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

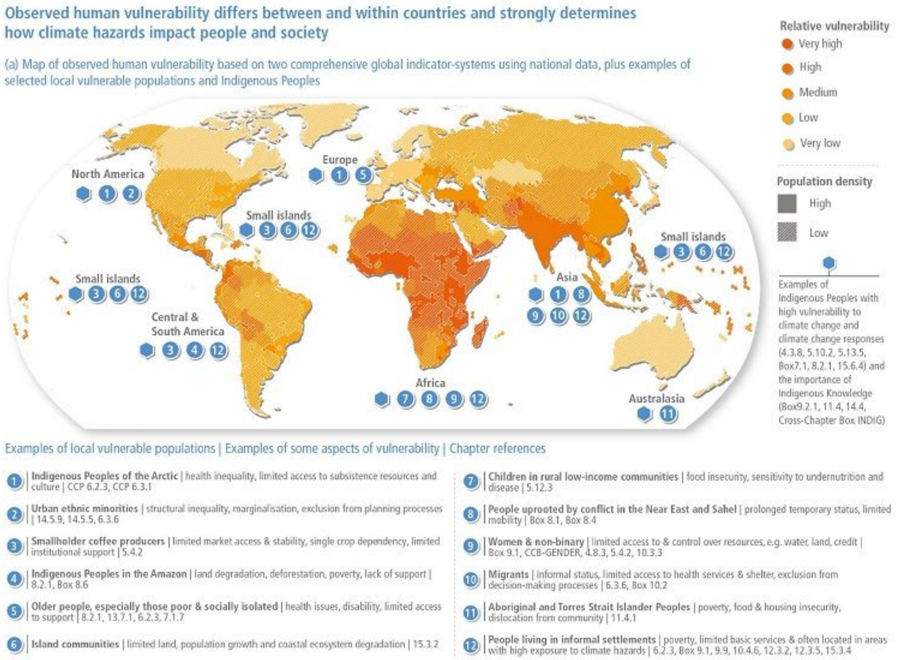
Tackling Climate Change with Machine Learning

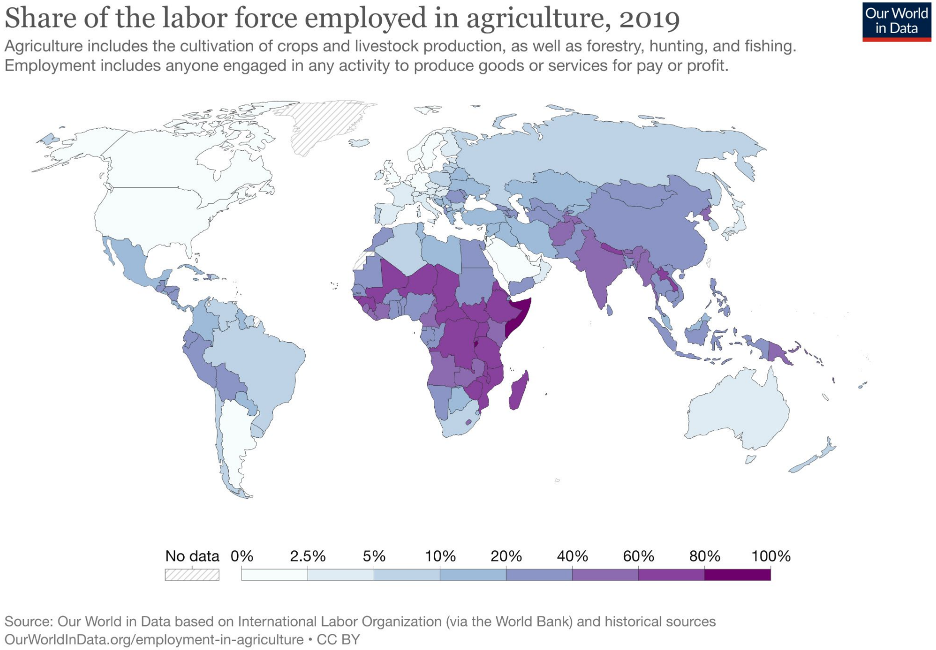
NeurIPS, Dec 2025

Andrew Hobbs Benson Adomako Addi Joof Jesse Anttila-Hughes

University of San Francisco, Department of Economics

Many of the world’s most climate vulnerable are farmers





Sources: IPCC; Our World in Data

Problem

- When crop yields are poor, farmers must rely on costly coping strategies, including reduced food consumption, removing children from school, or selling productive assets.
- Accurate, timely data on regional crop yields is essential to effective agricultural insurance, famine early warning, and government programs.
- Developing accurate crop yield measures requires costly ground-truth data
    - Precisely georeferenced farm yield data are scarce in developing countries
    - Existing measures are often not tested for accuracy.

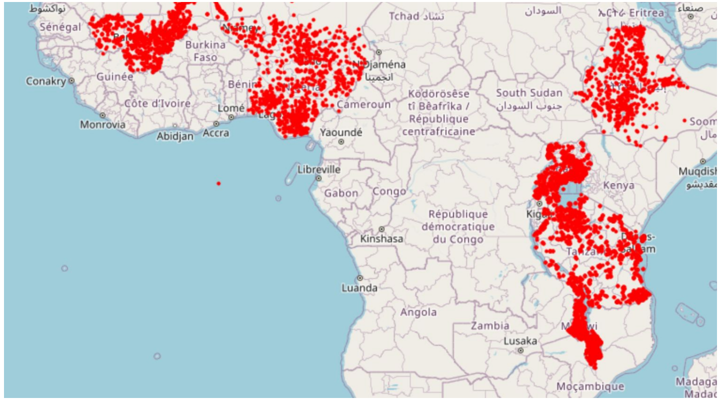
Our Approach

- We combine reported yields existing household survey data from the World Bank's Living Standards Measurement Survey (LSMS) with remotely sensed satellite data on climate and environmental features from six African countries
- Apply machine learning techniques to develop a predictive index of maize production within each country and across countries.
- Develop a promising new approach for generating accurate regional yield estimates for smallholder farmers.

Data

Crop Yields (Maize): World Bank LSMS-ISA Data from Nigeria, Uganda, Malawi, Tanzania, Mali, & Ethiopia
Croplands: NASA Global Food Security Analysis Support Data (GFSAD)
Vegetation Indices: NDVI, GCVI, and EVI (LANDSAT) monthly maximum.
Growing/ Killing Degree Days, Temperature + Soil Moisture: ECMWF ERA5
Precipitation: CHIRPS

Total of ~30,000 observations in ~7,500 enumeration areas.

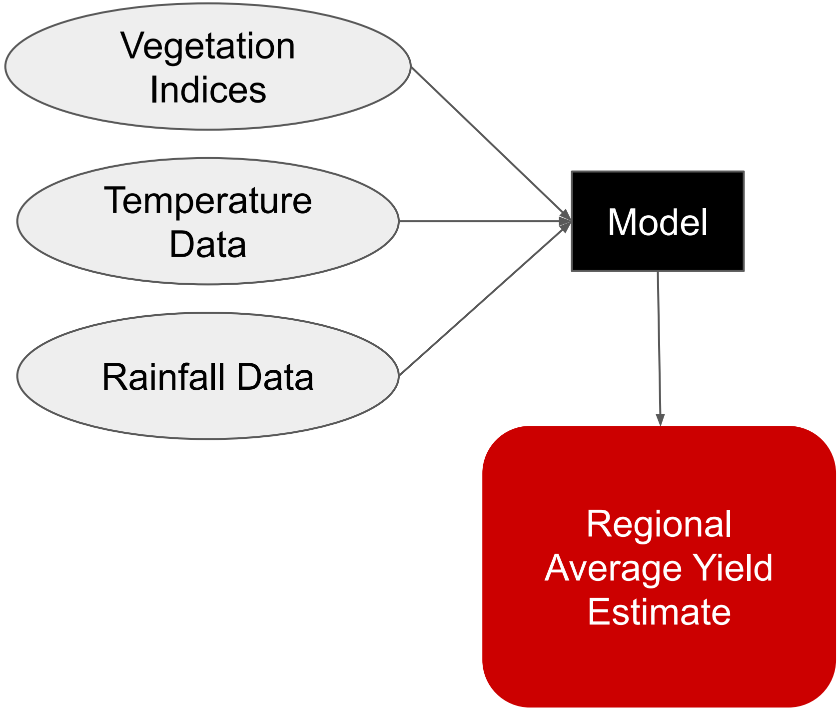


Why machine learning?

Multiple types of data tell us something about crop yields

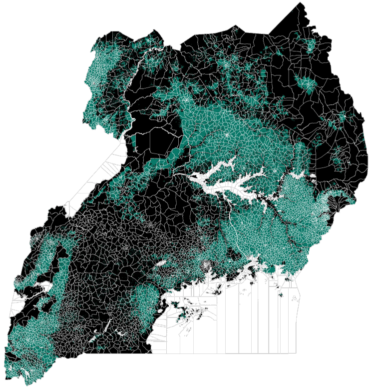
- Vegetation indices
- Precipitation
- Temperature
- Root/ soil moisture

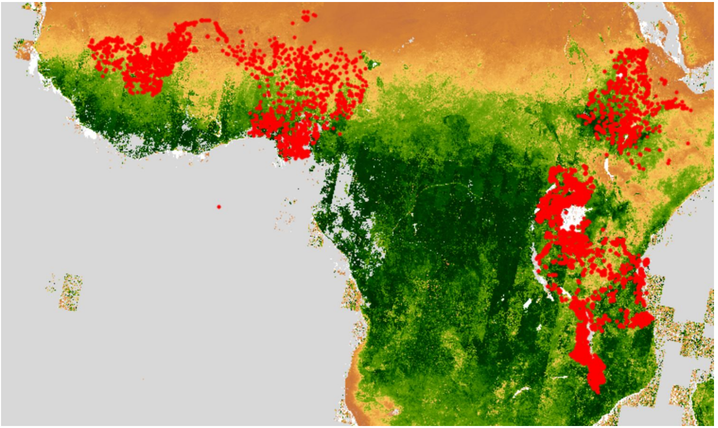
Combining these data into a single index of average crop yields in an area is complex and machine learning is well suited to this sort of task.



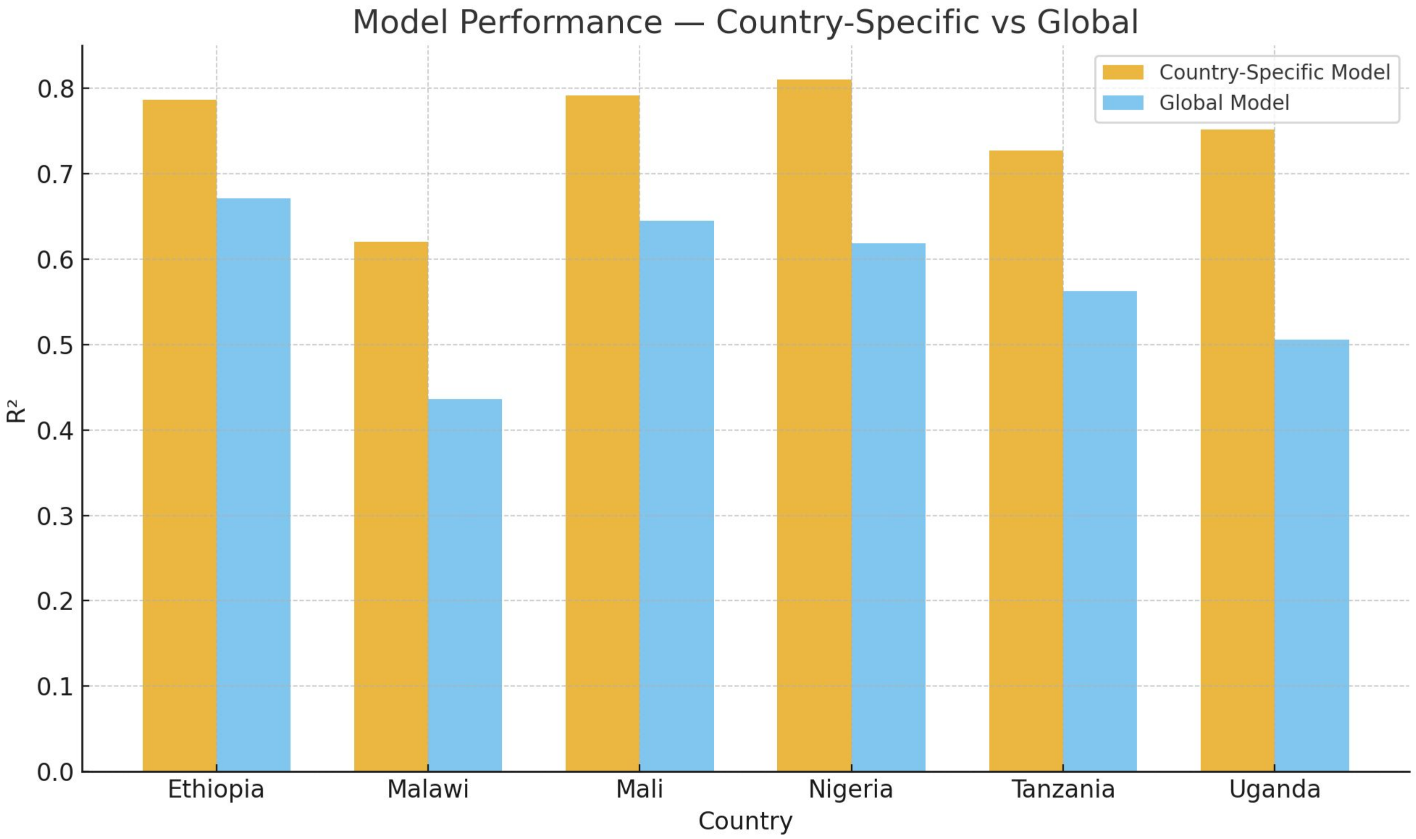
Linking satellite and survey data

1. Points indicate approximate center of ‘enumeration area.’
2. For each enumeration area, collect all data within a 10km circle on cropland pixels.
3. Calculate spatial mean.
4. Calculate monthly mean or max for 12 months leading to harvest.
5. Link those data with LSMS survey data on crop yields from the same location.





Results: Country-specific and Global models



Conclusions

- Early results suggest reasonably good accuracy in all countries.
- Global (or continent-wide) model does nearly as well as country-specific model in many cases.
- Next steps:
  - Risk averse hyperparameter tuning (emphasize correctness in ‘bad’ years).
  - Test models on smaller, higher accuracy datasets with field-specific data.
  - Link predicted regional yields to health and economic outcomes from other surveys.
  - Try CNNs, giving the model pixel-level + foundation model data rather than spatially aggregated data.