Sub-Field Crop Yield Predictions Using Satellite-Based Machine Learning: A Multi-Year Analysis with USDA-ARS Ground-Truth Data

Annelise Holguin[1], Eunsang Cho[2], Simon Kraatz[3], Samar Ranjit[2], Jisung Chang[3], Feng Gao[3], David Johnson[4], Martha Anderson[3], Haoteng Zhao[3], Richard Cirone[3], Michael Cosh[3]

[1]Department of Geography and Environmental Studies, Texas State University, Texas, USA; [2]Ingram College of Engineering, Texas State University, Texas, USA; [3]Hydrology and Remote Sensing Lab, USDA, Maryland, USA [4]USDA National Agricultural Statistics Service, Washington, DC, United States

Contact: vsh26@txstate.edu

AGU25 Poster # 0922

Summary

- Accurate yield prediction assists in promoting food security, optimizing agricultural management, and supporting economic planning.
- Machine Learning (ML) techniques are useful in yield prediction as they can calculate complex relationships between variables and rank them.
- The United States Agriculture Department – Agricultural Research Service (USDA-ARS) published a sub-field scale crop yield dataset (5m resolution raster) covering 2014-2024 at the Beltsville Agricultural Research Center (BARC) (Dulaney et al. 2024).
- Analysis is conducted on the Google Earth Engine (GEE) cloud-computing platform
- This project utilizes a Random Forest Machine Learning (RFML) classifier to analyze satellite imagery (Harmonized Sentinel-2 data and Landsat 8) and input climate, physical, and soil datasets to predict crop yield at a sub-field scale.

Research Questions

1. How accurately can corn, soybean, and winter wheat yield be predicted with a single year of data? Can unknown fields be accurately predicted?
2. Which input features are the most important in predicting yield for corn, soybeans, and winter wheat? Can yield be predicted with less input variables?
3. Can yield be predicted in-season/with less months of training data? And can multiple years of data be used to predict unknown/future years of yield?

Study Area

- The fields analyzed in this study are from the Beltsville Agricultural Research Center (BARC) in Beltsville, Maryland.

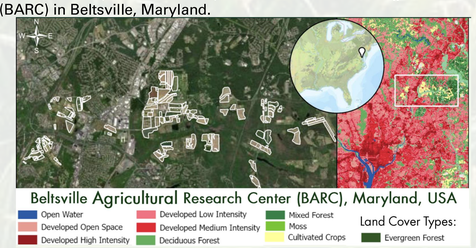

Figure 1: Study area at the Beltsville Agricultural Research Center (BARC) in Beltsville, Maryland

- The Atlantic Seaboard Fall Line crosses through the area, creating variable soils from east to west as the sandy, sedimentary Atlantic coastal plain transitions into the metamorphic and igneous Piedmont Plateau.

Data

Table 1: Data Sources

Type	Variables	Full Name	Values	Resolution (m^2)	Data Source
Vegetation Indexes	NDVI	Normalized Difference Vegetation Index	Median, Max, Range	10	Harmonized Sentinel 2 and Landsat 8
	GI	Greenness index	Median, Max, Range	10	Harmonized Sentinel 2 and Landsat 8
	EVI	Enhanced Vegetation Index	Median, Max, Range	10	Harmonized Sentinel 2 and Landsat 8
	NDWI	Normalized Difference Water Index	Median, Max, Range	10	Harmonized Sentinel 2 and Landsat 8
Climatic	LST	Land Surface Temperature	Median, Range	100	Landsat 9
	Precip	Precipitation	Time Period Sum	4,000	GRIDMET
	Arid	Aridity	Time Period Sum	4,000	GRIDMET
Physical	Slope	-	-	1	DEM
	Aspect	-	-	1	DEM
Soil	SSWC	Saturated soil water content	-	30	POLARIS
	SHC	Saturated hydraulic conductivity	-	30	POLARIS
	Clay %	Percentage of clay	-	30	POLARIS
	Tr_SWIR1*	Transformed SWIR Index 1	Median, Max, Range	20	Harmonized Sentinel 2 and Landsat 8
	Tr_SWIR2*	Transformed SWIR Index 2	Median, Max, Range	20	Harmonized Sentinel 2 and Landsat 8

*Tr_SWIR1 and Tr_SWIR2 are proxies of soil moisture based on shortwave-infrared bands 1 and 2

Methodology

- Composite images were computed over the time periods listed in Table 2 describing typical planting dates.

Table 2: Start and End Dates of the Analyses

Crop	Start Date	Harvest Date	In-Season Date
Corn	April 1	October 31	July 31
Soybean	April 1	November 15	August 15
Wheat	September 15	July 1	April 1

- Physiological differences between the crops account for different ranges of yield.
  - Avg yield: Corn 124.52 Mg/ha, Soybean 47.34 Mg/ha, Wheat 53.00 Mg/ha.

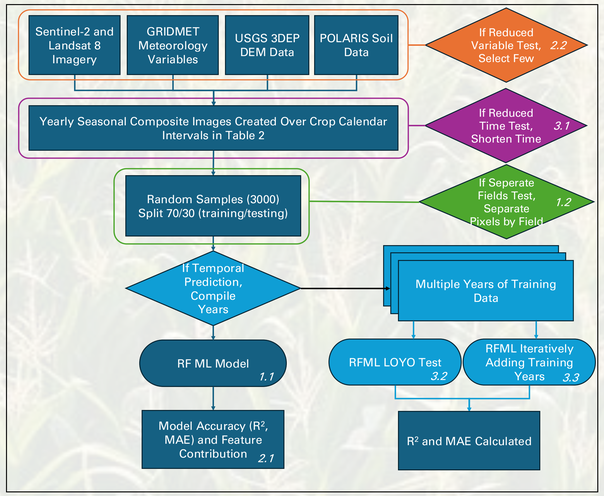

Figure 2: Methodology Flowchart

Results 1 – Spatial Predictions

1.1 Single Year Predictions with Mixed Training and Testing:

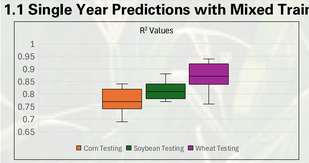

Figure 3: R^2 Result for Single Year Crop Yield Predictions

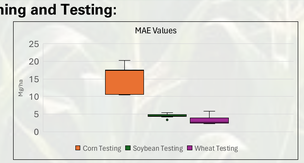

Figure 4: MAE Result for Single Year Crop Yield Predictions

- Corn had the lowest average testing R^2 of 0.78, soybean had an average of 0.81, and wheat had the highest R^2 of 0.87.
- However, predictions of soybean were the most consistent in both training and testing.

1.2 Single Year Predictions Training and Testing on Separate Fields:

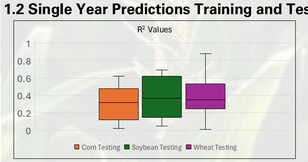

Figure 5: R^2 Result for Unknown Fields Predictions

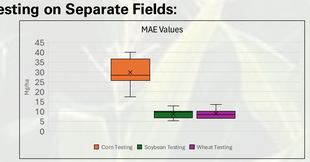

Figure 6: MAE Result for Unknown Fields Predictions

- Model performance for predicting unknown fields was variable, but this is a limitation of using R^2 as a metric. While it shows seemingly large differences in model performance, it is influenced by clustering based on the range of yield and which are used as training and testing. MAE is better able to show the stability of model performance.

Results 2 – Feature Variable Importance

2.1 Feature Variable Importance of Single Year Mixed Crop Yield Predictions:

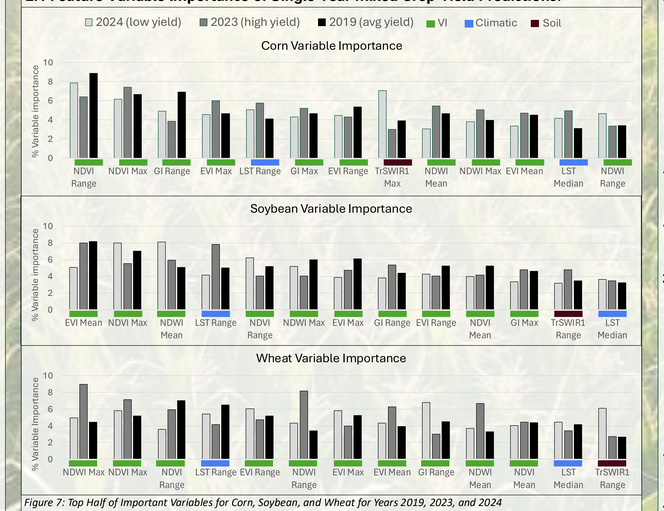

Figure 7: Top Half of Important Variables for Corn, Soybean, and Wheat for Years 2019, 2023, and 2024

- Out of all four variable categories, VIs were deemed the most important.
- Non-VI variables were spread, but LST range and median were in the top half of variables across the board.
- Precipitation, aridity, slope, and aspect were all the lowest ranked out of the 27 input variables. This is attributed to the small size of the study area, rendering these variables unchanging throughout the study area.

2.2. Yield Prediction with a Reduced Number of Inputs:

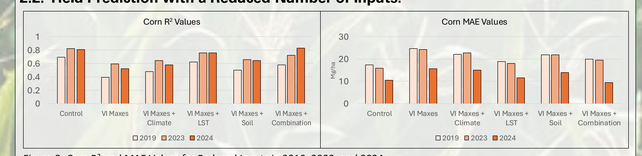

Figure 8: Corn R^2 and MAE Values for Reduced Inputs in 2019, 2023, and 2024

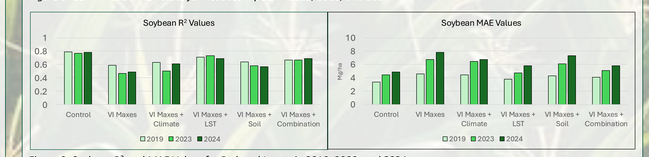

Figure 9: Soybean R^2 and MAE Values for Reduced Inputs in 2019, 2023, and 2024

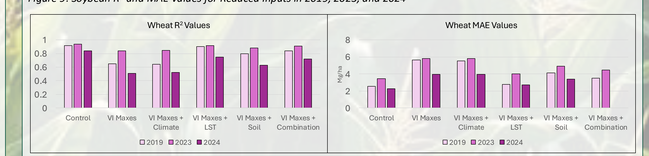

Figure 10: Wheat R^2 and MAE Values for Reduced Inputs in 2019, 2023, and 2024

- The best performing combination of variables was the VI + LST variables category, which can be seen in this result and the variable importance result.
- However, the addition of any set of environmental variables outperformed only using VIs, which is consistent with the literature.
- 2024, which was a low yield year due to drought, was also harder to predict for both corn and wheat, but did not seem to affect soybean predictions.

Results 3 – Temporal Predictions

3.1 In-Season Yield Mixed Training and Testing Yield Predictions:

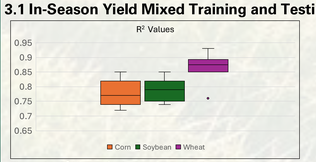

Figure 11: R^2 Values for In-Season Predictions

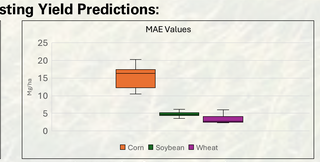

Figure 12: MAE Values for In-Season Predictions

- The correlation between the in-season and the single year model predictions was 0.92 for corn, 0.87 for soybean, and 0.97 for wheat, showing model performance was highly comparable to the model using the full amount of time.
- Shortening the composite interval by three months had negligible effect on model performance, except for soybean. This suggests utility for in-season yield prediction.

3.2 Leave One Year Out (LOYO) Testing:

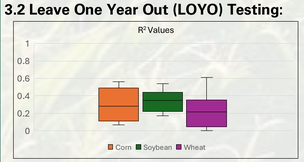

Figure 13: R^2 Values for LOYO Testing

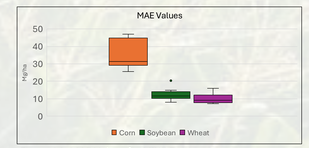

Figure 14: MAE Values for LOYO Testing

- The results indicate that the model captures yield trends reliably for soybeans but struggles with corn and wheat, where low R^2 values (often <0.3) and higher MAE suggest weaker generalization across years.

3.3 Incrementally Adding Training Data:

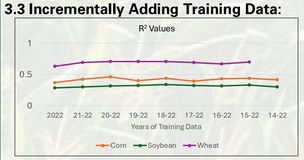

Figure 15: R^2 Values for Incremental Training Data

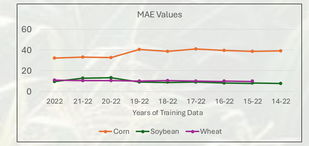

Figure 16: MAE Values for Incremental Training Data

- The results revealed that model performance does not improve when multiple years of training data are used. This is because of a flaw in RFML testing as it cannot adjust for differing ranges of input variables over the years.

Conclusions and Key Takeaways

- Crop yield was able to be predicted with an average R^2 of 0.78 for corn, 0.81 for soybean, and 0.87 for winter wheat based on yield measurements from a subset of local fields.
- Using the model before harvest or on untrained fields worked comparably well to the single year predictions.
- VIs were the most important types of variables, but performance was enhanced with any set of additional variables, especially LST variables.
- Predictions did not improve with more years of training data due to a flaw in RFML. This model and methodology can be viably used to predict crop yield

Future Work

- Future work should test transferability of this model by testing on other crops and in other climates
- Different types of ML and deep learning models should be experimented with as well, especially in applications using multiple years of training data.

Acknowledgements

This research is generously supported by the U.S. Department of Agriculture's Agricultural Research Service (USDA ARS) under Agreement Number 58-8042-4-179.

References

Dulaney, W. P., Anderson, M. C., Gao, F., Stern, A., Moglen, G., Meyers, G., Daughtry, C. S. T., White, W., Akumaga, U., & Showalter, J. (2024). Development of a gridded yield data archive for farm management and research at the USDA Beltsville Agricultural Research Center. Agrosystems, Geosciences & Environment, 7, e20474.https://doi.org/10.1002/agg2.20474.