Technical Documentation

UNIVERSITY OF CALIFORNIA
Santa Barbara

AN OPEN-SOURCE PIPELINE FOR REMOTE SENSING OF CROP YIELDS:
A ZAMBIA CASE STUDY

A Capstone Project submitted in partial satisfaction of the requirements for the degree
of Master of Environmental Data Science for the
Bren School of Environmental Science & Management

By:

Cullen Molitor
Grace Lewin
Juliet Cohen
Steven Cognac

Committee in charge:
Tamma Carleton
Allison Horst
Jonathan Proctor

JUNE 2022

Technical Documentation Signature Page

AN OPEN-SOURCE PIPELINE FOR REMOTE SENSING OF CROP YIELDS:
A ZAMBIA CASE STUDY

The Bren School of Environmental Science & Management produces professionals with unrivaled training in environmental science and management who will devote their unique skills to the diagnosis, assessment, mitigation, prevention, and remedy of the environmental problems of today and the future. A guiding principle of the School is that the analysis of environmental problems requires quantitative training in more than one discipline and an awareness of the physical, biological, social, political, and economic consequences that arise from scientific or technological decisions.

The Capstone Project is required of all students in the Master of Environmental Data Science (MEDS) Program. The project is a six-month-long activity in which small groups of students contribute to data science practices, products or analyses that address a challenge or need related to a specific environmental issue. This MEDS Capstone Project Technical Documentation is authored by MEDS students and has been reviewed and approved by:

_____
Cullen Molitor

_____
Grace Lewin

_____
Juliet Cohen

_____
Steven Cognac

This Capstone Technical Documentation is authored by MEDS students and has been reviewed and approved by:

_____

Tamma Carleton

_____

Jonathan Proctor

_____

Date

Table of Contents

1.0 Abstract

The environmental and social impacts of climate change are disproportionately distributed worldwide. Many highly impacted regions lack the assets to monitor and generate resource predictions, and therefore lack high-quality environmental and social data. As a result, it is difficult to make predictions about the impacts of climate change for these regions using conventional modeling. Recently, machine learning approaches applied to high-resolution satellite imagery have been successful in making predictions of a wide range of social and environmental variables. However, generating these predictions comes with significant barriers, including high computational, data storage, expertise, and financial resource costs. Reducing the financial and computational burden of machine learning approaches is essential to increasing the equity of environmental monitoring processes and outputs. Here, we demonstrate a pipeline to make predictions on ground-truthed data using the Random Convolutional Features method through a use case example of predicting crop yields in Zambia. These crop yield predictions can be used to analyze food security risk in the region. We apply the novel machine learning approach, MOSAIKS (Rolf et al., 2021), to create tabular features for Zambia using Landsat 8 and Sentinel 2 satellite imagery. We pair these generated features of Zambia with ground-truthed crop yield data to build a model that predicts crop yields over time and increases the spatial resolution of predicted crop yields. We then use this model to fill in a data gap of crop yield predictions in Zambia during the years 2020 and 2021, when crop yield data was not collected due to the COVID-19 pandemic. Beyond this use case, these tabular features of satellite imagery, and the intuitive Microsoft Planetary Computer featurization pipeline we developed to create them, provide a tool for others around the globe to create features, build models, and generate predictions of other social and environmental variables.

2.0 Executive Summary

The environmental and social impacts of climate change are unequally distributed worldwide (Porter et al. 2014). Future climate patterns are predicted to adversely affect agricultural productivity across many regions of the globe, posing a threat to large-scale food security, and yet, sub-national crop production data is sparse in regions of the globe projected to be most affected by climate change (Hultgren et al., 2022, in prep). Development of a comprehensive agricultural database on historical crop yields that reflects local climate conditions would be a major development in the field. This comprehensive database would enable an understanding of the historical relationship between climate conditions and agricultural output, as well as the development of forecasts under future climate change.

Recently, machine learning approaches applied to high-resolution satellite imagery have been successful in making predictions for a wide range of social and environmental variables (e.g., Burke et al. 2021, Jean et al. 2016), including agricultural yields (e.g., Quarmby et al., 1993). However, generating these predictions comes with significant barriers, including high computational costs, data storage, expertise, and financial resources. Reducing the financial and computational burdens of satellite imagery with machine learning approaches is essential to increasing the equity of environmental monitoring processes and outputs.

Here, we demonstrate a pipeline that pairs Random Convolutional Features (Rahimi and Recht, 2008) from two satellite imagery sources with ground-truthed data to make predictions on crop yields in the country of Zambia. Using crop yield data from 2013-2019, we predict the yields during the years of 2020 and 2021, when data was unable to be collected due to the COVID-19 pandemic. In addition to generating predictions to fill this data gap, we also increase the spatial resolution of the crop yield data that was collected. These crop yield predictions can contribute to analysis on food security risk in the region. Specifically, we apply the novel machine learning approach, Multi-task Observation using Satellite Imagery & Kitchen Sinks (MOSAIKS; Rolf et al. 2021), which encompasses both featurization and linear prediction. We use the MOSAIKS system to pair Landsat 8 and Sentinel 2 satellite imagery with machine learning (SIML) methods to create tabular features encoded with satellite information. Our methods to generate these features with publicly available satellite imagery are an important contribution to the MOSAIKS codebase, and can be used by the wider user base to generate features of images across many locations. Once features were made, we pair them with ground-truthed administrative crop data to build a supervised machine learning model that increases the spatial resolution of crop yield predictions and predicts crop yields over time to fill the data gap in Zambian crop yields due to data collection constraints during the COVID-19 pandemic. We implemented cross-validated ridge regression across all years of interest to make these crop yield predictions. Previous work has demonstrated that this unsupervised featurization can match the performance of deep learning methods across multiple tasks (Rolf et al., 2021). The features, predictions, and code base generated by application of the MOSAIKS system can be used by a variety of users, including those interested in analyzing current and future food security risk in Zambia.

These tabular features created from satellite imagery using MOSAIKS are task agnostic, meaning they can be used to predict any variable of interest. In future work led by the clients, the featurized satellite images and predicted crop yields will be integrated into a public-facing and freely accessible application programming interface (API). We hope that access to the preprocessed features and the methods to predict any variable of interest will enable researchers and decision-makers around the globe to generate predictions of a wide variety of other social and environmental variables.

3.0 Problem Statement

Climate change is predicted to adversely affect agricultural productivity across the globe, posing a threat to large-scale food security. Satellite imagery paired with machine learning can be used to make crop yield predictions for vulnerable regions. However, these satellite and machine learning methods require a high level of financial and computational resources. Sub-Saharan Africa is likely to suffer some of the largest impacts from climate change (Hultgren et al. 2022*,* in prep., Kurukulasuriya and Mendelsohn 2007). This region depends on local agricultural production to ensure food security across the continent, but there is currently no efficient way to generate estimates for historical crop yields and how they correlate with varying environmental factors. Additionally, the country of Zambia was unable to collect data on crop yields during the years of 2020 and 2021 due to the COVID-19 pandemic. This presents a critical data gap about food resources during the pandemic. This lack of data and computational resources in sub-Saharan Africa can be remedied with a pre-processed, generalized collection of encoded satellite data that would be stored in open-source archives. This database of features will enable policy makers, researchers, and all other users to execute diverse analysis and generate predictions of many domain specific tasks on basic laptops in a reasonable amount of time. In this use case, we use these features and an optimized model to predict maize yields during the years 2020 and 2021.

4.0 Specific Objectives

The objectives of this project are as follows:

1. **Featurize Satellite Imagery:** The primary objective of this project is to encode annual satellite imagery with random convolutional features over time for Zambia using the Kitchen Sink method of the MOSAIKS system.
2. **Predict Crop Yields:** The secondary objective is to pair ground truth administrative crop yield data for Zambia with featurized data summarized to the administrative boundary level. We will use this paired data to train a ridge regression model of crop yields on the features. This model will be applied to Zambia for the years 2020 and 2021 (due to the inability to collect data from COVID-19 restrictions) to make predictions of crop yields.
3. **Pipeline:** A pipeline that future users can apply to featurize monthly imagery for multiple satellites and to use those features to make crop yield predictions. (See future applications section).

5.0 Summary of Solution Design

The strategy to accomplish this project's objectives is listed below. A schematic outlining our approach is also provided in Appendix A.

1. Create a uniform grid of points spaced at even intervals from which to sample for the featurization step. These points are within the local coordinate reference system with

units in meters. Each selected point was set to 5 kilometers away from the other points. For computational efficiency, every n^{th} point was selected and each row was set to alternate starting points to create a checkerboard. A 1 square kilometer equal-area grid cell was buffered around each point to serve as evenly-spaced regions to match Landsat 8 and Sentinel 2 satellite imagery in the next step. After the grid cells were created, the grid was reprojected into the standard geodetic coordinate system, EPSG 4326.

2. Pulled geo-located Landsat 8 and Sentinel 2 satellite data using the Planetary Computer Spatio-Temporal Asset Catalog (STAC) API into Microsoft's Planetary Computer Hub, filtered for cloud cover. The least cloudy image that meets the cloud threshold for any given month was used. If no imagery was available for a point during a month due to cloud cover, the point is skipped and later interpolated (the exact method is documented in the notebook crop_modeling.ipynb). The spatial and temporal criteria were determined by the amount of flexibility over these dimensions that would provide more meaningful information to the model than a null value, while adhering to restrictions that allowed us to glean model performance over these dimensions. See the User Documentation for more information on this interpolation approach.
3. Compute random convolutional features (Rahimi and Recht, 2008) over all satellite imagery that matched the 1 kilometer² equal area grid cells. This was executed both on MPC as well as the Azure server. Random Convolutional Featurization is an unsupervised machine learning computation with the MOSAIKS system (provided by the client) on satellite data, which uses a featurization technique to convert the satellite images into tabular, georeferenced data. We test many iterations of the number of satellite bands, the number of points featurized, and the time range in the modeling process.
4. Summarized featurized satellite data to administrative boundary level (districts) to match spatial resolution of ground-truth crop yield data and merged data frames spatially (over the latitude and longitude points that serve as the center of the grid cells).
5. Used cross-validated ridge regression on the merged data.
    a. We made the informed assumption that each crop year does not impact the next, because maize is annual and therefore all the yield is harvested by the end of the growing season in August, and the fields are prepped for the next years' crop season each November (Baylis Lab, personal communication, April 11, 2022).
    b. In the modeling code, we did not include year as a term in the model because if you have a year dummy in the model you cannot predict for 2020 and 2021.
    c. When we *apply* the model, rather than when we train and test the model, we do not use the same dataframe sans the crop data, because then we would be predicting the features at just the district level, which was only necessary for model training because that is the level of resolution of the crop data. Instead, we refer back to the raw features dataframe to predict at higher resolution: at the feature level. This data frame has more rows, because there are multiple points in each district and each of those points is present for each year, just as they were for the summarized features.

6. Generated predictions for 2020 and 2021 for maize in Zambia at 1 kilometer² grid cell resolution.
    a. Metadata such as the sources of data and uncertainty associated with these predictions are included in the User Documentation and published alongside the code to ensure users know the degree of imperfection in these estimates.
    b. During model optimization, model performance for training years (all years prior to 2020) was measured using multiple statistical metrics: validation R^2, training R^2, training R, demeaned R^2, and demeaned R. Based on these R^2 metrics, 2 best models were chosen. The best model for overall performance and the best model for performance over time. Model training was done with ridge regression. During model optimization, we did not check the test set R^2 because this would have prematurely revealed the out-of-bag model performance that predicted into years for which we do not have ground truth crop data (2020 and 2021).
    c. We ran the two selected models on the test set, or the out-of-bag sample to produce two test R^2 values. We produced maps at the district level as well as the feature level for these predictions. We produced maps to show uncertainty over space to identify certain districts in which the model consistently underpredicted or overpredicted.
7. A notebook for equal angle gridding over Zambia to provide a pipeline for executing this broader featurization and modeling pipeline to fit different needs
    a. For example, a use case would be producing feature data for the MOSAIKS API that matches the gridding approach used for their archived data.
8. An alternative notebook for crop area sampling using 10% most cropped grid cells per district.

6.0 Products and Deliverables

Table 1. Capstone deliverables and applications.

Deliverable Overview (details below)	Application
Sentinel 2 & Landsat 8 features at 1 km² grid cell resolution. The temporal range of all features from Sentinel 2 is from 2015 to 2021 and the temporal range of all features from Landsat 8 is from 2013 to 2021. All features and predictions from both satellites are at annual temporal resolution.	These feature files are a contribution to the client's database of features for the MOSAIKS API. These features are a novel contribution as they were derived from public satellites, rather than a private satellite. Although these features are equal area and the existing features in the database are equal area, the notebooks and documented workflow to create the features is accessible to the MOSAIKS API team in order to be adjusted to match the existing features database. Additionally, these features are applied to the developed model in order to produce

	predictions for the years 2020 and 2021 at both 1 kilometer² grid cell resolution and district resolution.
Project documentation, Sentinel and Landsat featurization documentation, and model documentation (a separate README for each repository, as well as code comments markdown chunks throughout notebooks)	Main GitHub organization README informs clients and users about the overall goal of the project, how it was executed, and where to start in reproducing the product. Guides clients and users through the Sentinel and Landsat featurization and Sentinel and Landsat modeling processes with explanations for cropMOSAIKS' default code decisions and where to change variables to adjust the featurization or model processes (such as using different sets of bands, time ranges, or imputation approaches)
Outline of Landsat featurization code (model documentation applies to both satellites)	Guides clients and users in featurizing Landsat imagery and where to change variables to adjust the features
Notebook for pipeline using equal angle rather than equal area	Guides clients or users in achieving certain Future Work and Research Ideas and creating features to match the equal angle features that already exist in the MOSAIKS API

6.1 Client Deliverables

1. **Features**
    a. 1,000 features for Sentinel with red, green, blue, and near infrared bands, for 20,000 points sampled equal angle at the top 10% of crop land for each district for all months across all of Zambia for 2016-2021
        i. NIR increases spectral resolution (adding another band in general), which in turn increases model accuracy
        ii. NIR shares the same spatial resolution (10 meter) as RGB bands
    b. 1,000 features for Landsat (bands 1-7), for 20,000 points sampled equal angle at the top 10% of crop land for each district for all months across all of Zambia for 2016-2021
    c. 1,000 features for Sentinel with red, green, blue, and near infrared bands, for 15,000 sampled equal area for all months across all of Zambia for 2016-2021
    d. 1,000 features for Landsat (bands 1-7), for 15,000 sampled equal area grid cells for all months across all of Zambia for 2013-2021
    e. Notebook for pipeline for using equal angle to achieve Future Work and Research Ideas
    f. Stored on Taylor for the duration of the capstone project.

i. After the duration of the project, the client and team members transferred feature files to a shared Google Drive to retain shared access in perpetuity. Later on, the client will upload the features to the API or ensure that they are available upon request to the public with clear documentation on the API. If full API integration requires a more comprehensive set of features or adjustments to the features (such as converting them from equal area to equal angle), this will be completed by the client or team members after the conclusion of the capstone project.

Figure 1. Visualization of a remotely sensed image that has been distilled into numbers, as is executed during the process of random convolutional featurization. This figure was produced for the purpose of scientific communication; it is not a true satellite image that has been distilled into features.

2. Model

a. Documented exploration of performance across multiple dimensions in markdown chunks in Modeling notebook and User Documentation:
   i. Months of the year included in feature set
   ii. Sensors, and their combination (concatenate Landsat 8 and Sentinel 2)
      1. Equal area (1 kilometer² grid cell resolution), 15,000 points, 2016-2018
   iii. Ability to predict over time versus over space

b. "Final" (i.e., best-performing) model predictions for all featurized grid cells (the best-performing model for overall performance and the best-performing model for over time)
   i. Scatter plots for training, validation, demeaned, and test sets with R^2 metrics
   ii. Tabular results for training, validation, and demeaned R^2 (see Appendix B, Tables 1A-1B)
   iii. Dataframe containing predictions
   iv. Maps of predictions at 1 kilometer² grid cell level and district level (produced in the Modeling repository notebooks)

3. Documentation

a. Well-documented codebase with a README for the overall GitHub organization, a README for the Featurization repository, and a README for the Modeling repository
    i. Includes details of key decisions made in the featurization and modeling processes, such as how to handle clouds, choose various parameters from a set of options in the notebook, interpret graphs and maps, etc.
b. Executive Summary summarizing the project in this Technical Documentation document
c. Link to the MOSAIKS API in the GitHub organization README's

6.2 Academic Deliverables

Six academic deliverables will be produced as part of the Bren capstone requirements:

1. Design and Implementation Plan
2. Data and metadata
3. Faculty Review presentation
4. Technical Documentation Plan
5. Project repository
6. Capstone Project Final Presentation

7.0 Summary of Testing

A variety of tests have been employed in our featurization step and modeling process to ensure our pipeline has been optimized for maize yield predictions in Zambia. The purpose of testing is to ensure the accuracy of our results, measure uncertainty in our model, and allow for open-source, repeatable results. A summary of our testing procedures is outlined below.

7.1 Featurization Testing

Manual exploratory testing and debugging statements were applied throughout the random convolution featurization process. Manual exploratory tests used include plotting intermediate visualizations and verifying code chunk outputs. Visualizations help provide code checks to ensure the code is doing what we intend. Throughout our process we plot visualizations and code chunk outputs as often as possible and in as many different ways as possible to be sure the raw, intermediate, and output parameters are what we expect. For example, we plot activation maps for individual random convolutional feature sets. An activation map helps visualize where certain attributes are found within a defined space and time. In our activation maps a high activation means a certain attribute was found. We then plot these activation maps next to visual band (RGB) satellite imagery to inspect for outliers and to ensure that images are being featurized correctly. Additional visualizations created to test accuracy of outputs include plots of the uniform

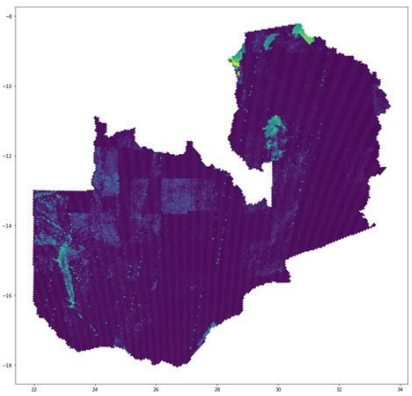


Figure 2. Activation Map

subset of the 1 km^2 grid of sample points, heatmaps, and 1 km^2 grid of Landsat 8 and Sentinel 2 imagery.

Debugging statements are internal self-checks in code to aid testing and troubleshooting errors. Within our code, we built in error messages to prevent silent errors from becoming bigger problems. The 'assert' statement is also used throughout our code for a similar purpose. An example of an assert statement used is that we require the number of points 'If statements' are implemented for a similar purpose, to only use images with a certain number of pixels to ensure that the images we use are full, complete images. In all of our coding steps, we implement peer code review. Each time an individual merges their branch into the main branch, it undergoes a code review before it is accepted and merged.

7.2 Modeling

7.2.1. Optimization

Our model has been tested on multiple subsets of satellites, cell size, bands, and months within the years that correspond to the growing season, to test which months contribute to the most accurate predictions. The model is optimized to predict over time (year to year) and through space (district to district). The dataset is split into a training dataset (80%) and a testing dataset (20%) and then run through the model.

A table of our model optimization metrics is provided in Appendix B: Optimization Spreadsheets for Zambia.

Cell Size and Number of Features

Model optimization includes the creation of features for Zambia at different spatial scales and number of features. Optimizing the ideal spatial scale translates to changing the density of the uniform subset of the spatial grid that which we sample, and training the model iteratively with this spectrum. With a lower-density uniform sample of the spatial grid, we retrieve fewer images of Zambia, and therefore lower the spatial resolution. However, lowering this resolution does not linearly translate to less accurate model performance. The same concept applies to the number of features used to train the model; choosing fewer features decreases the amount of image information we feed into the model, but this does not linearly translate to less accurate model performance. As such, we trained our model at different spatial scales and feature numbers to determine the threshold at which the accuracy is not significantly decreased and the image featurization can still be executed in a reasonable amount of time. A reasonable amount is defined as less than 1.5 hours for featurizing one year of images.

Bands

In addition to the cell size and the number of features, the bands used for both the Sentinel 2 and Landsat 8 satellite image featurization process is a factor in model optimization. Once the cell size and number of features is determined, the model is stacked with different groups of bands. These groups include:

1. Sentinel 2
    a. Red, green, and blue (2-3-4)
    b. Red, green, blue, and near infra-red (2-3-4-8)
2. Landsat 8
    a. Coastal aerosol, blue, green, red, near infra-red, shortwave infra-red 1, shortwave infra-red 2 (1-2-3-4-5-6-7)

The relative model performance for each band combination option determined the bands used for the final features that produce our capstone deliverables.

Crop Mask and Weights

The distribution of cropland within Zambia is not uniform. Although maize is grown to some degree in most or all districts, areas with a high density of crops are located in the central, eastern, and southern regions of Zambia. To target these areas, we use a 30 meter resolution global cropland dataset to remove all 1 km² grid cells that do not contain cropland. We take the remaining 1 km² grid cells and calculate the crop percentage within these cells to use as a weighted average. This can be used to mask our features to only cropland data for the evenly sampled points, and it can also be used to do a weighted average when summarizing the points to the district level. It also used to featurize only the top 10% of cropland for each district.

Time

Within years, the model for maize production is optimized based on the growing season in Zambia, which spans from November to July. The model performs differently when the input features are for the entire year, May through October, or April through September. Subsetting the data and extending the timeline to a few months beyond the end of harvest produced the most optimal model.

We check the model's ability to predict over time by demeaning our predictions and the observed crop yields by location. We then calculate an R^2 score demeaned by location for the model.

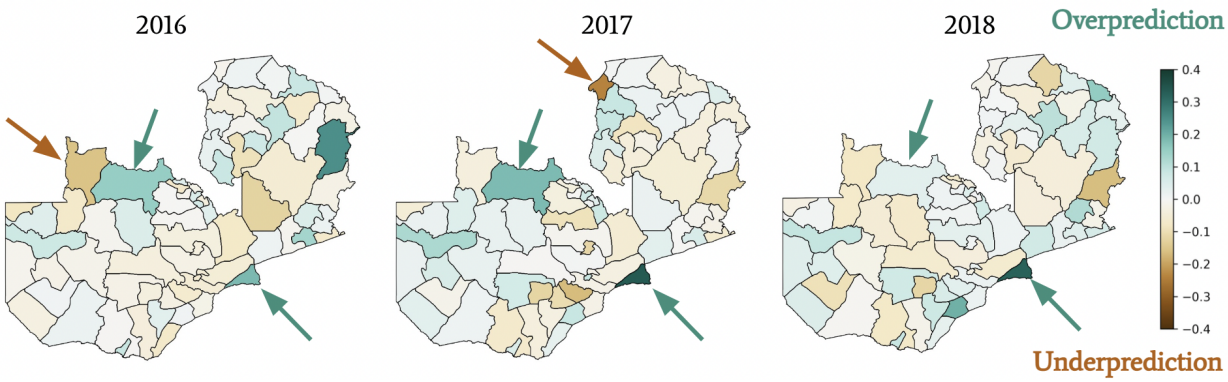


Figure 3. Demeaned maize yields in Zambia for the years 2016-2018 in units of log transformed metric tonnes of estimated maize production per hectare of planted cropland. Arrows highlight particularly underestimated and overestimated districts.

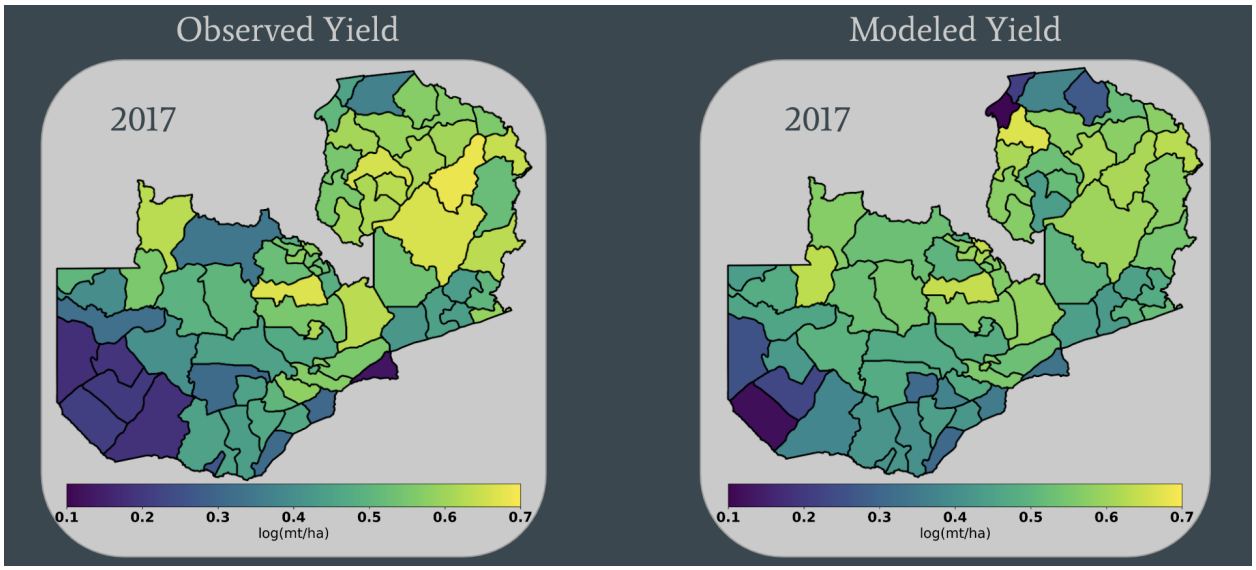


Figure 4. Comparison of observed maize yields in Zambia in 2017 and predicted maize yields in Zambia in 2017 with units of log transformed metric tonnes of estimated maize production per hectare of planted cropland. The similarity of the district colors in both maps represents that our optimized model produced results that are within the range of expected values that do not dramatically deviate from observed values.

7.2.2. Uncertainty

Our featurization procedure and model inputs all contain metrics of uncertainty. Uncertainty comes from modeled data and estimations in lieu of ground-truthed data. Each decision made in our pipeline introduces a degree of uncertainty. Here we describe the primary sources of uncertainty in our pipeline.

Maize Data

Zambia maize yields are forecast yields in units of metric tonnes of estimated maize production per hectare of planted cropland. These units are derived from the expected production reported by the farmers each year in units of metric tonnes, divided by the amount of hectares of farmland in that district. This data was collected by the Central Statistics Office of Zambia (2022). These data are forecasted from pre-harvest survey data collected in May preceding the harvest season (July-August). The forecast model is conducted in Stata, a general purpose statistical software, and adjusted with post-harvest season survey data. There is an unknown degree of uncertainty in this forecast data as the model process and parameters are unknown.

Crop Masks

The spatiotemporally consistent cropland mask for Zambia comes from a global dataset (Potapov et al. 2021). Of the global cropland estimate, Africa accounts for 16% of global cropland. Uncertainty metrics were calculated using a stratified random sampling approach with five strata chosen and reported at the 95% confidence interval.

Featurization

Our featurization process is an estimation of satellite imagery into a new feature space. A primary determinant of valid featurization is minimizing cloud cover percentage in satellite imagery. Minimizing cloud cover is critical to accurate featurization because the presence of clouds introduces bias into our model. While near and shortwave infra-red spectral bands can penetrate cloud to and extent, other bands cannot. In our process we whole-sale eliminate Landsat images that contain greater than 10% cloud cover.

Model

Estimation of model results is completed through multiple reporting metrics. We use a 5-fold cross-validated ridge regression on a training dataset to predict on a test dataset. We report the models' train, test, and validation R^2 and Pearson Correlation Coefficient scores on our datasets. We also report the R^2 score on our demeaned observations and predictions by location which provides a metric on how our model performs over time. Once we run our model, we plot residual maps at the district level. These plots display how accurate our model results are compared to actual ground-truthed data. We also visualize model predictions for all years with available ground-truthed crop yields as barplots (Figure 5.) and histograms.

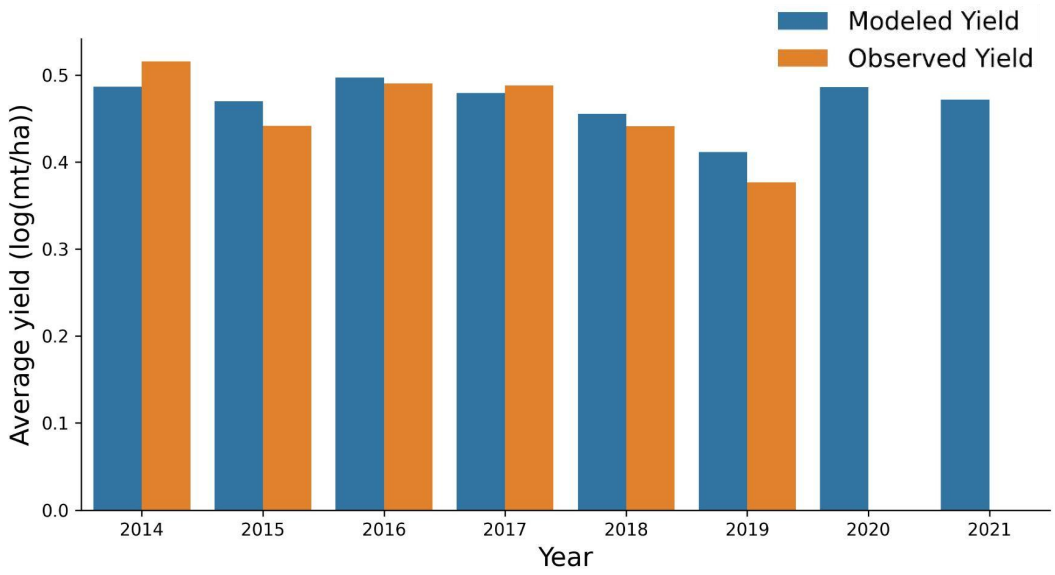


Figure 5. Predicted crop yields and observed crop yields in Zambia, 2014-2021 in units of log transformed metric tonnes of estimated maize production per hectare of planted cropland. These annual yields are the average of all districts for which we have crop data. This figure demonstrates that our model predictions for 2020 and 2021 are within a reasonable range of values compared to past years.

8.0 User Documentation

Expected User: Client and co-authors. Potential future applications for non-governmental organizations working in food security as well as agriculture policy-makers in Zambia. Potentially future MEDS students.

The CropMosaiks GitHub organization will be the primary documentation portal and distribution hub for our code and metadata documentation. The GitHub organization has a landing page README that directs users to the two primary repositories, Featurization and Modeling. Each repository has a detailed README describing the purpose, notebooks, and use instructions for getting started as well as potential ways to expand and contribute. Within each notebook, there will be detailed explanations of code embedded in markdown chunks and in comments as well as how users can adapt the code for other use cases.

In the Featurization repository, the README file explains how users can either extract features by utilizing the featurization notebooks, or by directing them to the MOSAIKS API. Feature data created by the cropMOSAIKS team will only be available to our clients through the MEDS server Taylor under the directory "/capstone/cropmosaiks/data/features/<satellite>." Further information on feature creation can be found in the two primary notebooks that take slightly different approaches to featurization.

In the Modeling repository, the README file explains the goal of the modeling process, datasets used, how to get started, and what the notebooks do. It is important to note that this notebook is best designed for utilizing monthly features for yearly data by pivoting the features wide by month. This means that 1,000 monthly features become 12,000 yearly features with the month information appended in each column with the feature number. Because the Zambia growing season spans the Gregorian new year (November planting, July harvest), we artificially change the year category such that October, November, and December are included with the following year.

The maize yield data that has been provided by our clients will not be made publicly available but the most recent versions will be available on the MEDS server Taylor under the directory "/capstone/cropmosaiks/data/crops." Similarly, the administrative boundary data that best matches the crop yield data (there have been recent district divisions creating more districts than is seen in the historic crop yield data), also provided by our clients, will not be made publicly available, but instead will be hosted on the MEDS server Taylor under the directory "/capstone/cropmosaiks/data/boundaries."

Additionally, there is a notebook for extracting land cover and land classification for 9 classes at 10 meter resolution. This notebook has a known problem and is not ready for use. Specifically, the workflow used creates data gaps at the UTM boundary delineations. If this problem is solved, it could potentially be used to provide weights to features for labels such as trees, flooded vegetation, crops, built area, bare ground, snow/ice, and rangeland.

9.0 Archive Access

- Generated features will be stored on the Taylor server, which the client Tamma Carleton has access to.
- Crop data is private and will not be shared per the request of the UC Santa Barbara Baylis Lab.
- Project code and user documentation will be located on the Github Project Organization. User documentation is provided for all repositories (Modeling, Featurization, and Crops).

10.0 Future Work and Research Ideas

- Test if the maize yield predictions for Zambia prior to 2020 can be used to detect crop yield fluctuations due to known climatic anomalies such as drought. If the crop predictions for these years show significant correlation with precipitation and temperature, this model can be improved upon and used as a tool for governments, community leaders, farmers, and food security initiatives to predict future crop yields for Zambia. This tool was demonstrated by producing predictions for all years used to train the model as well as 2020 and 2021. The ultimate goal is to provide more foresight regarding crop yields prior to harvest, so farmers and leaders can adjust crop imports, exports, and costs.

- A report presenting a correlational analysis between estimated crop yields and high-resolution, publicly available climate indicators (i.e., temperature and precipitation). This differs from the preceding suggestion because this correlational analysis relates to general temperature and precipitation data, not anomalies. Ideally, this report should be accessible in all dominant languages of Zambia in order to include local farmers and leaders who may not be fluent in English.
- Stack additional bands to Sentinel 2 in addition to visible spectrum (2,3,4). Examples include short wave infrared (12, 8, and 4), and red edge
- Utilizing notebook for 0.01 degree grid cells (an equal angle grid)
- Increasing the cloud cover limit from 10% to 15% or more to retain more images and therefore more points, increasing *n* of the training and test sets. This will likely improve model performance.
- Filtering cloud cover at the level of the resolution you are featurizing (0.01 degree for equal angle or 1 kilometer^2 grid cell for equal area) rather than at the image level.
- Maize is an annual crop and as such, time is not a term in our model. We account for time by grouping features by month and year, which allows us to predict over time. If we included time (i.e. year) as a term in our model, that would mean a single year's crop yields impact another year's crop yields. For our goal of predicting annual crop yields, using year as a term in the model is not helpful. However, this would be useful if you are trying to predict yield for perennial crops such as almonds or avocados.
- Include district as a term in the cross-validated ridge regression model.

11.0 References

Burke, Marshall, Anne Driscoll, David B. Lobell, and Stefano Ermon. “Using Satellite Imagery to Understand and Promote Sustainable Development.” *Science* 371, no. 6535 (March 19, 2021): eabe8628. https://doi.org/10.1126/science.abe8628.

Central Statistics Office. “Zambia Statistics Agency.” Portal. Zambia Data Portal, March 1, 2022. https://zambia.opendataforafrica.org/.

Gatti, Nicolas, Kathy Baylis, and Benjamin Crost. “The Effects of Market Access on the Agricultural Input Market Structure in Zambia.” *2020 Annual Meeting, July 26-28, Kansas City, Missouri*. 2020 Annual Meeting, July 26-28, Kansas City, Missouri. Agricultural and Applied Economics Association, July 2020. https://ideas.repec.org/p/ags/aaea20/304598.html.

Hultgren, Andrew, Tamma Carleton, Michael Delgado, Diana Gergel, Michael Greenstone, Trevor Houser, and Solomon Hsiang. “The Impacts of Climate Change on Global Grain Production Accounting for Adaptation.,” n.d.

Jean, Neal, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. “Combining Satellite Imagery and Machine Learning to Predict Poverty.” *Science* 353, no. 6301 (August 19, 2016): 790–94. https://doi.org/10.1126/science.aaf7894.

Kurukulasuriya, Pradeep, and Robert Mendelsohn. “Endogenous Irrigation : The Impact of Climate Change on Farmers in Africa.” Washington, DC: World Bank, July 2007. https://doi.org/10.1596/1813-9450-4278.

Niang, I., O.C. Ruppel, M.A. Abdrabo, A. Essel, C. Lennard, J. Padgham, and P. Urquhart, 2014: Africa. In: Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part B: Regional Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Barros, V.R., C.B. Field, D.J. Dokken, M.D. Mastrandrea, K.J. Mach, T.E. Bilir, M. Chatterjee, K.L. Ebi, Y.O. Estrada, R.C. Genova, B. Girma, E.S. Kissel, A.N. Levy, S. MacCracken, P.R. Mastrandrea, and L.L.White (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 1199-1265.

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library.” *ArXiv:1912.01703 [Cs, Stat]*, December 3, 2019. http://arxiv.org/abs/1912.01703.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. “Scikit-Learn: Machine Learning in Python.” *The Journal of Machine Learning Research* 12, no. null (November 1, 2011): 2825–30.

Porter, J. R., Xie L, Andrew J. Challinor, K. Cochrane, Howden Sm, M. M. Iqbal, David B. Lobell, and Maria Travasso. “Food Security and Food Production Systems.” Report. IPCC, 2014. https://cgspace.cgiar.org/handle/10568/68162.

Potapov, Peter, Svetlana Turubanova, Matthew C. Hansen, Alexandra Tyukavina, Viviana Zalles, Ahmad Khan, Xiao-Peng Song, Amy Pickens, Quan Shen, and Jocelyn Cortez. “Global Maps of Cropland Extent and Change Show Accelerated Cropland Expansion in the Twenty-First Century.” Nature Food 3, no. 1 (January 2022): 19–28. https://doi.org/10.1038/s43016-021-00429-z.

Quarmby, N. A., M. Milnes, T. L. Hindle, and N. Silleos. “The Use of Multi-Temporal NDVI Measurements from AVHRR Data for Crop Yield Estimation and Prediction.” *International Journal of Remote Sensing* 14, no. 2 (January 1993): 199–210. https://doi.org/10.1080/01431169308904332.

Rahimi, Ali, and Benjamin Recht. “Uniform Approximation of Functions with Random Bases.” In *2008 46th Annual Allerton Conference on Communication, Control, and Computing*, 555–61, 2008. https://doi.org/10.1109/ALLERTON.2008.4797607.

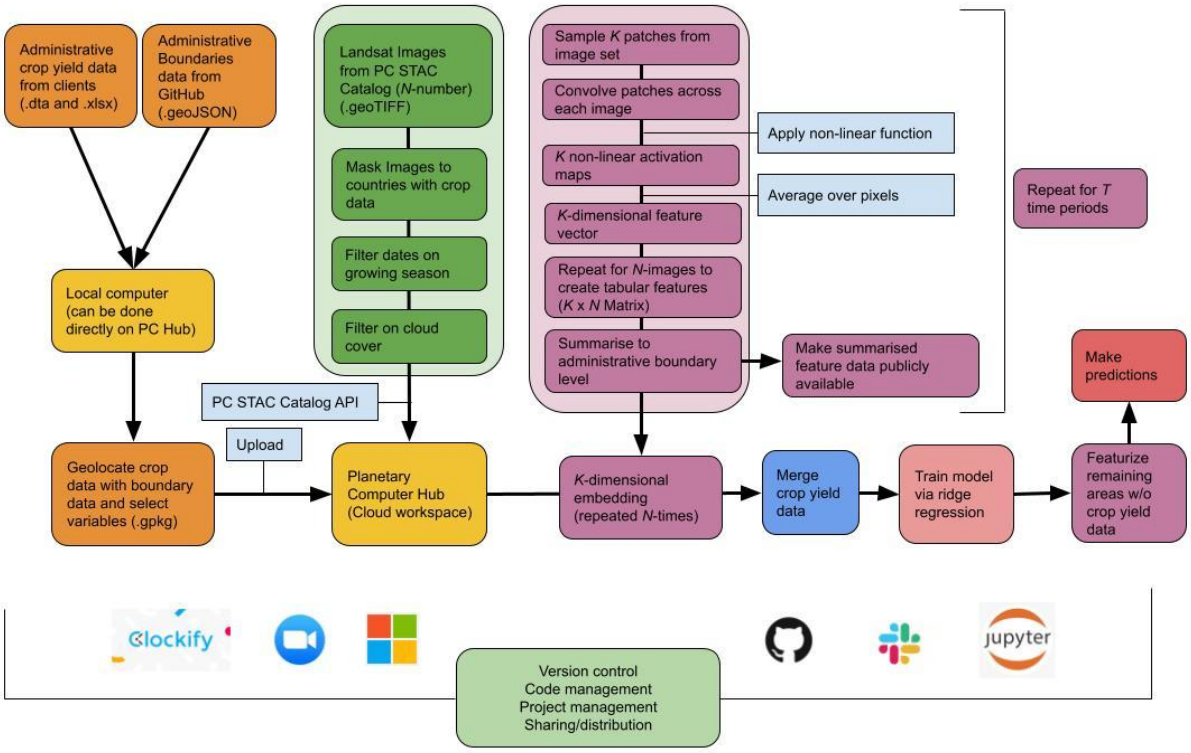
Rahimi A, Recht B (2008) Random features for large-scale kernel machines. In: Platt JC, Koller D, Singer Y, Roweis ST (eds) Advances in neural information processing systems, vol 20. Curran Associates Inc, Red Hook, pp 1177–1184

Rolf, Esther, Jonathan Proctor, Tamma Carleton, Ian Bolliger, Vaishaal Shankar, Miyabi Ishihara, Benjamin Recht, and Solomon Hsiang. “A Generalizable and Accessible Approach to Machine Learning with Global Satellite Imagery.” *Nature Communications* 12, no. 1 (December 2021): 4392. https://doi.org/10.1038/s41467-021-24638-z.

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python.” *Nature Methods* 17, no. 3 (March 2020): 261–72. https://doi.org/10.1038/s41592-019-0686

University of California, Santa Barbara, Department of Geography, Baylis Lab. Personal communication (April 11, 2022)

Appendix A: Workflow & Project Management Tools



Appendix B: Optimization Spreadsheets for Zambia

Table 1A. Initial Zambia Model Optimization Spreadsheet (Outdated in favor of Table 1B)

satellite	bands	country	num_points	features	year	month	crop_mask	weighted_avg	impute	validation_R2	train_R2	train_pearson_R	train_n	test_n
sentinel	2-3-4	ZMB	4000	1000	2016	all	no	no	simple	0.25	0.66	0.83	57	15
sentinel	2-3-4	ZMB	4000	1000	2017	all	no	no	simple	0.16	0.44	0.69	57	15
sentinel	2-3-4	ZMB	4000	1000	2018	all	no	no	simple	0.44	0.79	0.89	57	15
sentinel	2-3-4	ZMB	4000	1000	2016-2018	all	no	no	simple	0.19	0.43	0.66	172	44
sentinel	2-3-4-8	ZMB	4000	1000	2016	all	no	no	simple	0.15	0.54	0.78	57	15
sentinel	2-3-4-8	ZMB	4000	1000	2017	all	no	no	simple	0.05	0.41	0.7	57	15
sentinel	2-3-4-8	ZMB	4000	1000	2018	all	no	no	simple	0.13	0.71	0.86	57	15
sentinel	2-3-4-8	ZMB	4000	1000	2016-2018	all	no	no	simple	0.31	0.61	0.8	172	44
landsat	1-2-3-4-5-6-7	ZMB	4000	1000	2014	all	no	no	simple	0.01	0.17	0.53	57	15
landsat	1-2-3-4-5-6-7	ZMB	4000	1000	2015	all	no	no	simple	0.03	0.34	0.73	57	15
landsat	1-2-3-4-5-6-7	ZMB	4000	1000	2016	all	no	no	simple	-0.08	0.26	0.62	57	15
landsat	1-2-3-4-5-6-7	ZMB	4000	1000	2017	all	no	no	simple	0.04	0.19	0.54	57	15
landsat	1-2-3-4-5-6-7	ZMB	4000	1000	2018	all	no	no	simple	0.12	0.57	0.8	57	15
landsat	1-2-3-4-5-6-7	ZMB	4000	1000	2014-2018	all	no	no	simple	0.39	0.79	0.89	288	72
sentinel	2-3-4	ZMB	15000	1000	2016	all	no	no	simple	0.32	0.54	0.75	57	15
sentinel	2-3-4	ZMB	15000	1000	2017	all	no	no	simple	0.16	0.38	0.65	57	15
sentinel	2-3-4	ZMB	15000	1000	2018	all	no	no	simple	0.46	0.69	0.84	57	15
sentinel	2-3-4	ZMB	15000	1000	2016-2018	all	no	no	simple	0.25	0.65	0.81	172	44
sentinel	2-3-4	ZMB	15000	1000	2016-2018	all	yes	no	simple	X	X	X	172	44
sentinel	2-3-4	ZMB	15000	1000	2016-2018	4-9	no	no	simple	0.38	0.68	0.82	172	44
sentinel	2-3-4	ZMB	15000	1000	2016-2018	5-9	no	no	simple	0.34	0.64	0.8	172	44
sentinel	2-3-4	ZMB	15000	1000	2016-2018	4-8	no	no	simple	0.34	0.6	0.77	172	44
sentinel	2-3-4	ZMB	15000	1000	2016-2018	5-8	no	no	simple	0.31	0.54	0.74	172	44
sentinel	2-3-4	ZMB	15000	1000	2016-2018	4-10	no	no	simple	0.38	0.7	0.84	172	44
sentinel	2-3-4-8	ZMB	15000	1000	2016	all	no	no	simple	0.12	0.48	0.72	57	15
sentinel	2-3-4-8	ZMB	15000	1000	2017	all	no	no	simple	0.08	0.4	0.68	57	15

sentinel	2-3-4-8	ZMB	15000	1000	2018	all	no	no	simple	0.07	0.46	0.72	57	15
sentinel	2-3-4-8	ZMB	15000	1000	2016-2018	all	no	no	simple	0.37	0.72	0.85	172	44
sentinel	2-3-4-8	ZMB	15000	1000	2016-2018	all	yes	no	simple	0.56	0.81	X	172	44
sentinel	2-3-4-9	ZMB	15000	1000	2016-2019	4-9	yes	no	simple	0.5	0.76	0.87	172	44
sentinel	2-3-4-8	ZMB	15000	1000	2016-2018	4-9	no	no	simple	0.39	0.65	0.81	172	44
sentinel	2-3-4-8	ZMB	15000	1000	2016-2018	5-9	no	no	simple	0.5	0.74	0.86	172	44
sentinel	2-3-4-8	ZMB	15000	1000	2016-2018	4-8	no	no	simple	0.51	0.73	0.85	172	44
sentinel	2-3-4-8	ZMB	15000	1000	2016-2018	5-8	no	no	simple	X	X	X	172	44
sentinel	2-3-4-8	ZMB	15000	1000	2016-2018	4-10	no	no	simple	X	X	X	172	44
landsat	1-2-3-4-5-6-7	ZMB	15000	1000	2014	all	no	no	simple	0.02	0.18	0.53	57	15
landsat	1-2-3-4-5-6-7	ZMB	15000	1000	2015	all	no	no	simple	0.04	0.34	0.72	57	15
landsat	1-2-3-4-5-6-7	ZMB	15000	1000	2016	all	no	no	simple	0	0.61	0.82	57	15
landsat	1-2-3-4-5-6-7	ZMB	15000	1000	2017	all	no	no	simple	0.05	0.2	0.55	57	15
landsat	1-2-3-4-5-6-7	ZMB	15000	1000	2018	all	no	no	simple	0.11	0.59	0.8	57	15
landsat	1-2-3-4-5-6-7	ZMB	15000	1000	2014-2018	all	no	no	simple	0.46	0.81	0.9	288	72
landsat	1-2-3-4-5-6-7	ZMB	15000	1000	2014-2018	all	yes	no	manual	0.41	0.61	0.8	288	72
landsat	1-2-3-4-5-6-7	ZMB	15000	1000	2013-2018	4-9	yes	no	manual	0.57	0.82	0.91	288	72
landsat	1-2-3-4-5-6-7	ZMB	15000	1000	2013-2018	4-9	no	no	manual	0.57	0.77	0.88	288	72
landsat	1-2-3-4-5-6-7	ZMB	15000	1000	2014-2018	4-9	yes	no	manual	0.5	0.82	0.91	288	72
landsat	1-2-3-4-5-6-7	ZMB	15000	1000	2014-2018	4-9	no	no	simple	0.56	0.73	0.86	288	72
landsat	1-2-3-4-5-6-7	ZMB	15000	1000	2014-2018	5-9	no	no	simple	0.51	0.76	0.88	288	72
landsat	1-2-3-4-5-6-7	ZMB	15000	1000	2014-2018	4-8	no	no	simple	0.55	0.72	0.85	288	72
landsat	1-2-3-4-5-6-7	ZMB	15000	1000	2014-2018	5-8	no	no	simple	0.47	0.73	0.85	288	72
landsat	1-2-3-4-5-6-7	ZMB	15000	1000	2014-2018	4-10	no	no	simple	0.54	0.75	0.87	288	72
sentinel	2-3-4	ZMB	24000	1000	2016	all	no	no	simple	0.35	0.55	0.77	57	15
sentinel	2-3-4	ZMB	24000	1000	2017	all	no	no	simple	0.14	0.37	0.64	57	15
sentinel	2-3-4	ZMB	24000	1000	2018	all	no	no	simple	0.44	0.68	0.84	57	15
sentinel	2-3-4	ZMB	24000	1000	2016-2018	all	no	no	simple	0.27	0.78	0.89	172	44
sentinel	2-3-4	ZMB	42000	1000	2016	all	no	no	simple	0.18	0.63	0.81	57	15
sentinel	2-3-4	ZMB	42000	1000	2017	all	no	no	simple	0.12	0.41	0.67	57	15
sentinel	2-3-4	ZMB	42000	1000	2018	all	no	no	simple	0.46	0.8	0.9	57	15
sentinel	2-3-4	ZMB	42000	1000	2016-2018	all	no	no	simple	0.24	0.44	0.68	172	44

Table 1B. Final Zambia Model Optimization Spreadsheet

satellite	s2_bands	l8_bands	num_points	year	month	crop_mask	weighted_avg	rows_initial	rows_end	train_n	test_n	val_R2	train_R2	train_r	demean_R2	demean_r
sentinel	2-3-4	NA	15000	2016-2018	all	no	no	44865	39573	120	30	0.55	0.71	0.85	0.28	0.59
sentinel	2-3-4	NA	15000	2016-2018	all	yes	no	13914	11478	110	28	0.6	0.83	0.91	-0.08	0.53
sentinel	2-3-4	NA	15000	2016-2018	all	yes	yes	13914	11478	110	28	0.36	0.8	0.9	0.3	0.69
sentinel	2-3-4	NA	15000	2016-2018	4-9	no	no	44865	44865	172	44	0.36	0.67	0.82	0.31	0.63
sentinel	2-3-4	NA	15000	2016-2018	4-9	yes	no	13914	13914	172	44	0.43	0.69	0.83	0.08	0.54
sentinel	2-3-4	NA	15000	2016-2018	4-9	yes	yes	13914	13914	172	44	0.21	0.52	0.72	0.09	0.48
sentinel	2-3-4-8	NA	15000	2016-2018	all	no	no	44865	38610	115	29	0.52	0.75	0.87	0.19	0.52
sentinel	2-3-4-8	NA	15000	2016-2018	all	yes	no	13866	10908	103	26	0.45	0.85	0.92	0.26	0.61
sentinel	2-3-4-8	NA	15000	2016-2019	all	yes	yes	13866	10908	103	26	0.43	0.83	0.91	0.15	0.52
sentinel	2-3-4-8	NA	15000	2016-2018	4-9	no	no	44865	44865	172	44	0.39	0.65	0.81	0.19	0.49
sentinel	2-3-4-8	NA	15000	2016-2018	4-9	yes	no	13866	13866	172	44	0.5	0.76	0.87	0.08	0.49
sentinel	2-3-4-8	NA	15000	2016-2018	4-9	yes	yes	13866	13866	172	44	0.43	0.74	0.86	0.13	0.5
landsat	NA	1-2-3-4-5-6-7	15000	2014-2018	all	no	no	75114	48648	152	38	0.54	0.93	0.96	0.4	0.65
landsat	NA	1-2-3-4-5-6-7	15000	2014-2018	all	yes	no	23220	13882	148	37	0.39	0.92	0.96	0.37	0.64
landsat	NA	1-2-3-4-5-6-7	15000	2014-2018	all	yes	yes	23220	13882	148	37	0.47	0.73	0.86	-0.04	0.37
landsat	NA	1-2-3-4-5-6-7	15000	2014-2018	4-9	no	no	74971	74609	272	68	0.61	0.85	0.92	0.27	0.61
landsat	NA	1-2-3-4-5-6-7	15000	2014-2018	4-9	yes	no	23190	22881	272	68	0.48	0.81	0.9	0.1	0.48
landsat	NA	1-2-3-4-5-6-7	15000	2014-2018	4-9	yes	yes	23190	22881	272	68	0.42	0.72	0.85	-0.06	0.42
landsat	NA	1-2-3-4-5-6-7	15000	2013-2018	4-9	no	no	89926	89926	345	87	0.56	0.77	0.88	0.13	0.49
landsat	NA	1-2-3-4-5-6-7	15000	2013-2018	4-9	yes	no	27812	27812	345	87	0.56	0.82	0.91	0.09	0.51
landsat	NA	1-2-3-4-5-6-7	15000	2013-2018	4-9	yes	yes	27812	27812	345	87	0.46	0.82	0.91	-0.04	0.5
landsat	NA	1-2-3-4-5-6-7	15000	2016-2018	all	no	no	45088	24457	69	18	0.54	0.78	0.89	0.06	0.33
landsat	NA	1-2-3-4-5-6-7	15000	2016-2018	all	yes	no	13934	7088	64	17	0.42	0.74	0.87	-0.3	0.2
landsat	NA	1-2-3-4-5-6-7	15000	2016-2018	all	yes	yes	13934	7088	64	17	0.33	0.93	0.97	0.17	0.56
landsat	NA	1-2-3-4-5-6-7	15000	2016-2018	4-9	no	no	44960	44478	158	40	0.44	0.75	0.87	0.14	0.57
landsat	NA	1-2-3-4-5-6-7	15000	2016-2018	4-9	yes	no	13906	13491	158	40	0.1	0.56	0.76	-0.15	0.28
landsat	NA	1-2-3-4-5-6-7	15000	2016-2018	4-9	yes	yes	13906	13491	158	40	0.17	0.4	0.64	-0.15	0.11
combined	2-3-4	1-2-3-4-5-6-7	15000	2016-2018	all	no	no	45088	22937	60	15	0.72	0.93	0.97	0.15	0.47
combined	2-3-4	1-2-3-4-5-6-7	15000	2016-2018	all	yes	no	13934	6546	52	14	0.6	1	1	0.23	0.52
combined	2-3-4	1-2-3-4-5-6-7	15000	2016-2018	all	yes	yes	13934	6546	52	14	0.6	0.93	0.98	0.11	0.39
combined	2-3-4	1-2-3-4-5-6-7	15000	2016-2018	4-9	no	no	44960	44478	158	40	0.58	0.93	0.97	0.11	0.36
combined	2-3-4	1-2-3-4-5-6-7	15000	2016-2018	4-9	yes	no	13906	13491	158	40	0.48	0.94	0.97	0.03	0.18

combined	2-3-4	1-2-3-4-5-6-7	15000	2016-2018	4-9	yes	yes	13906	13491	158	40	0.38	0.92	0.97	0.04	0.2
combined	2-3-4-8	1-2-3-4-5-6-7	15000	2016-2018	all	no	no	45088	21960	55	14	0.7	1	1	0.1	0.33
combined	2-3-4-8	1-2-3-4-5-6-7	15000	2016-2018	all	yes	no	13934	5999	45	12	0.79	1	1	0.2	0.47
combined	2-3-4-8	1-2-3-4-5-6-7	15000	2016-2018	all	yes	yes	13934	5999	45	12	0.48	1	1	0.21	0.47
combined	2-3-4-8	1-2-3-4-5-6-7	15000	2016-2018	4-9	no	no	44960	44478	158	40	0.57	0.92	0.96	0.13	0.4
combined	2-3-4-8	1-2-3-4-5-6-7	15000	2016-2018	4-9	yes	no	13906	13491	158	40	0.47	0.94	0.98	0.05	0.23
combined	2-3-4-8	1-2-3-4-5-6-7	15000	2016-2018	4-9	yes	yes	13906	13491	158	40	0.42	0.92	0.97	0.06	0.24