Scalable Country-Level Crop Yield Modeling for Food Security and Risk Mitigation

Tackling Climate Change with Machine Learning

NeurIPS, Dec 2025

Andrew Hobbs, Jesse Anttila-Hughes, Benson Adomako, Addi Joof

Summary

Smallholder farmers in Sub-Saharan Africa face substantial risks from weather variability, including drought, heat stress, and shifting rainfall patterns. Traditional measurement of agricultural productivity relies on costly field surveys that are often infrequent and geographically limited. Scalable and timely yield prediction methods are urgently needed to support climate resilience, food security planning, and adaptation policy. This project develops a new machine learning framework that combines detailed household crop yield data from the World Bank's LSMS-ISA surveys with satellite-based indicators of vegetation health, precipition, soil moisture, and temperature stress. We train spatially validated random forest models to predict maize yields across multiple countries and agro-ecological zones.

We show that remotely sensed features can explain a large share of crop yield variation, and a unified global model generalizes well to new regions . This approach highlights the potential for satellite-based yield monitoring systems to enhance early warning capabilities and inform agricultural policy interventions at scale, particularly in data-scarce environments.

Data

Crop Yields: World Bank LSMS-ISA Data from Nigeria, Uganda, Malawi, Tanzania, Mali, Ethiopia,
Croplands: Global Food Security Analysis Support Data (GFSAD)
Vegetation Indices: NDVI (LANDSAT) and GCI (MODIS Terra) monthly maximum.
Temperature: ERA5 Dataset. Daily min/max converted to Growing Degree Days (GDD) and Killing Degree Days (KDD) and aggregated to monthly scale

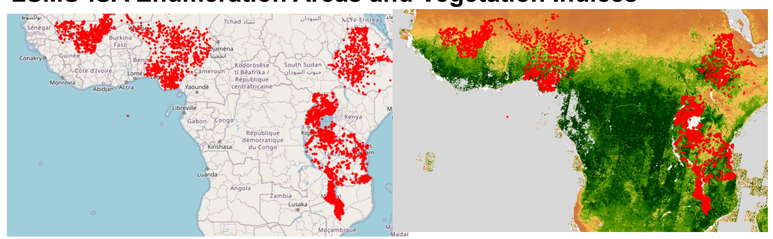
We obtain maize yield estimates for between 579 and 1260 households per year across the six countries. We restrict our sample to maize yields (kg/acre) because it is one of the primary staple crops in West and East Africa.

Method

The World Bank LSMS-ISA dataset contains about 30,000 survey reported household-level crop yields in 7,500 enumeration areas. For each enumeration area, we have a set of coordinates (latitude/ longitude) that roughly indicates where the households within the sample are located. Because our goal is to estimate regional average yields, this lack of precision fits our goal - we want a model that predicts average yields in the region, not farm-specific yields.

For each point, we extract remotely sensed vegetation, precipitation, soil moisture and temperature indicators for 12 months prior to the location-specific harvest month within a 10 km buffer in pixels that NASA's Global Food Security Analysis Data indicates are cropland. We then generate a temporal mean, sum or max for each variable by month (means for temperature, precipitation, maxes for vegetation indices, and sums for growing and killing degree days). We then log-transform yields.

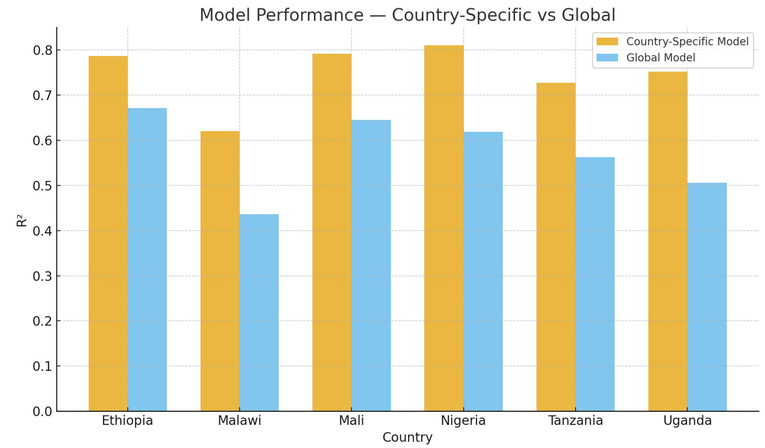
LSMS-iSA Enumeration Areas and Vegetation Indices



We train an array of machine learning models to predict log maize yields: a country-specific model trained and validated within each country and a unified global model trained on all six LSMS-ISA countries and tested separately on each to assess cross-country transferability. Because nearby farms are likely to share similar environmental and management conditions, we evaluate model performance using spatial GroupKFold cross-validation, where entire spatial clusters of observations (country–wave–GPS groups) are held out. After training, each model is tested on every country's held-out test set (30% of observations) to test how well yield prediction models translate between countries.

Results

Our global machine learning model trained on geospatial remote sensing data can explain over 50% of the spatial variation in average yields across several Sub-Saharan African countries. Our early models trained on country-specific data prediction 70-80% of the variation in yields. A general model trained using all the data in our sample does reasonably well in most countries, generally predicting over 50% of the variation in yields.



Next Steps

Thus far, we have trained only random forest models using aggregated data from throughout a 10 kilometer circle around enumeration area centroids. We next plan on training a range of alternative regression models, including gradient boosted trees, neural networks, and elastic nets. We also have in our dataset information on the administrative area an observation is in, which could allow us to focus our data collection on a more precise area in some circumstances by ignoring parts of the 10km circle that are outside the administrative area boundaries. Finally, we plan on using Google's Alpha Earth Model to attempt to specifically identify maize fields, allowing us to restrict our analysis to more relevant pixels and hopefully leading to even greater accuracy.