Deriving Vegetation Variables from Satellite Observations using a Data-driven Approach

Alan Wang
University of Maryland, December 2025
Mentor: Heshun Wang

August 2025

Section 1: Introduction

Section 1.1: Background

Fraction of vegetation cover (fCover), or *fractional vegetation cover* (FVC), provides a quantitative assessment of the spatial extent of green vegetation that is intrinsically linked to canopy structural attributes, such as *leaf area index* (LAI), leaf angle distribution, and clumping (Bacour et al., 2006). Due to its independence from illumination geometry, fCover is able to supplement classical vegetation indices, such as LAI and fraction of absorbed photosynthetically active radiation, in modeling vegetation dynamics and gradual vegetation cover changes. Although fCover has been estimated on a global scale using remote sensing data, optical remote sensing measurements are heavily hampered by atmospheric influence and cloud cover (Filipponi et al., 2018). As a result, these measurements require quality control, calibration, and validation with more reliable *ground-based observations for validation* (GBOV).

Section 1.2: Objectives

The objectives of the research conducted under Dr. Wang were as follows:

- Apply machine-learning (ML) frameworks/models to better understand the relationship between ground-measured fCover and *surface reflectance* (SR) data from the Visible Infrared Imaging Radiometer Suite (VIIRS) daily surface reflectance (VNP09GA) remote sensing product accessed from *Google Earth Engine* (GEE).
- Assess the performance of ML models in estimating ground-measured fCover from remote sensing data.
- Extrapolate the ML framework/model beyond the GBOV sites to a more global context.

Section 2: Accomplishments

Research primarily focused on tree-based models: XGBoost, Random Forest, and Cubist. These models were chosen for their built-in capability for ensemble learning, which would enable generalizing across distinct vegetation growth cycles and land covers. Work performed on these models fell into three categories:

- Data handling/preprocessing
- Feature engineering and model development
- Model assessment and preliminary satellite calibration

Section 2.1: Data Handling and Preprocessing

The feature engineering, model development, and application made use of three datasets across the 2022–2023 period: VNP09GA, National Ecological Observatory Network (NEON) fCover GBOV, and the Terra and Aqua combined Moderate Resolution Imaging Spectroradiometer (MODIS) Land Cover Type (MCD12Q1) Version 6.1 data product. The SR and land cover from the VNP09GA and MODIS datasets, respectively, were sampled at their native resolution of 500 m. VNP09GA was filtered to pixels with high-quality imagery band data unobstructed by cloud cover or shadows, aerosols, or snow. *The normalized difference vegetation index* (NDVI) was calculated as the normalized difference of the I1 (red) and I2 (near-infrared) bands and added to the dataset. For the ground measurements, pixels were aggregated from a 20-m to a 500-m spatial resolution to ensure a consistent scale across datasets. Aggregated GBOV pixels made from fewer than 150 pixels or less than 70% were filtered to ensure better representativeness of the 20-m data points. Datasets were then spatially joined and temporally aligned to the date of measurement.

For the satellite calibration, GBOV was applied at its native resolution of 20 m to calibrate the Sentinel 2 and Landsat 8 SR products. The remote sensing products were then filtered using their respective quality flags for cloudy pixels, snow and ice, and medium-high aerosol content. The *enhanced vegetation index* (EVI) and fCover were then

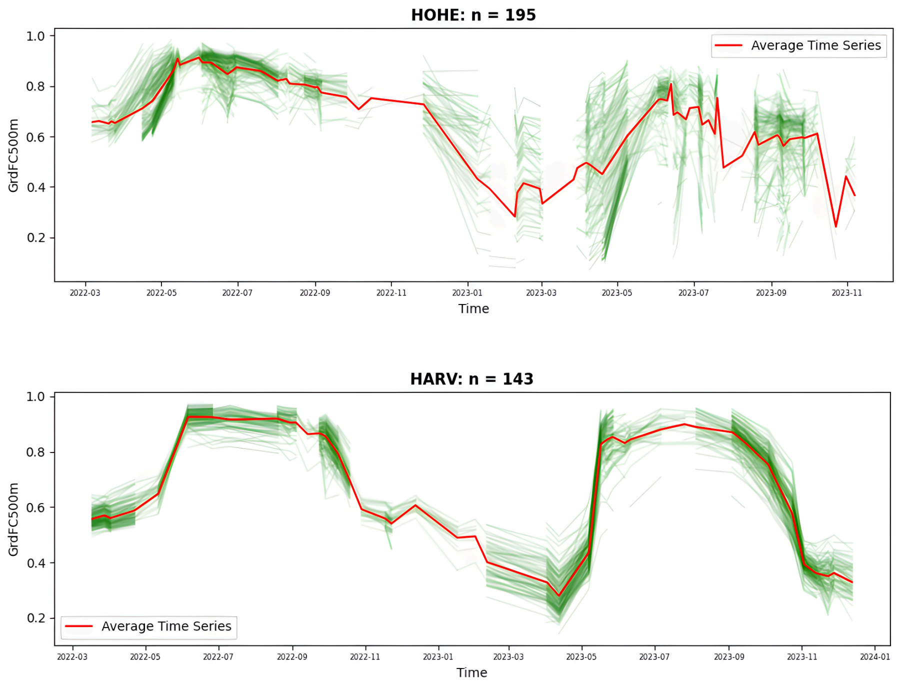

Figure 1. Time series clusters of aggregated GBOV data across two NEON ground measurement sites.

calculated and added to the datasets. EVI was given by

$$EVI = G * \frac{(NIR - Red)}{(Nir + C_1 * Red - C_2 * Blue + L)} ,$$

where L = 1, C_1 = 6, C_2 = 7.5, and G = 2.5. fCover was calculated using the pixel dichotomy model with

$$FVC = \frac{EVI - EVI_{min}}{EVI_{max} - EVI_{min}} ,$$

where EVI_{min} = 0 and EVI_{max} = 0.8 (Ma et al., 2023). A similar process was then used to align the Sentinel and Landsat data with the GBOV.

Section 2.2: Feature Engineering and Model Development

To facilitate feature engineering, a data pipeline was implemented using scikit-learn to apply data transformations and feed the transformed dataset into a regressor model. The feature engineering process included

- Transformation

Table 1: New features derived from VIIRS timestamp and angle information.

	Feature Name	Method of Derivation
Temporal	month_cos day_cos week_cos dayofweek_cos	$f(x) = cos\left(2\pi * \frac{x}{period}\right)$
	month_sin day_sin week_sin dayofweek_sin	$f(x) = sin\left(2\pi * \frac{x}{period}\right)$
Geometric	RelativeAngle	$cos(\lvert(Solar_{Azimuth} - View_{Azimuth})$ $(mod360) - 180\rvert * \frac{\pi}{180})$
	Cos_SolarZenith	$cos\left(Solar_{Zenith} * \frac{\pi}{180}\right)$
	Cos_ViewZenith	$cos\left(View_{Zenith} * \frac{\pi}{180}\right)$

- Selection

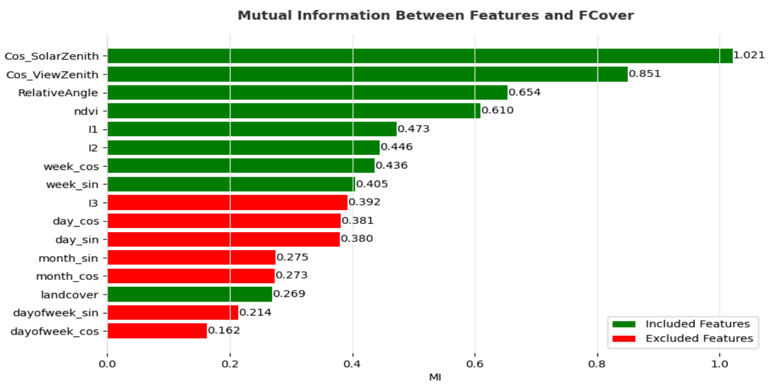

Figure 2: Mutual information used to indicate the initial set of features

- Normalization

The transformation of angle information is informed by its handling within the GEOV2 global product, which utilizes the cosine of the zenith and relative azimuth angles of solar and view directions (Verger et al., 2023). Within the selection phase, mutual information is applied to derive the initial subset of features, which is later refined by using permutation feature importance measured by *mean decrease in accuracy* (MDA). Since land cover had consistently high MDA across the three models, it was included in the final feature set.

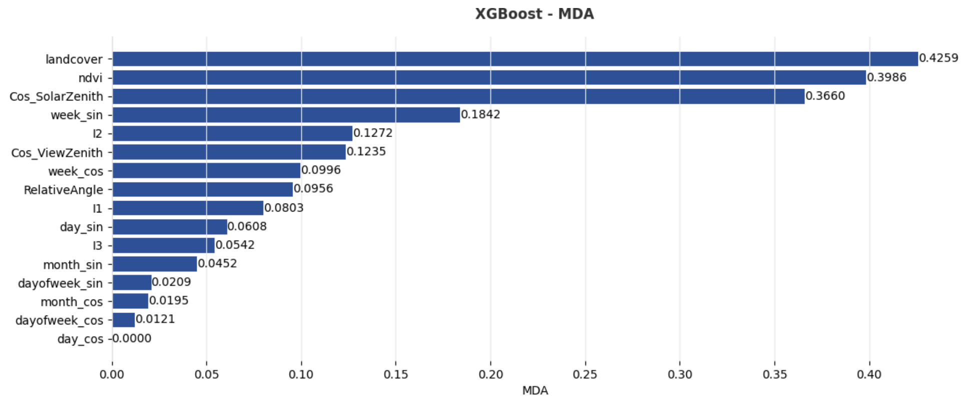

Figure 3: XGBoost mean decrease in accuracy corroborating use of land cover in the final feature set.

With this implementation, we were then able to proceed with our model development process, which consisted of iterations of

1. performing an 80–20 split to separate the aligned dataset into train and test datasets.

2. fitting the pipeline onto the train dataset and using it to transform both the train and test datasets.

3. stratifying the train dataset into folds based on land cover distribution and hypertuning over folds using 10–25 iterations of Bayesian optimization.

4. evaluating the performance of the XGBoost, Cubist regression, and random forest models on the test dataset using the *coefficient of determination* (R^2) and *mean absolute error* (MAE) as metrics.

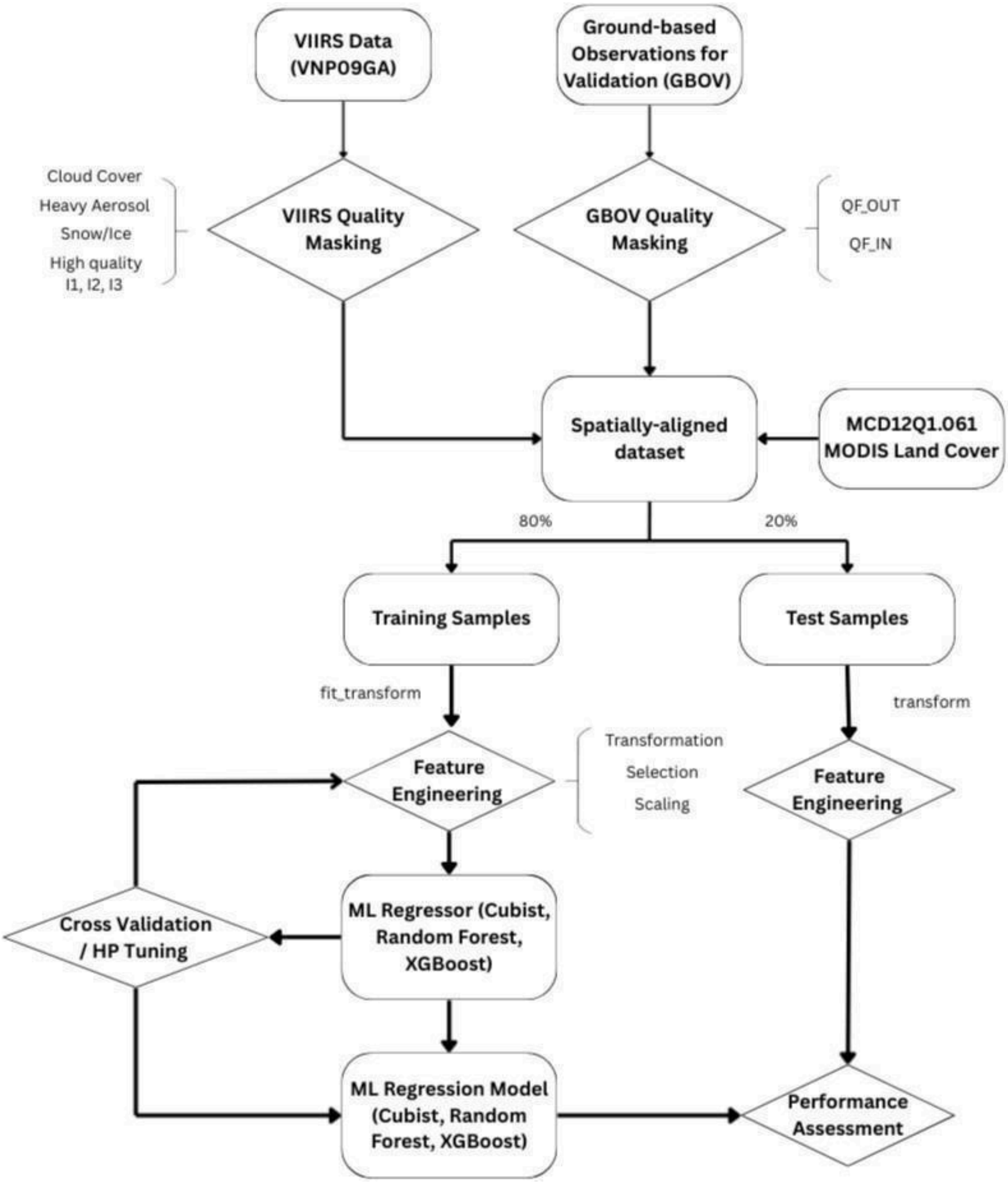

Figure 4: Diagram of data handling and the model development process.

Section 2.3: Model Assessment and Preliminary Satellite Calibration

Performances over the test dataset generally had an R^2 range of 0.90–0.915 and an MAE range of 0.04–0.05, suggesting high predictive accuracy on the test set.

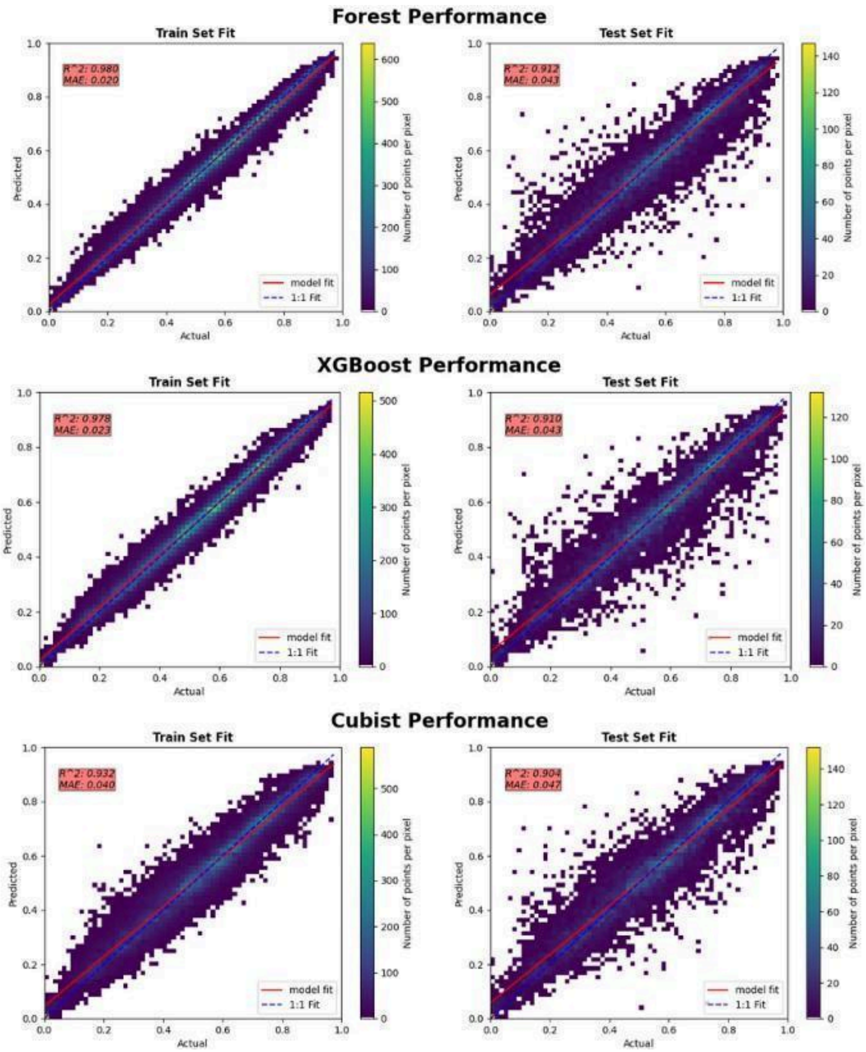

Figure 5: Density scatter plots of model performances over train and test datasets.

Based on these results, we extrapolated the random forest regressor, which performed best on the test dataset and cross validation, beyond the GBOV sites to generate predictions over the East Coast of the U.S. The East Coast was chosen due to the high presence of GBOV sites within the region. Figure 6 shows that the model predictions strongly align with the expected annual growth cycle for fCover: early growth, rapid growth, and senescence/harvest.

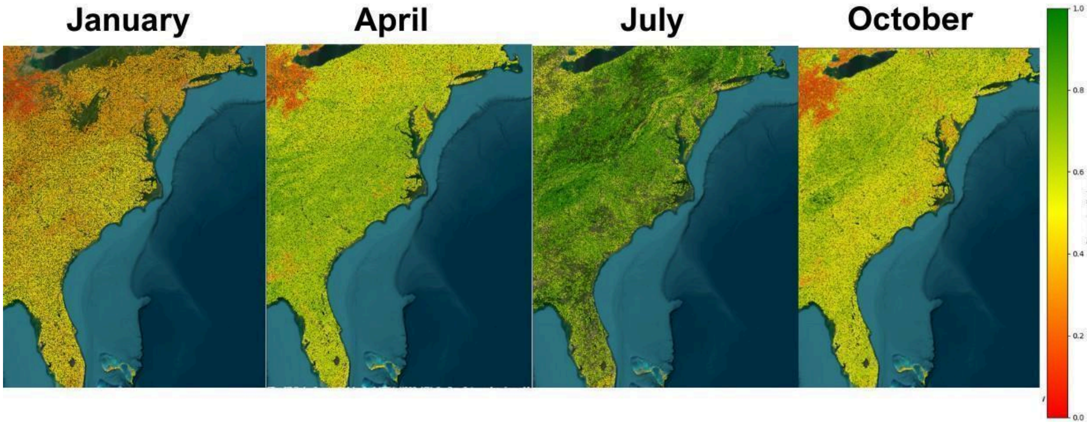

Figure 6: Median random forest predictions of ground fCover over the East Coast for 2023.

To quantify the performance of the model beyond the GBOV sites, we attempted to calibrate Sentinel 2 and Landsat 8 SRs by fitting a double logistic growth model to the remote sensing measurements and applying a scale factor and bias (y_0) to model the ground-measured fCover.

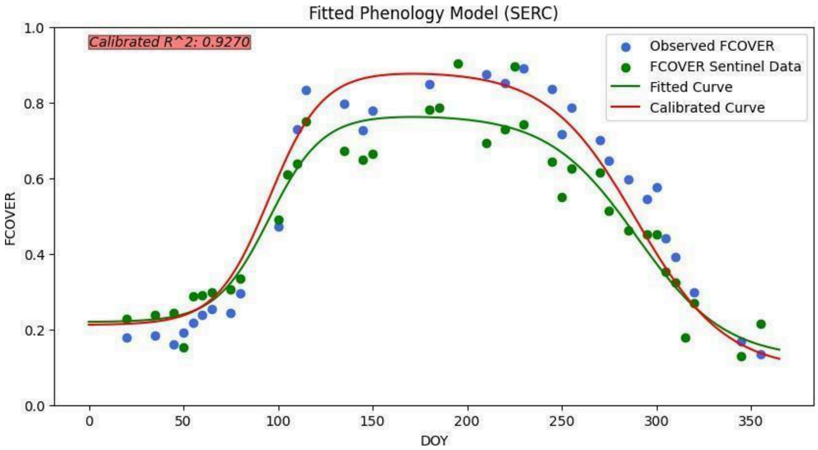

Figure 7: Result of the calibration process applied to a representative point in the Smithsonian Environmental Research Center (SERC) field site.

Towards this effort, we derived scale factors and biases across land cover types to determine if the parameters were consistent across land cover. We found that even within a single land cover type, a notable degree of variability was observed in scale factors and biases. This within-class variation precluded the

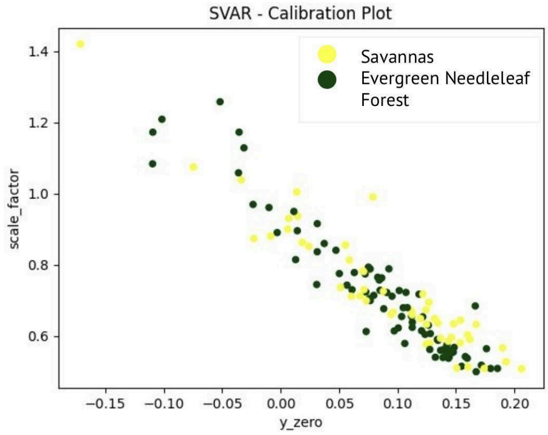

Figure 8: Distribution of scale factors and biases for calibration across points in SVAR for Sentinel 2 SR.

generalization of the calibration model to the entire set of GBOV sites and to points beyond the GBOV sites.

Section 2.4: Additional Figures/Tables

Table 2. Model performances over stratified k-fold cross validation of the train dataset.

	Fold 1 (R^2)	Fold 2 (R^2)	Fold 3 (R^2)	Fold 4 (R^2)	Fold 5 (R^2)	Mean (R^2)	SD (R^2)
Cubist	0.888633	0.889113	0.890394	0.891903	0.891903	0.890909	0.0013799
XGBoost	0.892615	0.887513	0.89036	0.892356	0.893786	0.891326	0.002202
Random Forest	0.895547	0.894318	0.894804	0.897094	0.900401	0.896433	0.002195

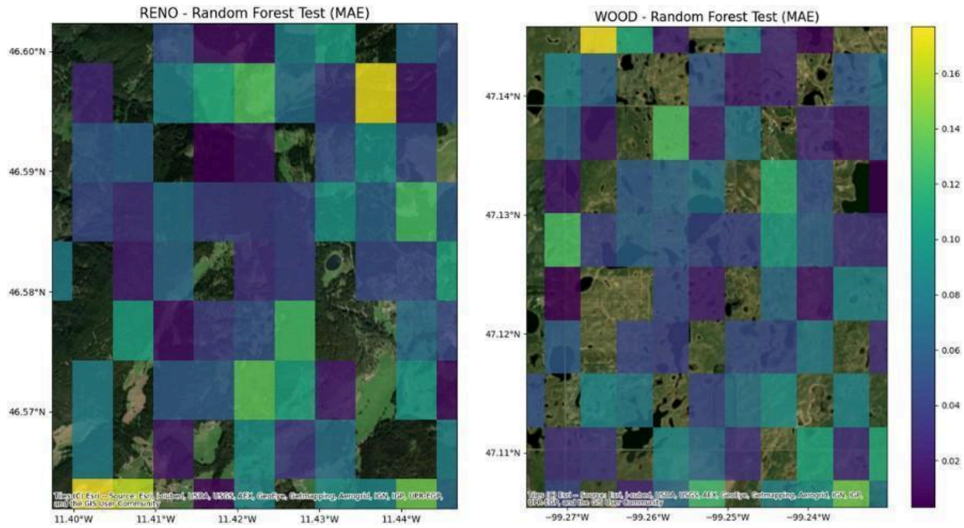

Figure 9. Visualization of random forest performance on the test dataset over two ground measurement sites.

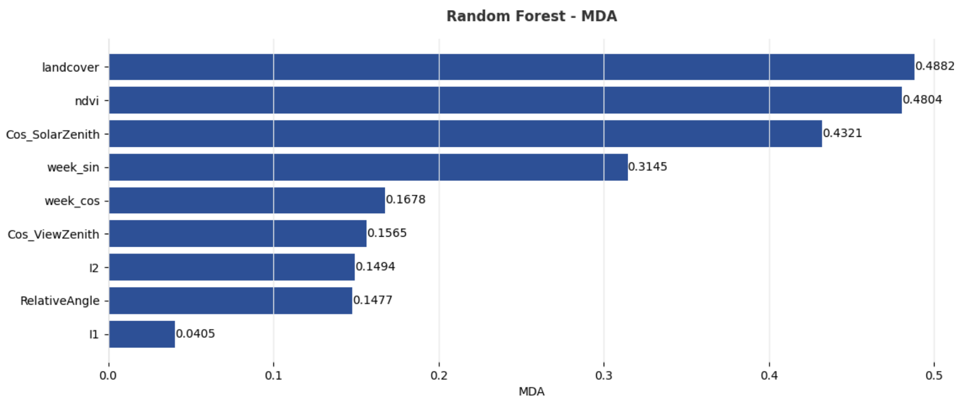

Figure 10. Permutation feature importance (MDA) of the random forest regression model, the best-performing model on the cross validation and test dataset.

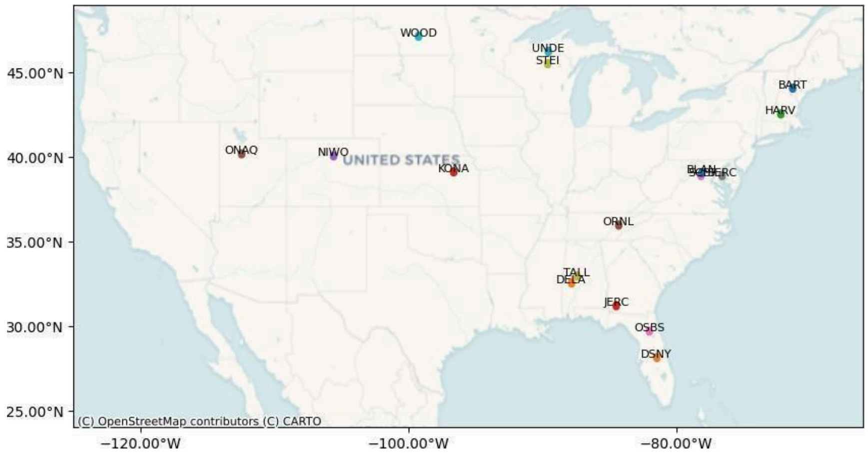

Figure 11. NEON GBOV sites across the continental U.S.

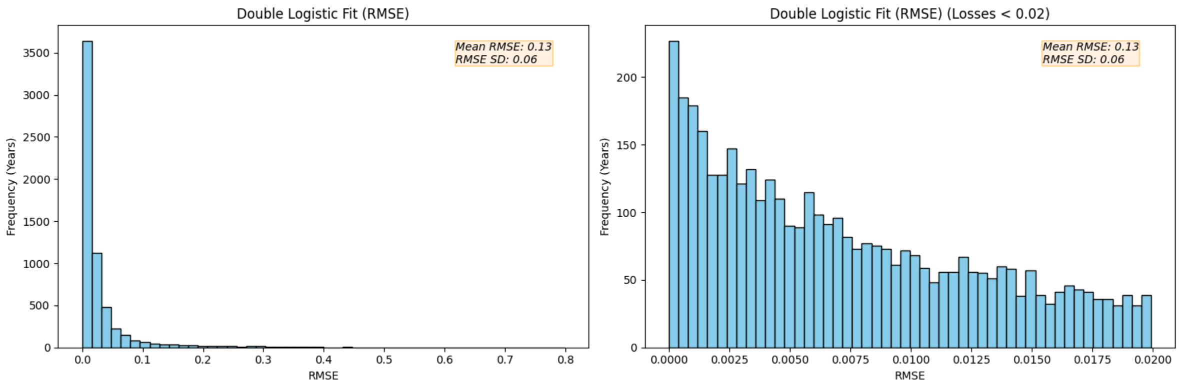

Figure 12. Performance of the double logistic curve fit on GBOV, supporting the expected annual fCover growth cycle.

Section 3: Conclusion

This research has demonstrated the effectiveness of tree-based models in predicting ground-measured fCover using feature information derived from VIIRS SR measurements. The land cover from the MODIS dataset and the VIIRS-derived features such as NDVI, week_sin, week_cos and Cos_SolarZenith were significant in improving the MAE and R^2 of the model predictions. The importance of NDVI and landcover in approximating fCover is consistent with previous work (Carlson & Ripley, 1997; Riihimäki et al., 2019). The work performed in developing the model provides a valuable framework for transforming GEE-supplied raster data and netCDF GBOV using scikit-learn's pipeline into a usable format for ML regression models. For the calibration, additional stratification features outside of land cover will be considered for a more representative set of parameters. Furthermore, calibration may benefit from investigating models such as the Whittaker Turner model or the asymmetric Gaussian model (Atkinson et al., 2012).

The limitations of the research arise from the significant loss of potential data resulting from the satellite quality control and data alignment process. As a result of this data loss, the aligned datasets contained an unequal distribution of data points across GBOV sites, which may have disproportionately biased both models and reduced generalizability. Future research will benefit from validating models using existing global products (e.g. MODIS, GEOV2/GEOV3, and CGLS-LC100) (Masiliūnas et al., 2021; Verger et al., 2023) and over the more globally representative BELMANIP2 sites (Baret et al., 2006).

Acknowledgements

I sincerely thank Dr. Heshun Wang for his invaluable support and feedback throughout the course of this work. This study was supported by NOAA grant NA24NESX432C0001 (Cooperative Institute for Satellite Earth System Studies - CISESS) at the University of Maryland/ESSIC.

References

1. Atkinson, P. M., Jeganathan, C., Dash, J., & Atzberger, C. (2012). Inter-comparison of four models for smoothing satellite sensor time-series data to estimate vegetation phenology. *Remote Sensing of Environment, 123*, 400–417. https://doi.org/10.1016/j.rse.2012.04.001

2. Bacour, C., Baret, F., Béal, D., Weiss, M., & Pavageau, K. (2006). Neural network estimation of LAI, fAPAR, fCover and LAI×Cab, from top of canopy MERIS reflectance data: Principles and validation. *Remote Sensing of Environment, 105* (4), 313–325. https://doi.org/10.1016/j.rse.2006.07.014

3. Baret, F., Morissette, J., Fernandes, R., Champeaux, J., Myneni, R., Chen, J., Plummer, S., Weiss, M., Bacour, C., Garrigues, S., & Nickeso, J. (2006). Evaluation of the representativeness of networks of sites for the global validation and intercomparison of land biophysical products: proposition of the CEOS-BELMANIP. *IEEE Transactions on Geoscience and Remote Sensing, 44* (7), 1794–1803. https://doi.org/10.1109/tgrs.2006.876030

4. Carlson, T. N., & Ripley, D. A. (1997). On the relation between NDVI, fractional vegetation cover, and leaf area index. *Remote Sensing of Environment, 62* (3), 241–252. https://doi.org/10.1016/s0034-4257(97)00104-1

5. Filipponi, F., Valentini, E., Nguyen Xuan, A., Guerra, C. A., Wolf, F., Andrzejak, M., & Taramelli, A. (2018). Global MODIS fraction of green vegetation cover for monitoring abrupt and gradual vegetation changes. *Remote Sensing, 10* (4), 653. https://doi.org/10.3390/rs10040653

7. Ma, X., Ding, J., Wang, T., Lu, L., Sun, H., Zhang, F., Cheng, X., & Nurmemet, I. (2023). A pixel dichotomy coupled linear kernel-driven model for estimating fractional vegetation Cover in arid areas from high spatial-resolution images. *IEEE Transactions on Geoscience and Remote Sensing, 61, 1-15.*

8. Masiliūnas, D., Tsendbazar, N., Herold, M., Lesiv, M., Buchhorn, M., & Verbesselt, J. (2021). Global land characterisation using land cover fractions at

100 m resolution. *Remote Sensing of Environment, 259*, 112409. https://doi.org/10.1016/j.rse.2021.112409

9. Riihimäki, H., Luoto, M., & Heiskanen, J. (2019). Estimating fractional cover of tundra vegetation at multiple scales using unmanned aerial systems and optical satellite data. *Remote Sensing of Environment, 224*, 119–132. https://doi.org/10.1016/j.rse.2019.01.030

10. Verger, A., Sánchez-Zapero, J., Weiss, M., Descals, A., Camacho, F., Lacaze, R., & Baret, F. (2023). GEOV2: Improved smoothed and gap filled time series of LAI, FAPAR and FCover 1 km Copernicus Global Land products. *International Journal of Applied Earth Observation and Geoinformation, 123*, 103479. https://doi.org/10.1016/j.jag.2023.103479