TETRA TECH

Enabling Crop Analytics At Scale

ENABLING CROP ANALYTICS AT SCALE (ECAAS)

Report on Final Results & Lessons Learned

Optimizing Crop Yield Data Collection for Supply Chain Enhancement, An Enabling Crop Analytics (ECAAS) at Scale Project

Contents

1

Project Summary

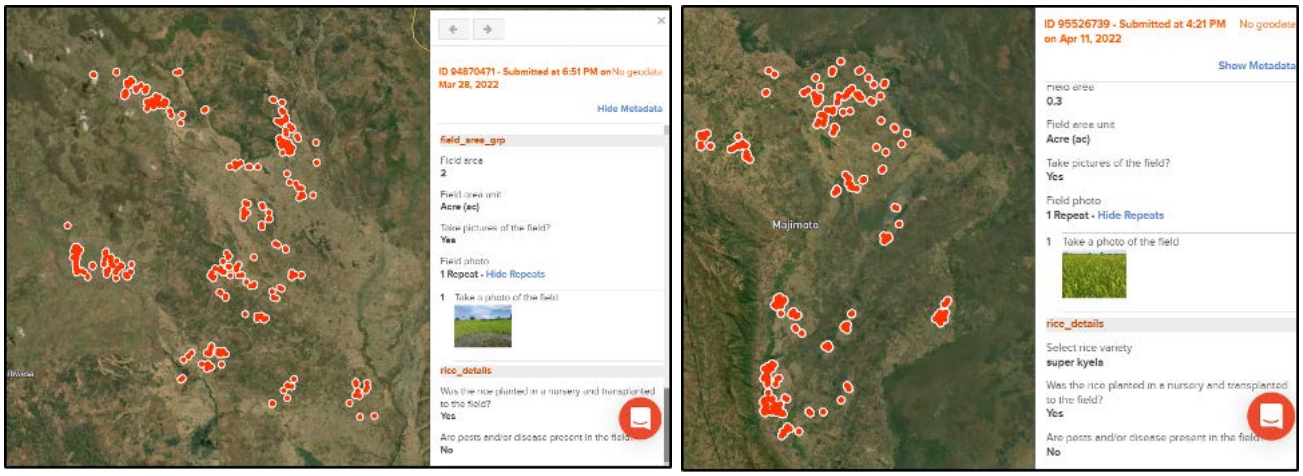
Food security is one of the most pressing social issues, if not the most pressing, facing many African countries today. In sub-Saharan Africa, agricultural system shocks in coming years will continue to have severe impacts on the food security of smallholder farmers. Analyzing the nature and extent of these impacts and assessing their significance on livelihoods are important in planning responses and mitigation efforts, but these can become overwhelming tasks with only conventional capabilities like on-the-ground observations and field surveys of farmers. Satellite-based Earth observations (EO), which provide crucial information about crops in near real time, can play a vital role in supplementing and enhancing such capabilities, enabling accurate estimates of production, earlier warnings of crop failures, and supporting response programs involving risk financing and other measures that reduce food insecurity.

Accurately determining crop yields at field scale can help farmers estimate their net profit and access services, including inputs, insurance, markets, and storage. This same information is critical for improving service providers' own operations, for example, to improve logistics (e.g., for buyers) and to ascertain amounts to purchase or payout. When aggregated, crop yield estimates are critical in monitoring food security at national and regional scales. While collecting ground-truth yield data is largely cost-prohibitive, it is the most reliable way to estimate yields at the field level.

The focus of this project was to collect yield data through a public-private partnership between NASA Harvest, Flamingoo Foods Limited, and theSokoine University of Agriculture in Tanzania at the end of the 2022 growing season, focusing on rice, the primary commodity for Flamingoo Foods. The project team surveyed 800 farmers in the Ikakla and Majimoto regions in Katavi, Tanzania. This project aimed to demonstrate the utility of machine learning models for optimizing yield data collection, ultimately reducing the cost associated with data collection. The project aimed to test the scalability of machine-learning-based crop yield forecasting models in estimating field-scale yield as well as the generalizability to other rice-growing regions.

This unique partnership not only explored a collaborative partnership in data collection but also applications and use of Earth observations in modeling crop yield. Moreover in the process, we collected critical data relevant to improving Flamingoo Foods' operations, ultimately improving farmers' access to storage and markets as well as better prices for their product. This project will enable the evaluation of the cost of data collection and potential to improve workflows and expansion to other crops. Our approach sought to improve Flamingoo Foods' business model by leveraging derived analytics to improve identification of surplus and deficit regions using satellite and weather data, use of real-time satellite observations to track the progress and performance of the harvest, and development of crop forecast products. This project provided a unique opportunity to augment their workflows by integrating advanced analytics made possible by collecting ground data and improving modeling approaches. The products from this project will support further identification of lucrative farmers, reduce post-harvest losses, increase access to storage (provided free of charge), and facilitate improved access to markets.

This project collected datasets, including **806 field boundaries and 617 yield measurements**, and other field characteristics from smallholder rice fields in Western Tanzania. Our local partners include the Sokoine University of Agriculture based in Morogoro and Flamingoo Food Limited. The project leveraged ECAAS's Open Data Kit (ODK) form and toolkit to collect yield data and evaluated the utility of these data by testing and applying multiple machine learning models at field scale on rice or the region. By demonstrating the utility of machine learning models for optimizing yield data collection, this project can inform and reduce the cost of collecting yield data critical for agricultural decision-making.



Images 1 & 2:
Screenshots of rice paddy fields participating in the first round of surveys in Ikaka (left) and Majimoto (right) in the Katavi region, along with examples of the meta and other data included with each plot point.

This report includes results from utilizing ground data collected to run crop-type and crop yield models and presents an analysis of the findings along with a summary of key lessons learned through this collaborative research project. A forthcoming technical paper will provide additional detail on the process of training the model and its performance.

1.1. Project Partners

This project is a program jointly implemented by NASA Harvest at the University of Maryland, the Sokoine University of Agriculture in Morogoro, Tanzania, and Flamingoo Foods Company Ltd. This project was conducted in two major paddy growing areas (Majimoto and Ikaka), located in Katavi, the western part of the United Republic of Tanzania, with the support from community leaders in those regions.

1.2. Project Objectives

The project mission was to collect rice paddy yield data in major growing regions in Tanzania to **inform production scalability and integration into business models.** As proposed, the first goal was to perform smallholder field-based interviews, surveying 800 paddy fields and collecting yield data from a subset of approximately 600 fields. The second goal was to use the data collected for training and validating remote sensing cropland and crop-type maps and apply the GEOCIF model at the field scale for Majimoto and Ikaka.

Given historical gaps in rice crop datasets, this collaboration sought to create a high-quality public dataset of field-scale yield and other relevant characteristics for rice in Tanzania. These data can inform the development of a field sampling approach to maximize the diversity of samples while minimizing the number of samples required. This approach can also be applied in future projects to other crops and regions in sub-Saharan Africa and beyond. The data provided an avenue to test already developed methods for crop type mapping, yield estimation, and crop condition monitoring and provide information on products needed for improved food security decision-making. By providing this open-dataset, we will enable the development of machine learning methods for enhancing smallholder agricultural systems and improving farmer outcomes.

2

Overall Project Approach

Survey data collection leveraged a network of Flamingoo Foods' field agents to connect with the rice farming community in the Katavi region to plan, time, and collect yield data with the farmers' support. The project adopted the (ODK) form and toolkit developed by ECAAS in collaboration with the Radiant Earth Foundation to map rice farms, conditions, and yield data. The data are readily accessible on Zenodo [https://doi.org/10.5281/zenodo.6824200][1] and Harvest's CropHarvest dataset.

Using the ground data, the team tested different machine learning models to map rice in the larger Katavi region and predict end-of-season rice yield from satellite data. **See Figure 1** for the overall project workflow.

Overall Project Approach

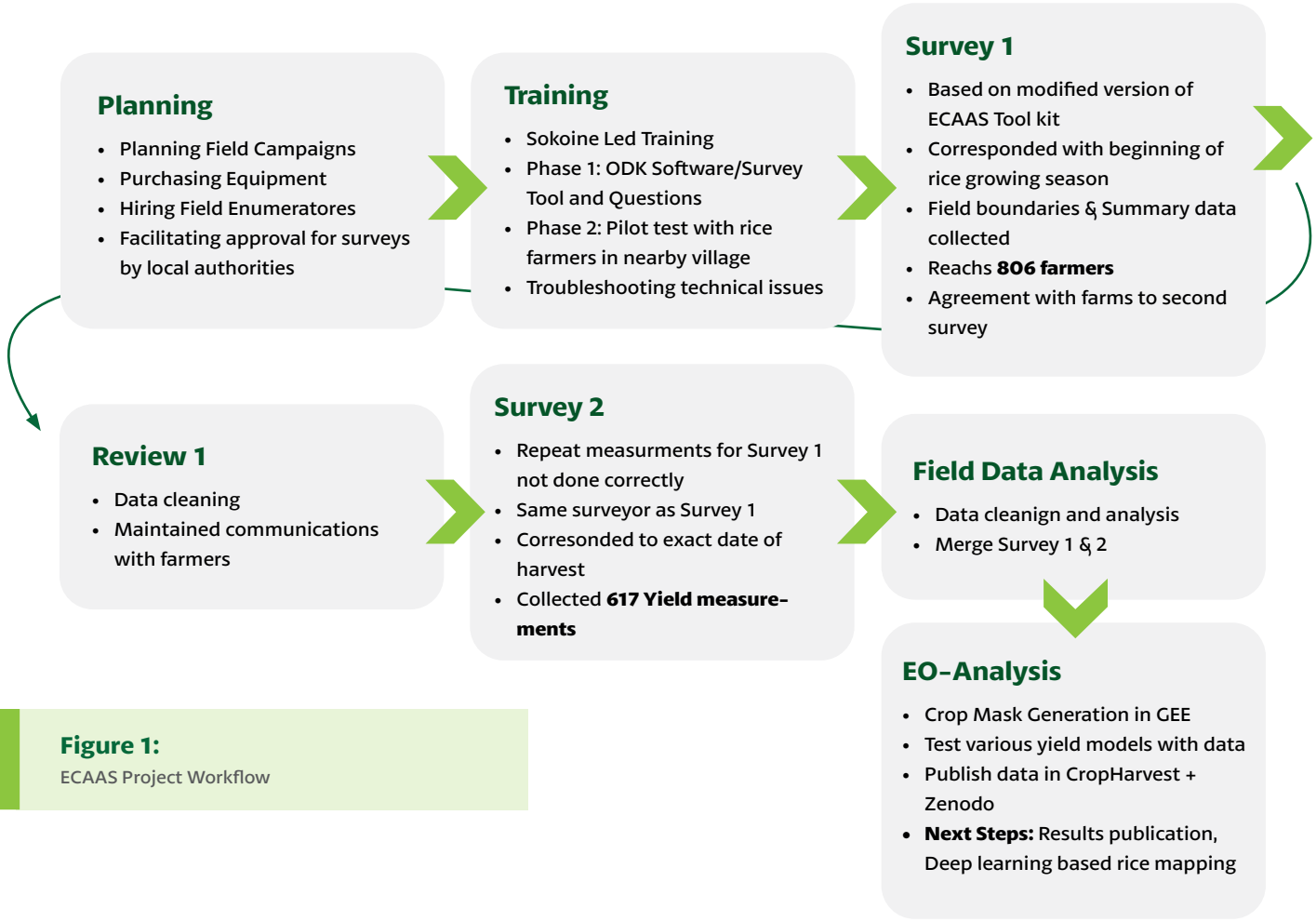


Figure 1:
ECAAS Project Workflow

Catherine Nakalembe, Andreas Schlueter, Sixbert Maurice, & Taryn Devereux. (2022). 2022 Rice Crop-type Data for Western Tanzania (Version 1) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.6824200

2.1 Regional Overview

The focus area of this project is the western part of the United Republic of Tanzania. The study sites were located in two major rice producing areas where Flamingoo Foods currently operates. The first study area lies near Majimoto in the northern Rukwa Valley which contains fertile alluvial soils. The total area of paddy production around Majimoto is roughly 20,000 ha (based on Flamingoo Foods' satellite-derived estimates). The other study area is around Ikaka, south-west of Mpanda, which is the capital of the Katavi region. In Ikaka, the entire paddy area is roughly 25,000 ha.

No exact figures are available on the numbers of rice farmers in both production regions. However, according to the 2007 Tanzania Agricultural Census, the average rice land holding size in Rukwa Region is 1.34 ha. Thus, we estimate that based on total paddy acreage, approximately 15,000 households around Majimoto engage in rice cultivation and roughly 18,500 farmers in the second production region around Ikaka.

National level seasonal rice performance data is not readily available for Tanzania. However, the start of the 2022 wet season with constant average precipitation gave favorable conditions for the growth of Maize crops, mostly concentrated in the Eastern area of the country, while the Pwani region on the Eastern coast experienced Exceptional conditions. Elsewhere, maize crops across southwestern Tanzania – Iringa, Katavi, Mbeya, Njombe, Rukwua, and Ruvuma – are currently under Poor conditions, with the rest of the country under Watch conditions.

2.2. Field Campaign

The following section summarizes the overall field campaign implemented by the project partners. For the full description of this approach, please refer to the Field Campaign Report.

The Field Campaign involved two phases: **(I) The Preparation Phase** to design, plan, and test the survey methodology, and **(II) The Data Collection Phase**, which incorporated field interviews, field plot measurement, field-based yield measurements, and the data processing and submission. **See Annex II** which summarizes the data collected as part of this project, accessible at https://zenodo.org/record/6824200

In March, Flamingoo Foods and the Sokoine University of Agriculture led a hands-on technical training of the field teams in Majimoto, Katavi, including a day of practical exercises and pre-testing the ODK tools and equipment. The initial field campaign and boundary delineation was completed in April during the rice growing season, while the yield data collection was done during May and June. The data collection timeline is based on the crop growing season calendar to ensure all necessary preparations were done for the harvest in May-June 2022.

Flamingoo Foods led and coordinated the identification of farmers and pilot villages through its network of agents in Western Tanzania working with local officials. The first survey on Field Boundary Delineation started at the beginning of the rice production season and ultimately reached **806** farmers. The second survey was more challenging as the surveyor needed to reach the farm at the exact harvest time to ensure that the full harvest quantity could be measured. The eight surveyors were in constant contact with the farmers from the first survey to note the harvest date, and with few exceptions, the surveyors were present at the time of harvest for the second survey, completing a total of **617** yield measurements **(Images 3, 4, 5)**. All local COVID-19 guidelines were followed, including wearing masks. Additionally, the field data supervisor facilitated phone calls with a sample set of farmers in the targeted areas (Majimoto and Ikaka) to verify information collected by research assistants as needed.

Image 3: Harvest taking place in Ikaka

Image 4: Weighing of the finished harvest in Ikaka

Image 5: Bags and drying sheets were given to smallholder farmers. These incentives greatly facilitated the collection of data.

2.3. Forecasting Model

Using the ground data, the team built a machine learning model to map rice in the larger region and predict end-of-season rice yield for future years from different Earth observation-derived vegetation indices. To test the utility of the data collected, we developed a rice mask in Google

Earth Engine and tested various machine learning models to estimate yield across the large area of interest. This section provides an overview of the yield estimation models.

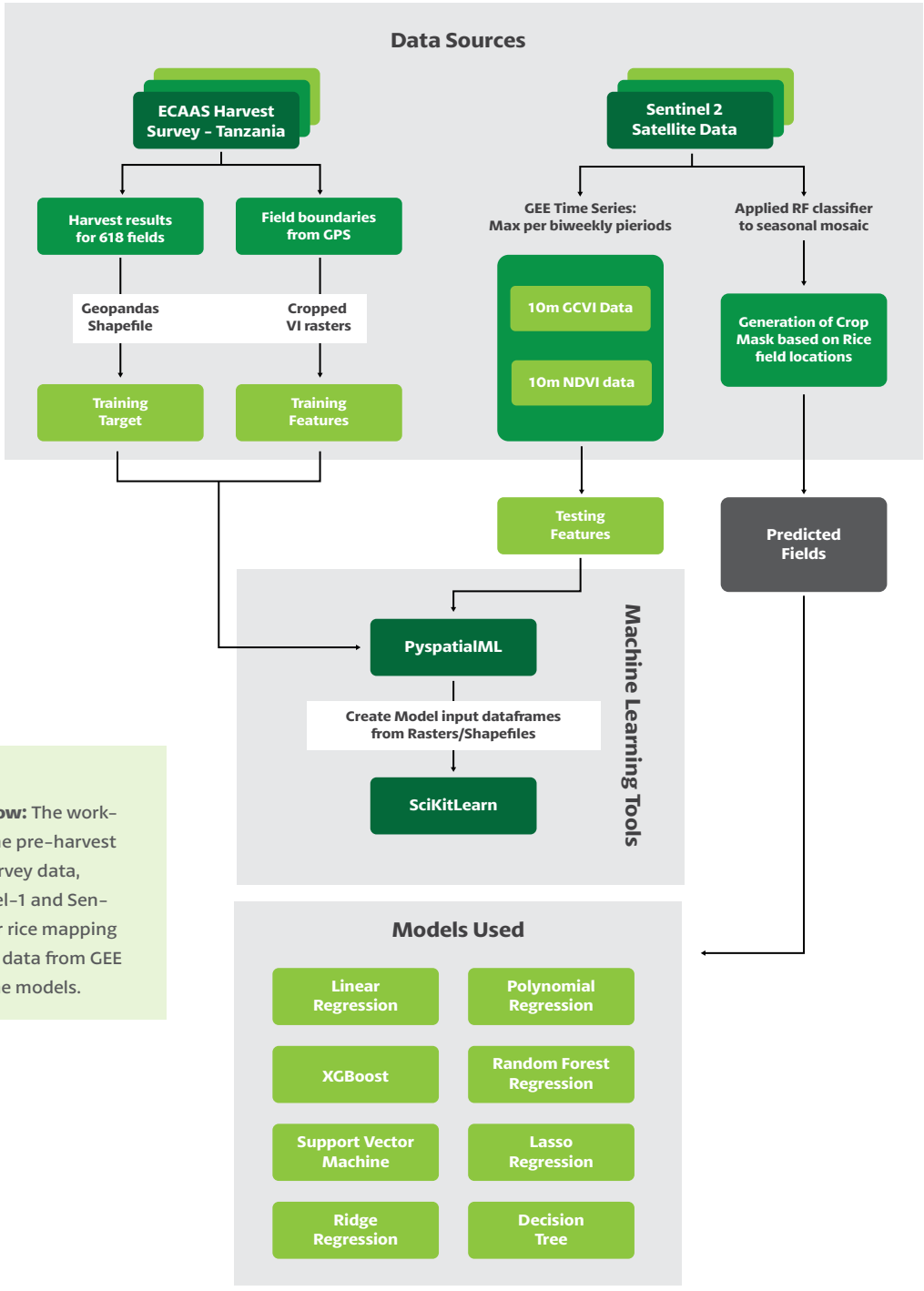


Figure 2:
Model Workflow: The workflow merged the pre-harvest and harvest survey data, utilized sentinel-1 and Sentinel-2 data for rice mapping and sentinel-2 data from GEE as inputs for the models.

Following the data collection campaign, the team created a shapefile to represent the geographical boundaries of each surveyed field. A GPS survey was used on-site to calculate the coordinates of the field's boundaries **(Figure 3)**. This data was then converted into Geometry data in Python.

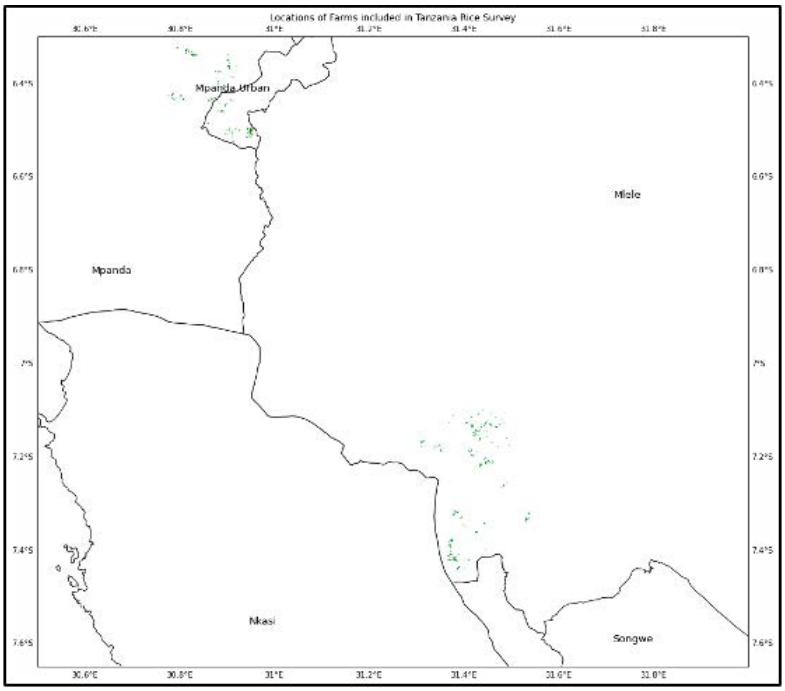


Figure 3: Data Visualization
The green areas represent rice farms located in Katavi. Administrative Level 2 borders

Crop Mask Generation

A single seasonal mosaic was created from Level-2A Sentinel-2 imagery in Google Earth Engine (GEE), excluding cloudy and shadow pixels, using the Sentinel2Cloud probability masks. The median pixel mosaic was then clipped to the area of interest. The random Forests (RF) model was selected after several experiments, including running the support vector machine and Regression Trees and Random Forest, all readily accessible in GEE. RF had the highest overall accuracy. The rice crop mask was only applied to the model following the generation of predicted yields.

Yield Mapping

Multiple time-series satellite datasets utilized for this task were imported from GEE including Sentinel-2A which was exported using a shapefile for the region of interest (Katavi, administrative level 1). The time series image collection was filtered to cover only the area of interest from January to June 2022 the growing season months.

Features

Monthly maximum Normalized Difference Vegetation Index (NDVI) and Green Chlorophyll Vegetation Index (GCVI) rasters were extracted from the GEE as bands with the tsraster library as input features. The target data, field-scale yield values, were overlaid on the rasters to extract corresponding feature raster values to create the training data set. The preprocessed data was then run through multiple prediction models to compare and analyze the results, performance metrics, and visualization.

A rice mask was applied to the model after the training process to limit predictions to those regions where there is known to be rice growing. **Figure 4** demonstrates differences in yields between the yield set used for training and testing the model taken from the survey, the predictions made for pixels based on the rice mask, and the predictions made for all pixels within the raster. The trend line for the predictions beneath the mask resembles the survey yields predictions.

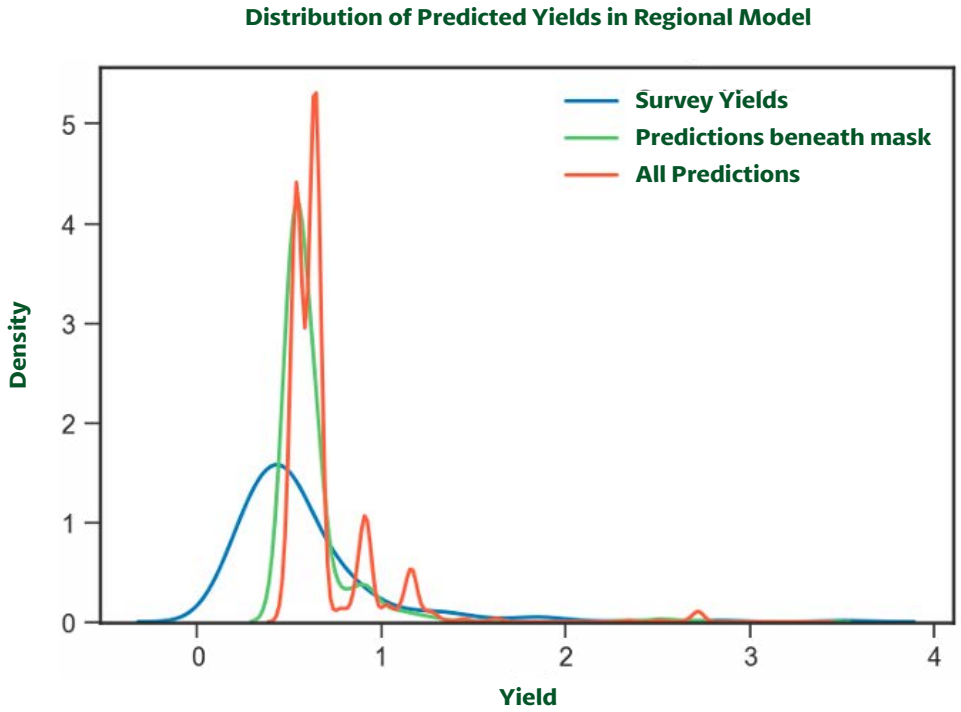


Figure 4: Data Visualization
The green areas represent rice farms located in Katavi. Administrative Level 2 borders

Preferred Model & Future Research

The Linear Regression algorithm performed the best on the maximum GCVI data for the 13 biweekly periods from January through June with the lowest Root Mean Squared Error Value of 0.47 and an R-squared value of 0.015 **(Table 1)**. This model has fewer outliers than the NDVI model, specifically for the SVM sigmoid model. However, most stat values were very similar across the two vegetation indices, especially when considering the values for cross-validation. The linear regression is a simple and fast algorithm that can be particularly effective when the relationship between the features (GVCI) and the target variable (yield) is relatively straightforward. In this case, the linear model is able to capture this relationship well with a small number of features and a small training sample, leading to good performance. While models like random model would perform better if additional features (increase complexity) were included in the model.

Future work will include additional EO features relevant for making predictions, for example soil moisture, rainfall and temperature, in addition to field management data from the first survey to improve results.

Model	Lin	Svm_rbf	svm_sig	svm_poly	svm_lin	Xgb	Dec. Tree	Poly	Lasso	Ridge
RMSE	00.472	0.709	0.709	0.712	0.708	0.730	1.104	0.773	0.692	0.690
R^2	0.015	-0.049	-0.048	-0.057	-0.046	-0.113	-1.546	-0.247	-0.0004	0.007
CV Accuracy	-0.595	-0.576	-0.580	-0.582	-0.579	-0.675	-0.864	-0.595	-0.570	-0.574

Table 1:
Performance Metrics comparison

3

Achievements and Scientific Contributions

Through this project we demonstrated the value of working with local stakeholders by partnering with Sokoine University to ensure the soundness of the data collection as well as rapid training and collection through working with a private company in the region.

We collected data in over 800 fields, with 100% of the farmers willing to participate in the data collection and able to consistently communicate with the field teams. Farmer willingness was critical for the success of the second campaign that had to be done at the exact harvesting time.

The data collected was utilized to develop a rice map/mask. The mask was then utilized to improve preliminary yield estimates for the focus region by focusing on regions/pixels where rice was growing. Moreover, the detailed field management data have already proven critical in explaining regional production differences, highlighting the importance of collecting farm management data when trying to model yield at field scale. The management practices data explain why crop fields with similar biophysical conditions may have different productivity that can not be explained with remote sensing alone.

The first survey included a question about the anticipated harvest date to facilitate the planning of the harvesting measurement. This data was extremely valuable in predicting harvest onset and peak timing. Knowing the anticipated delay in harvest is crucial for determining the right timing for purchasing paddy stocks. The survey helped the team anticipate which region would harvest first and when to expect the peak of rice paddy to flood the market. Future efforts can include building a test tool that enables the prediction of spatial variability in planting dates and yield outcomes.

While the project focused on the Katavi region, the workflows tested and developed can be applied and scaled to much larger regions. As more data become available, model results can be better evaluated. Another focus has been testing deep learning approaches, including a Task-informed meta-learning (TIML) model to run a crop-type model over a larger area toward developing future models that learn efficiently from sparse field data.

4

Discussion & Lessons Learned

Given historical gaps in rice crop datasets, this collaboration sought to create a high-quality public dataset of field-scale yield and other relevant characteristics for rice in Tanzania. These data can inform the development of a field sampling approach to maximize the diversity of samples while minimizing the number of samples required. This approach can also be applied in future projects to other crops and regions in sub-Saharan Africa and beyond. The data will provide an avenue to test already developed methods for crop type mapping, yield estimation, and crop condition monitoring and provide information on products needed for improved food security decision-making. By providing this open-access dataset [https://doi.org/10.5281/zenodo.6824200], we enable and accelerate the development of machine learning methods to enhance smallholder agricultural systems and improve farmer outcomes.

This project also formed a clear foundation between Harvest and Flamingoo Foods by improving analytics critical to our organizations' objectives and demonstrating a clear approach to addressing data gaps through a public-private partnership. This can ultimately lead to improved access to high-quality data and opportunities for more smallholder farmers as Flamingoo Foods expands to other countries. In addition to providing a consistent machine learning-ready dataset (including these rice labels in our CropHarvest Dataset) for smallholder rice fields, this project will spur research interest in Africa and globally.

This project revealed several critical lessons, both in data collection methodology and data processing and analysis, which can be applied to future efforts to improve workflows and data quality. This collaborative effort demonstrated that partnering with local organizations for training and supervision can be a cost-effective and successful approach to collecting high-quality yield data.

4.1. Field Campaign

During this project, despite the screening process and hands-on training (which included a three-day technical workshop and a day of testing the ODK tools and equipment), there were still notable issues, including inconsistencies in data entry. To address this issue, future work will explore hiring and training students as enumerators, for example, at the Sokoine University of Agriculture. This builds local institutional capacity and creates a sustainable pool of future enumerators and partners to lead field campaigns. The Field Supervisor should be comfortable working with these student enumerators and familiar with the larger farming region. In addition to the training at the beginning of the data collection process, there should be ongoing training and quality control throughout the field campaigns in which any input data that is immediately flagged (e.g., exceeding a predetermined threshold for that value) leads to a follow-up session conducted by the Field Supervisor with the enumerator responsible. This can determine if there is an equipment issue or whether additional training and supervision are needed.

Training should also continue emphasizing the importance of maintaining a positive and respectful relationship with the farming community. This is especially important given some of the challenges with recruitment. Depending on the region, farmers can often be overloaded by surveys and other forms of data collection which can use up valuable time and not produce any tangible benefits. Within this project, the first problem the surveying team encountered was low trust among the locals around Ikaka. In previous years, other companies had deceived and taken advantage of the farming population, leading many farmers to be suspicious about participating in the data collection. Cooperating with local authorities, village heads, and extension offers was critical. Future initiatives should continue to partner with the local community by onboarding community leaders and working with local institutions. The incentives provided through this project – bags and drying sheets – were in demand, welcomed and helped build trust among the survey participants. In the future, it is also worth exploring providing an extension-based service or consultation to those farmers who participate in all phases of the data collection process.

Coordinating the data collection during the harvest (Phase II) was easier than anticipated. This round needed to take place at the exact harvest date to ensure the full harvest quantity was measured, and there was concern that the Surveyor team would miss this day due to communication challenges. Between Phase I and II, the Surveyors maintained constant communication with the farmers to ensure the exact date, and in only a few cases, the Surveyors missed the harvest. It was important that the same Surveyor who conducted Phase I return to the farm during harvest in Phase II so that they had no trouble locating the same field and had already established rapport and trust with the farmer.

Another challenge was accessing the fields, as some were very remote, and it took a relatively long time to reach distant locations. Transportation costs ultimately were higher than previously budgeted. Access was particularly challenging during Phase I as heavy rains complicated the route: some paths were not passable, and it was, at times, too wet to walk the paddy fields and measure the boundaries. There was also one motorbike accident in which no one was hurt. Proper planning using weather forecasts was an important strategy for planning around wet conditions. In the future, it is recommended to start surveying after the wettest part of the rainy season has passed, particularly for surveying paddy rice fields, as they can be fully flooded during the rainy season, and the soil is typically loamy.

Finally, the measurement of the field boundaries turned out to be more difficult. Despite the testing, the accuracy of the smartphones was not as high as anticipated. Project coordinators purchased the latest smartphone devices to ensure the best accuracy that supported all major global navigation systems (A-GPS, Glonass, BDS, and Galileo). Recently, smartphones became available with dual-frequency receivers, which supposedly improve the accuracy of the GPS signal and minimize potential measurement errors caused by the ionosphere and troposphere. We purchased different phone models for testing whether there is a notable difference between smartphones with and without dual-frequency support. We could not measure any significant difference in the accuracy of the GPS measurements as reported by the ODK. One smartphone recorded inaccurate field boundaries, so the team needed to remeasure some of those field boundaries. The fields were too small to delineate using image segmentation programmatically via approaches with readily available satellite data such as Sentinel 2, Landsat, or Planet data. The last suggestion to improve the GPS measurements was manually digitizing field boundaries using high-resolution MAXAR data accessible in Google Earth Pro.

4.2. Data processing / Analysis / Yield Forecasting Model

This project sought to inform the development of a field sampling approach. Cleaning and analysis of the data revealed critical lessons for future efforts. Firstly, there is a need to define a priori clear measurement levels, and thresholds for all values included within the survey and to establish the maximum margin of error for yield data collection.

Field Boundary delineation is typically more straightforward to verify as most errors can be fixed through digitization. However, yield data has more measurement errors, meaning the problem might be with the field boundaries / GPS, the harvest yield, or both. A forthcoming paper will explore how to set a threshold for discarding data points/outliers (e.g., if the median yield for a given region is 5 tonnes/ha, 50 tonnes/ha is incorrect, but 12t/ha might be correct). Potential solutions include thresholding the values for the fields and including a duplicate field in the form so that the field agents enter the values twice.

Models are hungry for data. The high-quality field-level data collected in this study has enabled us to produce field-scale rice yield forecasts using machine learning models. Our results show that simpler models like multiple linear regression can outperform non-parametric tools like XGBoost when applied to satellite data to predict crop yields. Future work will explore comparing machine learning model results with mechanistic models such as DSSAT.

While the data collected in this project are valuable, their true value can be realized over multiple years and larger areas. The analysis would then benefit from the temporal dimension that satellite data offer, which is critical for modeling yield. Second, this project was also limited in spatial scale; while we have some idea of spatial variability in yield during 2022 for our area of interest, we do not have a strong basis for conducting this analysis over a much larger area, including other rice-growing regions in Tanzania. This scaling can be achieved by working with regional administrators or the Ministry of Agriculture. Linking with regional and national levels would enable data collection for multiple crops at a much larger scale, supporting initiatives such as the Tanzania Crop Monitor. The Food Security Division in the ministry of Agriculture is keen to improve the crop monitoring part of the Tanzania National Crop Monitor and to improve analytics that supports their food balance sheet summarized in the crop monitor.

In the future, in terms of building machine learning yield estimation models, feature extraction will be applied to the 10m resolution data (for example, with Sentinel-1 and Sentinel-2 data). Incorporating more composite features into the model may aid the model. In addition, rice crop masks will be incorporated more into the input data before the models are run. This will promote the model's efficiency and help create more reasonable predictions. analysis.

Future work will include the publication of a research paper on the data analysis and application of a task-informed meta-learning model and to expand the area of interest in addition to comparison with regional-level results. With the data collected, we illustrate the value of detailed management data in yield estimation and support the development of a national-level rice mask, ultimately developing a scalable framework that will allow expansion to other regions/ countries as field data becomes available.

1

Annex: Detailed Methology

This project leveraged ECAAS's Open Data Kit (ODK) form and toolkit to collect high-quality yield data and evaluated the utility of these data by testing and applying the Global Earth Observations for Crop Inventory Forecasting (GEOCIF) system at a field scale on rice. Using the ground data, the team developed a 2022 rice map for Katavi and estimated yield across the larger area.

This section summarizes the research methodology for 1) the Field Campaign for data collection and 2) building and testing the machine learning model. A forthcoming technical paper will provide additional information on training the model and its performance.

1. Overall Approach

The surveys were conducted with support from a network of Flamingoo's field agents to connect with the rice farming community in the Katavi region to plan, time, and collect yield data with the farmers' support. This pilot approach and the lessons learned can inform the adoption of future scalable, standardized approaches to collecting critical and time-sensitive data in remote regions, leading to unprecedented and invaluable rice datasets that can improve farmers' market access.

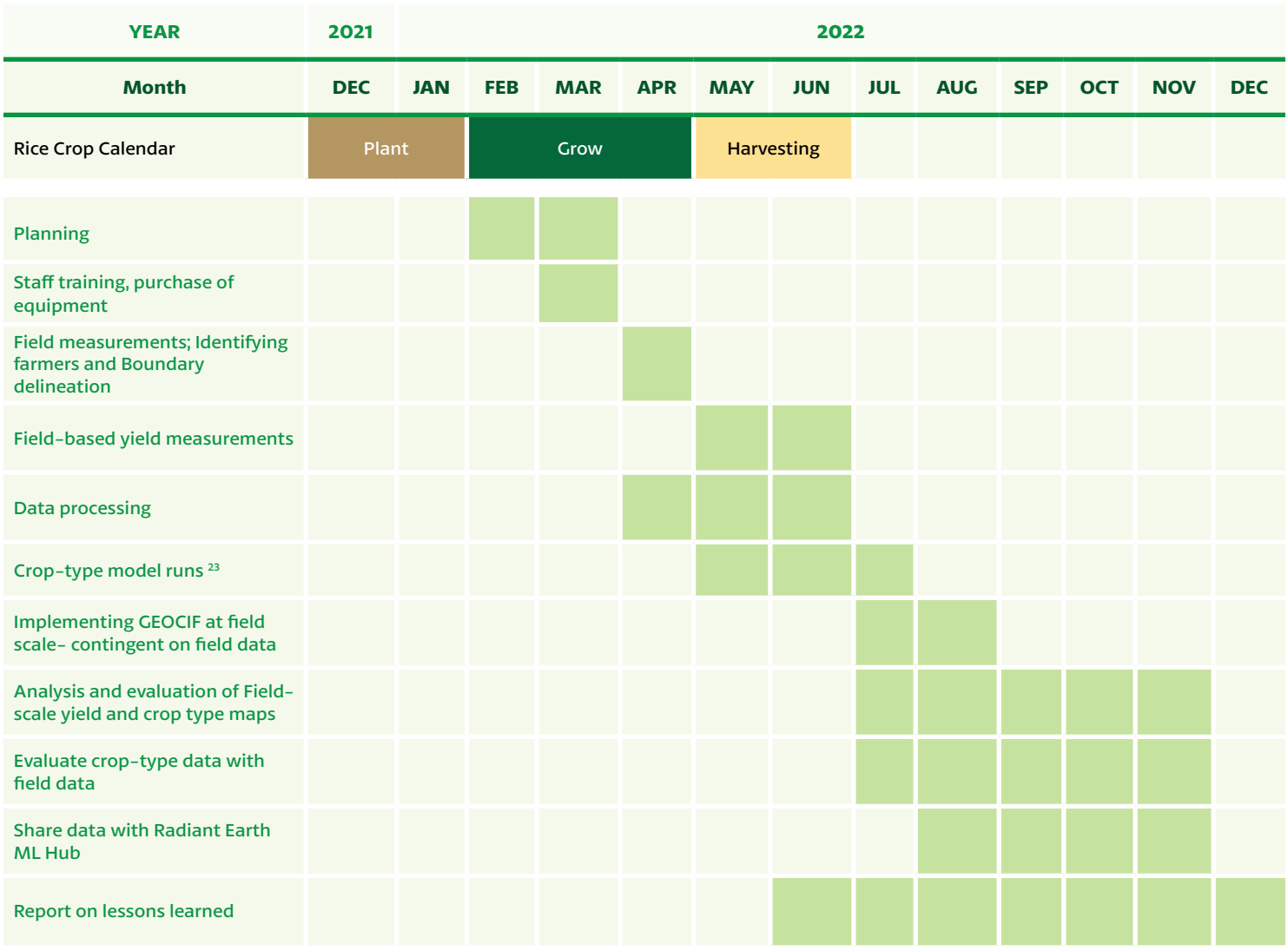
We adopted the ODK standard toolkit developed by ECCAS in collaboration with Radiant Earth Foundation to map rice farms, conditions, and yield data. In addition, we integrated qualitative and quantitative data into the analysis, which addressed the limitations of one singular approach and enriched and contextualized the findings. Using the ground data, the team built a machine learning model to map rice in the larger region and predict end-of-season rice yield for future years from different Earth observation features.

2. Field Campaign

The following section summarizes the overall field campaign implemented by the project partners. For a complete description of this approach, please refer to the Field Campaign Report.

The research activities involved two phases: **(I) The Preparation Phase** to design, plan, and test the survey methodology, and **(II) The Data Collection Phase**, which incorporated field interviews, field plot measurement, field-based yield measurements, and the data processing and submission. Below is the final timeline of activities **(Figure 1)**. The initial field campaign and boundary delineation was completed in April during the rice growing season, while the yield data collection was done during May and June. The data collection timeline is based on the crop growing season calendar to ensure all necessary preparations were done for the harvest in May–June 2022.

Figure 1:
Timeline of project activities



2.1. Phase I: The Preparation Phase

This phase incorporated staff recruitment, training, and related logistics such as purchasing equipment and facilitating approval for surveys by local authorities. The project hired and trained one Field Supervisor and eight (8) Field Surveyors to collect and conduct the crop-type data collection, field boundary mapping, and yield collection data. The training was completed in partnership with a technical team from the Sokoine University of Agriculture, a leading agricultural university in Tanzania. The objective of this phase was to improve our understanding of the field survey, build a good relationship among the survey team, refine the survey methodology to be rigorous and realistic, and ensure the data quality of the research. We also tested the equipment used during the data collection and introduced the project to local authorities and community leaders in Katavi.

From **March 16-18, 2022**, Flamingoo Foods and the Sokoine University of Agriculture led a hands-on, technical training of the field teams in Majimoto, Katavi, including a day of practical exercises. On **March 19, 2022**, the team pre-tested the ODK tools and equipment.

Flamingoo Foods led and coordinated the identification of farmers and pilot villages through its network of farmers in Western Tanzania and by working with local officials. The first survey on field boundaries started at the beginning of the rice production season and ultimately reached **806** farmers. The second survey was more challenging as the surveyor needed to reach the farm at the exact harvest time to measure the harvest quantity. The surveyors were in constant contact with the farmers from the first survey. With few exceptions, the surveyors were present at the time of harvest for the second survey, completing a total of **617** yield measurements.

2.3. Phase II: The Data Collection Phase

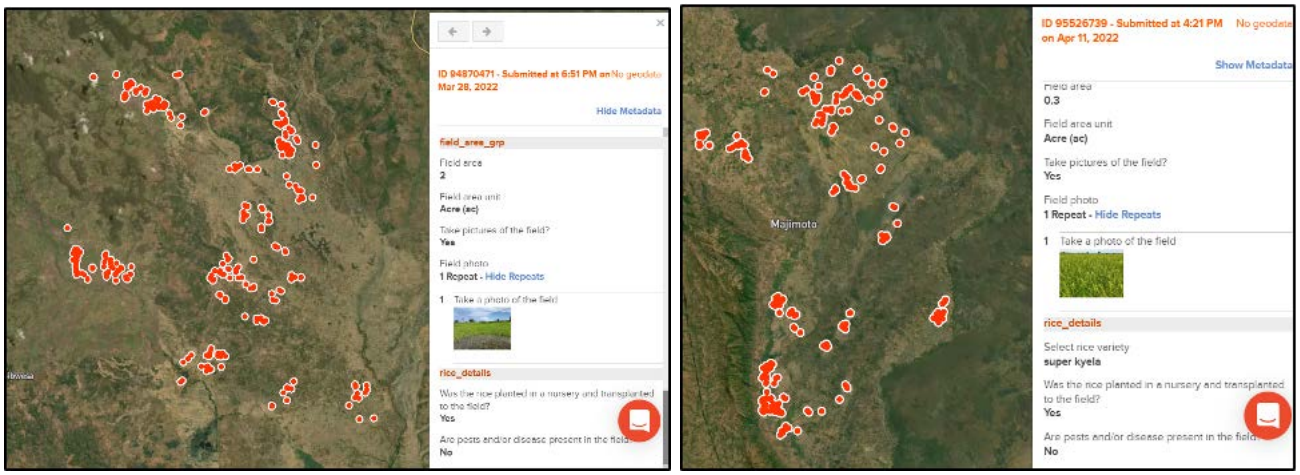
For this region, planting began in January for the rice crop season, and the harvest took place in May and early June. The planting date was recorded during the first survey, with additional data available via the data collection forms for each region. The surveyors also recorded the intended harvest date and remained in contact with the farmers to coordinate the date of the second survey. The same surveyor who completed the first survey returned for the second survey.

Access to the fields was challenging as some fields were very remote, and transport costs were higher than previously budgeted. In particular, access to several regions during the first survey was challenging due to heavy rains rendering some paths unpassable. It was too wet to walk the paddy fields and measure the boundaries. Therefore, future data collection efforts should start before or after the wettest part of the rainy season.

2.3.1 Field Farmer's Survey Part 1: Field Delineation

This survey included GPS-based delineation of **806** paddy fields. Field-based interviews followed a structured questionnaire format with smartphones with installed ODK software. The data collection tools included field location, field boundary and size; crop variety, presence and intensity of pests and diseases; management practices; planting, flowering, and harvest dates, and the distribution of marketing material.

Image 1: Photographs taken of the rice paddies by the Field Surveyors.



Images 2 & 3: Location of rice paddy fields mapping during the first round of surveys in Ikaka (left) and Majimoto (right) in the Katavi region, along with examples of the meta and other data collected at plot location.

Approximately one month was reserved for the first survey, which took place from **March 21 – April 15, 2022**, across two major paddy-growing areas (Majimoto and Ikaka), which are located in the Katavi region, the western part of the United Republic of Tanzania **(see Images 1 & 2 on previous page)**. This worked out to approximately 33 survey fields per day or about 4.1 fields per surveyor. One hundred percent of the participating farmers gave the Field Surveyors permission to walk inside the field boundary and indication of a strong willingness to participate.

Following the data collection procedure, the technical team identified and corrected any erroneous GPS measurements and replaced the field boundaries for more than 20 fields. This was necessary for inaccurate data. To ensure the best accuracy, the team purchased the latest smartphones that accommodated all major global navigation systems. Through the equipment testing process, we could not measure any significant difference in the accuracy of the GPS measurements as reported by the ODK. For the smartphone that recorded too inaccurate field boundaries, the team returned and re-measured those boundaries. A short break was scheduled between the two surveys to accommodate any potential delays.

2.3.2. Field Farmer's Survey Part 2: Yield Measurements

The second survey was anticipated to take longer, and the team planned for about one and a half months to reach the intended 600 yield measurements. The Field Surveyors identified these 600 farmers from the original 800 based on farmer interest in participating and their communication of the exact harvest date. This survey started on April 25 and concluded on June 29, 2022. The surveyors visited the farmers on the day of harvest and averaged about ten daily yield measurements. The team reported that the second survey was easier to coordinate as rapport had been developed between the farmers and the surveyors, and the farmers were eager for the surveyors to return to perform the measurement. The survey helped build trust with the local farming community, facilitating future business operations, including additional sampling efforts.

During the second half of the second survey, a preliminary review of the data revealed some inconsistencies in the yield totals, with about 10–20 yield data points flagged as falling outside the standard deviation for rice yield for the region, which averages about 5 tonnes/hectare. Some data points were erroneous (e.g., 50 tonnes/ha), which was attributable to Field Surveyor error, while others (e.g., 12 tonnes/ha) were less straightforward.

To supplement these points, the team collected additional data via crop cuts (approximately 12% of the fields where yield measurement was conducted) and some moisture content data, which will be verified in the Lab. Erroneous data was deleted. By extending the second survey period, the team collected additional data points to make up for the erroneous data and ultimately reached **620** paddy fields.

3. Yield Estimation Model

Using the ground data, the team tested multiple machine learning models to map rice and estimate end-of-season rice yield in the larger region using satellite data. This section summarizes the data, methods, results, evaluation, and next steps.

3.1. Datasets Utilized

Dataset	Dataset	Source/ Description	Application
Field/Ground Data	617 ground-based labels of Rice field location,	Field campaign utilizing an adapted version of the ODK standard toolkit developed by ECAAS in collaboration with Radiant Earth Foundation	Training data for crop-type mapping
	Field boundaries		Training yield model
	Yield Data		Training yield model
Satellite Data	Sentinel-1	Sentinel 1 VH channel and DEM data and temporal backscattering	Signature for rice and nonrice regions in VH channel
	Sentinel-2	Google Earth Engine	Crop-type Mapping
	Normalized Difference Vegetation Index (NDVI) from Sentinel-2	Normalized Difference Vegetation Index (NDVI) quantifies vegetation by measuring the difference between near-infrared (which vegetation strongly reflects) and red light (which vegetation absorbs)	Yield and crop-type modeling- biweekly from January through June from Google Earth Engine.
	Green Chlorophyll Vegetation Index (GCVI) from Sentinel-2	The Chlorophyll Index - Green (Clg) is a vegetation index used to estimate leaf chlorophyll content in plants based on near-infrared and green bands. In general, the chlorophyll value directly reflects the vegetation.	Yield modeling- biweekly from January through June from Google Earth Engine.

3.2 Test Data Boundaries

A shapefile was created to represent the geographical boundaries of each surveyed field. A GPS survey was used on-site to calculate the coordinates of the field's boundaries. This data was then converted into Geometry data in Python. These initial polygons underwent digitization before yield and area were calculated. The GEOCIF data for January through June comes with multiple features that provide information about the field's location, area, geometry, and harvest. Given a shapefile with a polygon for each field and an attribute for the weight of the harvest, we calculate a yield value for each field. Geopandas is used to convert the file to the UTM Zone 36S coordinate reference system, and then the area is calculated for each polygon in square meters. Geopandas also allows yield calculation by dividing the weight by the area in a new attribute.

3.3 Crop Mask Generation

A single seasonal mosaic was created from Level-2A Sentinel-2 imagery in Google Earth Engine (GEE), excluding cloudy and shadow pixels, using the Sentinel2Cloud probability masks. The median pixel mosaic was then clipped to the area of interest. Only 10-meters (Bands 2, 3, 4, and 8) and 20 meters (Bands 5, 6, 7, 8A, 11, and 12) were used for this analysis. The 20-meter bands were pan-sharpened to 10 meters before training the crop-type classifier. An NDVI band was added to the seasonal mosaic. NDVI data are commonly used in land cover mapping and have been shown to improve model accuracy when added as an input variable. The random Forests (RF) model was selected after several experiments, including running the support vector machine and Regression Trees and Random Forest, all readily accessible in GEE. RF had the highest overall accuracy. The rice crop mask was then applied to the yield estimation models following the generation of predicted yields. The data used to train models were all verified crop sites.

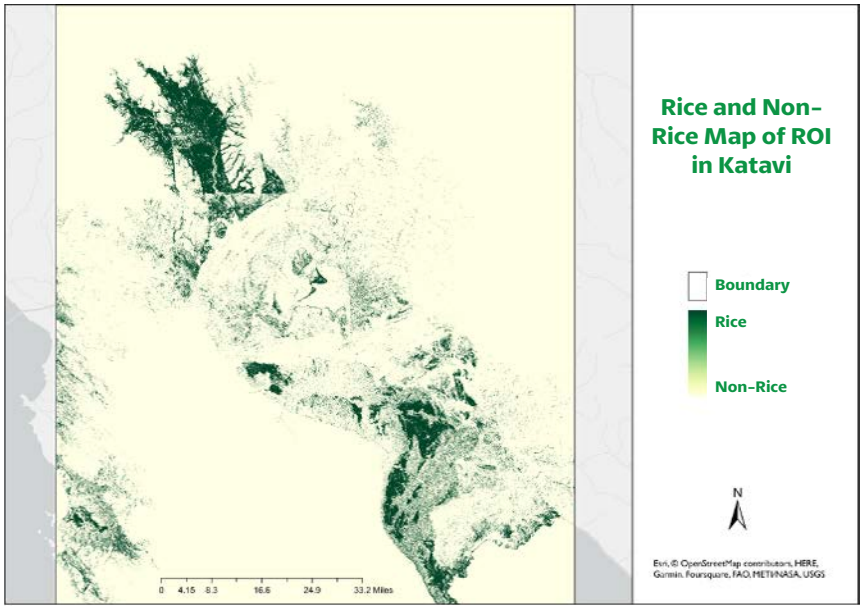


Figure 3:
Crop Mask Developed using sentinel-1 and Sentinel -2 Data in GEE

3.4. Yield Mapping

Multiple time-series satellite datasets utilized for this task were imported from GEE. The shapefile for the region of interest (Katavi, administrative level 1) to export data from the area of interest from January to June 2022. Features included Normalized Difference Vegetation Index (NDVI) **(Figure 2)** and Green Chlorophyll Vegetation Index (GCVI) from Sentinel-2 data **(Figure 3)** and run through multiple prediction models to compare and analyze the results, performance metrics, and visualization.

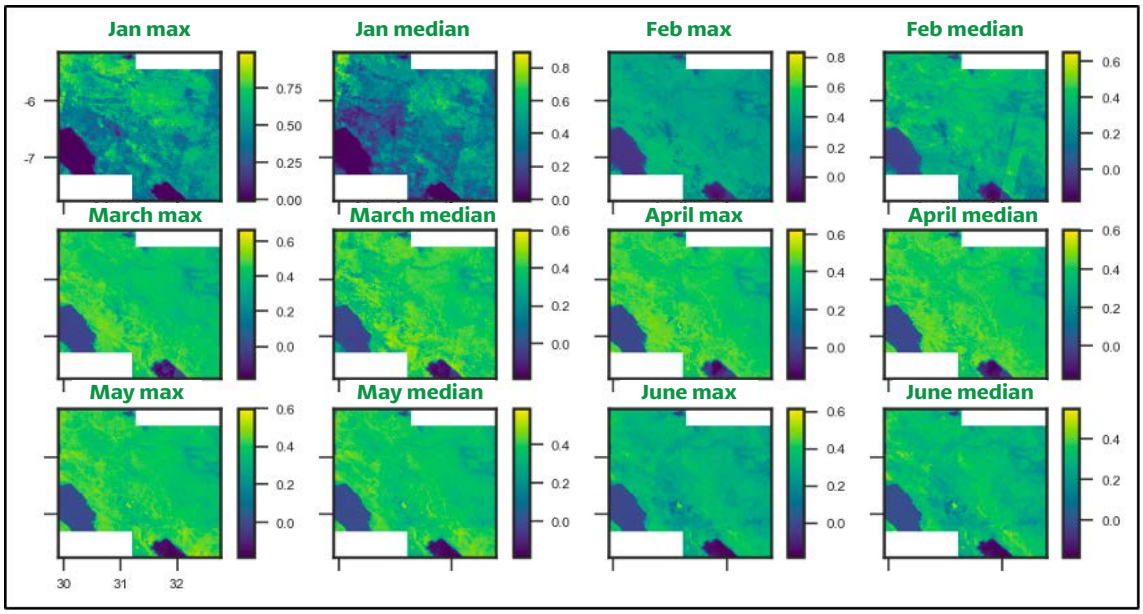


Figure 3:
Maps of NDVI max and median for each month of the rice growing season evaluated in the survey

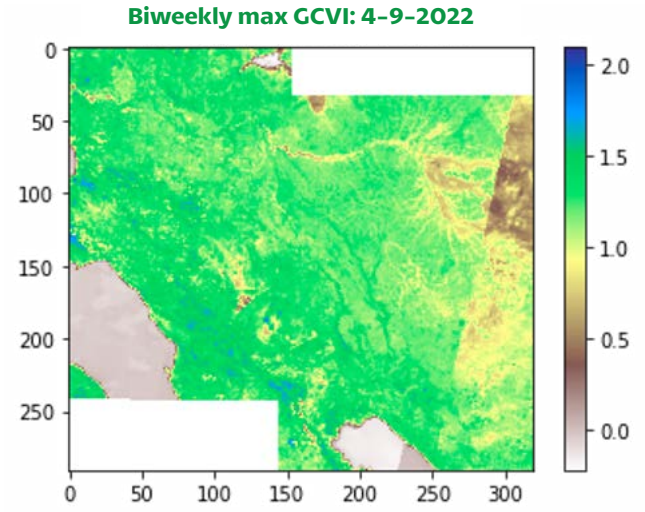


Figure 4:
Plot of raster containing max GCVI values over the two weeks (4/9/2022 – 4/23/2022).

3.5. Building the Regional Model

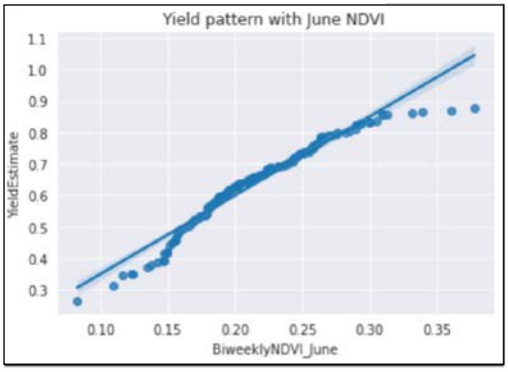
With the python library Pyspatialml, scikit-learn machine learning models can be applied to raster-based datasets. Here we used monthly maximum NDVI rasters extracted from the GEE products as bands with the tsraster library as input features.

The target data, field-scale yield values, were in a shapefile format that could be overlaid on the rasters to extract corresponding feature raster values to create the training data set. Pyspatialml took the target data in shapefile format and the feature data raster format to form a data frame that, for every field surveyed, matched the one-pixel value for each feature identified. The main two models run at 10m were done separately for each vegetation indices, and all features represented a biweekly period within the study's timeline. Some models run at 100m included both vegetation indices and used a combination of monthly maximum values and median values extracted by month. The training data frames that were created matched the yield for each. Pyspatialml also includes methods to plot the results and modify rasters. Pyspatialml outputs predictions as raster pixels, so the inputs for the testing data remain in their raster form.

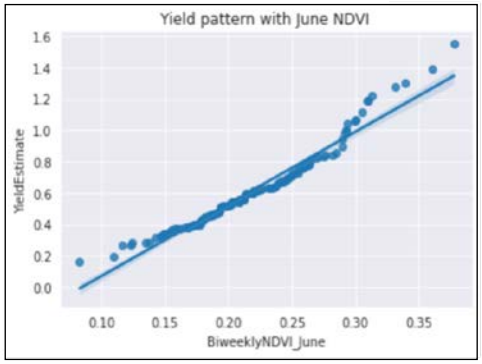
3.6. NDVI Data Modeling

The preprocessed data was run through multiple prediction models to compare and analyze the results, performance metrics, and visualization. The regression models assessed are shown in **Figure 5 below**, showing model predicted yield values against the Maximum Bi-weekly NDVI values for June.

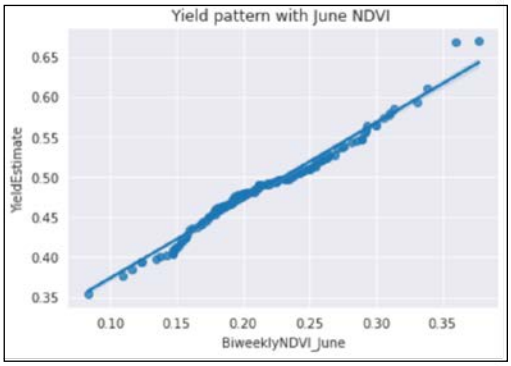
1. Linear regression:



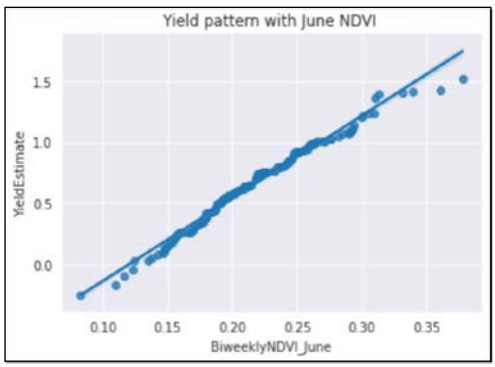
2.



3. Support Vector Machine: Regression:



4. Polynomial



5. Random Forest Regression:

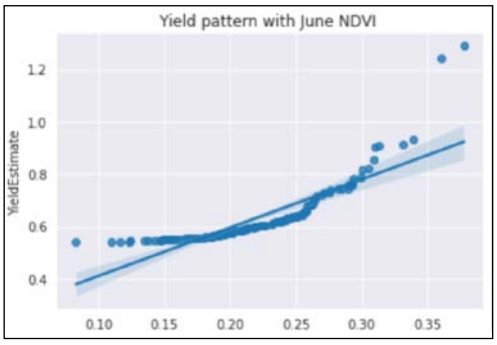


Figure 5:
Predicted values for yield vs. NDVI values during June. Each plot represents predictions generated with a different machine learning model. Input NDVI data at 10m resolution.

3.7. Performance Metrics and Model Evaluation

Evaluation Metrics:

1. RMSE - Root Mean Squared Error
2. R-squared
3. Cross-Validation Accuracy

Performance Metrics comparison:

Model	Lin	Svm_rbf	svm_sig	svm_poly	svm_lin	Xgb	Dec. Tree	Poly	Lasso	Ridge
RMSE	0.472	0.709	0.709	0.712	0.708	0.730	1.104	0.773	0.692	0.690
R^2	0.015	-0.049	-0.048	-0.057	-0.046	-0.113	-1.546	-0.247	-0.0004	0.007
CV Accuracy	-0.595	-0.576	-0.580	-0.582	-0.579	-0.675	-0.864	-0.595	-0.570	-0.574

Features (12)

- Max monthly NDVI (Jan, Feb, Mar, Apr, May, Jun)
- Median NDVI (Jan, Feb, Mar, Apr, May, Jun)

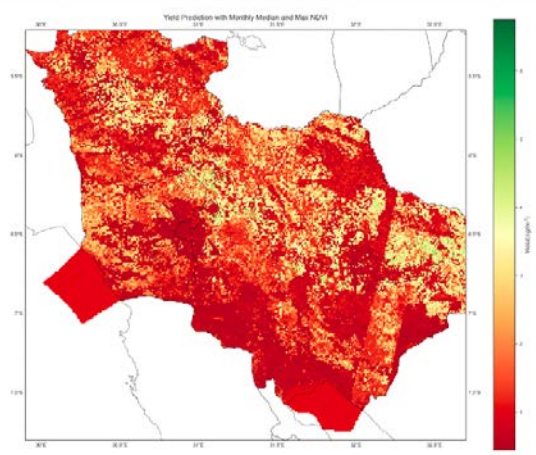
RMSE= 1.0693558359659092

R2 Score= -0.21230262885807005

- 12, NDVI max for biweekly periods from 1-01 to 6-18

RMSE= 0.03549804267800871

R2 Score= 0.6338055460535752



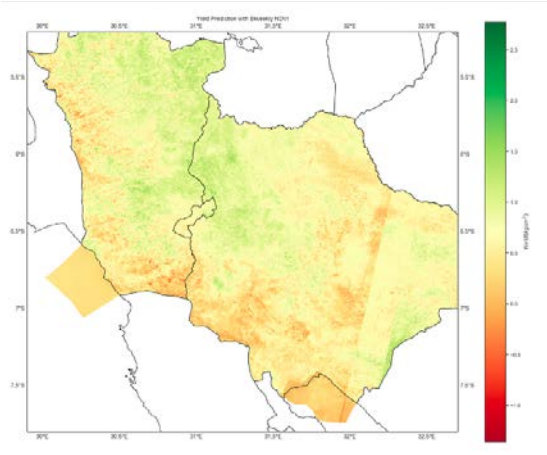


Figure 6: (left)

Output from Random Forest Regressor Model with monthly max and median NDVI as features. Masked to only show predictions within the region of interest. Borders represent Administrative Level 2 boundaries.

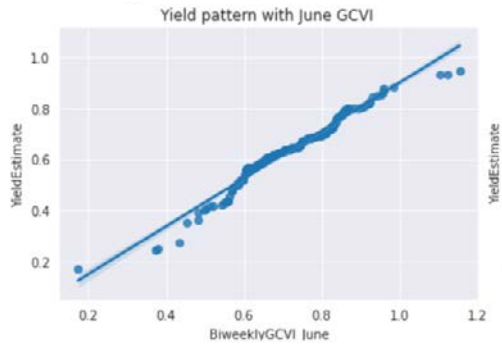
Figure 7: (right)

Output from Linear Regression Model with biweekly max NDVI as features. Masked to only show predictions within the region of interest. Borders represent Administrative Level 2 boundaries.

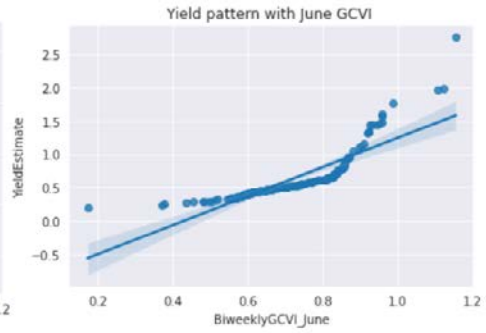
3.8. Preferred Model

From the above performance metrics, we note that the Linear Regression algorithm performed the best on the maximum NDVI data for the 13 biweekly periods from January through June with the lowest Root Mean Squared Error Value of 0.47 and an R-squared value of 0.015. None of the R-squared values showed strong correlation evidence, likely due to the limitations in data splitting and feature selection. The value for R-squared is within the typical range for Linear Regression yield models based solely on NDVI. The Linear Regression also had the smallest RMSE value across these models.

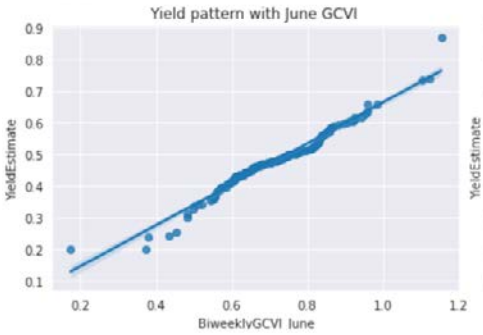
1. Linear regression:



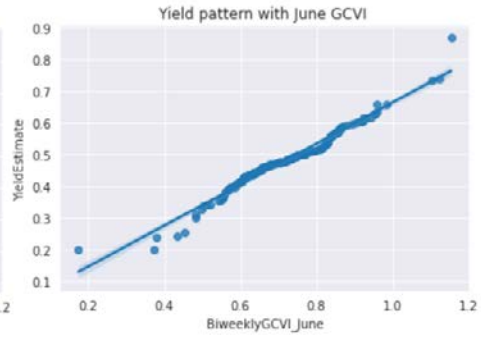
2. XGBOOST:



3. Support Vector Machine



4. Polynomial Regression



5. Random Regression:

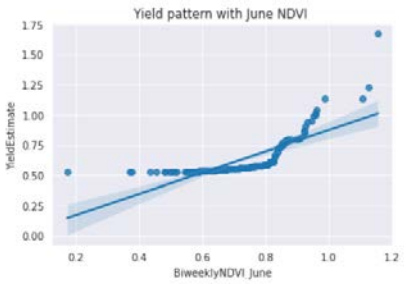


Figure 8:
Predicted values for yield vs. GCVI values from June. Runs were done with 10m resolution data. Each plot represents predictions generated with a different machine-learning model.

Run 2 – 100m resolution

Features 2

- Max seasonal NDVI
- Max seasonal GCVI

R2 score =1.001230770336234

RMSE/MSE = -0.06275935189089621

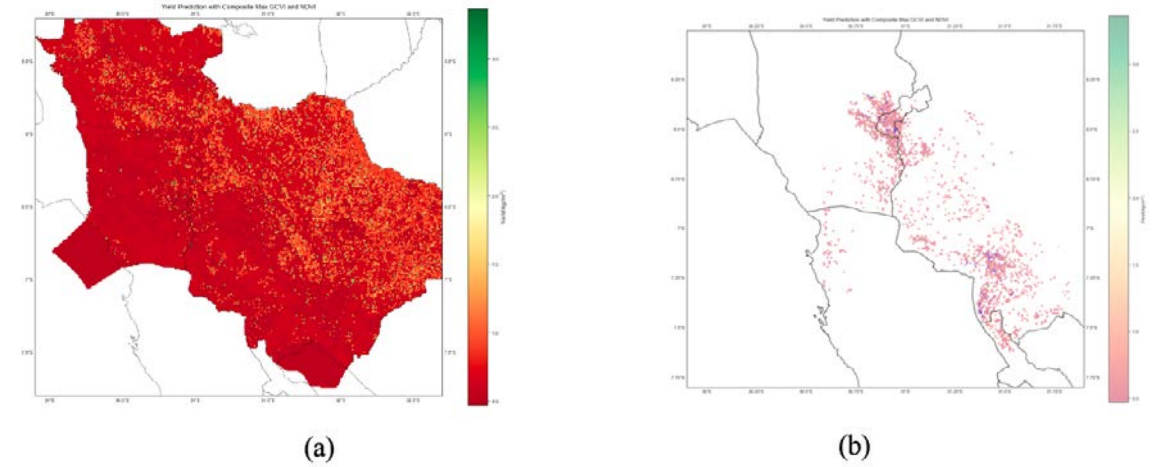


Figure 9: (a)

Output from Random Forest Regressor Model with seasonal max GCVI and NDVI as features. Masked to only show predictions within the region of interest. Borders represent Administrative Level 2 boundaries. (b) Output but masked only to show results for areas where rice is known to be growing based on crop mask. Blue polygons represent the locations of farms from the survey.

Run 4 – 10m resolution

Features extracted = 12:

GCVI max for biweekly periods from 1–01 to 6–18

R2 score = 0.033955251108837126

RMSE/MSE= 0.6287303658801712

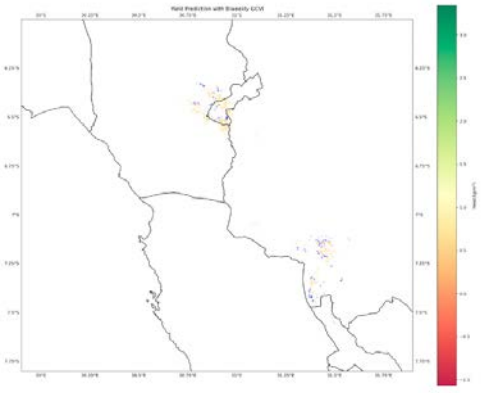


Figure 10:

Output from Linear Regression Model with biweekly max GCVI as features. Masked with crop mask to only show predictions where rice is growing. Borders represent Administrative Level 2 boundaries.

3.9. Observations

A rice crop mask was applied to the model after training the model to limit predictions to those regions where rice was growing. Below, **Figure 12** demonstrates differences in yields between the yield set used for training and testing the model taken from the survey, the predictions made for pixels based on the rice mask (e.g. shown in **Figure 10b and Figure 11**), and the predictions made for all pixels within the raster (e.g. **Figures 7, 8 and 10a**). The trend line for the predictions beneath the mask resembles the survey yields predictions.

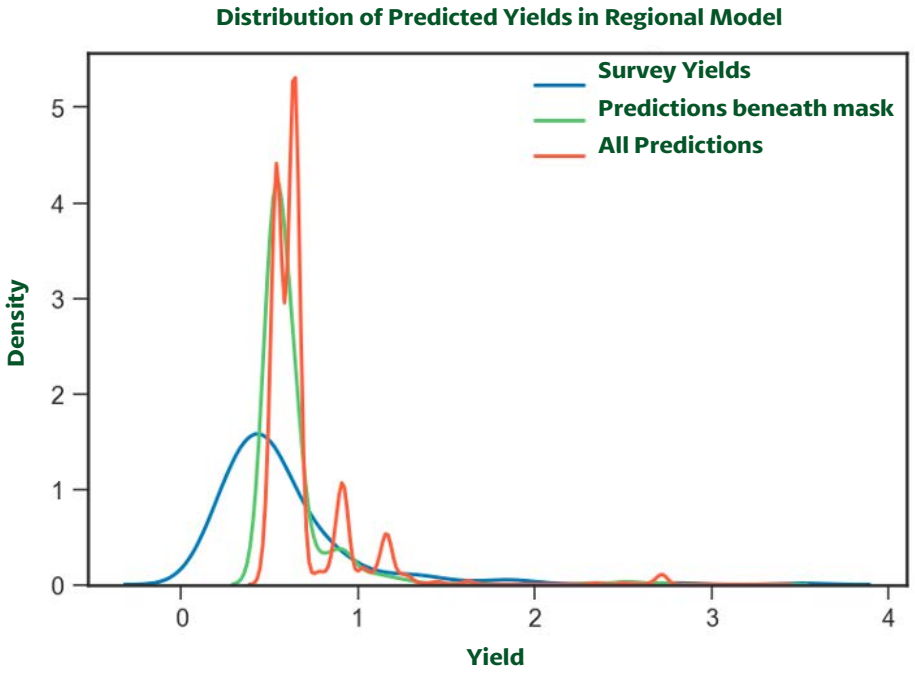


Figure 11:
Shows the distribution of yields used for training Run 2 of the model (100m) and the predictions. The blue line represents the distribution of the yields recorded in the farm survey. The green line represents yields for only pixels that were mapped as rice

3.10. Future Research Plans

Future work will aim to increase the resolution of the input features up to 10m resolution; incorporating more composite features into the model may aid the model in recognizing more useful trends within the NDVI and GCVI data. Potential features to apply include medians for monthly composites, quantiles, or mean absolute change, the field management data, and other EO data, including rainfall, temperature, and soil moisture. In addition, rice crop masks will be incorporated more into the input data before running the models. This will promote the model's efficiency and help create more reasonable predictions.

2

Annex: Field Data Summary

This section discusses relevant findings from the rice crop yield analysis of the surveyed region. Beyond the crop yield data, which was the primary purpose of this survey, the project collected various auxiliary data. As summarized in the Scalability Report, this data is useful as hen training models and provides a better understanding of the practices of local rice producers.

Image 1: Photographs taken of the rice paddies by the Field Surveyors.

Over 99% of the field areas were recorded as acres, and almost all Field Surveyors had photos of the rice paddies within their data collection **(Image 1)**. While rice was the primary crop for all surveyed farmers, about ten farms also reported maize as a secondary crop. Six major rice varieties were recorded, and yield outcomes ultimately varied significantly by variety, representing an even stronger predictive factor than other farm practices **(Figure 1)**, such as transplantation and irrigation.

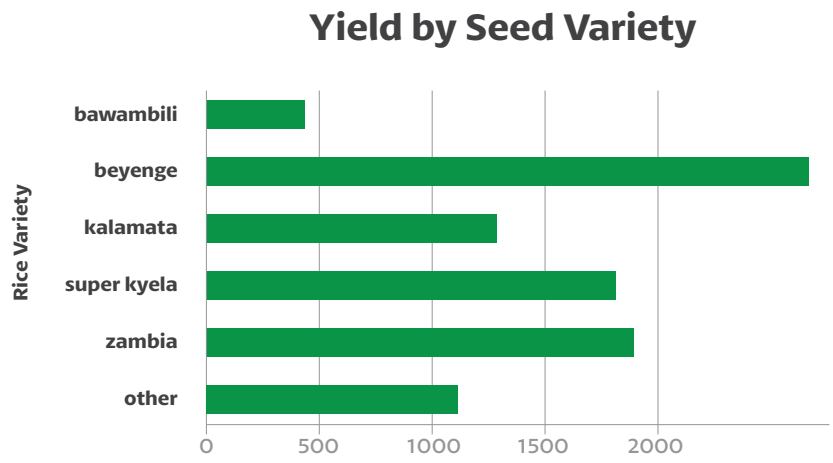


Figure 1.
Yield by seed variety

Information collected about the different planted rice varieties also has commercial relevance, as different varieties vary in their aroma and cooking properties. The survey revealed that Ikaka and Majimoto were reasonably distinct in the planted rice varieties **(Figure 2)**.

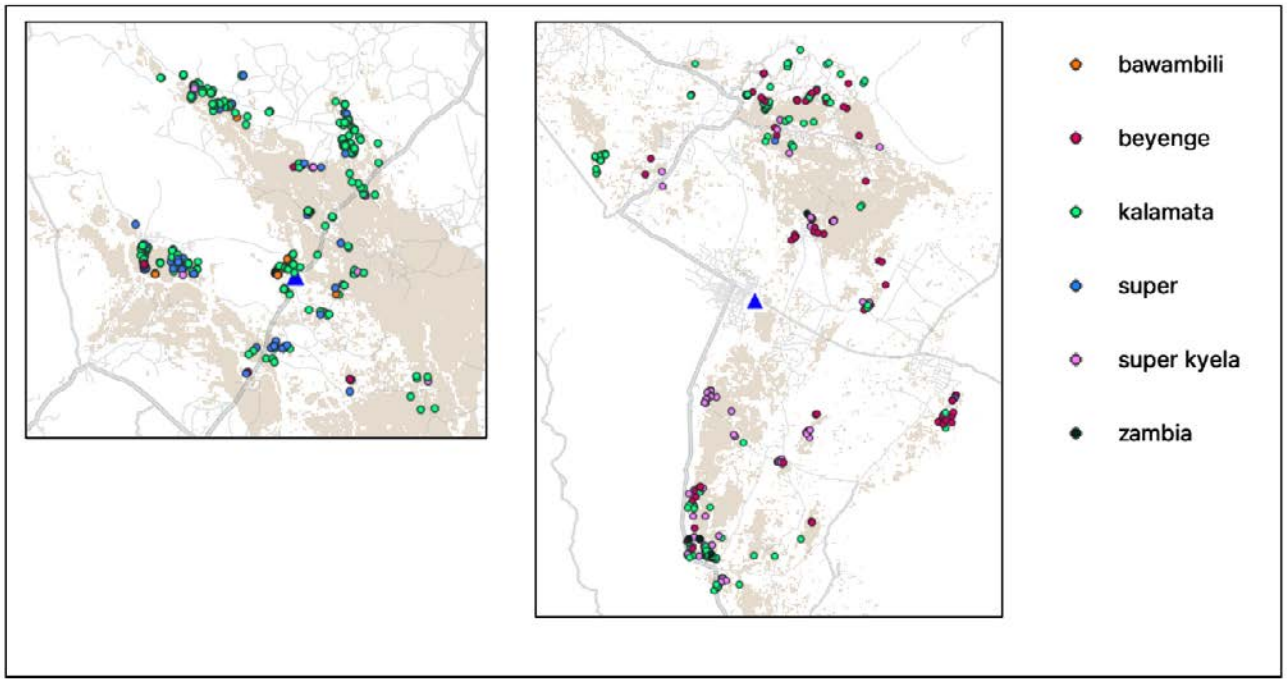


Figure 2:
Most rice fields in Ikaka (top panel) were planted with Kalamata and Super rice. In Majimoto (bottom panel), Benyenge and Super Kyela are the dominant varieties next to Kalamata.

Planting started at the end of November 2021, and while the majority of the surveyed farmers planted during January and February of 2022, there was ongoing planting through March **(Figure 3)**. The Katavi region of Tanzania is located in the unimodal rainfall zone, where the Msimu rainy season lasts from November through April. Ultimately there was significant yield variation due to the planting date, with farms that started in early January experiencing the highest yield performance.

This 2022 season, droughts across Tanzania led to crop failure and delayed planting and harvesting. During the first survey, the team included a question about the anticipated harvest date to facilitate planning the harvesting measurement. This data was extremely valuable in predicting harvest onset and peak timing. Knowing the expected delay in harvest is crucial for determining the right timing for purchasing paddy stocks. The survey helped the team anticipate which region would harvest first and when to expect the peak of rice paddy to flood the market. Future efforts can include building a test tool that enables the prediction of spatial variability in planting dates and yield outcomes.

The evaluation metrics on the test set showed that the model's accuracy was only marginally useful. Further research is needed to verify the operational use of this model.

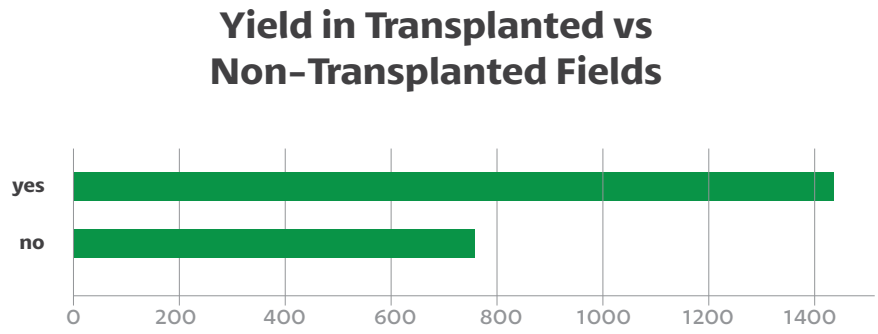


Figure 4:
Measured yield in transplanted versus non-transplanted rice paddy fields.

Most (89.8%) of the rice was planted in a nursery and then transplanted to the field **(Figure 4)**, though this corresponded somewhat to rice variety. Rice that had been transplanted had approximately double the amount of yield as non-transplanted rice. This might, in part, correspond to which rice varieties were transplanted. The *Beyenge* variety, the best yield performer of the six identified during the survey, was transplanted 99% of the time.

Figure 5:
Yield performance by rice variety

The majority of the farmers (542/806 = 67.2%) were rainfall dependent, though about 15% (121/806) either irrigated or used supplementary irrigation **(Figure 6)**. Water management significantly corresponded to rice variety, with Zambia rice being 100% rainfall dependent, while almost half of the farmers cultivating the super variety either irrigated or used supplementary irrigation.

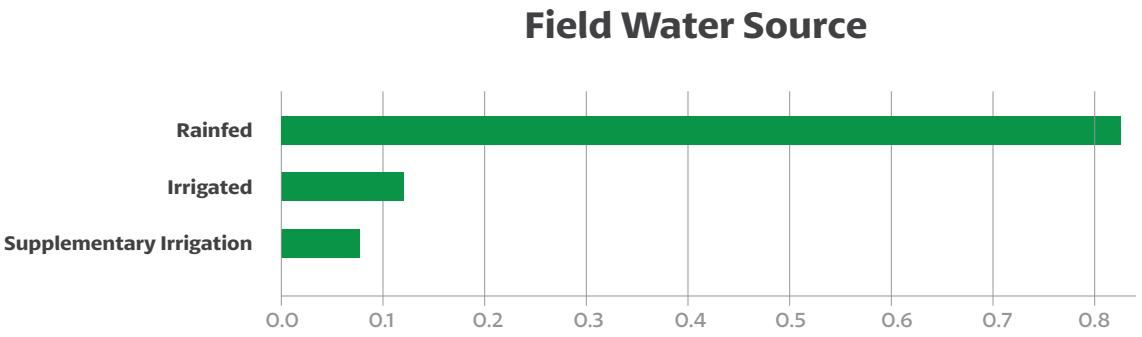


Figure 6:
Proportion of surveyed rice fields that were rainfed, irrigated, or received supplementary irrigation.

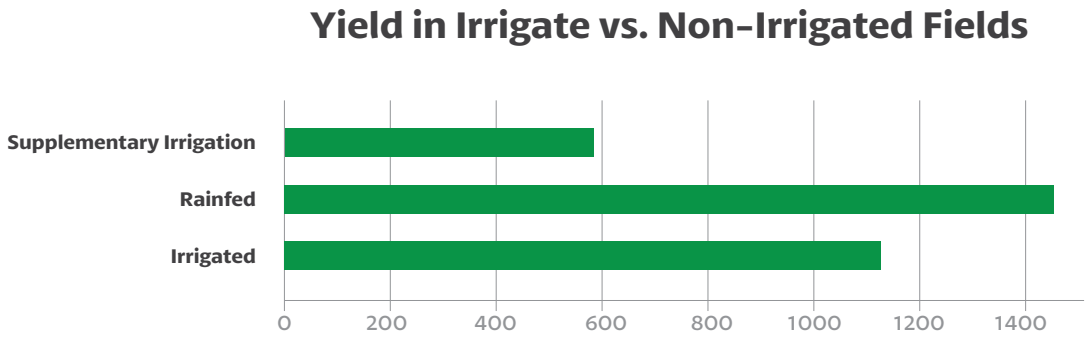


Figure 7:
Yield performance in irrigated vs. non-irrigated rice fields.

This can explain the significant variation in yield outcome between fields that were irrigated or used supplementary irrigation and those that were rainfed is likely also attributable to rice variety **(Figure 7)**. Zambia rice was one of the best-performing varieties, second only to Beyenge, while the super rice variety was second to last in terms of yield performance **(Figure 8)**.

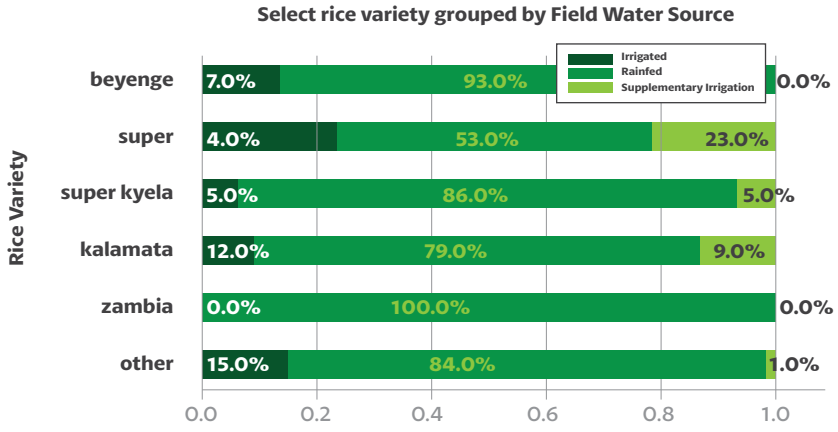
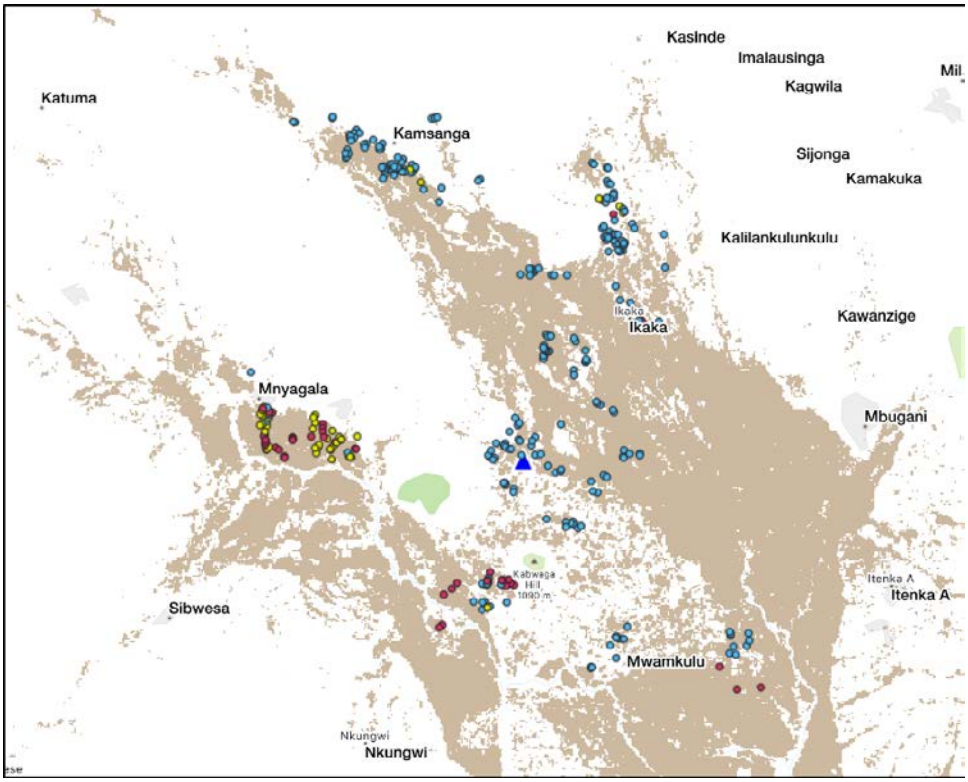


Figure 8:
Field water sources varied significantly by rice variety.

Surveying irrigation practices revealed interesting information about the region, including the discovery of a small irrigation scheme that was implemented in the villages of Mnyagala and Mwamkulu **(Figure 9)**. Irrigation schemes are a valuable asset in drought years. The presence of irrigation schemes is relevant for supply chain planning, particularly regarding seed variety, as noted above, since irrigation can allow favorable yields even in years with insufficient rainfall.

Figure 9.
The survey revealed the presence of full irrigation (red) and supplementary irrigation (yellow) between Mnyagala and Mwamkulu. Rainfed fields are in blue.



Enabling Crop Analytics At Scale

Report on Final Results & Lessons Learned

Optimizing Crop Yield Data Collection for Supply Chain Enhancement, An Enabling Crop Analytics (ECAAS) at Scale project

info.ecaas@tetratech.com
cropanalytics.net

Final report
January 2023

Prepared by:
University of Maryland/NASA Harvest Team