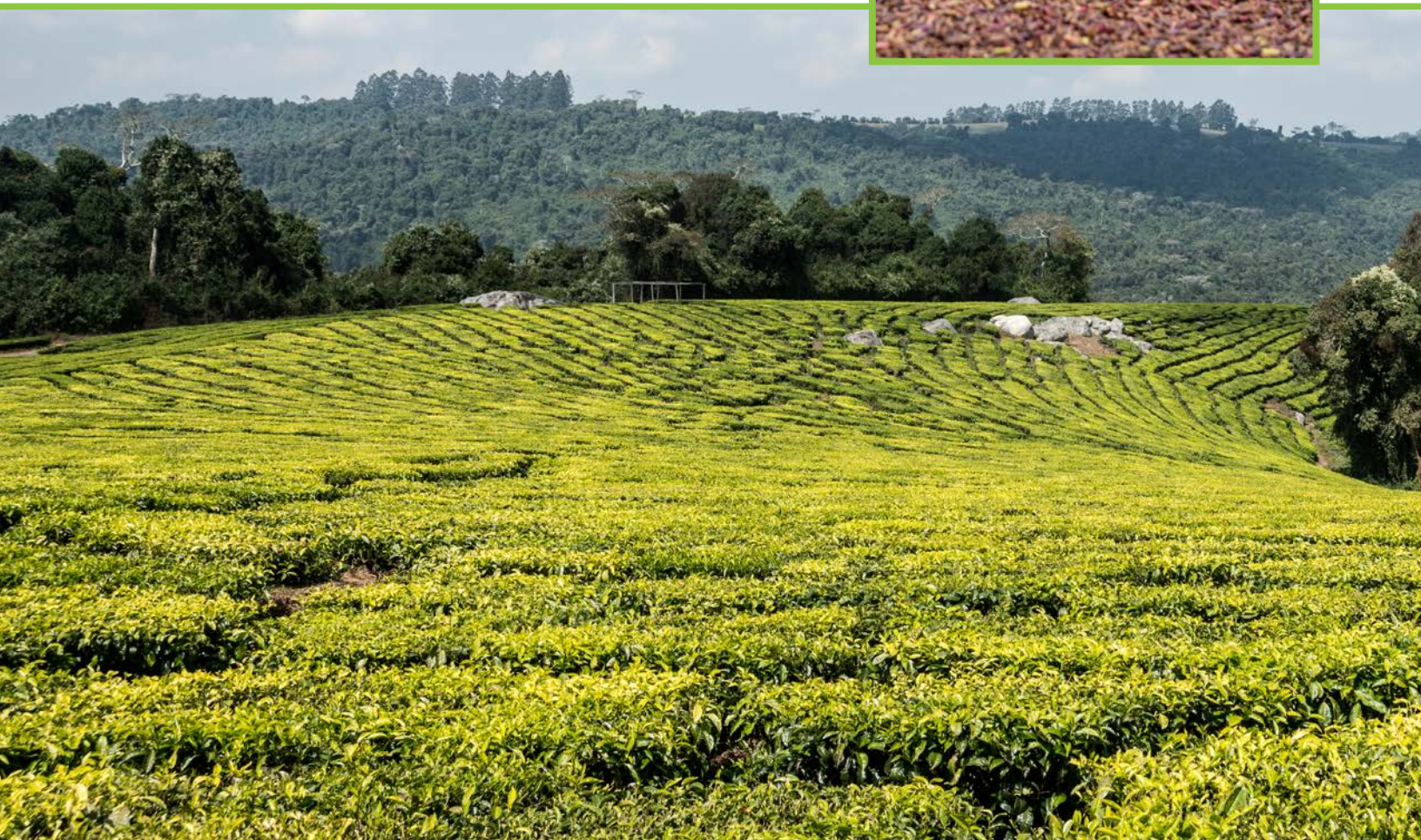
TETRA TECH

Enabling Crop Analytics At Scale

ENABLING CROP ANALYTICS AT SCALE (ECAAS)

Creating Next Generation Field Boundary and Crop Type Maps

Rigorous Multi-Scale Groundtruth Provides Sustainable Extension Services for Smallholders



Contents

1

Overview

A key challenge to providing improved extension services to smallholder farmers is the lack of accurate data on smallholders farms, including such basic information on where fields are and what crops they are growing. These data are necessary for developing and providing reliable, informed services. For example, an important set of recommendations that Farmerline makes to farmers is what seeds they should buy and how much fertilizer they should apply. As we often deliver such information through automated cell phone messages sent in response to SMS requests sent by farmers, it is important to know which crops are growing in each requester's region, otherwise the information we send can be irrelevant or misleading, resulting in lower demand for our services or even bad outcomes for farmers. Maps showing what crops were growing throughout our service region would enable us to target responses appropriately. Satellite-based analyses can provide such information, as new satellites and the growing capabilities of machine learning models make it increasingly possible to make accurate agricultural maps. However, a major obstacle to our ability to develop satellite-based agricultural maps that we can use to deliver reliable information to farmers, is the lack of ground-truth data, which are observations collected on the ground showing which crops were growing during a specific season in a representative sample of fields. These groundtruth data are essential for training the models that identify specific crops within satellite images, but are very hard and expensive to collect.

Our project sought to address the challenge of collecting ground truth while at the same time developing a capability to develop reliable and timely agricultural maps. To do so, we developed and tested innovative methods for collecting ground truth data, and integrated these with advanced machine learning and new satellite image sources to create improved maps of field boundaries and crop types. Our objectives were 1) to demonstrate the effectiveness of these methods at large-scale and over diverse geographies, 2) to make publicly available the data and methods to the broader crop analytics community, and 3) to use the resulting maps and methods to build improved services to our customers. This report provides a summary of our approach and the key methodological findings in this project.

Our open-source framework allows for future innovation in the development of EO solutions for increasing food security, which can be extended and scaled to other regions in Africa through the PDTT and our user network. Specific objectives of the framework include the demonstration of an end-to-end workflow comprising the following steps:

2

Approach

To achieve these objectives, our project had three major areas of methodological focus:

1. Creation of representative sampling strategies through the combination of the newly created continental DE Africa crop mask dataset and unsupervised machine learning.
2. Developing techniques that enable groundtruth to be collected over large areas for low cost, and which improve the resulting data's effectiveness for training and evaluating mapping models;
3. Developing models that can map crop types more accurately, and which are more transferable between regions and seasons, in order to minimize the amount of new labels that need to be collected;
4. Creating annual, country-scale maps of crop fields.

Our study included two regions in Ghana and one in Northern Tanzania **(Figure 1)**. In Ghana, our project areas covered 62,000 km^2 in northern Ghana and 21,600 km^2 in central Ghana, stretching between Ejura/Sekyedumase district in the east and Tain district in the west (the Ejura-Tain focal region). The study area in Northern Tanzania spans 213,000 km^2.

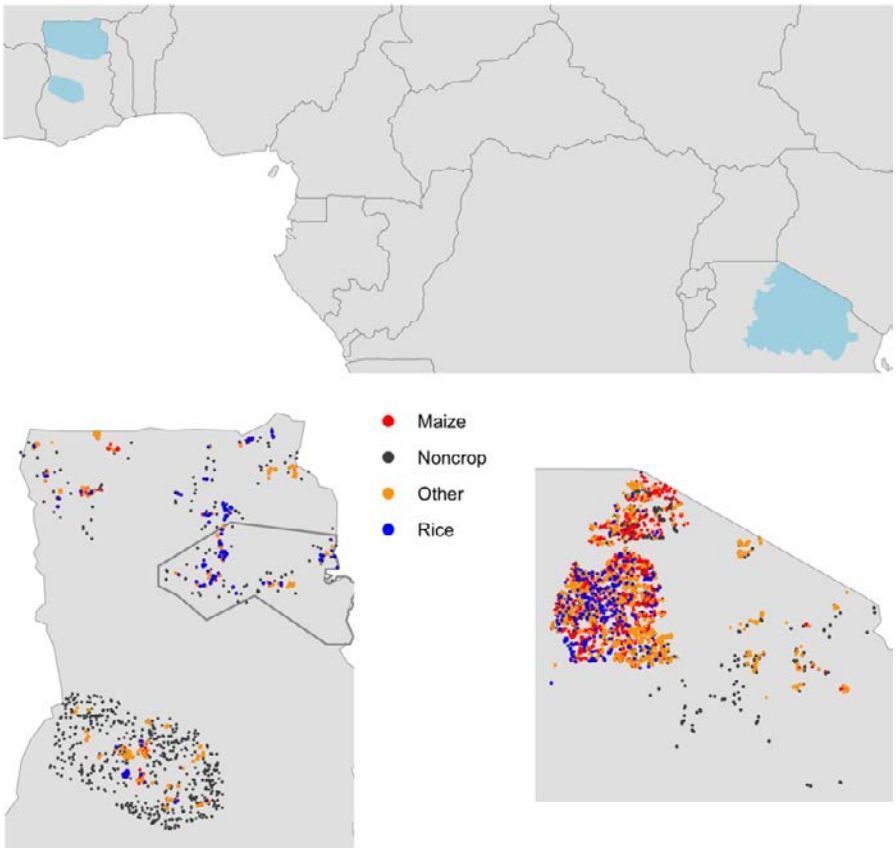


Figure 1:
The three primary focus areas of the project (blue regions in map on top row), with the distribution of collected labels in each region shown by crop type (bottom rows) . Plots in Ghana were collected as part of this project in 2021, points in Tanzania were gathered from public data sources, and represent crop observations collected in 2018 and 2019 **(see Table 1)**. The grey outline in Ghana shows the domain in which labels provided in a spatially de-referenced public dataset were collected in the year 2016 **(Table 1)**. Points on maps are a 10% sample of each dataset, to prevent visual crowding.

2.1. Groundtruth data and methods

We developed three methods for collecting groundtruth data for crop types (hereafter labels[1]). We grouped the resulting labels into 3 classes according to the method used to develop them, and which vary in their accuracy and suitability for use **(Table 1)**.

Class 1 labels were collected by Farmerline's agents on the ground during visits to farmers within our extension networks. Agents recorded the boundaries of individual fields and the crop types growing in them using the GPS-enabled Mergdata platform, following data collection and cleaning protocols established during our previous project. These labels were collected during 2021 in two regions of Ghana **(Figure 1)**, resulting in 3,065 cleaned labels, with 1,146 for maize and 470 for rice, 1,449 representing 24 other types **(Table A1)**. We also collected 2,221 non-cropland points through visual analysis of PlanetScope imagery.

Table 1:
Description of the crop type label datasets used in this project.

		Method/Source	Key Characteristics
Collected under this project	Class 1	Boundaries around fields and crop type collected on ground by Farmerline agents using the Mergdata platform	The most accurate and information rich class of label, but most expensive to collect, and does not conform to design requirements for reference samples
	Class 2	Labels interpreted from drone-collected imagery	Samples cover a large area and are collected following a probability design. Many crops can be accurately labeled in the imagery, but accuracy is lower than Class 1 labels and many smaller crops cannot be identified.
	Class 3	Model-generated labels, based on highest confidence crop type predictions from Random Forests model	Generates large numbers of additional training samples at low cost, but label quality can only be indirectly assessed through impact on model performance. Not suitable for validation/reference.
Public lables	-	Stanford University, accessed through Radiant MLHub	Ground-collected field boundary data from 2016 (Northern Ghana), rasterized with spatial references removed. Sentinel-1, Sentinel-2, and PlanetScope predictors are provided (Rustowicz et al, 2020)
	-	Great African Food Company, accessed through Radiant MLHub	Field center points providing crop type observations, converted to polygons using image interpretation (Great African Food Company, 2019)
	-	Tanzania Soil Information Service	Two datasets: 1) Crop scout data: point locations with details on crops growing within ~11 m radius; 2) geo-located maize trial data (Walsh et al, 2018)

[1] Labels are annotations (collected as GPS records on the ground, or as digitized polygons or points on an image) that define what crop is growing in that area, which are then combined with imagery to train and validate models.

To overcome limitations in the Class 1 labels (cost of collection, inadequate representation of crop types, limited spatial coverage), we developed a two-stage sample design in which drones (operated by Africa Surveys and Imaging Systems) were used to collected 3 cm resolution imagery over 194 pre-selected 550X550 m sites **(Figure 2)**. We then digitized the crop types visible in the imagery to create **Class 2 labels**. A subset of the data were collected over Class 1 labels **(blue points in Figure 2)**, and we used this overlap to assess the accuracy of Class 2 labels. We also tested whether convolutional neural networks could be used to effectively classify crop types in the drone imagery, and thereby automate the creation of Class 2 labels **(section 3.2 and B2)**.

In addition to Class 1 and 2 labels, we created synthetic **Class 3 labels** (**Sections 2.2 and B.1)** using the predictions from a crop type model. We also collected several freely available **public label** datasets, one covering Northern Tanzania for the years 2018 and 2019 and the other Northern Ghana for the year 2016 **(Figure 1)**.

We combined Class 1, 2, and 3 labels for model development and analysis. We used Class 1 labels primarily for model training and to assess the accuracy of Class 2 labels. We used Class 2 labels to develop validation samples for assessing model performance, as their design made them more representative of crop distributions. Class 3 samples were used exclusively to boost the size of model training samples, while the public labels were used to develop and evaluate deep learning-based crop type models.

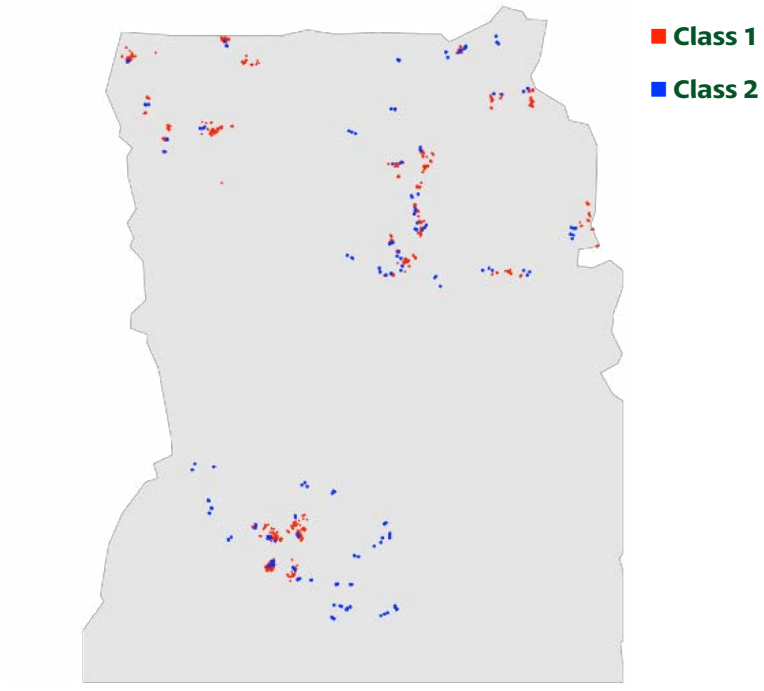


Figure 2:
The distributions of Class 1 and Class 2 labels collected during 2021 campaigns.

2.2. Crop type mapping

We developed and tested two different approaches for improving the quality of satellite-based crop type maps. In the first approach, we combined Class 1 and 2 labels to train Random Forests models with predictors derived from Sentinel-1, Sentinel-2, and PlanetScope imagery, and extracted the models' highest confidence predictions to create the Class 3 labels. We then tested if the models' performance improved after re-training with Class 3 labels. Our goal was to assess whether synthetic labels improve crop type maps, while reducing the need to collect labels.

In the second analysis, we evaluated two different neural networks that were trained to classify crop types within high frequency time series of Sentinel-1 and Sentinel-2 imagery. The two models, a Long Short-Term Memory Network (LSTM), and a temporal convolutional neural network (tempCNN), have been previously shown to outperform Random Forests for crop type classification. They may also be more easily transferred to new regions or time periods, which can help minimize future label collection effort. To evaluate model performance and transferability, we developed four datasets representing different regions (three in Ghana, one in Tanzania), as well as a global dataset containing labels from all four regions. We trained models using each of these datasets, and evaluated their performance 1) within their own region, 2) when applied to each of the other regions, and 3) within each region (other than itself) after fine-tuning on that region's labels.

2.3. Annual, large-scale mapping of field boundaries

In addition to mapping crop types, we also refined and implemented methods for mapping crop field boundaries. Field boundary maps provide crucial information on where crops are growing and in what type of fields they are being grown in (e.g. small-scale or large-scale commercial), and can improve the accuracy of crop type models by acting as a filter that focuses their predictions within likely fields. In Ghana and other smallholder-dominated agricultural systems, field boundaries shift frequently within and between seasons, therefore field boundary maps need to be updated on an annual basis to track active fields. A key goal in our project was to demonstrate the ability to make annual, fine-grained field boundary maps at large scales. To do this, we adapted a previously implemented version of a Unet model designed to work with high-resolution (~3.7 m) PlanetScope imagery, and used it to create updated maps of Ghana for the years 2019, 2020, and 2021, and for Tanzania for the years 2017 and 2018.

3

Key Findings

3.1. Drone-based labels improve model training and assessment

We found that using drones to collect imagery for labeling (Class 2) offered a number of advantages relative to purely ground-based (Class 1) label collection:

- Using drones enabled us to rapidly collect data over a larger region than we could cover on the ground **(Figure 2)**;
- The relatively large footprint (>30 ha) covered by each of the 194 collected images allowed us to generate a more labels covering a greater variety of crops than we could gather on the ground;
- As the drone imagery was collected using a probability design, the resulting labels were more representative of crop type distributions in our regions;
- Having a larger, more representative sample enabled us to more effectively train and assess the performance of crop type models.

However, Class 2 labels have several disadvantages compared to Class 1 labels:

- Their crop type definitions are less accurate **(see section A.2.)** than those of Class 1 labels, particularly for shorter crops, those with narrower canopies, or very fine leaves (e.g. carrots), which are hard to recognize in the drone images;
- Labeling drone images is time-consuming, requiring several hours per image.

Given these benefits and limitations, Class 2 labels should be used primarily for labeling larger, easier to recognize crops such as maize and rice.

Details

Class 2 labels were collected from sites that were pre-selected based on a two-stage, probabilistic sampling design. We digitized a total of 9,070 Class 2 polygons from the collected drone images (link to Class 2 label report), of which 4,600 had recognized crop types **(Table A2)**. These labels were collected over broader geographies than those cov-ered by Class 1 labels within our two focus regions **(Figure 2)**, within a relatively short period of time (3-4 weeks per region) by a small number of personnel. In addition to increasing our reach, we were also able to substantially increase the sample sizes for key crop species, more than doubling our maize and rice observations, while obtaining information on other crops that were not priorities for Class 1 collection (e.g. bananas, sorghum).

Class 2 labels provided a relatively unbiased sample of crop type distributions, which allowed us to create reference labels that give us a more accurate understanding of model performance. Ironically, this gave the appearance that our model became less accurate over time; our first Random Forests' map developed for Ejura-Tain, using only Class 1 labels for reference, achieved overall accuracies of 70-76%, compared to 62% in our latest version that we trained and evaluated with Class 1 and 2 labels. The reason

for the higher accuracy in the earlier version is almost certainly because reference labels were more clustered and in closer proximity to training labels, and therefore more correlated, thereby inflating accuracy estimates. The greater geographic reach of Class 2 labels resulted in maps that were less clustered and presumably better reflected actual crop distributions in each region.

Class 2 labels have several downsides compared to Class 1 labels. One of these is that the labels themselves are less accurate, as interpreting crop types, even in near-surface, high-resolution imagery, is difficult. This was made evident by a comparison between Class 1 and Class 2 labels in areas of overlap **(Figure 2)**, and by the uncertainty evident when multiple labellers mapped the same sites **(Figure 3)**. The uncertainty, and thus error, was highest for smaller crops (those growing lower to the ground and/or with narrow canopy cover).

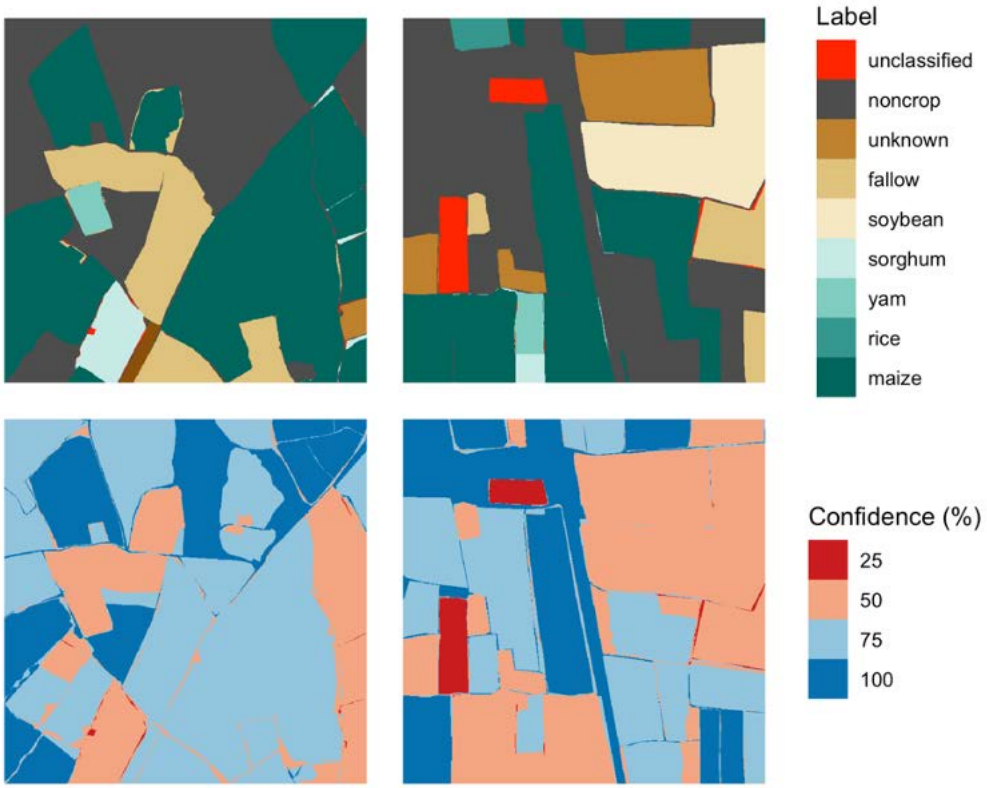


Figure 3:
Two of 8 sites where Class 2 labels were developed by 4 labelers, showing the consensus label (the most frequently labeled class) on the top row, and the between-labeler agreement expressed as the percentage choosing the class (e.g. 100% = all 4 labellers agreed; 25% = all labellers disagreed).

Class 2 labels are therefore best suited for developing labels for larger or more widely and uniformly planted crops that are easier to identify in imagery, such as maize and rice, rather than minor crops, such as tomatoes or leaf vegetables.

Another disadvantage of Class 2 labels is that they are extremely time-consuming to develop. A complete labeling of one orthophoto can take anywhere from 1-4 hours, depending on the complexity of the cropland in the scene. Assuming an average of 2 hours per scene, that equates to 380 person hours (9.5 weeks) to label all 190 scenes.

The accuracy and speed with which labeling can be completed can be improved (and there is some potential for automating label generation; see section 3.2), however, and the error rates, while preventing a full accounting of map accuracy, are unlikely to unduly bias accurate assessments because labeling errors are likely to be random within each crop type.

3.2. Drone-based labeling can be automated

We adapted Unet, the convolutional neural network we use for field boundary mapping, to develop models that could be used to automate the creation of Class 2 labels, and thereby reduce the large effort required for manual image labeling. Our initial experiments found that:

- The model showed moderate ability to classify maize, rice, and noncrops (f1 = 0.62-0.69) in 40 cm resolution drone imagery, but poor performance for other classes;
- Performance of the model and the coherence of predicted labels improved when simplifying the number of classes used to train the model from 9 to 4;
- Label error was a major factor limiting model performance.

Details

We used Class 2 labels to train three versions of Unet to predict crop types within the drone imagery we collected (Class 2 label modeling report). The two most promising versions were a full Unet trained on 9 classes (maize, rice, sorghum, legumes, other crops, tree crops, fallow, unidentified, and noncrop) derived from Class 2 labels, and one trained with 4 simplified classes (maize, rice, other crops, and noncrop). Both models achieved moderate performance for maize, rice, and noncrop. The 9-class model performed poorly in predicting all other classes. The 4-class model outperformed the 9-class model for the 3 common classes, with slightly higher scores for maize (F1 = 0.69 versus 0.63 in the 9-class model) and rice (F1 = 0.62 versus 0.60), but with substantially larger gains for the noncrop class (F1 = 0.65 versus 0.51).

The 4-class model also showed a large gain in the other crops class (F1 = 0.56 versus 0.26), but the classes were not identical between the models (other crops in the 4-class also included sorghum, legumes, unidentified, fallow, and tree crops). The 4-class model produced the highest quality maps with the greatest class coherence **(Figure 4)**.

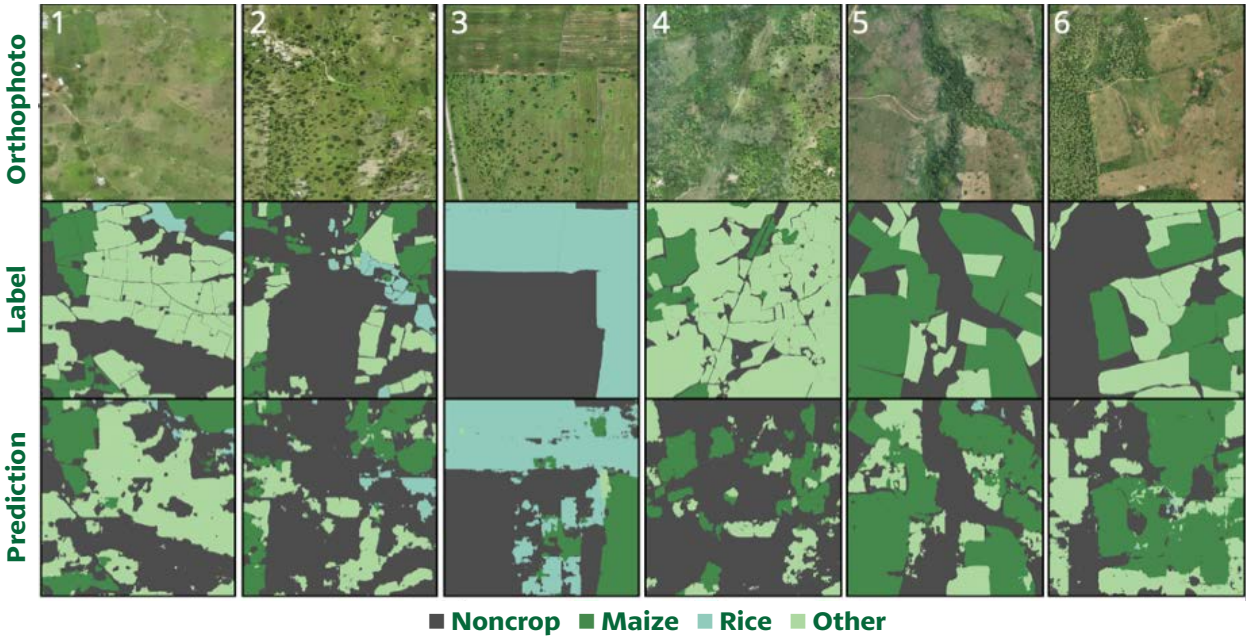


Figure 4:
A selection of 6 of 94 collected drone images (top row) used for model validation samples, their corresponding Class 2 labels (middle row), and the predictions for those labels generated by a Unet model (bottom row).

The low to moderate performance by these models is due in large part to errors in the Class 2 labels, rather than model inadequacies. The classes for which the model performed most poorly were the ones that were hardest to label (e.g. legumes, including groundnuts, soybean, etc) and thus most error-prone. Confusion within the training data undermined the 9-class model's performance, while errors within the validation/reference sample made it difficult to reliably assess model performance for certain classes. For example, areas where tree crops were either unlabelled or mislabelled, but were correctly predicted by the model, were counted as model commission errors. Combining these harder-to-label, more erroneous classes, into a single super class (the other crops class in the simplified 4-class model) helped to improve model performance.

Given these results and the capability demonstrated in a similar previous study that applied a CNN to map crop types in drone images (Chew et al., 2020, who seeded the idea for our analysis), we believe there is good potential to improve these results and develop the ability to automate Class 2 label creation for major crop classes.

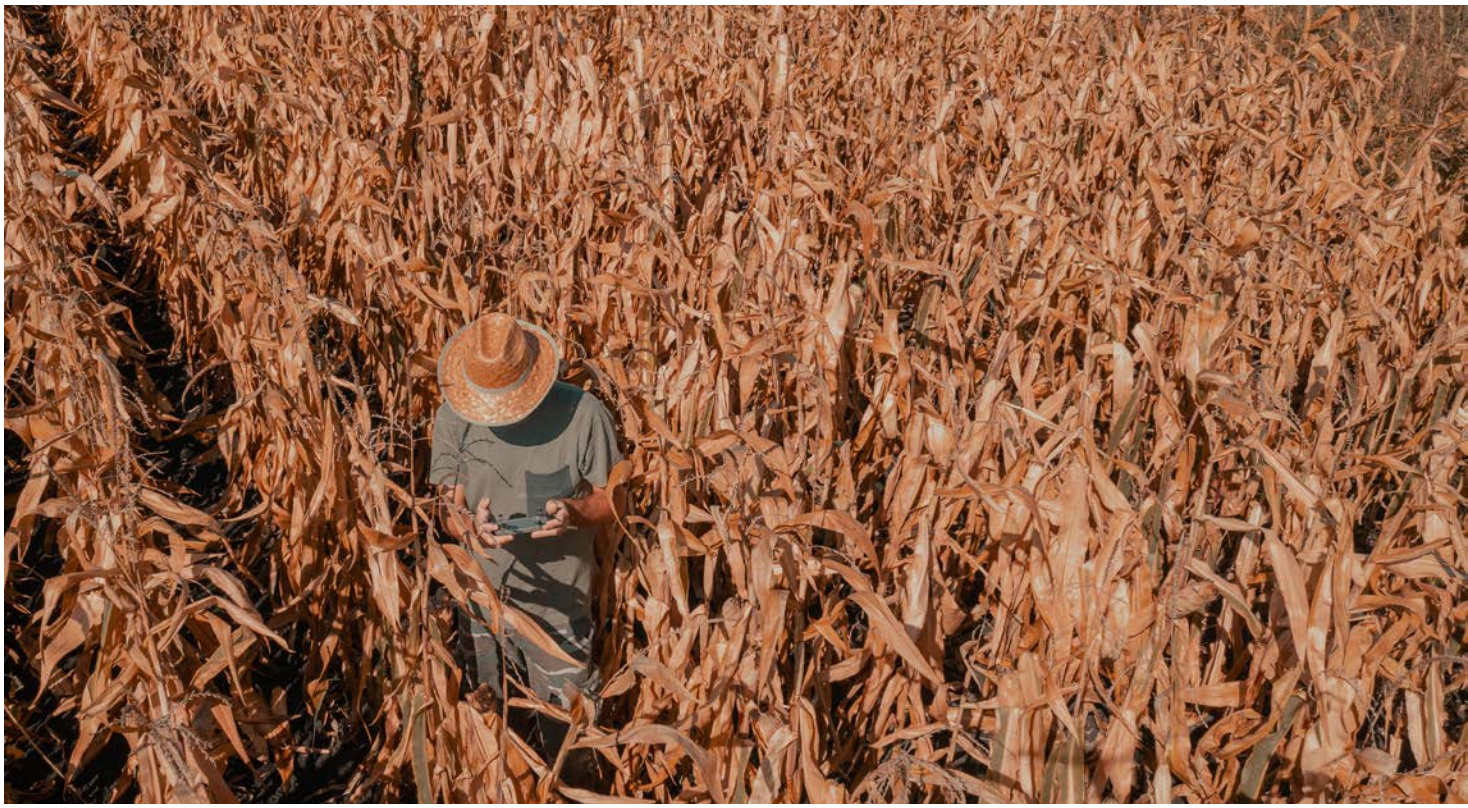
Improving the overall accuracy of Class 2 labels, even just a representative subset of the labels, may provide the most immediate potential gains. Reducing the labels into binary sets (e.g. maize/non-maize; rice/non-rice) and using these to train a series of separate models may further improve performance for each of the major crops, provided the process for merging results does not undermine the gains. Testing other semantic segmentation architectures (e.g. DeepLab; Chen et al., 2018), or adopting an image-classification approach, such as the one used by Chew et al (2020), may also yield better results.

3.2. Synthetic labels improve model performance

Comparing results from two Random Forests models, one trained with just Class 1 and 2 labels, the other trained with the same set of labels plus model-generated labels, demonstrated that synthetic Class 3 labels:

- Improved model performance in Northern Ghana, the largest focal region, while leaving it unchanged in Ejura-Tain;
- Improved the relative distributions of crop types in both regions, by reducing obvious false positive errors while making the relative share of each crop type more realistic **(Figure 5)**.

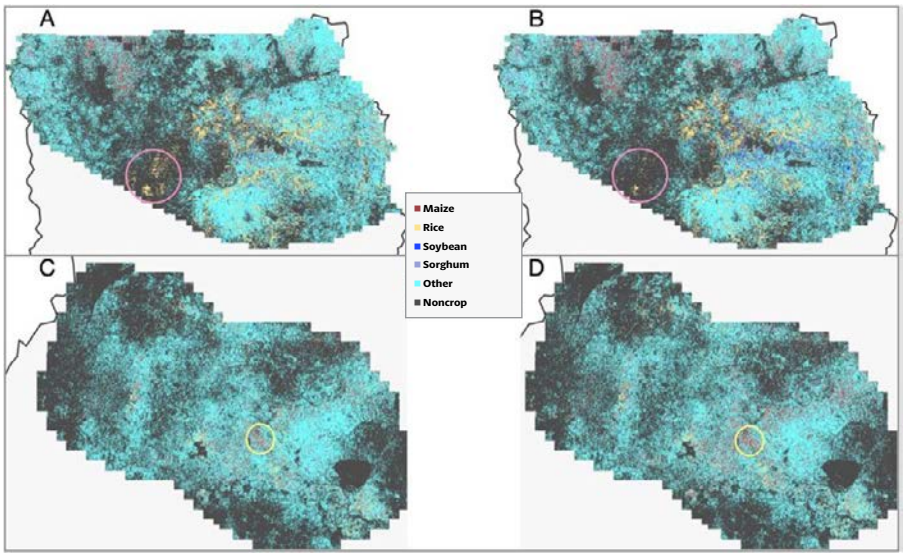
These results provide further evidence (Alemohammad, pers. comm., 2022) that synthetic labels provide a low-cost, effective means for improving crop type maps. However, because their identity cannot be verified, they should only be used in training samples, not for model validation.



Details

We developed Class 3 labels from crop type maps produced by Random Forests models (one for each region) trained using Class 1 and 2 labels, using predictors drawn from Sen-tinel-1, Sentinel-2, and PlanetScope imagery **(Class 3 Label Report)**. The model we developed for Northern Ghana showed gains in both User's accuracy (or precision, which factors in false positive error for a particular class) and Producer's accuracy (or recall, which accounts for false negative error) across nearly all 6 modeled crop types (maize, rice, soybean, sorghum, other crops, noncrop), raising the former by an average of nearly 3 percentage points (pp), and the latter by 1.6 pp, while overall accuracy increased by 2 pp. In addition to these gains in accuracy, the resulting crop type map showed noticeable improvements in the mapped distributions of crops, reducing, for example, an obvious over-prediction of rice **(Figure 5)**.

Figure 5:
Crop type maps resulting from the initial Random Forests models for northern Ghana (A) and Ejura-Tain (C), compared to those trained with additional Class 3 labels (B and D). Rose circles highlight an area of pronounced reduction in predicted rice crop distributions in the Class 3-enhanced model (B) relative to the initial model (A), while yellow circles indicate an area where the Class 3-enhanced model for Ejura-Tain (D) showed higher maize concentration compared to the initial model (C).



The Ejura-Tain model, which produced relatively few Class 3 labels, showed no net gain in accuracy—improvements in some classes were offset by declines in others—but nevertheless showed similar patterns of improvement in crop spatial distributions within the updated map, and brought the proportion of maize relative to rice more in line with expected values in the region **(Figure 5)**.

3.4. Neural networks improve the quality and transferability of crop type models

Of the two neural networks we developed for crop type mapping, the temporal CNN showed the best performance and greatest ability to generalize, and we therefore used it for all model experiments. Our results revealed several advantages of this modeling approach:

- The tempCNN had slightly higher performance than Random Forests models.
- Although models trained for one region could not be transferred directly to another, fine-tuning using training labels from the target region enabled successful transfer, resulting in performance that was close to that of a locally trained model.
- A fine-tuned global model (one initially trained with subsets of labels from all regions) showed the most promising results, as it slightly outperformed or achieved parity with the local model in most cases. The resulting maps were qualitatively better than those of the locally trained models.
- This improved transferability suggests that a "train global, refine local" strategy could minimize overall label collection efforts. Publicly available labels play a crucial role in enabling the development of global models.

The results also revealed several disadvantages to using these models:

- The image processing and model prediction pipelines are more complex and much slower than they are for Random Forests;
- The models were sensitive to randomization processes that occur during model training, producing substantially different performance metrics and maps when between otherwise identical runs.

Details

Initial comparisons of model performance showed that the LSTM overfit the data on all five datasets, while the tempCNN learned less deeply but generalized better, outperforming the LSTM in nearly every dataset and all four modeled classes (maize, rice, other, noncrop) (Crop Type Model Assessment Report). We therefore used the tempCNN for subsequent transfer experiments.

We found that region-specific models—those trained on a label set from a specific region—for Northern Ghana and Tanzania performed moderately well, with average F1 scores across the four modeled classes (maize, rice, other, noncrop) of 0.52–0.68 **(Table 2)**, with best performance for rice in Northern Ghana, and for maize, other crops, and noncrop in Tanzania. The model was most effective for Tanzania, followed by Northern Ghana (using labels we collected), and less effective for Ejura-Tain (mean F1 = 0.54), a cloudier region with a more imbalanced and clustered sample for rice, which the model predicted least accurately. Unexpectedly, given the overlap with our Northern Ghana focal region **(Figure 1)**, the model performed worst when trained with the public labels from Northern Ghana. The global model trained and validated with data drawn from all four regional data had the lowest performance (mean F1 = 0.38).

Table 2:

The performance results of experiments evaluating the tempCNN trained and evaluated with different combinations of the five label datasets. Row names indicate the dataset used to train the model; columns indicate the dataset each trained model was evaluated against. The top half of the table provides results for the initial trained models' performance against its own validation set (bold values on the diagonal) and the validation sets for the other three regional datasets (non-bold, off-diagonal values). The bottom half of the table contains results from the same models after "local" fine-tuning, where each model was fine-tuned on and assessed against another region's labels. Underlined bold values show where a fine-tuned model tied or outperformed the original local model. Values provided are the average F1 scores for maize, rice, other, and noncrop.

		N. Ghana	Ejura-Tain	N. Ghana*	Tanzania*	Global
initial	N. Ghana	**0.61**	0.25	0.10	0.14	-
	Ejura-Tain	0.23	**0.55**	0.12	0.22	-
	N. Ghana*	0.14	0.17	**0.52**	0.13	-
	Tanzania*	0.19	0.27	0.21	**0.68**	-
	Global	0.41	0.38	0.16	0.50	**0.38**
Fine-tuned	N. Ghana	-	0.54	0.45	0.67	-
	Ejura-Tain	0.58	-	0.41	0.65	-
	N. Ghana*	0.52	0.51	-	0.64	-
	Tanzania*	**0.62**	0.53	0.46	-	-
	Global	**0.61**	0.54	0.48	**0.68**	-

*Indicates the two publicly sourced label datasets.

Directly transferring models trained in one region to another had fairly poor results (mean F1 = 0.23), as might be expected, although the global model showed moderate effectiveness against the Tanzania dataset (mean F1 = 0.50).

Experiments showed that fine-tuning a model trained for one region with labels from the target region achieved performance that came close to, but was generally slightly lower than, that of the local model. The one exception to this was the Tanzania model fine-tuned for Northern Ghana, which scored 0.01 higher than the local model. However, local fine-tuning of the global model produced the most promising results **(Table 2)**, slightly outperforming the local model for Northern Ghana, achieving parity with the Tanzania model, and being just 0.01 lower than the Ejura-Tain model.

Refining the global model appears to be most promising not only because of the validation scores, but also because of better quality in the prediction maps, which we assessed at two scales: a broader tile scale 5 X 5 km; **Figure 6**), using qualitative assessment, and the finer scale of our standard labeling grid (550 X 550 m; **Figure 7**), where they could be assessed against reference labels. Although the quality of prediction are poorer at this finer resolution (as with most maps), the global model, both refined and unrefined, corresponds slightly better than the others to the labels (the F1 of this five-site comparison is highest for the two global models), and is better than our Class 3-enhanced Random Forests model at distinguishing non-cropland from crop types.

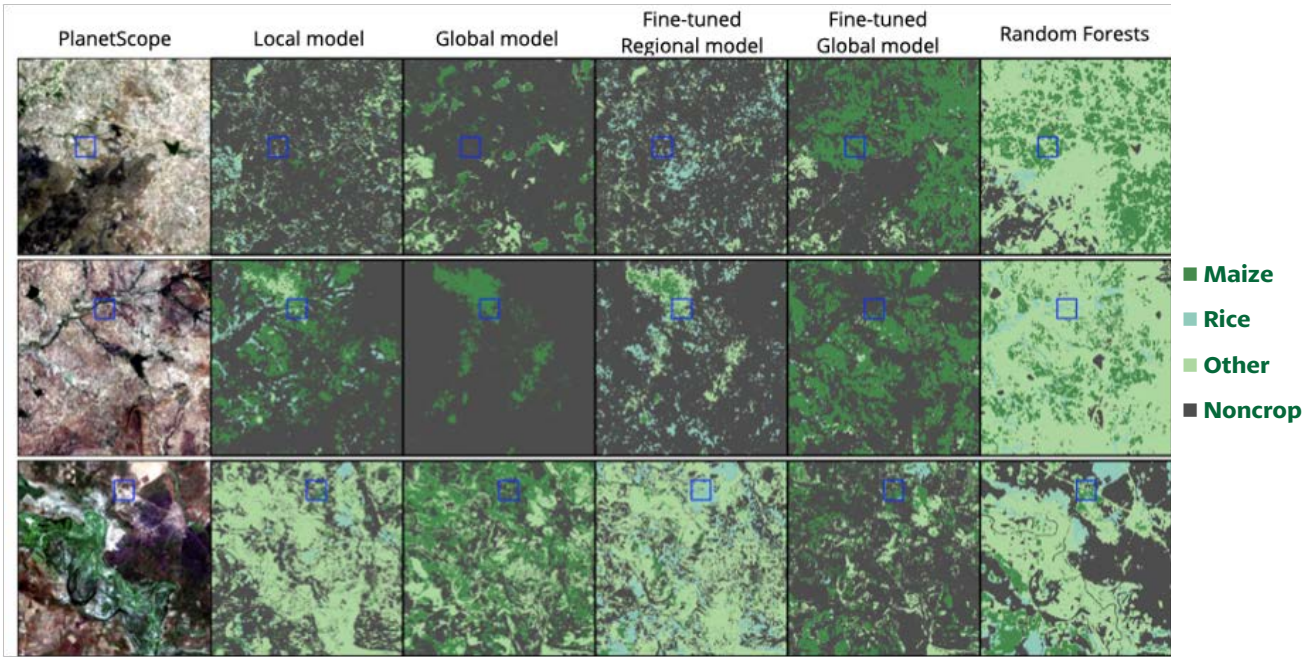


Figure 6:
Mapped crop class predictions for three selected sites in Northern Ghana, shown at the tile scale (5 X 5 km) with a November, 2021 Planet-Scope basemap image provided to give landscape-level context. Column headers refer to different training approaches: Local = tempCNN trained on Northern Ghana labels; Global = model trained with the global label set; Fine-tuned Regional = Tanzania model refined on Northern Ghana; Fine-tuned Global = global model refined on Northern Ghana; Random Forests = predictions for the same locations produced by a Random Forests enhanced with Class 3 labels **(see 3.2)**. The blue squares indicate the location of grid-scale comparisons shown in **Figure 7**.

These results therefore suggest that the temporal CNN improves the quality of crop type predictions, and because of its transferability, has the potential to reduce label collection efforts. In terms of the former, although the performance scores are modest, they represent a slight improvement on earlier work. The tempCNN scored better than our Class 3 enhanced Random Forests models, which achieved average F1 scores of 0.56 and 0.53 in Northern Ghana (although predicting more classes) and Ejura-Tain, respectively, compared to 0.61 and 0.55 for our corresponding local tempCNNs. Our tempCNN for Northern Ghana also scored higher than a 2D-Unet+ConvLSTM that the authors of the Ghana public label dataset applied to those data (average F1 = 0.57; Rustowicz et al, 2019), which falls within the Northern Ghana region, and the tempCNN achieved close to that level of performance (mean F1=0.52) when applied to the same public labels.

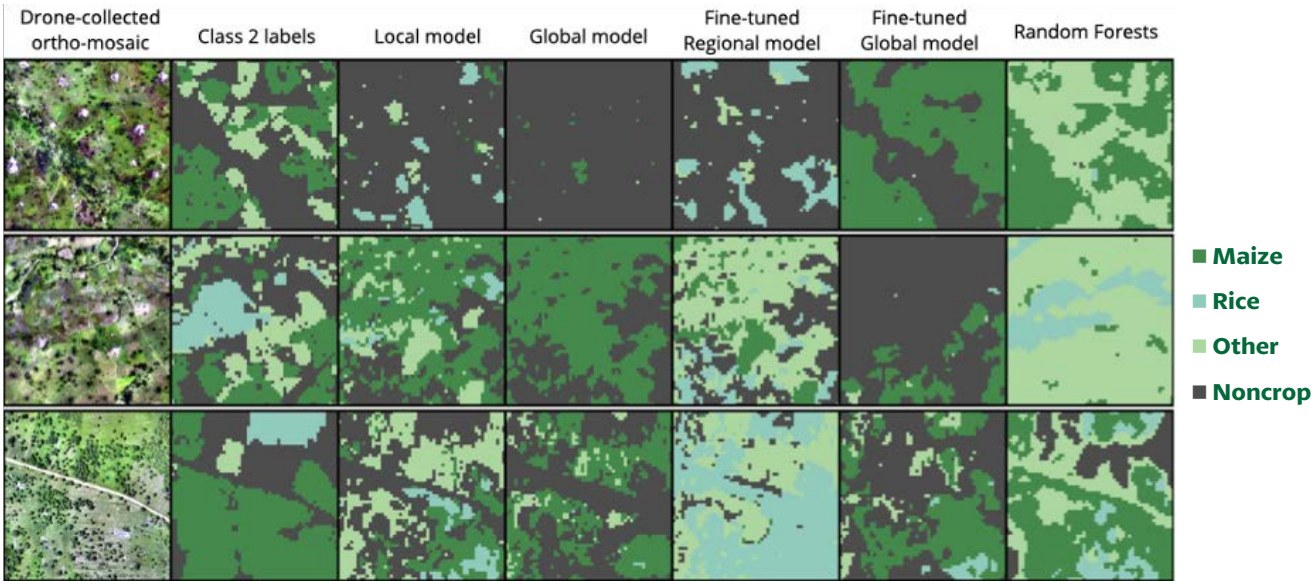


Figure 7:
Class 2 labels collected within a 550 X 550 m portion (blue box in left column) of each of the five selected 5X5 km map tiles in Northern Ghana, shown in relation to overlapping model predictions for the tempCNN trained using different label sets, as well as a Random Forests model (see **Figure 6** for definition of column headers).

In terms of reducing the amount of labels, the results from local fine-tuning of the globally model suggest that this model can be fine-tuned on new labels each year, with potentially fewer labels than are needed to train a new model from scratch. Although we did not test for required minimum sample sizes, other work with CNNs has shown that a pre-trained model refined locally can achieve good performance while greatly reducing the number of labels to map a new location (Wang et al., 2022), which is supported by our own prior results on this project.

Weighing against this modeling approach are practical concerns about implementation. The data processing required to make maps with the tempCNN is more complex than it is for Random Forests, and the speed of inference tasks much slower—as currently coded, 15-30 minutes is required to generate a single 5X5 km map tile with a GPU computing instance. Mapping a large region is therefore much slower and more expensive than for Random Forests, which runs quickly on much cheaper CPUs (all of Northern Ghana takes a few hours to map). Further effort is therefore needed to automate the image processing pipeline, and to improve speed of inference. On top of this, the models are sensitive to randomization effects that occur during training. Running the same model with the same parameters but a different random seed can produce substantially different results, with average mean absolute F1 differences of 0.029 across crops and maps with markedly different crop distributions.

A final major finding from our results is that they demonstrate the value of public datasets for model development. Although it required substantial pre-processing to use, the Radiant MLHub-hosted Stanford dataset afforded us the ability to start building models using a dataset from our region, at a time when we were still collecting our own labels, affording us more time for model development. The public labels also provided important building blocks for developing the more effective global model.

3.5. Updating field boundary maps updated

Using a variant of the Unet we used previously to map Ghana for the year 2018, we produced maps for all of Ghana for the years 2019, 2020, and 2021, and for Tanzania for the years 2017 and 2018 within the space of a few weeks, including the time needed to develop the image catalog. These results demonstrate the capacity to develop annually updated, high resolution field boundary maps over large areas.

However, despite initial high performance metrics, the model we initially used had a tendency to underpredict fields in Ghana and over-predict in certain areas in Tanzania. Switching back to an earlier version of the model reduced these problems, but over-prediction artifacts remain in some areas due to image brightness shifts, particularly in maps for earlier years. These remaining discrepancies are likely due to the image normalization procedure, and demonstrate the challenges that can arise when scaling up from test datasets to large area maps.

Details

We used a globally-trained, locally-refined densely-fused Unet to undertake the mapping, based on the model's higher performance than the standard Unet we had used to generate prior maps.

The refined models had a mean accuracy of 87.6% (standard deviation = 0.05%; range = 72.2% - 94.2%), an average true positive rate of 66.1% (sd = 9.1%, range = 46% - 81.3%), and an average false positive rate of 10% (sd = 4.9%, range = 2.7% to 23.9%). This performance was generally comparable to or exceeded that of the global model. However, the maps themselves showed several inconsistencies, notably a tendency to underpredict fields in Ghana, particularly in the year 2020 **(see Figure 8)**, and to over-predict in Tanzania (where accuracy was also lowest) on uncultivated barelands, along with image-related artifacts (patches of extreme false or false negatives positives). The over-prediction in Tanzania was primarily due to the relatively small number of labels used to refine the model relative to the very large mapping region (northern Tanzania is nearly as large as all of Ghana, which we divided into 16 sub-regions for refining), as well as the extensive barelands that occur in the region's grassland areas, which have similar reflectance to croplands. Under-predictions in Ghana were likely due to the choice of layers left free to vary when refining the model, while image artifacts appear related to image normalization.

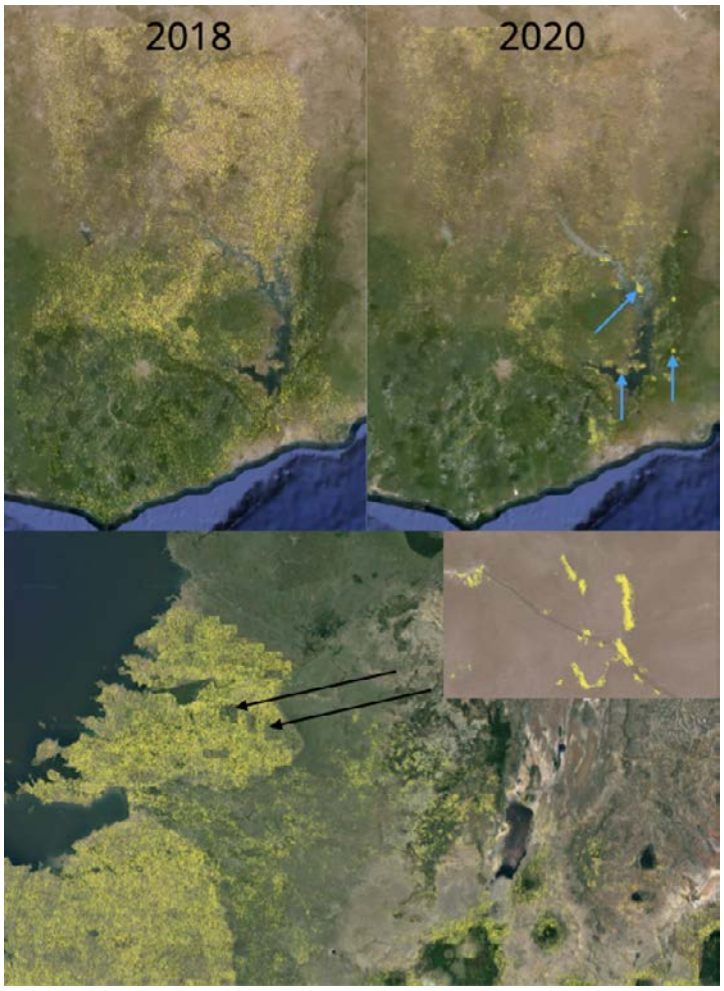


Figure 8:
A comparison of the 2018 Unet-based map of Ghana with the refined DFUnet for 2020 (top row), highlighting the relative under-prediction as well as patches of over-prediction due to image artifacts (indicated with blue arrows). The predictions for Northern Tanzania (bottom row) detect croplands effectively, but over-predict in sparsely vegetation grasslands (see inset). Black arrows highlight image under-prediction artifacts.

To address these issues, we tried freezing all but the last layer of the model when refining, but that did not measurably improve results. We have since reverted to using our original Unet, which substantially improved the under-prediction results **(see Figure 9)**, but over-prediction artifacts remain in several areas, particularly in earlier years (2019), where image brightness varies more often between tiles.

We are addressing these residual errors through changes to the image normalization technique, and will update the maps after these fixes are applied.

Our previous work has demonstrated the feasibility of making annually updated cropland maps at scale, therefore these findings do not show that the task is infeasible. Instead, they offer a key lesson, which is that algorithmic changes that appear effective on smaller test datasets often do not scale up to production.

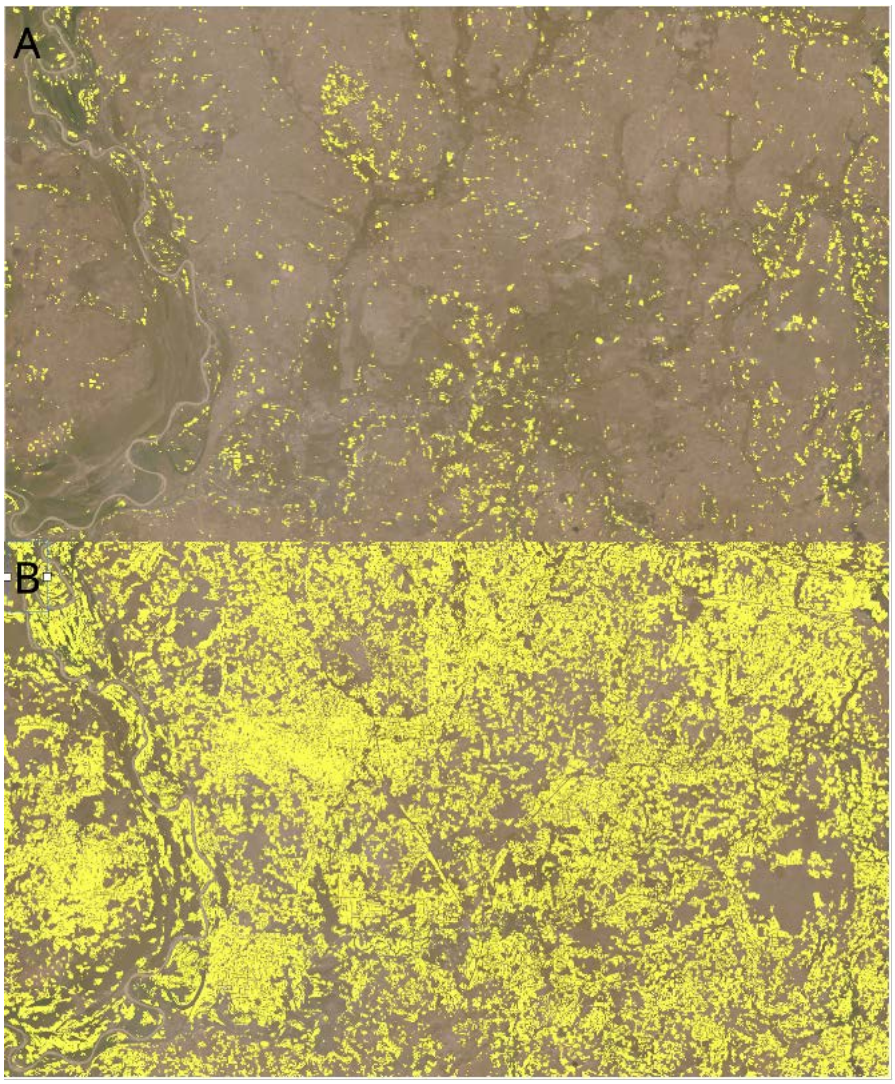


Figure 9:
A comparison between locally refined versions of DFUnet (A) and original Unet (B) on PlanetScope imagery for 2020. The DFUnet substantially under-predicted fields in dense cropland areas that were successfully detected by the original Unet. The refined Unet shows overprediction in certain areas due to image normalization problems.

4

Recommended improvements and next steps

Our findings suggest a number of improvements that could be made to the existing methods, as well as areas that merit further investigation and investment of effort.

4.1. Class 2 labels

Labeling imagery collected by drones helps reduce the cost of collecting groundtruth data, while improving its statistical properties. The following recommendations can further improve the quality of Class 2 labels:

- Capture imagery at the peak of the growing season, when crops are likely to be most recognizable in the images;
- Coordinate the collection of Class 1 labels and drone operations so that imagery is captured within 1–2 weeks after Class 1 labels are collected. Ensure that the overlapping Class 1 labels have sufficient examples of each key crop (at least 30 each) so that the accuracy of Class 2 labels can be reliably assessed;
- Collect imagery at two different levels. In addition to the current approach, in which the drone acquires complete image coverage of the 550X550 m sample cells at 3 cm resolution, several transects across the cell can be flown at lower altitude, providing partial scene coverage at higher resolution. The lower-level imagery will help labellers identify crops in the higher level imagery;
- Collect multi-spectral imagery that includes near-infrared bands, to improve both visual image interpretation and the ability of CNNs to predict Class 2 labels.
- Focus on improving the quality of existing Class 2 labels in a subset of scenes, in order to create a high accuracy reference label set and a smaller, high quality training sample that can be used to refine label-predicting CNNs;
- If and when independent and reliable crop survey data become available, compare the frequency distribution of crop types from Class 2 collected for the same season, in order to verify their representativeness.

4.2. Synthetic labels

Using models to generate groundtruth shows promise, both for labeling drone imagery (Class 2 labels), and to create additional training labels for satellite-based crop type models (Class 3 labels). To improve the capacity to create model-generated Class 2 labels:

- Test different training strategies for the existing Unet-based approach. For example, evaluate whether separate binary models for each crop (e.g. maize/non-maize) improve performance;
- Evaluate different model architectures and modeling approaches tested, such as image classification approaches (e.g. Chew et al, 2020)

For satellite-based mapping, further effort should be made to test whether Class 3 labels can help to improve the performance of the tempCNN.

4.3. Crop type models

Neural network-based crop type models demonstrate the potential to be more effective than Random Forests, and may help reduce overall label collection efforts. To fully understand this potential, the following additional tests should be undertaken:

- Re-evaluate the model on data collected from the coming season;
- Perform label reduction tests to identify the minimum sample needed for model fine-tuning, and compare that to the minimum sample needed to train a Random Forests model of comparable effectiveness;
- Improve the quality and composition of labels used to train the model, by rebalancing the global training sample, and filtering labels so that they represent purer examples of each crop type.
- Make the image processing pipeline and inference procedures for the tempCNN more efficient;
- Explore further architectural improvements, including enhancements that improve the model's ability to learn in the presence of clouds (e.g. , and the potential improvements that can be obtained by variants that learn from both spatial and temporal information (e.g Rustowicz et al, 2020).

4.4. Field boundary maps

The field boundary maps are in the process of being updated to deal with the image artifacts, and the maps will be released once these are satisfactorily addressed. Subsequent improvements will include:

- Developing labels on imagery for each of the years being mapped, in order to improve fine-tuning performance;
- Similar to the Class 3 labeling approach, we are using the model's predictions to help generate new labels.

5

Conclusion

This project developed several methods that have advanced our ability to map and analyze croplands, enabling us to develop improved services we provide to our customers. Our findings also make several contributions to the broader field of crop analytics. First, our results for Class 2 labels demonstrate the value of combining drones and innovative in-field data collection to create groundtruth that is less expensive and more effective for training and evaluating crop type models. Our key innovation, which builds on the drone-based labeling methods developed by Hegarty-Craver et al (2020), was the design and implementation of a probabilistic sampling technique, along with procedures for quantifying the reliability of the image-interpreted labels. By testing the ability of a CNN to map crop types in the drone imagery, we further show the potential for automating Class 2 label creation, complementing the work of Chew et al (2020) and the growing literature that explores the ability of deep neural networks to classify drone imagery.

Second, we show that a model's predictions (Class 3 labels) can be used to improve its performance, allowing more accurate crop type maps to be developed for little additional cost. This finding contributes to work on synthetic label generation led by Radiant MLHub (Alemmohammad, personal communication, 2022), and to prior research on model-generated labels (Wang et al., 2019).

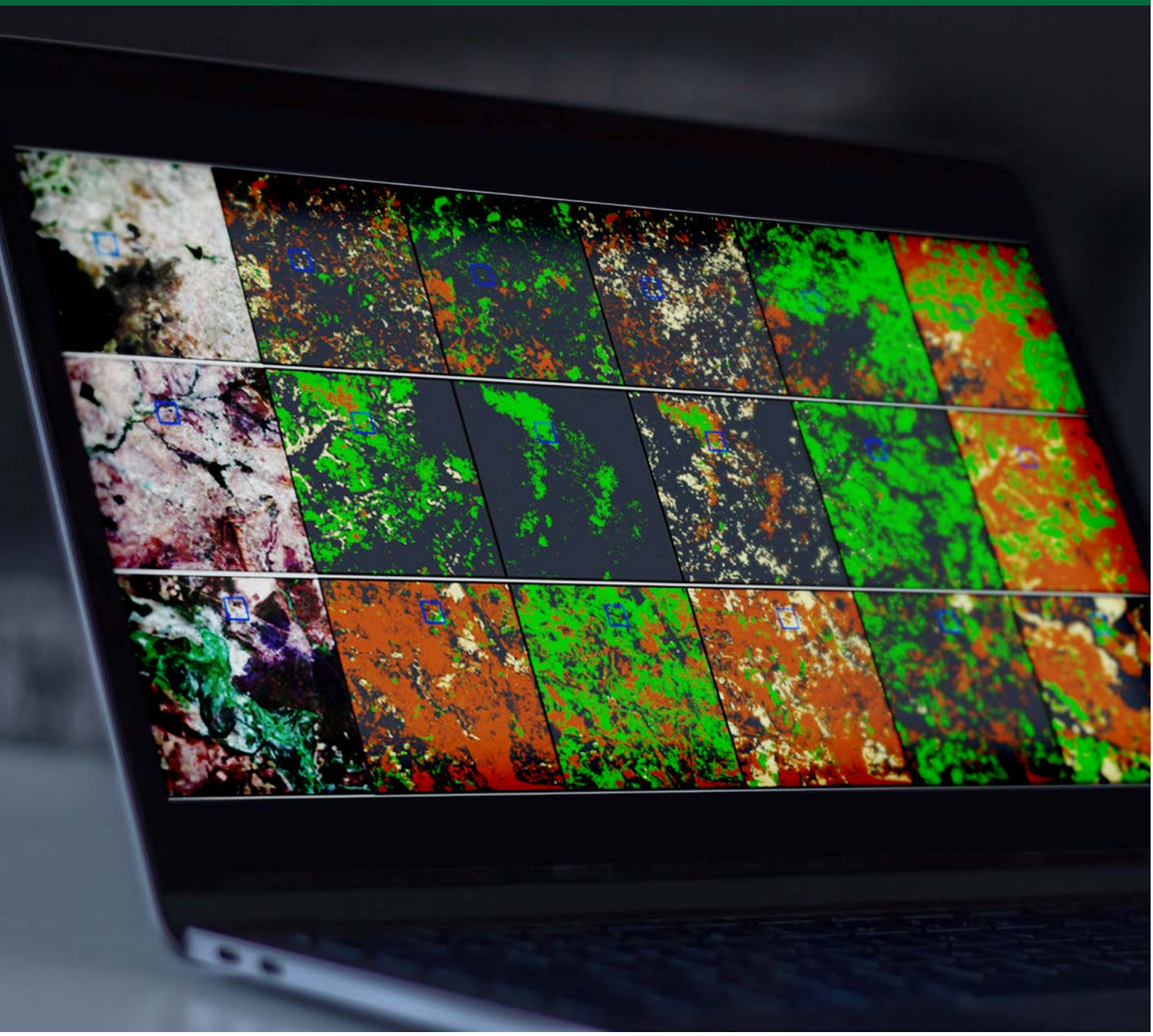
Third, our findings support existing work that shows that temporal CNNs outperform Random Forests for mapping crop types in smallholder-dominated agricultural regions (Wang et al, 2020), and further show that a "train global, refine local" modeling strategy may help reduce label requirements. Although our experiments did not identify a minimum sample size, the potential for this approach to reduce the need for new labels is supported by prior studies that use CNNs for field boundary mapping (including our own), which have shown that fine-tuning pre-trained models with a small number of labels achieves good performance when mapping new locations (Wang et al., 2022).

Finally, we demonstrated the capability to rapidly develop annual, country-scale, high resolution maps of crop fields. Although the resulting maps have inconsistencies that we are correcting, they are more accurate than those produced by the prior study that informed our approach (Estes et al, 2022), and cover a much larger area and longer period of time. Furthermore, the Unet-based approach required <10% of the computational resources and <20% of the labels used by the previous method (based on Random Forests).

When combined, these approaches substantially increase the ability to reliably and sustainably collect high resolution, spatially extensive data on crop type and field dynamics in hard-to-map agricultural systems.

6

Data Availability



The following datasets are available as this project, accessible through links provided in opencropmaps repository to a public AWS bucket account as well as a Box folder:

- Class 1, 2, and 3 labels
- Orthophotos (cloud-optimized geotiffs)
- Class 3 enhanced Random Forests maps for Ejura-Tain and Northern Ghana for 2021
- Cropland maps for Ghana for 2018-2021 and for Tanzania for 2017 and 2018 (when fixes completed)

The Class 1 and 2 labels are also being prepared for submission to Radiant MLHub, and portions will be released through 6Grain's Data Sharing Platform.

Code (currently private while being scrubbed for committed credentials, will be opened when complete):

- LSTM and tempCNN code: https://github.com/agroimpacts/croptypemapper
- Image processing code: https://github.com/agroimpacts/cscdc

Methods

A. Groundtruth Data

We developed and integrated three methods for generating crop type labels, which we categorized into three classes (Class 1, 2, and 3) that vary in their accuracy and suitability for use (see Table 1). In addition to these, we also collected and processed several sets of publicly available labels, and evaluated their usefulness for model development and evaluation **(Table 1)**.

A.1. Class 1 labels

Class 1 labels represent the gold standard in terms of verifying crop identity and providing additional details that can aid model development (e.g. crop planting dates).

We conducted two campaigns to collect labels on the ground in Ghana during this project, during which our agents visited farmers within our networks. One campaign was conducted in Ejura-Tain between October-December, 2021, the other in Northern Ghana between September and December, 2021. During each campaign, agents captured field boundaries and recorded crop types using the GPS-enabled Mergdata platform, following data collection and processing protocols established during our previous project. After processing, we had a total of 3,065 usable Class 1 labels for both regions, representing 26 different crop species, 7 of which had more than 100 records each **(Table A1)**.

Table A1:
The total cou¬nt of Class 1 labels collected in Ghana across both regions in 2021, by crop species. The other crops category comprises 19 species that had <100 samples each.

Crop	Count
Maize	1146
Rice	470
Groundnuts	230
Soybean	198
Sorghum	159
Millet	137
Other crops	725

An additional 2,221 non-cropland points were collected by visually interpreting and annotating high resolution PlanetScope imagery captured in November, 2021.

A.2. Class 2 labels

While Class 1 labels are the most reliable for verifying crop identity, they can be unrepresentative of actual crop distributions in our regions of interest. The reason for this is that they are hard to collect according to a probability design, due to factors such as limited access to farms and transportation logistics, which can result in geographically clustered samples with skewed class distributions. Such samples undermine the training and assessment of crop type models.

Although field-based crop type sampling can be designed to correct these shortcomings, such designs are substantially more expensive to implement and thus harder to sustain. Therefore, to address the limitations in our Class 1 labels in a more economical way, we followed Hegarty-Craver et al (2020) in implementing a drone-based label collection strategy, building on that approach by developing a probabilistic, two-stage sample design for collecting the imagery, which allowed us to collect a sample that was likely to be more statistically representative of region crop type distributions.

We implemented this sample design within a campaign conducted in Northern Ghana in September, 2021, and another in Ejura-Tain in November-December, 2021. The campaigns were conducted by Africa Surveys and Imaging Systems (ASIS), a drone services provider. ASIS collected imagery within 194 550 X 550 m grid cells that provide the basic sampling unit for our agricultural mapping framework **(Figure 3)**, and converted the resulting images into 3 cm orthomosaics that covered each sample cell with 5 m horizontal accuracy. A team of 4 then labeled each of 188 usable orthomosaics following an established set of labeling rules, which included use of an image reference library that provided examples for key crop species.

The Class 1 and Class 2 sample designs were partially overlapped so that both sets of labels were collected within a subset of sites (10 in Northern Ghana, 33 in Ejura-Tain; blue points in **Figure 2**).

We used this overlap to assess the accuracy of Class 2 labels, which was generally low (<50%) for small crops (crops growing close the ground or with narrow canopy cover) but higher for major crops such as rice and maize (55-90% User's accuracy), which has since been improved through a second round of editing. We also assessed how well labellers agreed by having each label the same 8 images (cyan points in **Figure 2**), which showed 79-97% agreement on maize, rice, and noncrop classes, with lower agreement on other classes. The results of these assessments and further detail on label development protocols are provided in our detailed report on Class 2 labels. The current version of our Class 2 labels contains 9,070 polygons (Table A2), of which nearly 2,500 contained crops that could not be identified and 1,930 had no visible crops but had been recently harvested or newly prepared (we loosely describe these as "fallow").

Table A2:
The total count of Class 2 labels collected in 2021. Unidentified indicates labels where the crop species could not be determined. Other comprises 7 species with <100 samples each. Legumes include soybean, bambara beans, cow peas, and groundnuts. Noncrop are areas between the labeled classes.

Crop	Count
Maize	2766
Unidentified	2498
Fallow/newly prepared	1930
Rice	498
Sorghum	415
Legumes	207
Noncrop	188
Other	188
Yam	157
Cabbage	118
Banana	105

A.3. Class 3 labels

We developed a third class of labels by extracting the highest confidence crop type predictions from Random Forests models. We refer to these synthetic, model-generated samples as Class 3 labels, and their development is described in more detail in **Section B.1**.

A.4. Public datasets

We collected four open datasets to use in this project **(Table 1)**, three from Tanzania and one from Ghana. Two of the Tanzania datasets (geolocated crop scout and maize trial points) were provided by the Tanzania Soil Information Service (Walsh et al, 2018) and accessed through the Open Science Foundation portal, and the third was provided by the Great African Food Company and obtained from Radiant MLHub (Great African Food Company, 2019). These three datasets provide crop types for the years 2018 and 2019. A dataset covering a portion of our northern Ghana focal region, submitted by Stanford University to Radiant MLHub, provided crop type observations for the year 2016 along with image time series data (Rustowicz et al, 2020).

Each dataset contained partially overlapping sets of crop types and came in different formats, which required substantial pre-processing to create a common format. The Stanford dataset for Ghana is provided as image chips without spatial coordinates, which we converted to individual pixel arrays to make them compatible with the Tanzania data, which primarily consisted of point observations. Neither dataset of labels contained noncrop observations, therefore we collected noncrop samples for each using image interpretation. For the Stanford dataset, we interpreted noncrop locations directly in the Sentinel-2 images provided with the data, while for Tanzania we placed samples within a 5 km buffer around the provided points and identified the points that fell outside of crop fields in both 2018 and 2019.

We simplified the crop types in these datasets to four common classes: maize, rice, other, and noncrop. The distribution of the samples, which represent counts of individual 10 m resolution image pixels, are shown in **Table A3**. Further details on the preparation of these datasets can be found in our report on crop type model evaluation.

Crop	Ghan	Tanzania
Maize	186,807	15,391
Rice	36,670	3,534
Other	428,772	37,626
Noncrop	56,512	7,134

A.5. Label usage

Each of the three classes of labels have different strengths and weaknesses, which governed how we used them for developing and analyzing crop type models **(see Table 1)**. Each class was used for one or more of the following purposes: training; validation; label quality assessment. Training data are used to develop models, and are typically developed to provide balance across classes and to represent the range of variability in the population. Validation data are used to independently assess model performance and generalizability, and should be collected using a probability design to ensure statistical rigor (Stehman and Foody, 2019). Label quality assessment data[2] is to estimate the amount of error in training and validation labels, which is particularly important for image-interpreted labels.

We used Class 1 labels primarily for training satellite-based crop type mapping models and for label quality assessment. With the exception of a small number of labels, we did not use Class 1 labels for validation because they were not collected using a probability design. We also used Class 1 labels to assess the accuracy of Class 2 labels, where they overlapped.

We constructed validation samples for the satellite-based models from Class 2 labels, since they were collected using a probability design. We randomly selected labels from each class, ensuring that there were at least 30 samples per class, and reserved these for validation. The remaining selected labels were then used for model training, excluding those that fell within the same orthophotos as validation samples.

Class 3 labels were only used for model training samples, as their class identity cannot be verified.

The two public label datasets were used for both model training and validation. As with the Class 2 labels, we first randomly selected a validation sample from each dataset, and then used the remainder for training.

To train the Random Forests models used to generate Class 3 labels, we combined the Class 1 and 2 training samples, and used the Class 2 validation sample to assess model accuracy **(B.1)**. We used the same combined Class 1 & 2 sample design to create two of the four label sets used to develop and analyze the neural network-based crop type models **(B.3)**. The other two were provided by the public label datasets. We constructed a "global" training sample by randomly selecting up to 5000 labels per crop from each of the four dataset's validation samples, and 10,000 pixels per crop from each of their training samples.

To train and evaluate the CNNs used to predict crop types in drone data **(B.2)**, we only used Class 2 labels, first selecting a validation sample, and assigning the remaining labels to the training sample. As before, we removed labels that fell within the same orthophoto as validation samples.

[2] This term is synonymous with training reference data, which is defined by Elmes et al, 2020. We chose this term to avoid confusion with training data.

Methods

B. Crop type and field boundary mapping

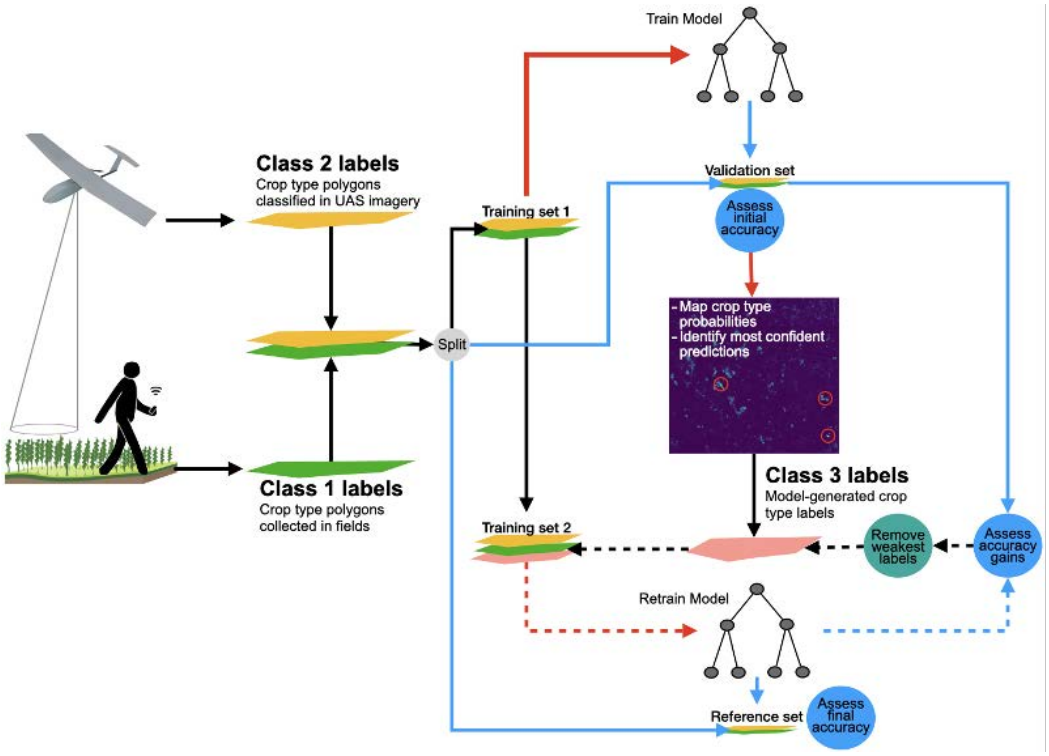
During the course of the project, we undertook a series of analyses aimed at improving our ability to map the agricultural systems in our region. These included two separate efforts to improve satellite-based crop type maps, another aimed at mapping crop types in drone images, as well as an effort to develop multi-year, country-scale crop field maps.

B.1. Using synthetic labels to improve crop type maps

Given the lack of labels for crop types, and the significant costs in collecting them, we undertook an analysis to see whether predictions from a crop type model could help to improve its performance. This effort builds on the concept of synthetic label generation advanced by Radiant Earth Foundation, which has shown promise for improving crop type models (Alemohammad, personal communication, 2022).

In this project, we developed two Random Forests (Breiman, 2001) models, which we trained and evaluated with our Class 1 and 2 labels, using image predictors drawn from Sentinel-1, Sentinel-2, and PlanetScope basemap imagery (see our earlier reports here and here for more details on image processing). We used these models to create two crop type probability maps, one for Ejura-Tain and one for Northern Ghana. We then identified the sites within each map that represented the highest confidence predictions for each crop type and extracted their boundaries to create model-generated, Class 3 labels. We then added these Class 3 labels to the models' original training sample, and retrained each model. We evaluated the updated models for performance gains, measured in terms of improvements against the reference sample, and in distributional improvements in the resulting maps. The full details of the methods behind this approach are provided in our detailed **Class 3 Label Report**. A graphical overview of the approach is shown in **Figure A1**.

Figure A1:
An overview of the workflow used to develop Class 3 labels and evaluate their ability to improve model performance.



B.2. Developing a model to predict Class 2 labels

We evaluated a second approach for generating synthetic labels, drawing on our Class 2 labels and their corresponding drone imagery. In this case we adapted the convolutional neural network we use for cropland mapping, a Unet (Ronneberger et al., 2015), and trained it using our Class 2 labels to predict crop types within the drone imagery (see our separate **drone report** for more details. The goal of this analysis was to determine whether the effort needed to label drone imagery, which requires up to 4 hours per 550 X 550 m orthophoto, can be reduced by using a model to develop the labels.

B.3. Neural networks to improve crop type maps and model transferability

In addition to testing whether synthetic labels could improve conventional Random Forests' crop type maps, we developed and evaluated two models based on neural networks, a Long Short-Term Memory Network (Hochreiter & Schmidhuber, 1997), and a 1-dimensional convolutional neural network (Hu et al., 2015). Our goal in developing these models was to leverage 1) the superior capabilities of neural networks for classification tasks, including for crop type mapping in this (Rustowicz et al., 2019) and other smallholder-dominated agricultural regions (Wang et al., 2020), 2) their ability to learn in the presence of cloud and atmospheric contamination (Rußwurm & Körner, 2017; Wang et al., 2020), and 3) their ability to be transferred to new regions or time periods (e.g. Wang et al., 2022), which offers the possibility of reducing the number of labels that have to be collected.

In our case, we developed each model (the LSTM and the 1d CNN, which we refer to as the temporal CNN or tempCNN) to work with high density, annual time series data from the Sentinel-1 and Sentinel-2 satellites (36-52 time points per year), and trained them to develop crop type predictions for individual pixels based on these time series. Both models had a separate branch for each sensor, which were trained separately to create two separate predictions, which were then fused to create a single output prediction.

As the model architectures and their inputs are substantially different than the Random Forests model, we developed a separate image processing processing pipeline to create the necessary time series data, which included point-based extraction of image time series from Google Earth Engine, as well as the development of 5 X 5 km predictor tiles, using imagery obtained from Google Cloud Platform and the European Space Agency's servers.

We developed a series of experiments to evaluate model performance and the ability of each model to transfer between our focal regions **(see Figure A2 for overview)**. Using our Class 1 and 2 labels together with the public datasets, we developed five sets of labels representing each of our three focal regions (Northern Ghana, Ejura-Tain, and Tanzania), the non-spatial public dataset for Northern Ghana, as well as a global dataset comprised of subsets randomly drawn from each of the four regional datasets. We then trained each model on these five datasets, and evaluated their performance 1) within their own region, 2) when applied to each of the other regions, and 3) within each region (other than itself) after fine-tuning on that region's labels.

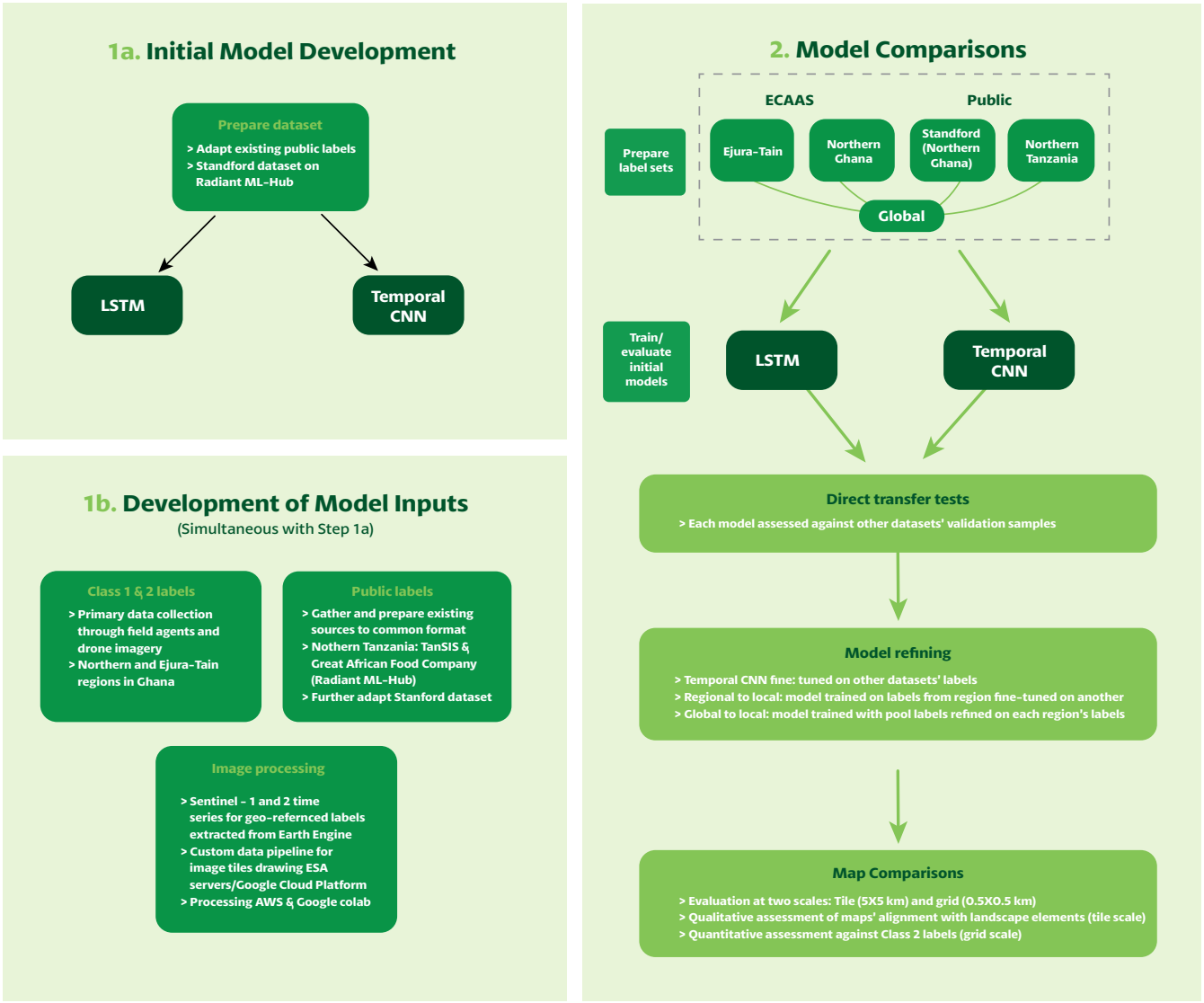


Figure A2:
A diagram of the workflow and experiments used to assess the neural network-based crop type maps.

B.4. Annual, large-scale mapping of field boundaries

To demonstrate the ability to make annual high resolution, country-scale maps of crop fields, we adapted an existing Unet model that we had already developed to improve on prior efforts to map Ghana's crop field boundaries (Estes et al., 2022). The adapted model we used was a "densely-fused" Unet (DFUnet), a variant of the ordinary Unet that we used to map all of Ghana for the year 2018, and the Ejura-Tain region for 2020 and 2021. The DFUnet provides connections between each node of the decoder (after upsampling) with each node in the encoder that has the same spatial extent or larger. We adapted this approach because it improves the ability of the model to make use of image features derived from multiple scales, and it showed initial promise in reducing omission errors relative to the standard Unet. We used it to create updated maps of Ghana for the years 2019, 2020, and 2021, and for Tanzania for the years 2017 and 2018.

To map field boundaries, we trained the model on high resolution PlanetScope imagery to recognize three classes: the edges of fields, the interior of fields, and non-field areas. The predicted field interiors can then be used to segment the resulting map into instances of individual fields.

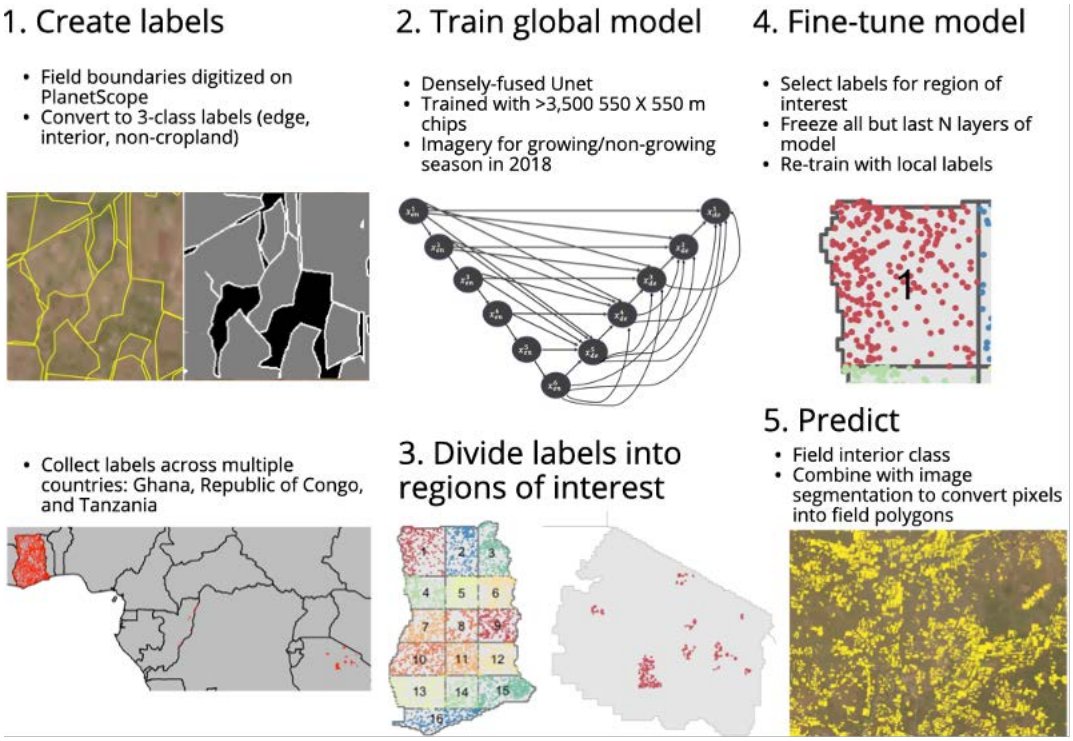


Figure A3:
An overview of the cropland mapping framework.

We first trained a "global" model using an extensive set of field boundary polygons collected within 4,598 550 X 550 m grid cells spread across Ghana, Northern Tanzania, and the Republic of Congo, resulting in a model that had 82.7% accuracy on the field interior class, with true and false-positive rates of 71% and 14.5%. We then fine-tuned the model to make several region-specific variants (16 for Ghana, 1 for Northern Tanzania) to improve model accuracy. To fine-tune the model, we froze the trained weights on all but the last two layers, and then retrained the model on the labels specific to each region.

We used the resulting fine-tuned models to make maps in each region for each of the years. To enable this, we developed an image catalog from PlanetScope's monthly basemap archive, made freely available through Norway's International Climate and Forest Initiative. More details of our modeling approach can be found in our prior cropland mapping report. A graphical overview of the approach is illustrated in **Figure A3**.

References

Alemohammad, H. (2022) Improvement gains in crop type model using synthetic labels. Personal communication.

Breiman, L. (2001) Random Forests. *Machine Learning*, 45, 5–32.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A.L. (2018) DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40, 834–848.

Chew, R., Rineer, J., Beach, R., O'Neil, M., Ujeneza, N., Lapidus, D., Miano, T., Hegarty-Craver, M., Polly, J. & Temple, D.S. (2020) Deep neural networks and transfer learning for food crop identification in UAV images. *Drones*, 4, 7.

Elmes, A., Alemohammad, H., Avery, R., Caylor, K., Eastman, J.R., Fishgold, L., Friedl, M.A., Jain, M., Kohli, D., Laso Bayas, J.C., Lunga, D., McCarty, J.L., Pontius, R.G., Reinmann, A.B., Rogan, J., Song, L., Stoynova, H., Ye, S., Yi, Z.-F. & Estes, L. (2020) Accounting for training data error in machine learning applied to Earth Observations. *Remote Sensing*, 12, 1034.

Estes, L.D., Ye, S., Song, L., Luo, B., Eastman, J.R., Meng, Z., Zhang, Q., McRitchie, D., Debats, S.R., Muhando, J., Amukoa, A.H., Kaloo, B.W., Makuru, J., Mbatia, B.K., Muasa, I.M., Mucha, J., Mugami, A.M., Mugami, J.M., Muinde, F.W., Mwawaza, F.M., Ochieng, J., Oduol, C.J., Oduor, P., Wanjiku, T., Wanyoike, J.G., Avery, R.B. & Caylor, K.K. (2022) High resolution, annual maps of field boundaries for smallholder-dominated croplands at national scales. *Frontiers in Artificial Intelligence*, 4, 744863.

Great African Food Company (2019) "Great African Food Company Tanzania Ground Reference Crop Type Dataset", Version 1.0, Radiant MLHub. https://doi.org/10.34911/RDNT.5VX40R

Hegarty-Craver, M., Polly, J., O'Neil, M., Ujeneza, N., Rineer, J., Beach, R.H., Lapidus, D. & Temple, D.S. (2020) Remote Crop Mapping at Scale: Using Satellite Imagery and UAV-Acquired Data as Ground Truth. *Remote Sensing*, 12, 1984.

Hochreiter, S. & Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, 9, 1735–1780.

Hu, W., Huang, Y., Wei, L., Zhang, F. & Li, H. (2015) Deep Convolutional Neural Networks for Hyperspectral Image Classification. *Journal of Sensors*, 2015, e258619.

Ronneberger, O., Fischer, P. & Brox, T. (2015) *U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* Lecture Notes in Computer Science. (ed. by N. Navab), J. Hornegger), W.M. Wells), and A.F. Frangi), pp. 234–241. Springer International Publishing, Cham.

Rußwurm, M. & Körner, M. (2017) Multi-temporal land cover classification with long short-term memory neural networks. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-1/W1, 551–558.

Rustowicz, R.M., Cheong, R., Wang, L., Ermon, S., Burke, M. & Lobell, D. (2019) Semantic Segmentation of Crop Type in Africa: A Novel Dataset and Analysis of Deep Learning Methods. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 75–82.

Rustowicz R., Cheong R., Wang L., Ermon S., Burke M., Lobell D. (2020) Semantic Segmentation of Crop Type in Ghana Dataset, Version 1.0, Radiant MLHub. https://doi.org/10.34911/rdnt.ry138p

Stehman, S.V. & Foody, G.M. (2019) Key issues in rigorous accuracy assessment of land cover products. *Remote Sensing of Environment*, 231, 111199.

Walsh, M., Meliyo, J., Scott, B., Walsh, B., Macmillan, B. (2018). Tanzania Soil Information Service https://doi.org/10.17605/OSF.IO/HR8E6.

Wang, S., Azzari, G. & Lobell, D.B. (2019) Crop type mapping without field-level labels: Random forest transfer and unsupervised clustering techniques. *Remote Sensing of Environment*, 222, 303–317.

Wang, S., Di Tommaso, S., Faulkner, J., Friedel, T., Kennepohl, A., Strey, R. & Lobell, D.B. (2020) Mapping Crop Types in Southeast India with Smartphone Crowdsourcing and Deep Learning. *Remote Sensing*, 12, 2957.

Wang, S., Waldner, F. & Lobell, D.B. (2022) Unlocking large-scale crop field delineation in smallholder farming systems with transfer learning and weak supervision. https://export.arxiv.org/abs/2201.04771

Prepared by

In Collaboration with

FARMERLINE

Enabling Crop Analytics At Scale

Creating Next Generation Field Boundary and Crop Type Maps Rigorous Multi-Scale Groundtruth Provides Sustainable Extension Services for Smallholders

info.ecaas@tetratech.com
cropanalytics.net

Final report

August 2022

Contributors

Farmerline
Amos Olertey Wussah, Mary Dziedzorm Asipunu, Worlali Senyo, Eric Gbekor, Emmanuel Kwaku Duah

Clark University
Lyndon Estes, Sam Khallaghi, Lei Song, Sitian Xiong, Ismail Alatise, Cat Mai, Nguyen Ha, Sai Vishal Muda, Ravi Thapaliya, Adriana Chamorro, Sreeja Vinod, Garren Kalter, Aandishah Samara, Zeyu Zhang, Boka Luo, Su Ye, Priscilla Ahn, Zexing Zeng, Zihan Chen, Zhenhua Meng, Qi Zhang