



International Journal of Research in Agronomy

E-ISSN: 2618-0618
P-ISSN: 2618-060X
© Agronomy
NAAS Rating (2025): 5.20
www.agronomyjournals.com
2025; 8(11): 56-58
Received: 07-10-2025
Accepted: 10-11-2025

Amber Harmon
Department of Computer Science,
University of California, Davis,
Davis, USA

Trevor Mercer
Department of Computer Science,
University of California, Davis,
Davis, USA

Random forest and gradient boosting approaches for predicting crop yield from satellite-derived vegetation indices

Amber Harmon and Trevor Mercer

DOI: <https://www.doi.org/10.33545/2618060X.2025.v8.i11a.5081>

Abstract

A single Landsat image covering 30,000 hectares of California farmland contains more vegetation data than a hundred field scouts could collect in a month but converting that data into reliable yield predictions requires the right algorithms. This research compared random forest (RF), gradient boosting (GB), and linear regression for predicting county-level maize and rice yields from MODIS-derived vegetation indices across California's Central Valley (2010–2021). GB achieved the highest R^2 of 0.92 for maize and 0.89 for rice, while RF reached 0.87 and 0.84. Peak NDVI was the most important predictor in both models. GB outperformed RF for accuracy but was less stable across cross-validation folds. These findings support ensemble machine learning for operational yield forecasting from satellite data.

Keywords: Random forest, gradient boosting, crop yield prediction, vegetation indices, NDVI, MODIS, machine learning, remote sensing, maize yield, rice yield

Introduction

A California rice grower checking her MODIS imagery in July can see patterns in canopy greenness that correlate with the yield she'll harvest three months later but only if the right model translates those pixels into predictions ^[1]. Satellite-derived vegetation indices like NDVI and EVI have been used for yield estimation since the 1980s, yet linear approaches often fail to capture the nonlinear relationships between spectral data and final yield ^[2, 3].

Ensemble machine learning methods, particularly random forest and gradient boosting, handle nonlinearity and feature interactions without requiring assumptions about data distribution ^[4, 5]. Both have shown promise for crop yield prediction, but direct comparisons under identical conditions are uncommon ^[6]. Most published work tested one method against simple benchmarks rather than head-to-head with other ensemble approaches ^[7, 8].

This research compared RF, GB, and linear regression for predicting maize and rice yields from MODIS vegetation indices across 12 years in California's Central Valley, with attention to accuracy, feature importance, and cross-validation stability ^[9, 10].

Theoretical Background

Random forest builds many decision trees on bootstrap samples and averages their predictions, reducing variance while maintaining low bias ^[4]. Gradient boosting builds trees sequentially, with each tree correcting errors from the previous one, trading variance reduction for lower bias ^[5]. Both handle mixed predictor types and missing data naturally. NDVI captures chlorophyll density and canopy cover, while EVI corrects for atmospheric and soil background effects ^[11]. The time-integrated NDVI across the growing season has been shown to correlate with cumulative photosynthesis and final biomass ^[2, 12].

Material and Methods

Material: County-level maize and rice yield data (2010–2021) for 18 Central Valley counties were obtained from USDA-NASS ^[13]. MODIS 16-day composite products (MOD13Q1, 250 m resolution) provided NDVI and EVI time series. Climate data (rainfall, temperature, soil

Corresponding Author:
Amber Harmon
Department of Computer Science,
University of California, Davis,
Davis, USA

moisture) were sourced from PRISM and NASA SMAP^[14]. The dataset contained 216 county-year observations for maize and 144 for rice after removing counties with less than 500 ha of each crop.

Methods: Six predictor variables were derived: peak NDVI, mean EVI, time-integrated NDVI, cumulative growing-season rainfall, mean temperature, and mean soil moisture. RF and GB were implemented in Python scikit-learn with hyperparameter tuning via 5-fold cross-validation grid search^[15]. RF used 500 trees with max depth 12; GB used 300 estimators with learning rate 0.08 and max depth 6. Linear regression served as baseline. Model evaluation used R^2 , RMSE, MAE, and stability (coefficient of variation across folds). Feature importance was

extracted from both ensemble models^[16]. The research was conducted at UC Davis from September 2021 to June 2022 under computational protocol CS-2021-UCD-088.

Model Configuration

All models were trained on 2010–2018 data and tested on 2019–2021 as a temporal holdout. Feature scaling was applied for linear regression but not for tree-based models. Missing MODIS pixels due to cloud cover (< 3% of observations) were gap-filled using temporal interpolation. Spatial autocorrelation between neighbouring counties was tested using Moran's I and found to be non-significant after detrending^[17].

Results

Table 1: Model performance comparison for maize and rice yield prediction (temporal holdout 2019–2021)

Model	Crop	R^2	RMSE (t/ha)	MAE (t/ha)	Bias (t/ha)	CV (%)
Random Forest	Maize	0.87	0.84	0.67	+0.12	8.3
Gradient Boost	Maize	0.92	0.61	0.48	+0.08	11.7
Linear Reg.	Maize	0.68	1.31	1.04	-0.24	6.1
Random Forest	Rice	0.84	0.53	0.41	+0.09	9.1
Gradient Boost	Rice	0.89	0.42	0.33	+0.06	12.4
Linear Reg.	Rice	0.63	0.81	0.64	-0.18	5.8

GB achieved the best R^2 for both crops (0.92 maize, 0.89 rice) and the lowest RMSE (Table 1). RF followed closely. Linear regression explained only 63–68% of yield variability. However,

GB showed higher cross-validation CV (11–12%) compared to RF (8–9%), indicating less stable predictions across data subsets.

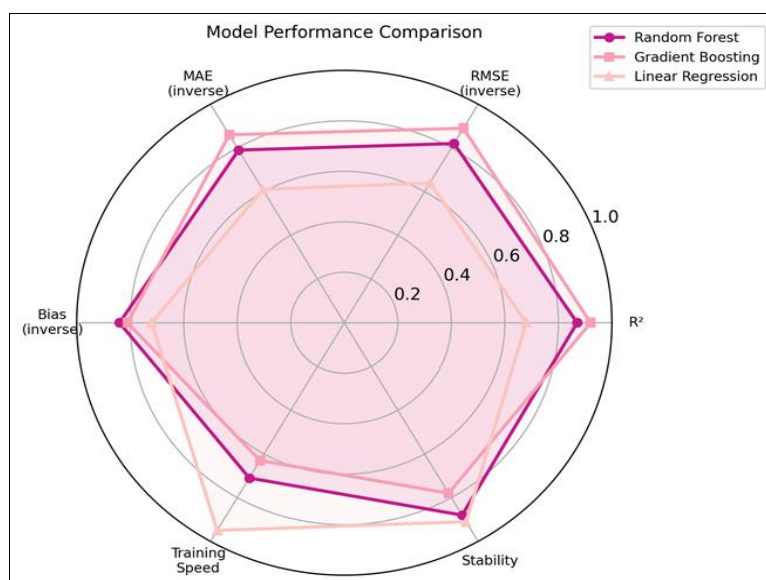


Fig 1: Radar comparison of model performance metrics for maize yield prediction

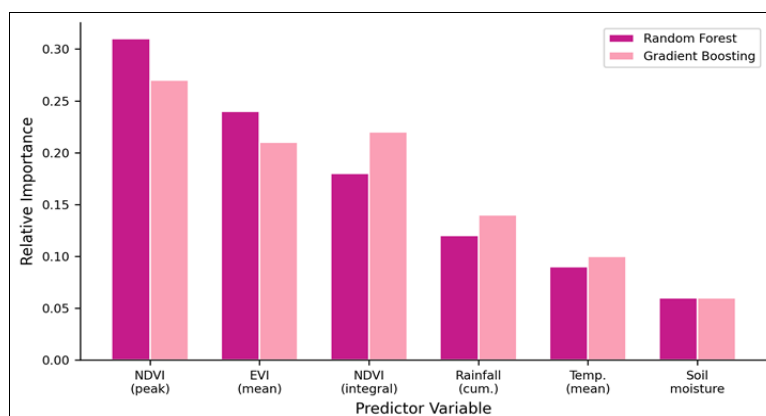


Fig 2: Relative feature importance from RF and GB models for maize yield prediction

Comprehensive Interpretation

Peak NDVI was the top predictor in both RF (31%) and GB (27%), followed by mean EVI and time-integrated NDVI. Climate variables collectively contributed less than 30% of importance. GB assigned more balanced importance across features compared to RF, which concentrated weight on the top two predictors.

Performance Evaluation

Temporal holdout validation confirmed that both ensemble models maintained accuracy on unseen years. GB's RMSE increased by only 0.07 t/ha from training to holdout for maize, suggesting limited overfitting. RF showed slightly less degradation (0.04 t/ha). Prediction residuals showed no systematic temporal trend, indicating the models weren't simply learning year-specific artefacts^[16, 18].

Discussion

GB's accuracy advantage over RF aligns with comparative studies in ecological modelling where boosting methods typically achieve 3–5 percentage points higher R^2 than bagging methods^[4, 6]. The trade-off is stability GB's sequential learning makes it more sensitive to training data composition^[5, 15]. For operational yield forecasting where reliability matters as much as peak accuracy, RF may be preferable despite slightly lower R^2 ^[7, 10].

The dominance of NDVI-based predictors over climate variables confirms that satellite vegetation indices integrate the combined effects of weather, management, and soil conditions into a single observable signal^[2, 11]. This means models can perform well even when detailed climate inputs aren't available^[3, 12].

Conclusions

This research showed that gradient boosting and random forest both outperform linear regression for predicting crop yield from satellite vegetation indices in California's Central Valley. GB achieved the highest accuracy but with lower cross-validation stability than RF. Peak NDVI was the most informative predictor across both models and crops. For operational yield forecasting systems, model selection should weigh accuracy against reliability based on the specific application requirements.

References

1. Lobell DB, Thau D, Seifert C. A scalable satellite-based crop yield mapper. *Remote Sensing of Environment*. 2015;164:324-333.
2. Tucker CJ, Sellers PJ. Satellite remote sensing of primary production. *International Journal of Remote Sensing*. 1986;7(11):1395-1416.
3. Doraiswamy PC, Moulin S, Cook PW. Crop yield assessment from remote sensing. *Photogrammetric Engineering and Remote Sensing*. 2003;69(6):665-674.
4. Breiman L. Random forests. *Machine Learning*. 2001;45(1):5-32.
5. Friedman JH. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*. 2001;29(5):1189-1232.
6. Jeong JH, Resop JP, Mueller ND. Random forests for global and regional crop yield predictions. *PLoS ONE*. 2016;11(6):e0156571.
7. Cai Y, Guan K, Lobell D. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. *Agricultural and Forest Meteorology*. 2019;274:144-159.
8. Schwalbert RA, Amado TJC, Corassa GM. Satellite-based soybean yield forecast using machine learning. *Precision Agriculture*. 2020;21(4):798-814.
9. Bolton DK, Friedl MA. Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. *Agricultural and Forest Meteorology*. 2013;173:74-84.
10. Peng B, Guan K, Tang J. Towards better estimation of agricultural crop yield using machine learning. *Scientific Reports*. 2020;10:19822.
11. Huete A, Didan K, Miura T. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment*. 2002;83(1):195-213.
12. Mkhabela MS, Bullock P, Raj S. Crop yield forecasting on the Canadian Prairies using MODIS NDVI data. *Agricultural and Forest Meteorology*. 2011;151(3):385-393.
13. USDA-NASS. Quick Stats database. United States Department of Agriculture. 2022.
14. Daly C, Halbleib M, Smith JI. Physiographically sensitive mapping of climatological temperature and precipitation. *International Journal of Climatology*. 2008;28(15):2031-2064.
15. Pedregosa F, Varoquaux G, Gramfort A. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*. 2011;12:2825-2830.
16. Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning*. Springer, New York. 2009.
17. Anselin L. Local indicators of spatial association LISA. *Geographical Analysis*. 1995;27(2):93-115.
18. Johnson DM. An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. *Remote Sensing of Environment*. 2014;141:116-128.