that still had significant haze remaining in the imagery. Our results highlight the ability of Sentinel-2 imagery to map field-level yields in smallholder systems, though accuracies are limited in regions with high cloud cover and haze.

1. Introduction

Efforts to meet global food demand in the coming decades will be challenged by population growth and climate change (Godfray et al. 2010, Tilman et al. 2011). One way to meet this growing demand is to increase agricultural production. Yet, increasing agricultural production through agricultural extensification is associated with various environmental costs, such as greenhouse gas emissions and biodiversity loss (Godfray et al. 2010). Instead, closing yield gaps, or narrowing the gap between current agricultural yields and potential agricultural yields on existing land, could be a way to meet future food demand more sustainably (Lobell et al. 2009). This is especially important in regions such as sub-Saharan Africa (SSA), where yield gaps are large (Mueller et al 2012), climatic impacts are severe (IPCC, 2022), and population growth is large (FAO 2022). In particular, Ethiopia, which is the second most populous country in SSA, is one of the countries in SSA that is most food insecure and vulnerable to climate change (Mohamed 2017, Di Falco et al. 2011). Closing yield gaps for maize will be especially important in this region, given that maize provides nearly 20% of the nation's calories (Abate et al. 2015) and is one crop that is projected to be the most negatively impacted by climate change (FAO 2022).

In order to identify the causes of and potential solutions to close yield gaps, we must be able to reliably estimate yields across large spatial and temporal scales. Yet, to date this has been challenging through on-the-ground data collection efforts. This is because such on-the-ground surveys are expensive, difficult to conduct at larger spatial scales, and tend to rely on often-inaccurate, self-reported data (Carletto et al. 2013, Paliwal and Jain, 2020). A potential low-cost way to produce agricultural statistics at scale is to use remote sensing. However, mapping field-level yields in smallholder systems such as Ethiopia can be challenging given small field sizes (<

1 ha) and a high degree of between-field variability due to heterogeneity of management practices and environmental conditions. The launch of high spatial and temporal resolution satellites, such as the public Sentinel constellations and other private satellites, such as PlanetScope, have helped overcome such challenges (Azzari et al. 2017). For example, several recent studies have demonstrated the potential of optical Sentinel-2 data to map field-level yields in heterogeneous smallholder systems (Jin et al. 2019, Sweeney et al. 2015, Hunt et al. 2019). Sentinel-2 imagery also has frequent revisit times (5-day), which has been linked with a higher degree of accuracy when mapping yields (Jain et al. 2016).

Various vegetation indices have been used to map crop yields, and efficacy may depend on crop type and local conditions (Zhang et al. 2021). While the Normalized Difference Vegetation Index (NDVI) has been used extensively to map yields (Zhang et al. 2021), vegetation indices that use the green rather than the red band to optimize for chlorophyll sensitivity may be more reliable for crop yield estimation (Clevers & Gitelson 2013, Houborg & McCabe 2018). Specifically, previous studies have shown that the Green Chlorophyll Vegetation Index (GCVI), which is more sensitive to moderate to high levels of canopy chlorophyll, can outperform more traditional vegetation indices such as NDVI (Burke & Lobell 2017, Nguy-Robertson et al. 2014, Zhang et al. 2021). In addition, the Medium Resolution Imaging Spectrophotometer (MERIS) Terrestrial Chlorophyll Index (MTCI), which uses red-edge reflectance, is sensitive to canopy chlorophyll and nitrogen content, and has also been shown to outperform NDVI and GCVI in previous comparisons (Jin et al. 2017, Nguy-Robertson et al. 2014, Clevers & Gitelson 2013). However, MTCI indices generated from Sentinel-2 imagery have a coarser spatial resolution than NDVI and GCVI, owing to the reliance on the red edge (RE) band (band 5), which is provided at 20 m instead of 10 m spatial resolution. Thus, the potential of MTCI to map crop yields in small fields should be further investigated.

Previous studies have used ground-based yield measurements, such as crop cuts, to train linear regression models that translate vegetation indices into yield estimates (Lobell et al. 2015). Linear regression models have been favored for their simplicity and ease of implementation, especially when applied in cloud computing platforms such as Google Earth Engine (GEE), and because of the observed linear relationships between vegetation indices and yield (Lobell et al. 2015, Jain et al. 2016). However, it is possible that machine learning models, such as random forest, may outperform such linear regression models as they can account for interactions among explanatory variables (Jain et al. 2017). They are also considered to be resistant to overfitting (Farmonov et al. 2023).

Finally, there is concern that models that require ground calibration are limited in their scalability (Sibley et al. 2014). This is because the relationship between vegetation indices and yields varies depending on region, crop, and management practices (Lobell et al. 2015). Thus, it is suggested that these models should be recalibrated using new ground data before they are applied to other geographic areas (Jin et al. 2017). It is unclear to what extent such models can be used to estimate spatial patterns in yield outside the region in which they are calibrated. Approaches that utilize crop growth models have been advanced as a possible solution. However, these approaches are computationally intensive, and require various agro-

meteorological inputs (Desloires et al. 2023). Satellite based crop yield models that rely on ground data should thus be validated on data outside of their geographic area before extrapolating over larger regions.

This study adds to the growing body of work that assesses the ability of using high spatio-temporal resolution satellite data to map field-level yields in smallholder systems at scale. We compare multiple methods to estimate maize yield in two regions within Oromia District, Ethiopia during the 2021 growing season. In this study, we specifically examine:

(1) How well can we map field level yields in smallholder maize systems in Ethiopia using Sentinel-2 imagery?
(2) Which vegetation index results in the highest yield prediction accuracies: NDVI, GCVI, or MTCI?
(3) Which model leads to higher prediction accuracies: multiple linear regression or random forest regression?
(4) Is it possible to create a generalizable model that accurately estimates yields across multiple regions using limited ground data for training?

Our results will provide important insights into the ability of Sentinel-2 imagery to map field-level yields in heterogeneous smallholder systems. This is critically important as smallholder systems are projected to face some of the largest increases in food demand over the coming decades, and such yield information can help identify where yield gaps are the largest and potential interventions that may help close these yield gaps.

2. Study Area

Our study area spans a 30,000 km^2 region in Oromia district, Ethiopia (Figure 1), with data collected in two distinct sub-regions. The first sub-region comprises an approximately 8,800 km^2 area straddling parts of the East Shewa and Garaghe Zones. The second sub-region comprises a 1,700 km^2 area in the western Jimma Zone. The greater region is dominated by smallholder agriculture, with cropland covering over 72% of the land area. There are two agricultural growing seasons in Ethiopia when grain is grown: the long rainy season (Meher) and the short rainy season (Belg) (Hadado et al. 2009). The focus of this study is the long rainy season of Meher, when the majority of grain is grown, which spans from April to September (Wakjira et al. 2021). Agricultural management practices and soil conditions are highly heterogeneous over the study region. Most of the fields surveyed (81%) were less than 5 hectares, and approximately one quarter of all fields (26%) were less than two hectares. Most farmers applied fertilizer (93%), such as DAP and urea, but inputs varied from 0 to 300 kg throughout the season. There was likewise a diversity of maize varieties planted, with the most common varieties being BH661, Limu, Damote, and Shone. Soil texture was either silt, clay, or sand, with silt being the most common (40%). A majority of fields (67%) had fruiting trees, but only 20% of fields were intercropped with other species, of which beans were the most common. None of the fields surveyed used irrigation. Although sow date and harvest date varied slightly, most fields (90%) were sown in April or May, and all fields were harvested in November.

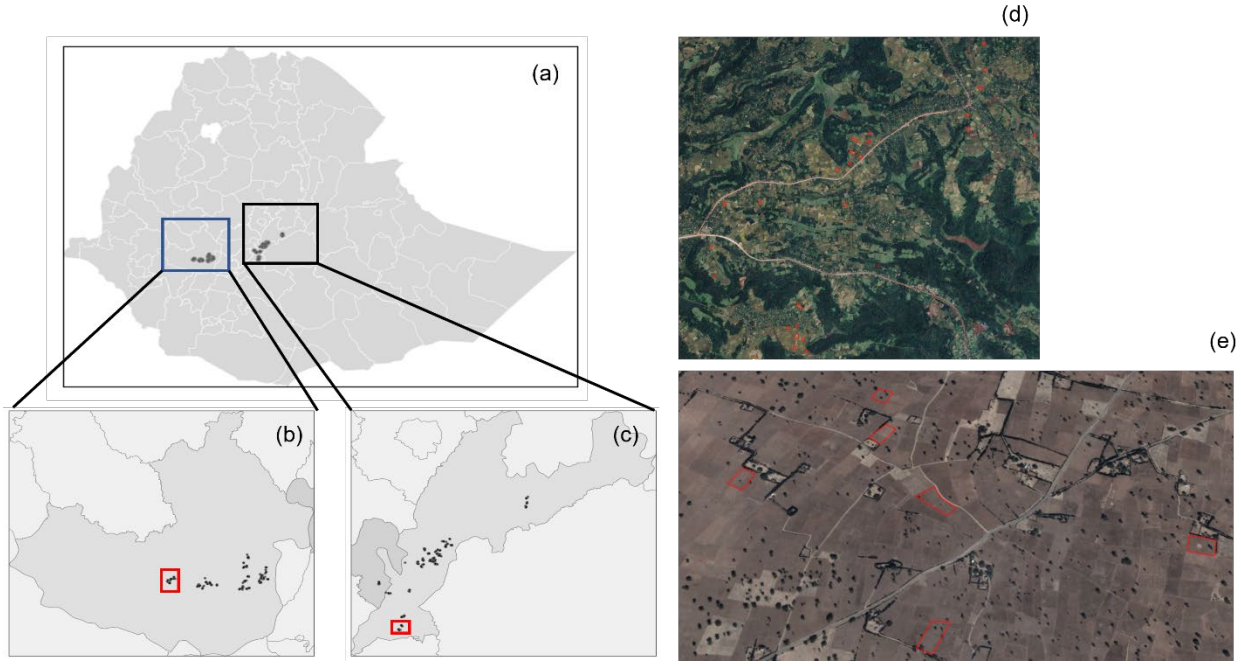

Figure 1: Map of study area with (a) countrywide map of Ethiopia with field locations, (b) Jimma Zone with field locations, (c) East Shewa and Guraghe Zones with field locations, (d) detail of panel (b) showing field boundaries over high resolution aerial imagery, and (e) detail of panel (c) showing field boundaries over high resolution aerial imagery. Aerial imagery via Google Earth 7.3.6 (2023) CNES/Airbus [Accessed 4/18/2023].

3. Methods

We processed Sentinel-2 imagery (Section 3.2) to estimate maize yields across fields where we collected on-the-ground crop cut yield information (Section 3.1). To answer our main research questions, we created and compared several different vegetation indices (Section 3.2.). We also compared two different yield estimation models, linear regression and random forest regression (Section 3.3). Finally, we examined how generalizable our yield estimation algorithms may be across the full study region by training a model using data from only one region and then applying and validating it in the other region (Section 3.4).

3.1. Crop Cut Data

Most methods for estimating crop yield use field level yield data to calibrate remote sensing models (Jain et al. 2016). The gold standard for yield estimation in field is through collecting crop cuts, which we used to estimate yield at the end of the growing season (Tiedeman et al. 2022). Agricultural surveys and crop cuts were administered by collaborators working with the International Maize and Wheat Improvement Centre for 600 fields across our study area in 2021. Each maize field was split into four quadrants. At the center of each quadrant, a 5 m x 8 m area was harvested. The maize was dried and threshed before it was weighed in-field. We

averaged all subplot yields to obtain field-level yields. GPS coordinates were collected for each crop cut sub-plot, as well as for the larger field boundaries. Field boundary polygons were constructed in Python using the Shapely package version 2.0.1 (Gillies et al. 2007). Boundaries were then manually corrected in Google Earth Pro by aligning field boundaries with visible boundaries from the latest available high-resolution aerial imagery in Google Earth Pro. Field boundaries were described as 'low', 'medium', or 'high' confidence based on how closely they corresponded to visible boundaries in the aerial imagery. We retained only the 321 'high' confidence fields for our analysis. These were fields which had raw GPS coordinates that lined up well with field boundaries in the latest available very high-resolution (VHR) aerial imagery and required little to no manual adjustment. This comprised 158 fields in the Jimma sub-region and 163 fields in the East Shewa-Garaghe sub-region.

3.2. Sentinel-2 Imagery

We accessed Level 2A, atmospherically corrected surface reflectance Sentinel-2 imagery through the Google Earth Engine (GEE) platform (Gorelick et al. 2017). As the growing season coincided with the rainy season, many of the Sentinel-2 images had a high number of cloudy pixels. Therefore, pixel-wise cloud masking was performed in GEE using the cloud probability score generated by the Sentinel Hub Cloud Detector algorithm using the s2Cloudless Python library (Zupanc 2017). Sentinel Hub Cloud Detector is a readily available machine learning cloud and cloud shadow masking algorithm for use in conjunction with Sentinel-2 surface reflectance imagery. Based on visual inspection of cloud removal, we set the cloud probability threshold parameter at 40% and the cloud filter threshold parameter to 100%. As the result, every image from the Sentinel-2 surface reflectance image collection from 3/15/2021 to 12/05/2021 was retained for our analysis, regardless of cloud cover percentage, and pixels with a cloud probability greater than 40% were masked.

Three vegetation indices commonly used in satellite-based yield estimation studies, namely NDVI, GCVI, and MTCI, were calculated for each Sentinel-2 image following the application of the cloud mask. The relevant Sentinel-2 bands used to calculate each vegetation index are recorded in Table 1, and the formula for each index is listed in Table 2. We then created maximum vegetation index composites at the finest temporal resolution that allowed for wall-to-wall coverage of imagery using the 'qualityMosaic' function in GEE. Specifically, maximum values were computed on a pixel-by-pixel basis for all available, cloud-free pixels. For the East Shewa-Garaghe sub-region (Figure 1b) we were able to create bi-weekly mosaics. However, for the Jimma sub-region (Figure 1c), there was more cloud cover and we were only able to create monthly mosaics. At the regional level that spanned both sites, we created monthly mosaics as this was the finest temporal resolution over which we could produce wall-to-wall mosaics across the study area. The specific compositing windows and the number of Sentinel-2 tiles that contributed to each composite are recorded in the supplementary materials (Table S1). The mean value of each vegetation index was calculated for each field and for each compositing window using the 'reduceRegions' function in GEE.

Band	Spectral Range (nm)	Resolution (m)
Green	543-578	10
Red	650-680	10
Red Edge (RE)	690-730	20
Near Infrared (NIR)	760-850	10

Table 1: Sentinel-2 bands used for the computation of vegetation indices (VIs)

Vegetation Index	Formula	Reference
Normalized Difference Vegetation Index (NDVI)	(NIR - Red)/(NIR + Red)	Rouse et al. 1973
Green Chlorophyll Vegetation Index (GCVI)	(NIR/Green - 1)	Gitelson et al. 2003
MERIS Terrestrial Chlorophyll Index (MTCI)	(NIR - RE)/(RE - Red)	Dash & Curran 2004

Table 2: Three Vegetation indices (VIs) used in this study and their formulae

3.3. Model Parameterization and Validation

We compared two models to estimate yield, a linear regression model and a random forest regression model. The linear regression model was estimated using Equation 1.

$$Yield = \beta_0 + \beta_1 VI_1 + \beta_2 VI_2 + \ldots + \beta_n VI_n + \epsilon \quad (1)$$

where *Yield* is observed yield at the field scale calculated via crop cuts, B_1VI_1 represents the coefficient for the mean VI value for each field for the first composite window, B_2VI_2 represents the coefficient for the mean VI value for each field for the second composite window, and B_nVI_n represents the coefficient for the mean VI for each field for the n^{th} composite window. The linear regression models that we developed used VIs from all possible compositing windows, as preliminary analyses and previous studies suggested that models that used all image dates had higher prediction accuracies in general. Separate linear regression models were run for each vegetation index and each region, resulting in three linear regression models for each region. The linear regressions were performed in Python using the 'sklearn.linear_model.LinearRegression' routine from the Scikit-Learn version 1.2.2 machine learning library (Pedregosa et al. 2011).

The second model we created was a random forest regression that used the same predictor variables as our linear regression model (Equation 1). Random forest is an ensemble learning method which creates multiple decision trees using randomly drawn subsamples of the

data (Breiman 2011). The average is calculated across all decision trees to output a final model. We developed the random forest regression models using the 'sklearn.ensemble.RandomForestRegressor' routine from the Scikit-learn Python module with default hyperparameters. Separate random forest regression models were run for each vegetation index and each region, resulting in three random forest regression models for each region.

To validate our models, a 70:30 train-test split was used for each model, meaning that 70% of the fields were used to train our algorithms and 30% of the fields were used to validate the accuracy of our algorithms. The linear regression model and random forest regression model for each region and VI parameterization were trained and evaluated using the same train-test split polygons. Models were scored on the basis of their coefficient of determination (R^2) and Root Mean Squared Error (RMSE) when compared to the observed crop cut yield estimates for each field.

3.4. Comparison of Models by sub-region:

We assessed the spatial generalizability of our modeling approaches by validating the model developed in one sub-region on observations from the other sub-region. Contrary to our initial approach in which a train-test split was used to validate the models, here, for each sub-region, models were generated using all available observations in the sub-region. Observations from the remaining sub-region were then used to validate the models, and R^2 and RMSE values were recorded. This analysis was performed for both the East Shewa-Guraghe sub-region and the Jimma sub-region.

4. Results

Considering our main research question which was to evaluate how well we could map yield in smallholder maize systems in Ethiopia, we found that we were able to map field-level yields fairly well, with R^2 values reaching 0.24 to 0.47. This compares well with previous studies in smallholder systems, which generally find R^2 values between 0.4 and 0.6. for the best fit models in each region (Table 3). Comparing the three different indices used, GCVI, NDVI, and MTCI, we found that overall MTCI led to the highest prediction accuracies across all models and locations, as evidenced by higher R^2 values and lower RMSEs. This was true across all models except for the random forest regression model in the eastern region, where GCVI slightly outperformed MTCI. On average, MTCI improved R^2 values by 0.065 and reduced RMSE by 79 kg/ha compared to the next best performing vegetation index. After MTCI, GCVI led to the highest prediction accuracies, outperforming NDVI across all models and locations.

We next examined which model type led to the highest yield prediction accuracies, linear regression models, or random forest regressions. We found that when we developed localized models, linear regressions outperformed random forest regression models, though the difference in R^2 was generally small (< 0.05). Specifically, in the East Shewa-Guraghe sub-region, the sub-region with highest model performance on average, the linear regression models improved R^2 by 0.0167 and reduced RMSE by 65.8 kg/ha over the random forest regression models. The difference in performance was larger in the Jimma sub-region, where the linear regression models improved R^2 by 0.09 and reduced RMSE by 103.2 kg/ha on average compared to the

random forest regression models. However, for the regional models, we found that random forest regression was a slight improvement over the linear regression model, though the difference was small. Random forest regression models improved R^2 by 0.0167 and reduced RMSE by 17.7 kg/ha compared to the linear regression models.

Finally, considering the generalizability of our models, we find that our results are not very generalizable. One indicator of poor generalizability was the variability in model performance between the two sub-regions of our study. We observed much higher accuracies in the East Shewa-Guraghe sub-region, where the best performing model had a high R^2 value of 0.47 and lower RMSE equal to 1607.5 kg/ha (linear regression model with MTCI). In the Jimma sub-region, the best performing model had an R^2 of only 0.24 and a higher RMSE value of 1879.2 kg/ha (linear regression model with MTCI). One potential reason for this difference in accuracy was the length of the temporal mosaic, given that the East Shewa-Guraghe sub-region had less cloud cover, which allowed us to create biweekly mosaics; in the Jimma sub-region, there was more cloud cover and haze allowing us to only create monthly mosaics. To test how much of a reduction in accuracy was due to differences in the temporal mosaic length, we reran the best performing model (linear regression model with MTCI) in the East Shewa-Guraghe sub-region using monthly mosaics. We found that the R^2 dropped to 0.33 and the RMSE increased to 1815.2 kg/ha. This suggests that some of the difference in accuracy is due to the presence of cloud cover and the reduction in available imagery, though the monthly model in the East Shewa-Guraghe sub-region still outperformed a similar model in the Jimma sub-region.

Region/Sub-region	Regressor	Vegetation Index (VI)	Parameterization	Coefficient of Determination (R^2)	Root Mean Squared Error (RMSE)
East Shewa-Guraghe	Linear	GCVI	Biweekly	0.42	1679.5
East Shewa-Guraghe	Linear	MTCI	Biweekly	0.47	1607.5
East Shewa-Guraghe	Linear	NDVI	Biweekly	0.37	1754.8
East Shewa-Guraghe	Random Forest	GCVI	Biweekly	0.45	1643.1
East Shewa-Guraghe	Random Forest	MTCI	Biweekly	0.43	1673.3
East Shewa-Guraghe	Random Forest	NDVI	Biweekly	0.33	1805.2
Jimma	Linear	GCVI	Monthly	0.18	1954.0
Jimma	Linear	MTCI	Monthly	0.24	1879.2
Jimma	Linear	NDVI	Monthly	0.09	2061.9
Jimma	Random Forest	GCVI	Monthly	0.11	2039.5
Jimma	Random Forest	MTCI	Monthly	0.21	1919.7
Jimma	Random Forest	NDVI	Monthly	-0.08	2245.5
Regional	Linear	GCVI	Monthly	0.12	2039.4
Regional	Linear	MTCI	Monthly	0.22	1919.4
Regional	Linear	NDVI	Monthly	0.09	2071.7
Regional	Random Forest	GCVI	Monthly	0.16	1994.4
Regional	Random Forest	MTCI	Monthly	0.26	1876.5
Regional	Random Forest	NDVI	Monthly	0.06	2106.4

Table 3: Coefficient of Determination (R^2) and associated Root Mean Squared Error (RMSE) for each regression model

We also developed general, regional models across both sub-regions to test generalizability (Table 3). The best performing regional model was the random forest regression model using monthly MTCI composites (R^2 = 0.26, RMSE = 1876.5 kg/ha, Table 3). This model performed comparably to the best performing model in the Jimma sub-region. However, when compared to the best performing model in the East Shewa-Guraghe sub-region, R^2 was reduced

by 0.21 and RMSE was increased by 269 kg/ha. This suggests that localized models can lead to higher prediction accuracy in this study region.

Finally, we tested generalizability by validating the models produced in one region on test data from the other region. Overall, we found that our models were not very generalizable, particularly when taking models developed in the East Shewa-Guraghe sub-region and applying them to the Jimma sub-region (Table 4). When we did this, we found that most models resulted in negative R^2 values and large RMSE values (> 2000 kg/ha). The highest scoring model was the random forest regression model with MTCI composites (R^2 = 0.13, RMSE = 1834.25 kg/ha). The Jimma sub-region models, however, were more generalizable to the East Shewa-Guraghe sub-region. For the most part, these results had a moderate R^2 values around 0.2, with the best performing linear regression MTCI model reaching R^2 values of 0.31 (RMSE = 1796.5 kg/ha).

Training Sub-region	Validation Sub-region	Regressor	Vegetation Index (VI)	R^2 (training data)	R^2 (validation data)	RMSE (training data)	RMSE (validation data)
East Shewa-Guraghe	Jimma	Linear	GCVI	0.42	-1.20	1679.5	2911.9
East Shewa-Guraghe	Jimma	Linear	MTCI	0.47	-0.52	1607.5	2419.6
East Shewa-Guraghe	Jimma	Linear	NDVI	0.37	-0.61	1754.8	2490.9
East Shewa-Guraghe	Jimma	Random Forest	GCVI	0.45	-0.33	1643.1	2264.1
East Shewa-Guraghe	Jimma	Random Forest	MTCI	0.43	0.13	1673.3	1834.3
East Shewa-Guraghe	Jimma	Random Forest	NDVI	0.33	-0.40	1805.2	2321.4
Jimma	East Shewa-Guraghe	Linear	GCVI	0.18	0.21	1954.0	1920.4
Jimma	East Shewa-Guraghe	Linear	MTCI	0.24	0.31	1879.2	1796.5
Jimma	East Shewa-Guraghe	Linear	NDVI	0.09	0.00	2061.9	2163.0
Jimma	East Shewa-Guraghe	Random Forest	GCVI	0.11	-0.07	2039.5	2240.8
Jimma	East Shewa-Guraghe	Random Forest	MTCI	0.21	0.22	1919.7	1907.4
Jimma	East Shewa-Guraghe	Random Forest	NDVI	-0.08	-0.13	2245.5	2297.8

Table 4: Results of regression analysis for local (sub-regional) yield models. The East Shewa-Guraghe sub-region models were validated on data from the Jimma sub-region and vice-versa

5. Discussion

This study adds to the growing body of work that assesses the use of high spatio-temporal resolution Sentinel-2 satellite data to map field-level yields in smallholder systems. Using over 300 crop cut field measures of yield, we developed field-level maize yield estimates for the main growing season in 2021. We examined which vegetation indices (NDVI, GCVI, and MTCI) and which yield model (linear regression and random forest) led to the highest yield prediction accuracies. Finally, we analyzed the generalizability of our yield estimation models over larger spatial scales, outside of the region in which the model was originally trained. Overall, we found that we were able to map yields with high accuracy (with R^2 values up to 0.47), principally when using linear regression models with MTCI. We did find, however, that our results were not very generalizable and models developed using local training data led to substantial increases in R^2 and reductions in RMSE. These results suggest that Sentinel-2 satellite data can be used to successfully map field-level yields even in smallholder systems, but more generalizable algorithms need to be used if these models are to be applied across large spatial scales.

Considering which vegetation index led to the highest accuracy, we found that models using MTCI led to the highest accuracies, followed by models using GCVI. There are several reasons MTCI likely outperformed the other vegetation indices considered in our study. First, MTCI and GCVI are optimized for chlorophyll detection, while NDVI has been shown to be related to leaf area index (LAI) (Dash & Curran, 2007). Previous studies have suggested that indices with increased sensitivity to chlorophyll concentrations are better able to account for the effects of nutrient stress on yields (Burke & Lobell 2017), which could be particularly important in our study region where fertilizer application rates were low and fields were likely nitrogen limited. Second, NDVI may have performed poorly because it is more likely than MTCI or GCVI to become saturated at high levels of biomass (Tucker 1977, Sellers 1985, Gu et al. 2013, Ulfa et al. 2022, Dash et al. 2008). To support this theory, we found low correlation values between NDVI and yield compared to that of other VIs during late season periods of peak biomass (Figures S1, S2, and S3). Third, the three vegetation indices considered in our study are differentially impacted by haze and atmospheric scattering. Specifically, MTCI is less sensitive to haze and atmospheric effects than GCVI and NDVI because it is computed using two nearby spectral bands that are affected similarly by atmospheric scattering (Dash & Curran 2007, Lobell et al. 2020). This could be particularly important for our study region where we observed patches of haze in various images (Figure S4) despite cloud masking and image compositing. These results suggest that MTCI may do especially well compared to other VIs during the rainy growing season, when cloud cover and haze are extensive. Finally, we found that MTCI still performed best despite its availability at a coarser spatial resolution compared to the other VIs considered in this study (20 m vs 10 m). This is likely because the fields considered in our study were still relatively large; the mean plot size across all fields was 3.5 ha, and only 11 fields (3% of all fields) were smaller than one hectare. Future work should examine if MTCI still outperforms other indices in locations with very small field sizes (< 0.5 ha).

We found that the linear regression model outperformed the random forest regression in most cases. However, in the East Shewa-Guraghe sub-regional and regional scale analyses, this difference was very slight. This is likely because the relationship between VIs and yield is largely linear, and the linear regression model better captured this relationship while still remaining generalizable. Even so, other studies have suggested that random forest regressions have great potential for crop yield estimation, as they easily allow for the integration of additional predictor variables, such as temperature, precipitation, solar radiation, and soil moisture (Sakamoto et al. 2020). For example, Hunt et al. (2019) show that random forest regressions can allow for the inclusion of environmental variables, such as the soil water index (SWI), that have no direct relationship with yield, but do have an underlying relationship with spectral reflectance. Future work should examine how much yield estimation can be improved by including additional environmental data, and whether random forest regression models are better able to handle potential complex and non-linear relationships than linear regression models.

Differences in model performance between the two sub-regions in our study area were notable. Models performed much better overall in the East Shewa-Guraghe sub-region than in the Jimma sub-region across all VIs and model types. This difference in accuracy could not be fully explained by the differences in the temporal availability of imagery, as we found that a

linear regression model trained using monthly MTCI values performed better in the East Shewa-Guraghe sub-region (R^2 = 0.33, RMSE = 1815.2) compared to in the Jimma sub-region (R^2 = 0.24, RMSE = 1879.2). One reason for this may be because, despite our extensive cloud masking and mosaicking, there was still significantly more haze seen in imagery in the Jimma region compared to the East Shewa-Guraghe sub-region (Figure S4). Future work should examine whether including additional imagery that is less sensitive to cloud cover and haze, such as radar Sentinel-1 imagery, may improve yield prediction accuracies in regions plagued by high cloud cover during the rainy growing season.

Considering the generalizability of our model, we found that a regional model that was trained using data from both sub-regions performed similarly for the Jimma sub-region but poorer for the East Shewa-Guraghe sub-region, with the model only explaining 26% of the observed variation in yield (Table S3). We also found that the models trained in one region and applied to another region, where no data were used for training, performed relatively poorly, particularly when applying models to Jimma that were trained in the East sub-region. This may be because the two sub-regions are located in different agroecological zones (Amede et al. 2015) with differences in overall yields and yield variation; The Jimma sub-region, which is located in a cool/humid zone, had higher observed yields on average with less variability (mean = 6235 kg/ha, sd = 1967 kg/ha) compared to the East Shewa-Guraghe sub-region (mean = 5461 kg/ha, sd = 2168 kg/ha), which is located in a cool/subhumid zone. Such differences likely affect model generalizability as the relationship between VIs and yield vary across space, and VIs are not always able to capture yield variability due to environmental stress (Sakamoto 2020, Jain et al. 2017). Future work should examine how other approaches that may be more generalizable, such as those that use crop model simulations to train algorithms instead of localized ground data, perform when mapping yield across disparate regions.

In conclusion, we found that we were able to use Sentinel-2 satellite imagery to map field-level maize yields accurately, particularly in the eastern portion of our study region, where we achieved an R^2 of 0.47 in our best performing model. Across the East Shewa-Guraghe sub-region, model accuracies were comparable to those found in other studies that used high-resolution satellite imagery to map field-level grain yields in smallholder systems (Jain et al. 2019). We found that linear regression models that used MTCI data largely led to the best performing models, though these models were not very generalizable. One of the main reasons for this difference in performance across regions may be extensive cloud cover and haze that was difficult to remove for the western Jimma study region. Our results broadly show the promise of Sentinel-2 for mapping field-level yields, even during the rainy season in regions with heterogeneous smallholder fields.

6. References

1. Abate, T.; Shiferaw, B.; Menkir, A.; Wegary, D.; Kebede, Y.; Tesfaye, K.; Kassie, M.; Bogale, G.; Tadesse, B.; Keno, T. Factors that transformed maize productivity in Ethiopia. Food Sec. **2013**, 7, 965–981.

2. Amede, T.; Auricht, C.; Boffa, J.M.; Dixon, J.; Mallawaarachchi, T.; Rukuni, M.; Deneke, T. The Evolving Farming and Pastoral Landscapes in Ethiopia: A Farming System Framework for Investment Planning and Priority Setting. ACIAR, **2015**; ISBN: 1056875700.
3. Azzari, G.; Jain, M.; Lobell, D. Towards fine resolution global maps of crop yields: Testing multiple methods and satellites in three countries. Remote Sens. Env. **2017,** 202, 129-141
4. Bezner Kerr, R., T. Hasegawa, R. Lasco, I. Bhatt, D. Deryng, A. Farrell, H. Gurney-Smith, H. Ju, S. Lluch-Cota, F. Meza, G. Nelson, H. Neufeldt, and P. Thornton, 2022: Food, Fibre, and Other Ecosystem Products. In: Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [H.-O. Pörtner, D.C. Roberts, M. Tignor, E.S. Poloczanska, K. Mintenbeck, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem, B. Rama (eds.)]. Cambridge University Press, Cambridge, UK and New York, NY, USA, pp. 713–906.
5. Breiman, L. Random Forests. *Machine Learning. **2001,*** 45, 5–32.
6. Burke, M.; Lobell, D.B. Satellite-based assessment of yield variation and its determinants in smallholder African systems. Proc. Natl. Acad. Sci. USA **2017,** 114, 2189–2194.
7. Carletto, C.; Jolliffe, D.; Banerjee, R. From Tragedy to Renaissance: Improving Agricultural Data for Better Policies. J. Dev. Stu. **2013**, 51, 133-148.
8. Clevers, J.G.P.W.; Gitelson, A.A. Remote estimation of crop and grass chlorophyll and nitrogen content using red-edge bands on Sentinel-2 and -3. **2013**, 23, 344-451.
9. Curran, P.J.; Dash, J. Algorithm Theoretical Basis Document ATBD 2.22 Chlorophyll Index; University of. Southampton: Southampton, UK, 2005.
10. Dash, J.; Curran, P.J. The MERIS terrestrial chlorophyll index. International Journal of Remote Sensing. **2004**, 2523, 5403–5413.
11. Dash, J.; Lankester, T.; Hubbard, S.; Curran, P. J. Signal-to-noise ratio for MTCI and NDVI time series data. Proceedings of the 2nd MERIS/(A)ATSR User Workshop, 2008.
12. Desloires, J.; Ienco, D.; Botrel, A. Out-of-year corn yield prediction at field-scale using Sentinel-2 satellite imagery and machine learning methods. Computers and Electronics in Agriculture. 2023, 209, 107807.
13. Di Falco, S.; Veronesi, M.; Yesuf, M. Does adaptation to climate provide food security? A micro-perspective from Ethiopia. Am. J. Agric. Econ. **2011**, 93, 829–846.
14. Farmonov, N.; Amankulova, K.; Szatmári, J.; Urinov, J.; Narmanov, Z.; Nosirov, J.; Mucsi, L. Combining PlanetScope and Sentinel-2 images with environmental data for improved wheat yield estimation. International Journal of Digital Earth. 2023, 16, 847-867.
15. Gillies, S., & others. Shapely: manipulation and analysis of geometric objects. Retrieved from https://github.com/Toblerity/Shapely (2007).
16. Godfray, C.; Beddington, J. Crute, I.; Haddad, L.; Lawrence, D.; Muir, J.; Pretty, J.; Robinson, S.; Toulmin, C. Food Security: The Challenge of Feeding 9 Billion People. Science. **2010**, 327, 812-823.

17. Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment. 2017, 202, 18-27.
18. Hadado, T. T.; Rau, D.; Bitocchi, E.; Papa, R. Genetic diversity of barley (Hordeum vulgare L .) landraces from the central highlands of Ethiopia: comparison between the Belg and Meher growing seasons using morphological traits. Genetic Resource Crop Evolution. **2009**, 56, 1131-1148.
19. Hunt, M.L; Blackburn, G.A.; Carrasco, L; Redhead, J.W.; Rowland, C.S. High resolution wheat yield mapping using Sentinel-2, Remote Sens. Env. **2019,** 233, XXX-XXX.
20. Houborg, R.; McCabe, M. A Cubesat enabled Spatio-Temporal Enhancement Method (CESTEM) utilizing Planet, Landsat and MODIS data. Remote Sensing of Environment. **2018,** 209, 211-226.
21. Jain, M.; Srivastava, A.K.; Singh, B.; Joon, R.J.; McDonald, A.; Royal, K.; Lisaius, M.C.; Lobell, D.B.; Mapping Smallholder Wheat Yields and Sowing Dates Using Micro-Satellite Data, Remote Sens. **2016**, 8, XXX-XXX.
22. Jain, M.; Singh, B.; Srivastava, A.A.K.; Malik, R.K.; McDonald, A.J.; Lobell, D.B. Using satellite data to identify the causes of and potential solutions for yield gaps in India's Wheat Belt. Environ. Res. Lett. **2017**, 12, 094011.
23. Jin, Z.; Azzari, G.; Burke, M.; Aston, S.; Lobell, D. Mapping smallholder yield heterogeneity at multiple scales in Eastern Africa. Remote Sens. **2017**, 9, XXX-XXX.
24. Jin, Z.; Azzari G.; You, C.; Di Tommaso, S.; Aston, S.; Burke, M.; Lobell, D.B. Smallholder maize area and yield mapping at national scales with Google Earth Engine. Remote Sens. Env. **2019,** 228, 115-128.
25. Lobell, D. B.; Cassman, K. G.; Field, C. B. Crop Yield Gaps: Their Importance, Magnitudes, and Causes. Annual Review of Environmental and Resources **2009**, 34, 179-204.
26. Lobell, D.B.; Thau, D.; Seifert, C.; Engle, E.; Little, B. A scalable satellite-based crop yield mapper, Remote Sens. Env. **2015,** 164, 324-333.
27. Lobell, D.B.; Azzari, G.; Burke, M.; Gourlay, S.; Jin, Z.; Kilic, T.; Murray, S. 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. 2020, 102, 202-219.
28. Mohamed, A. Food Security Situation in Ethiopia: A Review Study. International Journal of Health Economics and Policy. **2017**, 2, 86-96.
29. Nguy-Robertson, A. L.; Peng, Y.; Gitelson, A. A.; Arkebauer, T. J.; Pimstein, A.; Herrmann, I.; Karnieli, A.; Rundquist, D. C.; Bonfil, D. J. Estimating green LAI in four crops: Potential of determining optimal spectral bands for a universal algorithm. Agricultural and forest meteorology. 2014, 192, 140-148.
30. OECD/FAO, 2022, OECD-FAO Agricultural Outlook 2022-2031, OECD Publishing, Paris.
31. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

32. Rouse, J.W.; Haas, R.H.; Schell, J.A.; Deering, D.W. Monitoring Vegetation Systems in the Great Okains with ERTS. In Proceedings of the Third Earth Resources Technology Satellite-1 Symposium, Washington, DC, USA, 10–14 December 1973; Freden, S.C., Mercanti, E.P., Eds.; NASA: Washington, DC, USA, 1974.
33. Sellers, P. J. Canopy reflectance, photosynthesis and transpiration. International Journal of Remote Sensing. 1985, 6, 1335–1372.
34. Sibley, A.M.; Grassini, P.; Thomas, N.E.; Cassman, K.G.; Lobell, D.B. Testing Remote Sensing Approaches for Assessing Yield Variability among Maize Fields. Agronomy Journal. 2014, 106, 24-32.
35. Sweeney, S.; Ruseva, T.; Estes, L.; Evans, T. Mapping Cropland in Smallholder-Dominated Savannas: Integrating Remote Sensing Techniques and Probabilistic Modeling. Remote Sens. **2015**, 7, 15295-15317.
36. Sakamoto T. Incorporating environmental variables into a MODIS-based crop yield estimation method for United States corn and soybeans through the use of a random forest regression algorithm ISPRS Journal of Photogrammetry and Remote Sensing. **2020**, 160, 208-228.
37. Tilman, D.; Balzer, C.; Hill, J.; Befort, B. L. Global food demand and the sustainable intensification of agriculture. Proceedings of the National Academy of Sciences. **2011**, 108, 20260-20264.
38. Tucker, C. J. Asymptotic nature of grass canopy spectral reflectance. Applied Optics. 1977, 16, 1151–1156.
39. Wakjira, M. T.; Peleg, N.; Anghileri, D.; Molnar, D.; Alamirew, T.; Six, J.; Molnar, P. Rainfall seasonality and timing: implications for cereal crop production in Ethiopia. Agricultural and Forest Meteorology. **2021**, 310, 108633.
40. Zhang, L.; Zhang, Z.; Luo, Y.; Cao, J.; Xie, R.; Li, S. Integrating satellite-derived climatic and vegetation indices to predict smallholder maize yield using deep learning. Agricultural and Forest Meteorology. **2021**, 311, 108666.
41. Zupanc, A. Improving Cloud Detection With Machine Learning, 2017.

Supplementary Information

Composite #	East Shewa-Guraghe (# of images)	Jimma (# of images)	Regional (# of images)
1	12	12	72
2	12	12	72
3	12	12	72
4	12	12	72
5	12	14	80
6	12	12	72
7	12	12	72
8	12	12	77
9	12	8	48
10	12		
11	12		
12	12		
13	12		
14	12		
15	15		
16	12		
17	16		
Total	211	106	637

Table S1. The number of Sentinel-2 Scenes per composite window in each sub-region.

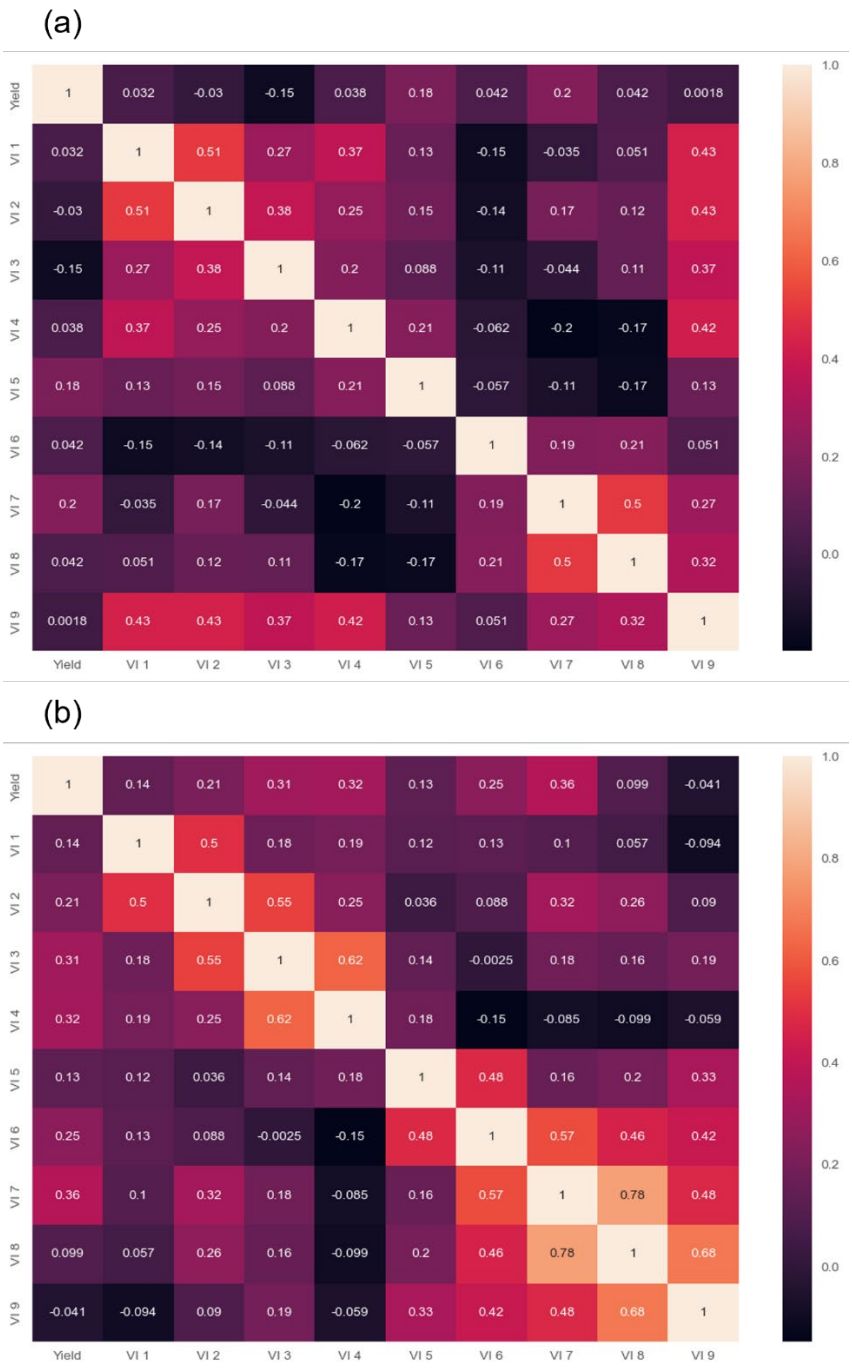

Figure S1. Pairwise correlation heatmap of yield and monthly NDVI composites for (a) the Jimma Subregion and (b) the East Shewa-Guraghe sub-region during the 2021 Meher growing season.

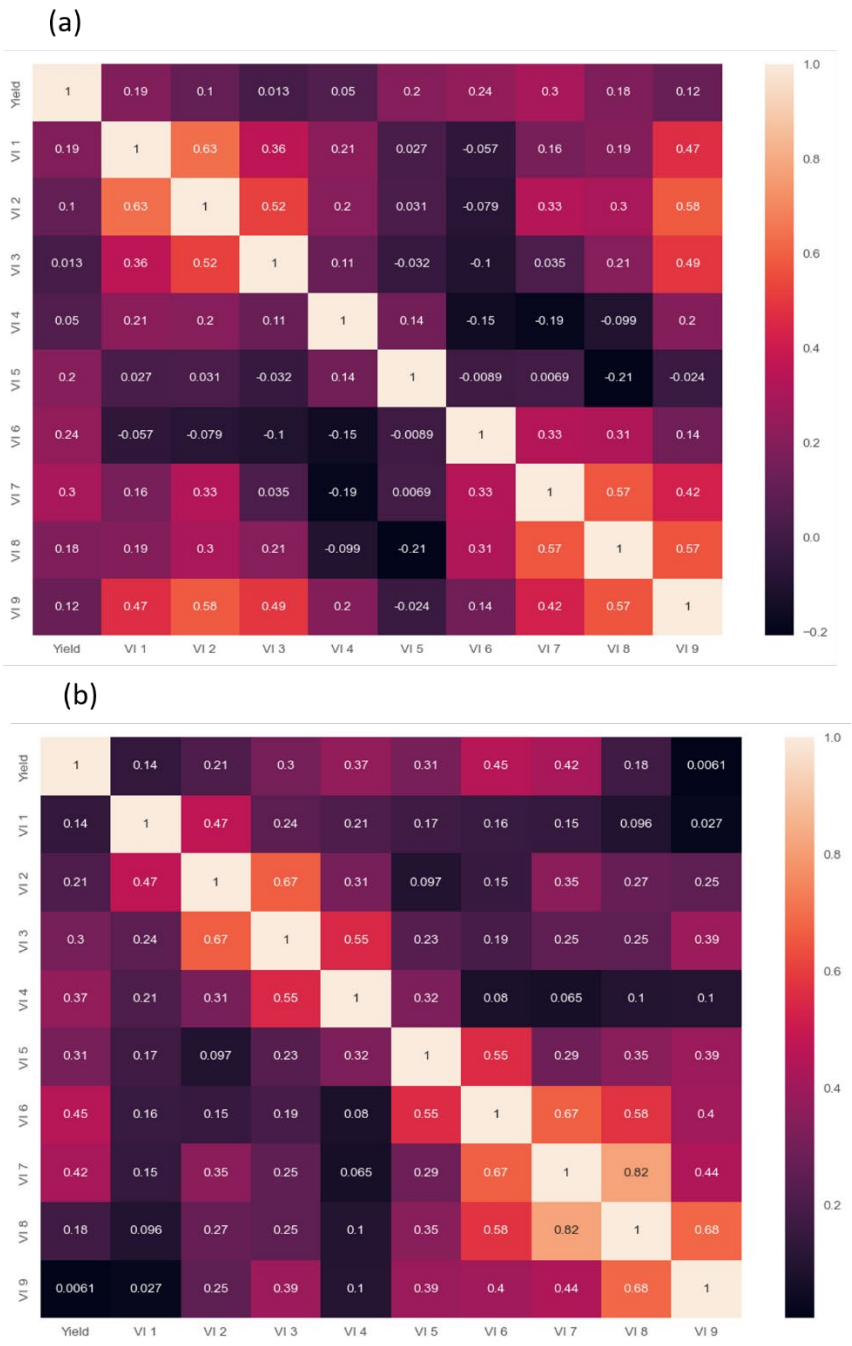

Figure S2. Pairwise correlation heatmap of yield and monthly GCVI composites for (a) the Jimma Subregion and (b) the East Shewa-Guraghe sub-region during the 2021 Meher growing season.

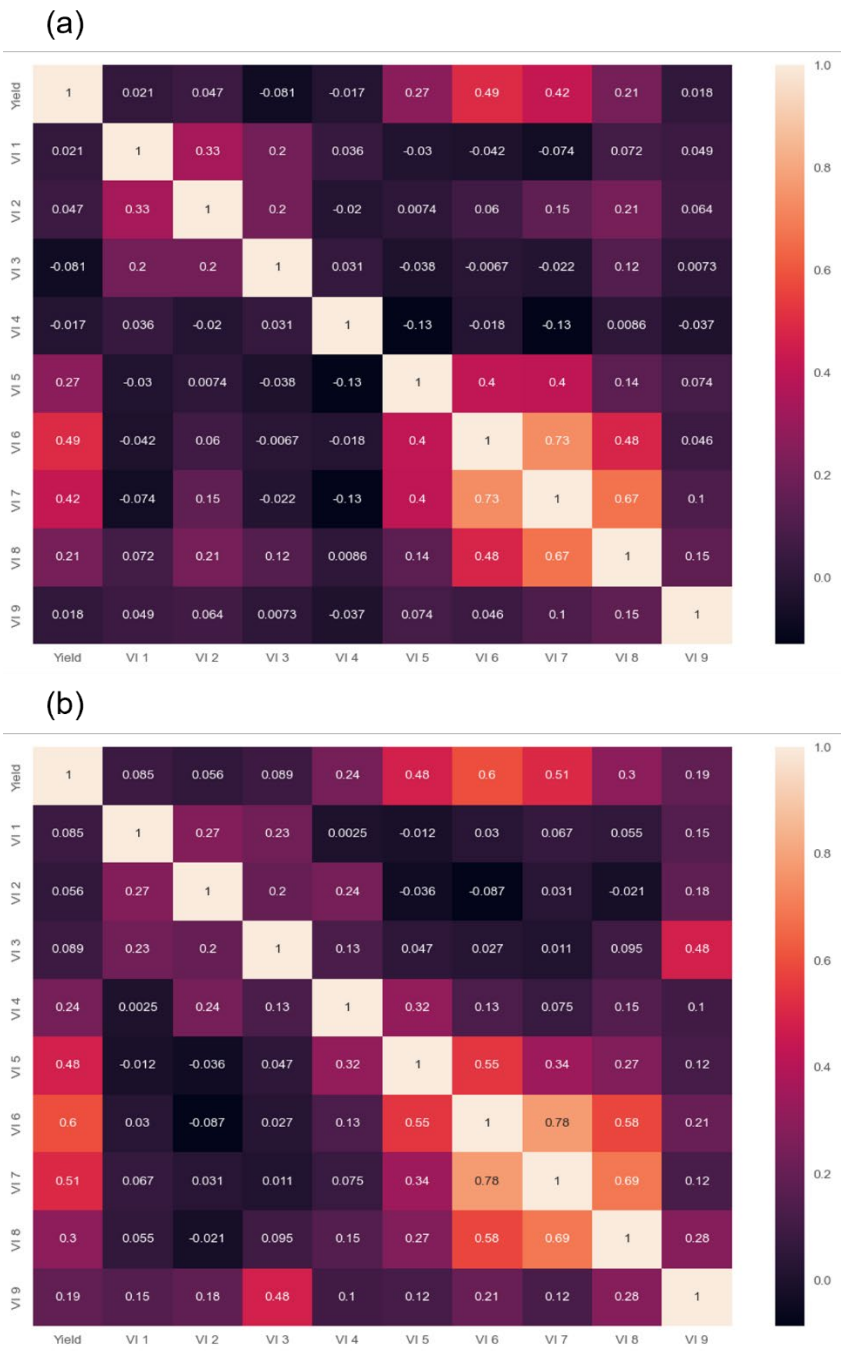

Figure S3. Pairwise correlation heatmap of yield and monthly MTCI composites for (a) the Jimma Subregion and (b) the East Shewa-Guraghe sub-region during the 2021 Meher growing season.

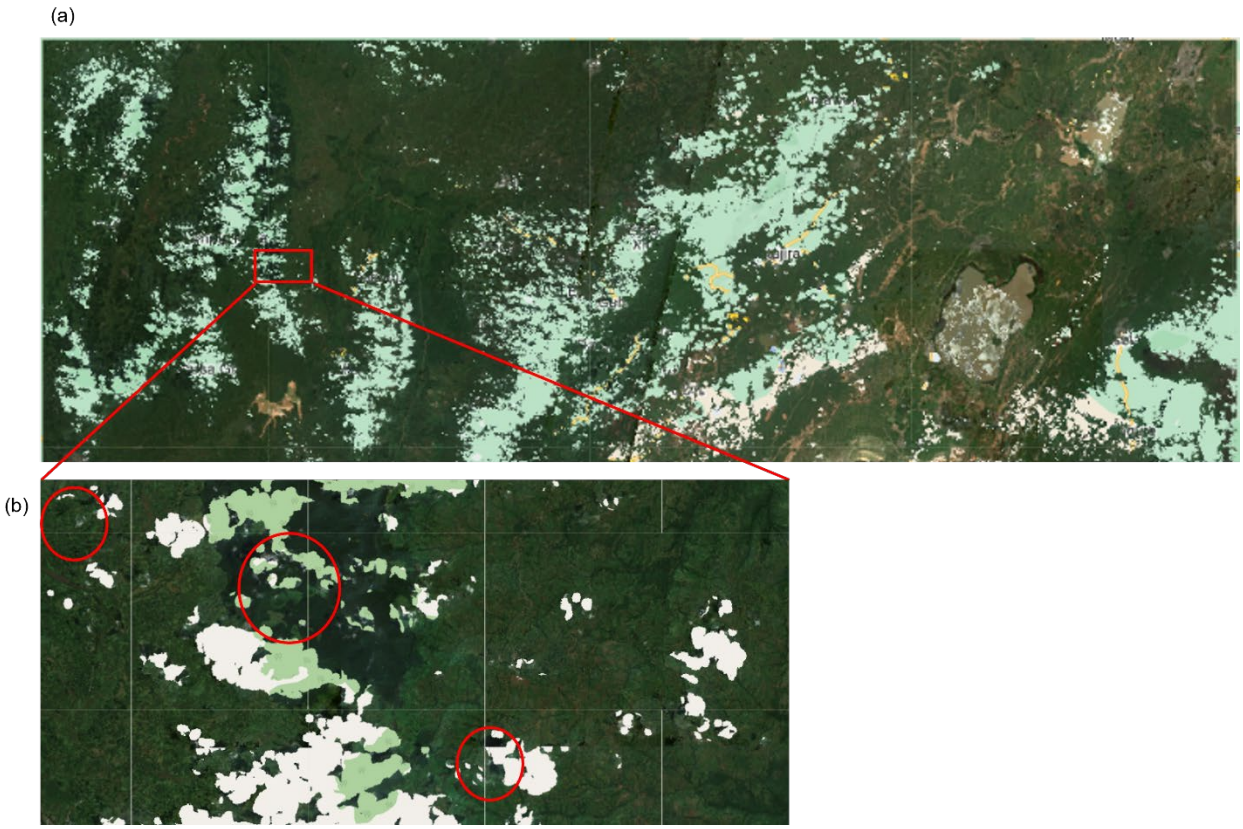

Figure S4. (a) Sentinel-2A 30-day image composite (9/15/2021-10/15/2021) rgb display (B4, B3, B2) and (b) detail with haze and cloud omissions circled in red.