

Predicting forest biodiversity from spaceborne GEDI LiDAR and machine learning: structural drivers and regional transferability across Pacific Northwest forests

Satyam Shah^{1,*} ORCID id: 0009-0009-8451-2305

¹School of Geography, Geology and the Environment, University of Leicester, University Road, LE1 7RH, Leicester, United Kingdom

Abstract

SHAH, S., 2026. Predicting forest biodiversity from spaceborne GEDI LiDAR and machine learning: structural drivers and regional transferability across Pacific Northwest forests. *Folia Oecologica*, 53 (3): 213–226.

Remote sensing offers important opportunities for large-scale biodiversity monitoring, yet the capacity of spaceborne laser scanning to predict forest biodiversity remains insufficiently understood. This study evaluates relationships between forest structural complexity and tree species diversity using data from the GEDI mission. Structural metrics describing canopy height, vegetation density and vertical foliage distribution were linked to ground-based forest inventory data from 9,692 plots across the Pacific Northwest region of the United States. Tree biodiversity was quantified using species richness and two diversity indices. Structural complexity showed consistent positive relationships with biodiversity, with vegetation density and vertical foliage heterogeneity outperforming maximum canopy height as predictors. Machine-learning models explained approximately 17–22% of biodiversity variation under standard cross-validation. In contrast, spatial cross-validation revealed strong reductions in predictive performance, indicating limited geographic transferability. Including broad-scale climate variables produced modest improvements, while regionally calibrated models substantially increased accuracy, achieving up to 40% explained variance for species richness. Overall, the results demonstrate that spaceborne laser-derived forest structure contains meaningful biodiversity information, but effective large-area application requires regional calibration to account for spatial variation in forest composition and environmental conditions.

Keywords

canopy structure, forest biodiversity, GEDI LiDAR, spatial transferability, species richness

Introduction

Forests support the majority of Earth's terrestrial biodiversity and provide critical ecosystem services including carbon sequestration, water regulation, and habitat for countless species making them a cornerstone of both conservation and sustainable development (FAO and UNEP, 2020; UNEP-WCMC and IUCN, 2020). At the same time, forest ecosystems face accelerating pressures from climate change, land-use conversion, and other anthropogenic disturbances, creating an urgent need for moni-

toring approaches that can deliver broad spatial coverage and repeatable measurements to inform conservation and management actions (FAO and UNEP, 2020). Traditional ground-based inventories remain the gold standard for describing species composition and structure but are expensive, time-consuming and necessarily spatially limited, which constrains their ability to provide wall-to-wall biodiversity information over large or remote forested landscapes (STANKE et al., 2020). These practical constraints have driven extensive interest in remote sensing as a scalable complement to *in-situ* inventories for biodiversity monitoring.

*Corresponding author:

e-mail: satyamshah444@gmail.com

© 2026 Authors. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

Ecological theory links habitat structure and heterogeneity to species coexistence: more structurally complex environments tend to support greater niche diversity and thus higher species richness (MACARTHUR and MACARTHUR, 1961; TEWS et al., 2004). Vertical stratification of forest canopies creates gradients of light, moisture and microhabitats that favor species partitioning and coexistence; consequently, three-dimensional structural descriptors that capture vertical complexity are expected, in principle, to be informative predictors of biodiversity. Empirical work using active remote sensing has repeatedly shown that LiDAR-derived structural metrics can explain substantial variation in diversity across taxa and habitats, demonstrating that airborne and terrestrial LiDAR provide powerful proxies for many biodiversity attributes (SIMONSON et al., 2014; MOESLUND et al., 2019).

LiDAR technology is especially well suited to capture three-dimensional forest structure because it directly samples vertical profiles of vegetation, providing estimates of canopy height, vertical distribution of foliage, canopy density and other structural descriptors that optical sensors cannot measure directly (DUBAYAH et al., 2020). Airborne LiDAR studies have linked measures such as foliage-height diversity, canopy cover, canopy height distributions and

leaf/plant area indices to species richness and composition for plants, birds and other taxa (SIMONSON et al., 2014; ATKINS et al., 2023). However, airborne campaigns are costly and spatially patchy, which has limited their utility for consistent, large-extent biodiversity monitoring.

The Global Ecosystem Dynamics Investigation (GEDI) lidar, deployed to the International Space Station, represents a major step toward a spaceborne solution for measuring canopy vertical structure at continental scales (DUBAYAH et al., 2020). GEDI provides footprint-level estimates (~25 m) of relative height percentiles (RHx), canopy cover, plant area index (PAI) and foliage height diversity (FHD), enabling standardized, repeatable measurements of vertical structure across much of the globe (GEDI team documentation). Early comparisons between GEDI and airborne lidar show GEDI can recover the large-scale patterns of canopy height and vertical structure, although accuracy and representativeness depend on factors such as shot density, terrain and canopy closure (BURNS et al., 2024; DHARGAY et al., 2022).

Despite this promise, several critical knowledge gaps remain. It is not yet clear how strong and consistent structure–biodiversity relationships are when extracted from spaceborne LiDAR (as opposed to airborne lidar), which

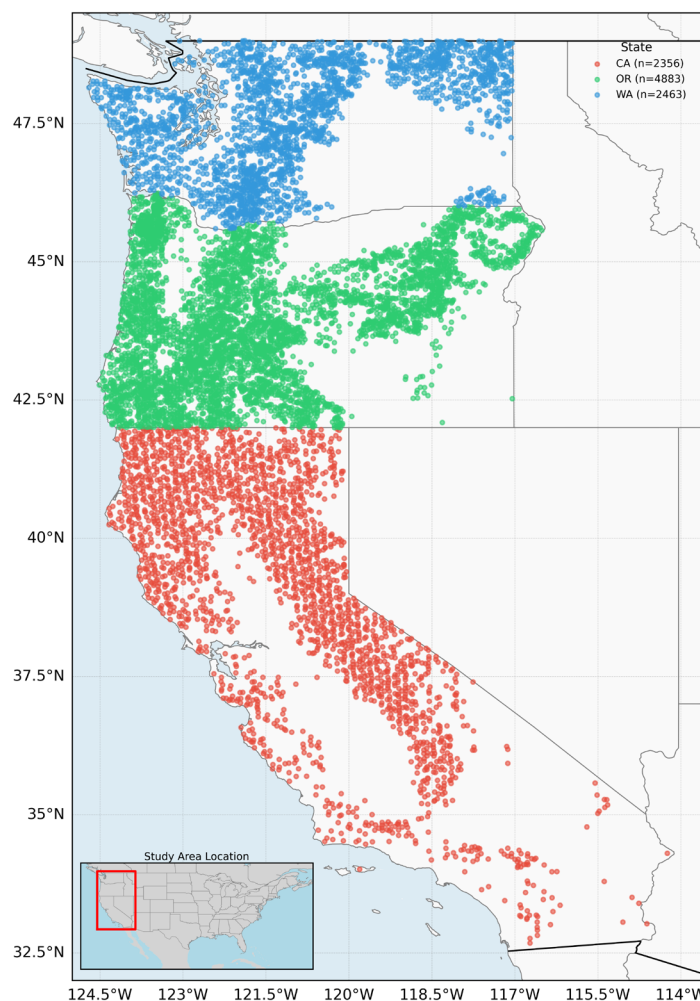


Fig. 1. Distribution of forest inventory plots across the Pacific Northwest study area. CA – California; OR – Oregon; WA – Washington.

structural metrics are most predictive of diversity across broad environmental gradients, and how transferable models trained in one region are when applied elsewhere (TORRESANI et al., 2023; MANDL et al., 2023). Resolving these questions is important for operationalizing GEDI for routine biodiversity monitoring, because reliable, spatially explicit biodiversity mapping would enable identification of biodiversity hotspots, tracking of temporal changes following disturbance, and support for conservation prioritization at scales that ground inventories alone cannot achieve.

The Pacific Northwest (California, Oregon, Washington) is an ideal testbed for exploring these questions, as it contains diverse forest types from coastal temperate rainforests to montane mixed-conifer and Mediterranean oak woodlands and benefits from systematic ground-truth data via the U.S. Forest Service Forest Inventory and Analysis (FIA) program (STANKE et al., 2020). By linking GEDI structural metrics to FIA-derived biodiversity indices and testing model transferability across regional gradients, we can assess both the promise and the limits of spaceborne LiDAR for scalable forest biodiversity monitoring.

Methods

Study area

The analysis focused on forested landscapes across three Pacific Northwest states (California, Oregon and Washington), encompassing a broad range of forest types from coastal temperate rainforest to montane mixed-conifer and Mediterranean-climate woodlands. Latitude varied from 32.68° N to 49.00° N, capturing large environmental gradients in temperature, precipitation and forest composition; the region was selected because of extensive GEDI coverage and systematic ground inventory data from the U.S. Forest Service Forest Inventory and Analysis (FIA) program (WOUDENBERG et al., 2010). These forests exhibit strong gradients in climate, elevation, productivity, and species composition, making the region well suited for evaluating relationships between forest structural complexity and biodiversity across heterogeneous landscapes. The final analysis dataset comprised 9,692 forest inventory plots with complete GEDI structural data: Oregon (n = 4,883; 50.4%), Washington (n = 2,454; 25.4%) and California (n = 2,355; 24.3%) (Fig. 1).

Forest Inventory and Analysis (FIA) data

Ground-based biodiversity data were obtained from the United States Forest Service Forest Inventory and Analysis (FIA) program database. The FIA program provides a nationally consistent forest inventory based on a systematic grid of permanent sample plots distributed approximately every 2,400 ha across forested lands in the United States (BECHTOLD and PATTERSON, 2005).

Each FIA plot consists of four fixed-radius subplots arranged in a standardized configuration, within which all live trees exceeding 12.7 cm diameter at breast height are

measured and identified to species level (USDA Forest Service, 2022). Tree-level attributes include species identity, diameter, status (live or dead), and plot condition information. Each subplot has a fixed radius of 7.32 m (24 ft), with the four subplots arranged in a standardized cluster spanning approximately 0.4 ha, following the national FIA sampling design (BECHTOLD and PATTERSON, 2005).

For this study, FIA plot data collected between 2019 and 2022 were used to align temporally with the operational period of the GEDI mission. Plots lacking valid geographic coordinates or containing no live trees were excluded. Species-level tree data were aggregated at the plot level to compute biodiversity metrics.

Biodiversity metrics

Three commonly used alpha-diversity indices were calculated for each FIA plot to represent forest biodiversity:

Species richness (S), defined as the total number of unique tree species observed within a plot;

The Shannon diversity index:

$$H' = -\sum_{i=1}^S p_i \ln(p_i)$$

where p_i represents the proportional abundance of species i ;

The Simpson diversity index:

$$D = 1 - \sum_{i=1}^S p_i^2$$

which emphasizes dominance and evenness within species assemblages.

These diversity indices are widely applied in forest ecology and provide complementary measures of community composition and structure (MAGURRAN, 2004; CHAO et al., 2014).

GEDI LiDAR data

Forest structural information was derived from the Global Ecosystem Dynamics Investigation (GEDI) LiDAR mission aboard the International Space Station. GEDI is a full-waveform lidar instrument designed to measure three-dimensional vegetation structure globally, with a nominal footprint diameter of approximately 25 m (DUBAYAH et al., 2020).

GEDI Level 2A and Level 2B products were used in this analysis. Level 2A provides relative height (RH) metrics describing the vertical distribution of canopy returns, while Level 2B includes estimates of canopy cover, Plant Area Index (PAI), and Foliage Height Diversity (FHD). RH metrics represent the height below which a given percentage of cumulative waveform energy is returned and are commonly used indicators of canopy height and vertical complexity (DUBAYAH et al., 2020; HANCOCK et al., 2019).

The structural variables extracted for this study included RH50, RH75, RH98, RH100, canopy cover fraction, PAI, and FHD. RH50, RH75, RH98 and RH100 denote the height (in meters above ground) below which 50%,

75%, 98% and 100% of the cumulative waveform energy is returned, respectively, providing percentile-based descriptors of the vertical canopy profile. Canopy cover fraction represents the proportion of ground area obscured by vegetation when viewed from directly above. Plant Area Index (PAI) is a dimensionless measure of the total one-sided area of photosynthetic and non-photosynthetic plant material per unit ground area, integrating structural information across the entire vertical canopy profile and reflecting overall vegetation density. Foliage Height Diversity (FHD) is calculated as the Shannon diversity index applied to the vertical distribution of plant area volume density across height bins, quantifying how evenly foliage is distributed throughout the canopy rather than concentrated within a single stratum (MACARTHUR and MACARTHUR, 1961; DUBAYAH et al., 2020).

Spatial matching of GEDI and FIA plots

To associate GEDI observations with FIA plots, a spatial buffering approach was implemented to account for intentional spatial perturbation applied to FIA plot coordinates for confidentiality purposes (fuzzing and swapping). A 500 m buffer was generated around each FIA plot location, and all GEDI footprints intersecting the buffer were extracted. To evaluate the sensitivity of this choice, GEDI structural metrics were additionally extracted at 250 m and 1,000 m buffer radii for a representative subsample of 800 plots. Model performance under 5-fold cross-validation varied only modestly across buffer sizes (species richness R^2 : 0.158–0.172; Shannon index R^2 : 0.106–0.154; Simpson index R^2 : 0.086–0.139 across the three radii; full results in Supplementary Table S2), indicating that biodiversity predictions were not strongly sensitive to buffer radius within this range and supporting the use of a 500 m buffer as a reasonable balance between capturing adequate GEDI coverage and minimizing over-smoothing of local structural variation. This footprint-averaging approach is also consistent with established practice for mitigating GEDI's discontinuous sampling geometry and the positional uncertainty introduced by FIA coordinate fuzzing (DUNCANSON et al., 2022; TORRESANI et al., 2023; MANDL et al., 2023).

For each plot, GEDI structural metrics were aggregated by calculating mean values across all intersecting footprints. This aggregation approach reduces noise associated with individual laser shots and has been widely applied in studies integrating GEDI with ground-based forest inventories (DUNCANSON et al., 2022; TORRESANI et al., 2023).

Climate variables and statistical modelling

To represent broad-scale environmental gradients, climate covariates were derived from the WorldClim Version 1 dataset at 30-arc-second spatial resolution (HUMANS et al., 2005). Four bioclimatic variables were selected based on ecological relevance to forest growth and species distribution: annual mean temperature (bio01), temperature seasonality (bio04), annual precipitation (bio12), and precipitation seasonality (bio15).

Climate variables were extracted at FIA plot locations using the Google Earth Engine platform, which enables large-scale geospatial processing and consistent spatial sampling (GORELICK et al., 2017). Statistical relationships between GEDI-derived structural metrics, climate variables, and biodiversity indices were evaluated using a combination of correlation analysis and predictive modelling. Pearson correlation coefficients were computed to assess bivariate associations between individual structural metrics and biodiversity indices.

Predictive models were developed using multiple regression and machine-learning approaches, including ordinary least squares linear regression, Ridge regression to address multicollinearity (HOERL and KENNARD, 1970), Random Forest regression (BREIMAN, 2001), and Gradient Boosting machines (FRIEDMAN, 2001). These methods were selected to capture both linear and non-linear relationships between forest structure and biodiversity. Model performance was evaluated using five-fold cross-validation, with predictive accuracy quantified using the coefficient of determination (R^2) and mean absolute error (MAE). Predictor variables were standardized for linear and penalized regression models, while tree-based methods were fit using unstandardized inputs. Random Forest models used 100 trees with a maximum tree depth of 10 (`n_estimators` = 100, `max_depth` = 10); Gradient Boosting models used 100 boosting stages with a maximum tree depth of 5 (`n_estimators` = 100, `max_depth` = 5); both used a fixed random seed (`random_state` = 42) for reproducibility. Ridge regression used a regularization strength of α = 1.0. All models were implemented in scikit-learn (PEDREGOSA et al., 2011). Plot elevation was additionally tested as a supplementary predictor alongside the seven core GEDI structural variables (Section 3.4). A full overview of all model configurations, dependent variables, predictors, sample sizes and validation approaches is provided in Supplementary Table S1.

To assess model transferability and reduce spatial autocorrelation bias, spatially structured cross-validation was implemented using latitudinal blocking, whereby models were trained on subsets of the data and evaluated on geographically distinct blocks (ROBERTS et al., 2017). Permutation-based variable importance was calculated for tree-based models to quantify the relative contribution of each predictor to model performance. This approach measures the reduction in predictive accuracy following random permutation of individual variables and is less biased than impurity-based importance metrics (ALTMANN et al., 2010). Specifically, plots were divided into five approximately contiguous latitudinal bands spanning the full study extent (32.68–35.94° N, n = 250; 35.95–39.20° N, n = 843; 39.21–42.47° N, n = 1,796; 42.47–45.73° N, n = 4,127; 45.74–49.00° N, n = 2,686), and models were trained on four bands and evaluated on the held-out fifth band in a leave-one-band-out scheme. This spatial cross-validation procedure was applied to all four candidate models (linear regression, Ridge regression, Random Forest and Gradient Boosting) to determine whether limited geographic transferability was spe-

Table 1. Summary statistics of forest biodiversity metrics across the study area

Metric	Mean	SD	Min	25th	Median	75th	Max
Species Richness	3.18	1.63	1.00	2.00	3.00	4.00	11.00
Shannon Index	0.71	0.46	0.00	0.34	0.71	1.06	2.10
Simpson Index	0.39	0.25	0.00	0.18	0.45	0.59	0.87

Table 2. Summary statistics of GEDI-derived structural metrics

Structural Metric	Description	Mean	SD	Min	Max
rh50 (m)	Median canopy height	4.50	3.65	-0.13	23.43
rh75 (m)	75th percentile height	7.77	4.81	0.36	33.36
rh98 (m)	98th percentile height	12.79	6.02	1.26	44.36
rh100 (m)	Maximum canopy height	13.81	6.21	1.68	46.30
Cover	Canopy cover fraction	0.45	0.22	0.02	0.86
PAI	Plant Area Index	1.74	1.09	0.04	4.36
FHD	Foliage Height Diversity	2.61	0.42	1.22	3.28

cific to any one algorithm or reflected a broader property of the structure–biodiversity relationship (full results in Supplementary Table S5; see Section 3.5).

All analyses were conducted using Python, primarily employing the pandas and numpy libraries for data processing and the scikit-learn framework for model development and evaluation (PEDREGOSA et al., 2011).

Results

Dataset characteristics

The analysis encompassed 9,692 forest inventory plots across three Pacific Northwest states, matched with space-borne LiDAR structural metrics and climate variables. The dataset exhibited complete coverage with no missing values for GEDI-derived structural metrics, indicating robust satellite data availability across the study region. Oregon contributed the largest proportion of plots (50.4%, $n = 4,883$), followed by Washington (25.4%, $n = 2,454$) and California (24.3%, $n = 2,355$). Plot elevations ranged from near sea level to subalpine zones, capturing substantial environmental heterogeneity across the Pacific Northwest's diverse forest ecosystems. Species richness was moderately right-skewed (skewness = 0.656), reflecting a majority of low-diversity plots alongside a smaller number of higher-richness outliers; the Shannon index was approximately symmetric (skewness = -0.024); and the Simpson index showed mild left skew (skewness = -0.368). Distributional shapes for all three metrics, including boxplots highlighting outlier structure, are shown in Supplementary Figure S1 (Table 1; Table 2).

The structural metrics exhibited substantial variability,

with maximum canopy heights ranging from 1.68 m in low-stature woodlands to 46.30 m in mature coniferous stands. Plant Area Index (PAI) and Foliage Height Diversity (FHD) demonstrated particularly wide ranges, providing sufficient variance for predictive modelling (Table 2).

Structure–Biodiversity Relationships

Correlation analysis revealed consistent positive associations between GEDI-derived structural metrics and all three biodiversity indicators (Supplementary Table S3; Fig. 2). Foliage Height Diversity (FHD) exhibited the strongest correlations with biodiversity, followed by canopy cover and Plant Area Index (PAI).

Notably, vertical structural complexity (FHD) and canopy density metrics (cover, PAI) consistently outperformed height-based metrics (RH50–RH100, GEDI relative height percentiles as defined in Section 2.4) as biodiversity predictors. This pattern was consistent across all three biodiversity indices, with correlation coefficients for FHD ranging from 0.383 to 0.427, compared to 0.291 to 0.355 for maximum canopy height (RH100). These findings suggest that forests with vertically complex, dense canopies support greater tree species diversity, regardless of maximum canopy height attained.

Predictive Model Performance

Four machine learning algorithms were evaluated for biodiversity prediction using 5-fold cross-validation (Table 3). Random Forest consistently outperformed other approaches across all biodiversity metrics, though improvements over linear models were modest (Table 3).

Random Forest achieved R^2 values of 0.222, 0.192,

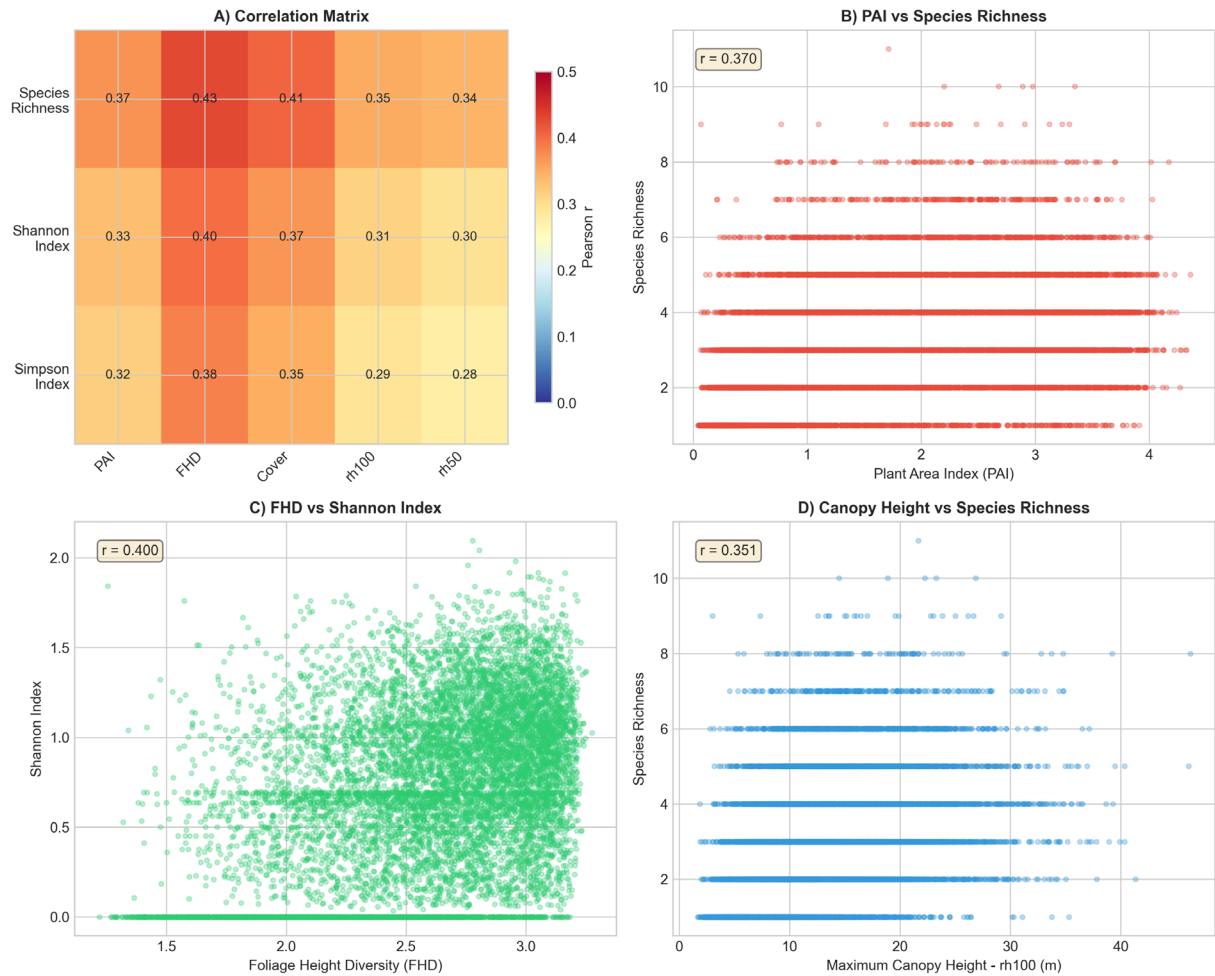


Fig. 2. Relationships between GEDI-derived structural metrics and forest biodiversity indices. (A) Correlation matrix showing Pearson coefficients between three biodiversity metrics (species richness, Shannon index, Simpson index) and five structural variables (PAI, FHD, cover, RH100, RH50). Warmer colors indicate stronger positive correlations. (B) Plant Area Index (PAI) versus species richness ($r = 0.370$). (C) Foliage Height Diversity (FHD) versus Shannon index ($r = 0.399$). (D) Maximum canopy height (RH100) versus species richness ($r = 0.350$). Vegetation density (PAI) and structural complexity (FHD) exhibited stronger correlations with biodiversity than canopy height metrics.

Detailed description of Fig. 2: Panel A is a correlation heatmap (Pearson r 0.28–0.43) between five GEDI metrics (PAI, FHD, cover, RH100, RH50) and three biodiversity indices; foliage height diversity shows the strongest correlations throughout. Panel B plots Plant Area Index against species richness ($r = 0.370$). Panel C plots foliage height diversity against Shannon index ($r = 0.400$). Panel D plots maximum canopy height against species richness ($r = 0.351$), a similar but weaker pattern than Panel B.

Table 3. Model comparison using 5-fold cross-validation (mean $\pm$ standard deviation)

Model	Species Richness – R^2	Species Richness – MAE	Shannon Index – R^2	Shannon Index – MAE	Simpson Index – R^2	Simpson Index – MAE
Linear Regression	0.210 $\pm$ 0.013	1.140 $\pm$ 0.012	0.179 $\pm$ 0.011	0.344 $\pm$ 0.004	0.162 $\pm$ 0.012	0.189 $\pm$ 0.003
Ridge Regression	0.210 $\pm$ 0.013	1.140 $\pm$ 0.012	0.180 $\pm$ 0.011	0.344 $\pm$ 0.004	0.162 $\pm$ 0.011	0.189 $\pm$ 0.003
Random Forest	0.222 $\pm$ 0.010	1.131 $\pm$ 0.007	0.192 $\pm$ 0.008	0.340 $\pm$ 0.005	0.172 $\pm$ 0.008	0.187 $\pm$ 0.003
Gradient Boosting	0.213 $\pm$ 0.016	1.139 $\pm$ 0.010	0.184 $\pm$ 0.010	0.342 $\pm$ 0.004	0.168 $\pm$ 0.010	0.187 $\pm$ 0.003

Note: Best performance per metric shown in bold. MAE = Mean Absolute Error.

and 0.172 for species richness, Shannon index, and Simpson index, respectively. The marginal improvement of ensemble methods over linear regression ($\Delta R^2 = 0.010$ – 0.012) indicates that structure-biodiversity relationships are predominantly linear, with limited non-linear interactions captured by tree-based algorithms (Fig. 3).

Feature Importance Analysis

Permutation importance analysis revealed that vegetation density (PAI) and vertical structural complexity (FHD) were the dominant predictors of forest biodiversity, substantially outweighing height-based metrics (Supplementary Table S4; Fig. 4).

Plant Area Index (PAI) emerged as the single most important predictor across all biodiversity metrics, with importance values 1.5 to 1.9 times greater than the second-ranked variable (FHD). Contrary to common

assumptions, maximum canopy height metrics (RH98, RH100) ranked among the least important predictors. This counterintuitive finding suggests that vegetation density and structural complexity rather than height alone drive the relationship between remotely sensed forest structure and tree species diversity. The dominance of PAI likely reflects its integration of both horizontal and vertical canopy characteristics, capturing overall forest productivity and niche availability. Plot elevation was also evaluated as a supplementary predictor alongside the seven core GEDI structural variables, yielding small but consistent gains in explained variance across all three biodiversity metrics ($\Delta R^2 = +0.010$ for species richness, $+0.009$ for the Shannon index, and $+0.007$ for the Simpson index). This modest improvement suggests that GEDI structural metrics already capture most elevation-related variance in forest structure, since elevation strongly shapes the vegetation structure

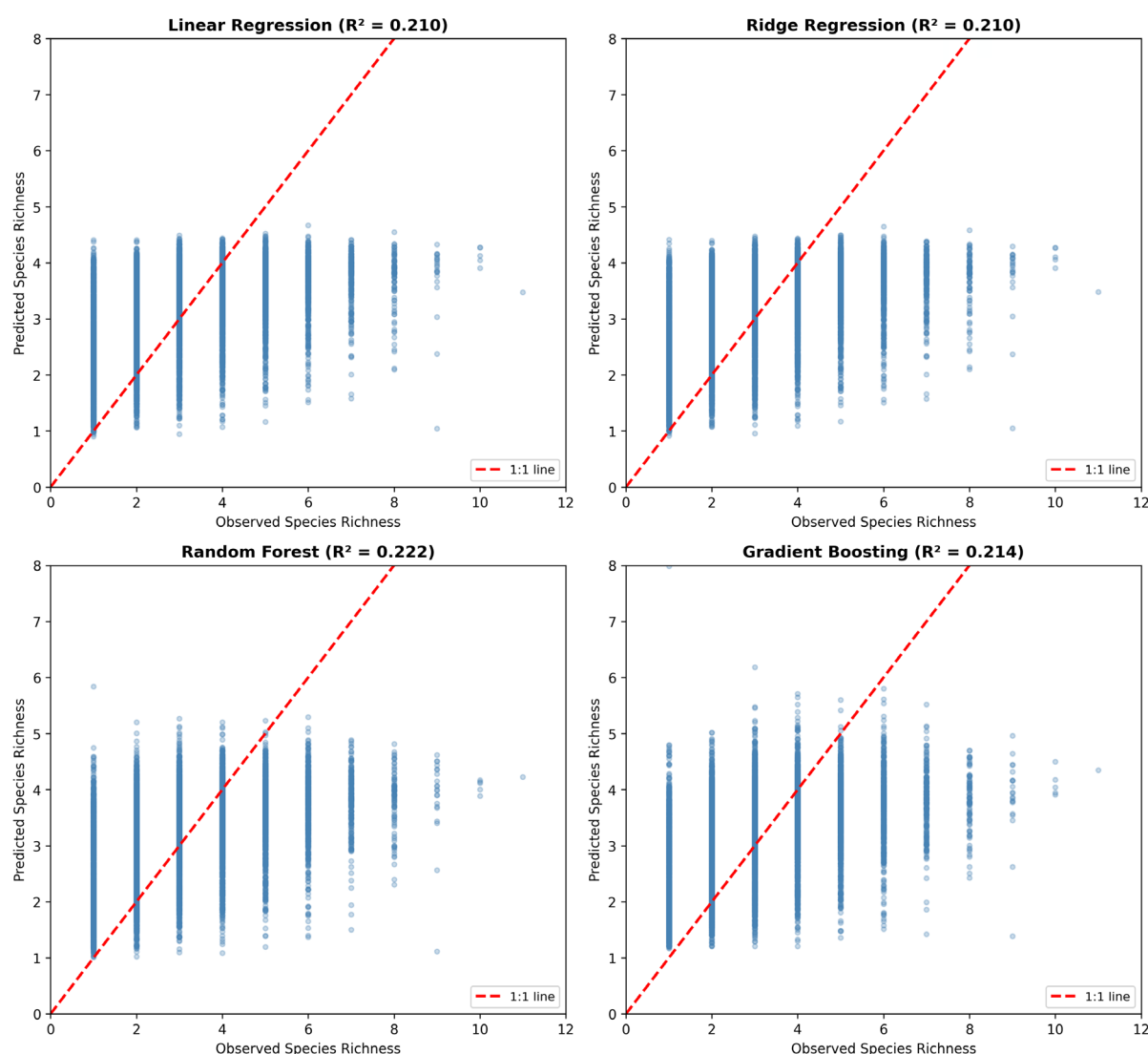


Fig. 3. Predicted versus observed species richness under 5-fold random cross-validation for all four models (Linear Regression, Ridge Regression, Random Forest, Gradient Boosting; $n = 9,692$ plots). The dashed red line represents the 1:1 relationship indicating perfect prediction. Random Forest achieved the highest R^2 (0.222) of the four models.

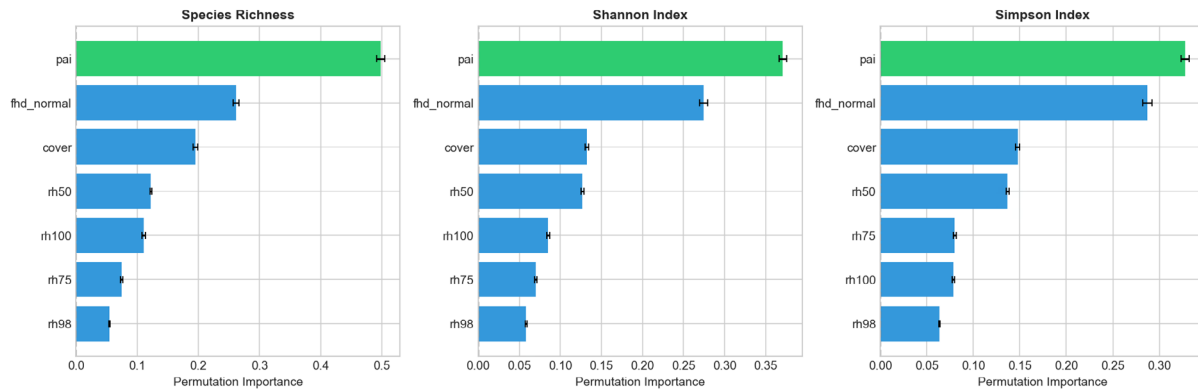


Fig. 4. Permutation importance of GEDI structural metrics for predicting forest biodiversity using Random Forest models. Importance values represent mean decrease in R^2 when each feature is randomly permuted. Results shown for (A) species richness, (B) Shannon index, and (C) Simpson index. Plant Area Index (PAI) and Foliage Height Diversity (FHD) consistently emerged as the dominant predictors (shown in red), substantially outperforming canopy height variables (RH50–RH100, shown in blue) across all biodiversity metrics.

and composition that GEDI measures directly; given the limited magnitude of this gain, elevation was not retained as a core predictor in the primary models. Other environmental covariates that may influence forest biodiversity, including soil properties, land-use history and distance to coast, were not available at consistent resolution across the full study extent and represent priorities for future work (Section 5.6).

Spatial Transferability Assessment

To evaluate model generalizability across geographic regions, latitude-based spatial cross-validation was applied (Supplementary Table S5; Fig. 5). This approach tests whether models trained in one region can accurately predict biodiversity in spatially distinct areas.

Model performance degraded substantially under spatial cross-validation, with R^2 values declining by 0.10–0.14 compared to random cross-validation. This degradation indicates that structure-biodiversity relationships vary across latitude bands, likely reflecting differences in regional species pools, climate regimes, and forest community composition. The high standard deviation in spatial

CV results (0.11–0.19) further indicates heterogeneous model performance across geographic blocks. Critically, this degradation pattern was not specific to Random Forest: spatial cross-validation reduced species richness R^2 from 0.210 to 0.048 for Linear Regression, from 0.210 to 0.052 for Ridge Regression, from 0.222 to 0.082 for Random Forest, and from 0.213 to 0.069 for Gradient Boosting, with comparable degradation observed for the Shannon and Simpson indices across all four models (Supplementary Table S5 and Fig. 5). Random Forest showed the smallest average degradation and the highest spatial cross-validation R^2 of the four models, suggesting marginally greater robustness to geographic extrapolation, but the broader pattern of substantial performance loss under spatial blocking was consistent across all four modelling approaches, indicating that limited geographic transferability is a general property of the structure–biodiversity relationship in this dataset rather than a limitation specific to any single algorithm.

Climate Covariates and Spatial Transferability

The effect of incorporating climate variables on spatial

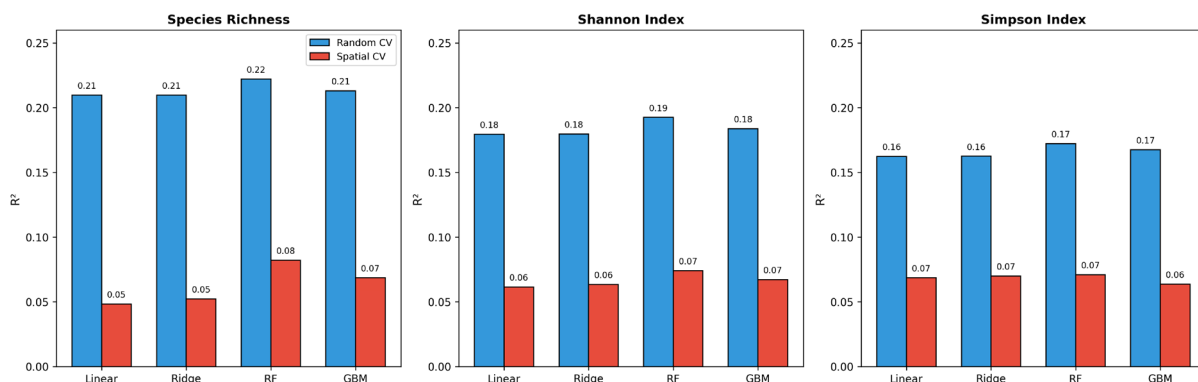


Fig. 5. Comparison of model performance under random versus spatial (latitude-block) cross-validation. Spatial cross-validation resulted in 59–63% performance degradation across all biodiversity metrics, indicating limited geographic transferability of structure-biodiversity relationships.

Table 4. Effect of climate covariates on spatial cross-validation performance

Biodiversity Metric	GEDI Only R^2	GEDI + Climate R^2	Improvement
Species Richness	0.079 ± 0.195	0.116 ± 0.211	+0.037
Shannon Index	0.070 ± 0.143	0.088 ± 0.175	+0.017
Simpson Index	0.070 ± 0.108	0.092 ± 0.124	+0.022

transferability was evaluated by comparing models that included WorldClim bioclimatic variables (annual temperature, temperature seasonality, annual precipitation, and precipitation seasonality) with structure-only models (Table 4).

Adding climate covariates improved spatial transferability by 0.02–0.04 R^2 units across all biodiversity metrics. While statistically meaningful, this improvement was insufficient to fully resolve the spatial transferability limitation. The persistent performance gap between random CV ($R^2 \approx 0.22$) and climate-augmented spatial CV ($R^2 \approx 0.12$) suggests that factors beyond climate such as biogeographic history, disturbance regimes, and regional species pools contribute to spatial heterogeneity in structure-biodiversity relationships.

Regional Model Performance

Given the limited spatial transferability of pooled models, region-specific calibration was evaluated by training and validating separate Random Forest models within each state using 5-fold cross-validation (Supplementary Table S6).

Regional calibration substantially improved model performance, with California achieving $R^2 = 0.402$ for species richness nearly double the spatial CV performance ($R^2 = 0.082$) and 25% higher than the pooled model ($R^2 = 0.321$). Oregon demonstrated similar patterns with $R^2 = 0.334$. However, Washington exhibited notably weaker performance ($R^2 = 0.166$), potentially reflecting lower structural heterogeneity or different species-structure relationships in its forest ecosystems (Fig. 6; Fig. 7). The regional models compared in this section (Supplementary Table S6, Fig. 6, Fig. 7) include climate covariates in addition to GEDI structural metrics, distinguishing them from the core structural-only models reported in Table 3 and Fig. 3.

The substantial improvement from regional calibration confirms that structure–biodiversity relationships are region-specific. California’s superior performance may reflect its diverse forest types (coastal redwoods, mixed-conifer, oak woodlands), providing greater variance in both structural and biodiversity gradients for model training.

Discussion

Structural complexity as a predictor of forest biodiversity

The results demonstrate that GEDI-derived measures of

forest structural complexity are significantly associated with plot-level tree diversity across the Pacific Northwest. Structural metrics capturing canopy heterogeneity showed consistent positive relationships with species richness and diversity indices, supporting ecological theory that links habitat complexity with niche availability and species coexistence. Forests with greater vertical stratification provide a wider range of microclimatic conditions, light environments and resource gradients, thereby facilitating higher species diversity (MACARTHUR and MACARTHUR, 1961; TEWS et al., 2004).

These findings align with extensive evidence from airborne LiDAR studies showing that structural heterogeneity is a strong proxy for biodiversity across forest ecosystems (SIMONSON et al., 2014; MOESLUND et al., 2019). Importantly, the present study extends these insights to spaceborne LiDAR, confirming that GEDI captures ecologically meaningful variation in canopy structure despite its coarser footprint size and sparser sampling relative to airborne systems.

Dominance of PAI and foliage height diversity

Among all predictors evaluated, Plant Area Index (PAI) and Foliage Height Diversity (FHD) consistently emerged as the most influential variables. Both metrics integrate information across the full vertical profile of the canopy rather than relying solely on upper-canopy height. PAI reflects overall vegetation density and productivity, while FHD quantifies the distribution of foliage across vertical strata, making them direct indicators of structural niche space.

Similar dominance of vertical-complexity metrics has been reported in previous airborne and GEDI-based studies, which found that canopy density and foliage distribution outperform maximum height metrics when predicting species diversity and ecosystem functioning (ATKINS et al., 2023; TORRESANI et al., 2023). This convergence across platforms and regions underscores the ecological relevance of three-dimensional structure over simple height-based measures and supports prioritizing PAI and FHD in biodiversity modelling frameworks.

Model performance and comparison with previous studies

This study is not the first to link GEDI-derived structure to biodiversity: TORRESANI et al. (2023) previously demonstrated that GEDI canopy height heterogeneity relates to tree diversity patterns, and MANDL et al. (2023) evaluated GEDI structural retrievals in topographically complex ter-

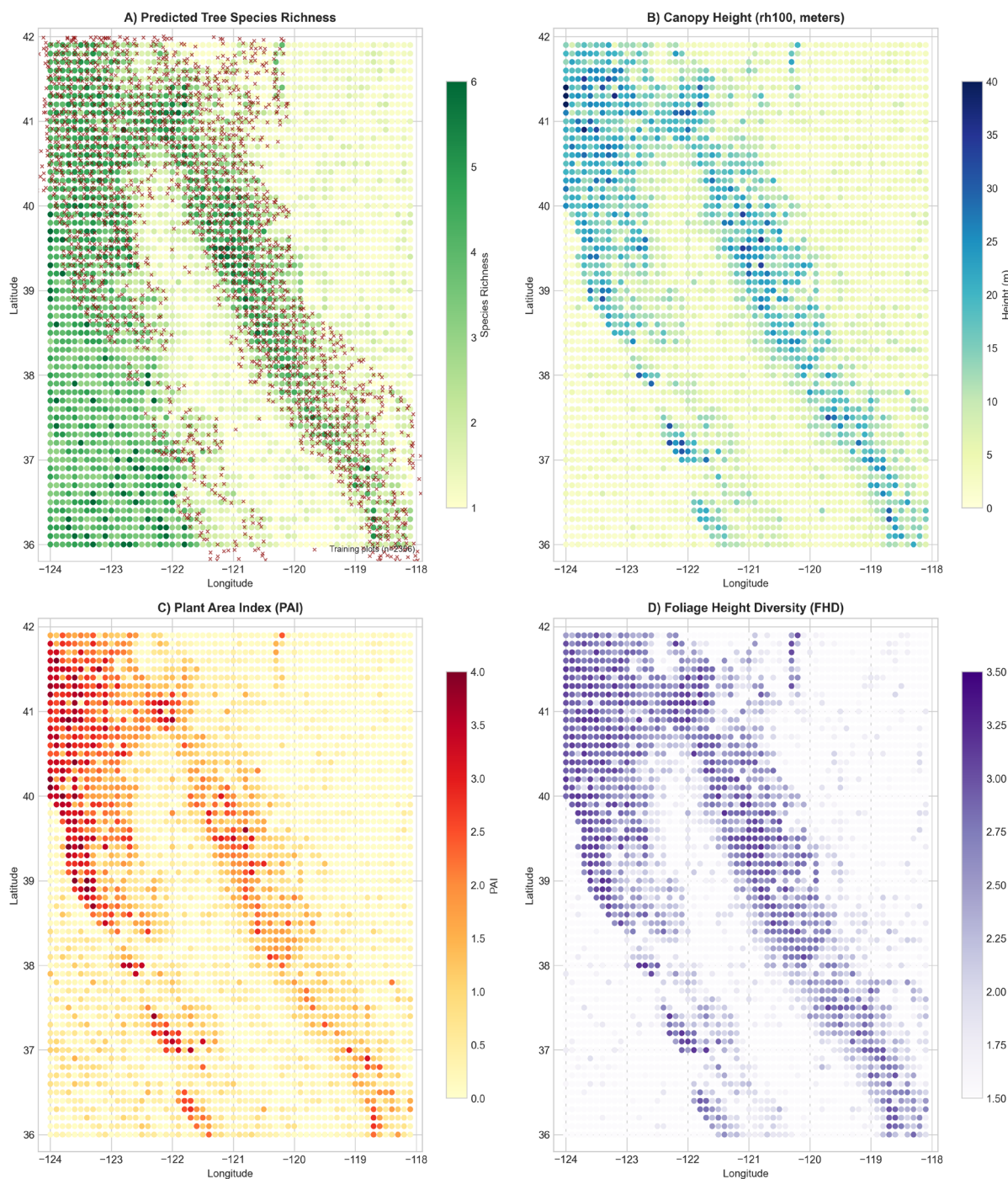


Fig. 6. Wall-to-wall biodiversity prediction for Northern California forests using a regionally calibrated Random Forest model ($R^2 = 0.40$). (A) Predicted tree species richness, with dark red markers indicating FIA training plot locations ($n = 2,355$). (B) GEDI-derived maximum canopy height (RH100) in meters. (C) Plant Area Index (PAI), the strongest predictor of biodiversity. (D) Foliage Height Diversity (FHD). Predictions were generated at 0.1° (~ 10 km) resolution across 3,600 grid points. Species richness predictions ranged from 1.02 to 6.12, with a mean of 2.53 species per plot. For each of the 3,600 grid points, GEDI structural metrics were extracted using the same 500 m buffering and averaging procedure applied to FIA training plots (Section 2.5), ensuring methodological consistency between training and prediction data; grid cells lacking valid GEDI coverage (e.g., non-forested areas) were excluded from the prediction surface.

Detailed description of Fig. 6: Panel A shows predicted species richness across Northern California, highest along the coastal and northern mountain ranges and declining toward the Central Valley and southeast. Panel B shows canopy height following the same coastal-to-interior gradient, tallest in the northwest. Panel C shows Plant Area Index, highest in the northwest coastal zone. Panel D shows foliage height diversity, highest in the same coastal and northern forested areas, fading toward the drier interior.

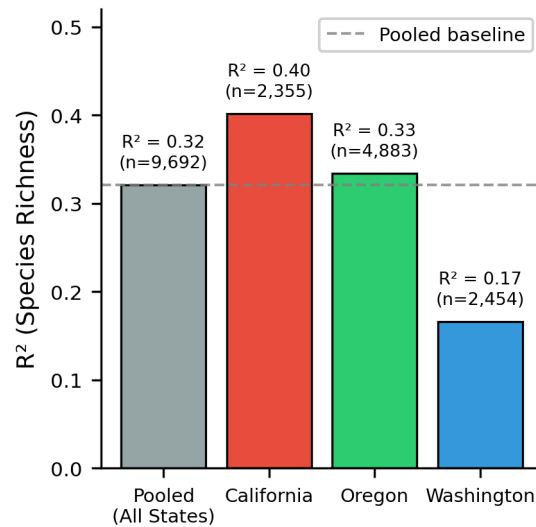


Fig. 7. Regional model performance demonstrating improved predictive accuracy with state-specific calibration. California achieved the highest performance ($R^2 = 0.40$), substantially exceeding the pooled baseline model ($R^2 = 0.32$, dashed line). Sample sizes shown for each region.

rain. The present study extends this line of work by directly predicting species richness and diversity indices across a much larger, ground-truthed inventory (9,692 FIA plots across three states) and by explicitly testing spatial transferability, which prior GEDI-biodiversity studies have not addressed in comparable depth. Direct numerical comparison with prior work is complicated by differing validation frameworks, but useful benchmarks exist from airborne LiDAR studies, which operate at finer spatial resolution than GEDI. A systematic review of airborne laser scanning biodiversity studies (TOIVONEN et al., 2023) reports R^2 values for structure–diversity models ranging from 0.30 (LEUTNER et al., 2012, species richness) to 0.89 (best-case Shannon index model); HERNÁNDEZ-STEFANONI et al. (2014) report $R^2 = 0.39$ – 0.49 for airborne LiDAR-based species richness prediction in tropical dry forest; and KAMOSKE et al. (2022), fusing LiDAR with hyperspectral imagery, report $R^2 = 0.46$ for taxonomic tree diversity. Against this range, the explained variance obtained here for the core structural model ($R^2 \approx 0.17$ – 0.22 under random cross-validation) sits at the lower end of what is typically reported for airborne LiDAR studies. This is plausibly explained by GEDI’s substantially coarser spatial resolution and sparser, discontinuous footprint sampling relative to airborne LiDAR, which achieves near-continuous canopy coverage; this trade-off between spaceborne GEDI’s global coverage and airborne LiDAR’s finer structural resolution is an important consideration for future work comparing the two platforms directly on the same ground plots.

Although machine-learning models slightly outperformed linear regression approaches, gains were relatively modest, suggesting that structure–diversity relationships within temperate forests are largely monotonic across the observed gradients. Comparable performance levels have been reported for both airborne LiDAR and GEDI-based biodiversity models, particularly when ground plots and lidar footprints are not perfectly co-lo-

cated (SIMONSON et al., 2014; DUNCANSON et al., 2022).

Spatial transferability and regional variability

A key finding of this study is the pronounced decline in predictive accuracy under spatially blocked cross-validation. This reduction indicates limited geographic transferability of structure–diversity models when applied beyond the regions used for training. Such behaviour is consistent with spatial ecology theory and with recent methodological work demonstrating that random cross-validation substantially overestimates model performance for spatial prediction tasks (ROBERTS et al., 2017).

Several mechanisms likely contribute to this pattern. Species pools differ across biogeographic regions, meaning similar structural configurations may support different levels of diversity. Additionally, disturbance regimes, land-use history and edaphic conditions vary across the Pacific Northwest and influence forest composition independently of canopy structure. These results emphasize that structure–biodiversity relationships are context dependent and that regionally stratified or hierarchical modelling approaches are preferable for large-area applications. This conclusion is strengthened by the finding that spatial cross-validation degradation was consistent across all four modelling approaches tested (Section 3.5), rather than being confined to Random Forest, indicating that the transferability limitation reflects genuine spatial heterogeneity in structure–biodiversity relationships rather than an idiosyncrasy of a particular algorithm.

Role of climate in improving model generalization

Incorporating coarse bioclimatic variables led to modest improvements in spatial transferability, indicating that climate partly mediates regional differences in structure–diversity relationships. Climate strongly influences species

pools, productivity and successional trajectories, and thus provides complementary explanatory power beyond canopy structure alone (HIJMANS *et al.*, 2005).

Nevertheless, the limited magnitude of improvement suggests that climate alone cannot fully resolve transferability challenges. This finding is consistent with recent multi-scale studies that highlight the need to integrate structural, climatic and historical drivers when modelling biodiversity across heterogeneous landscapes (DE CONTO *et al.*, 2024). Future approaches combining GEDI with imaging spectroscopy, SAR or disturbance history data may further enhance generalization.

Methodological limitations and uncertainties

Several limitations must be considered when interpreting these results. First, FIA plot coordinates are intentionally perturbed for confidentiality, introducing spatial uncertainty when matching plots to GEDI footprints. Although buffering approaches mitigate this issue, residual positional error remains and may attenuate observed relationships (COULSTON and REAMS, 2004). A sensitivity analysis using 250 m and 1,000 m buffer radii on a representative subsample ($n = 800$) showed that model performance varied only modestly across buffer sizes (Section 2.5; Supplementary Table S2), suggesting that results are not highly sensitive to this methodological choice within a reasonable range.

Second, scale mismatch between small FIA subplots and GEDI's ~25 m footprints introduces aggregation effects that can weaken correlations between structural metrics and localized species counts. Third, GEDI's sampling strategy results in discontinuous spatial coverage, limiting representativeness at fine spatial scales relative to airborne lidar. Recent GEDI-derived gridded products partially address this limitation and may improve future biodiversity mapping efforts (BURNS *et al.*, 2024).

Finally, biodiversity was represented solely by tree alpha diversity. While trees are foundational to forest ecosystems, extending analyses to functional traits, phylogenetic diversity and multi-taxon assemblages would provide a more comprehensive assessment of ecosystem biodiversity (SCHNEIDER *et al.*, 2017).

This study focused on structural and broad-scale climatic predictors; other environmental covariates that may influence forest biodiversity including soil properties, land-use history, and proximity to coastal or geological features were not incorporated due to inconsistent data availability across the full study extent, and represent a priority for future work. A supplementary test including plot elevation as an additional predictor produced small but consistent gains in explained variance (Section 3.4), suggesting further environmental covariates may offer modest additional benefit. Similarly, while this study relied on global and locally blocked (latitude-based) modelling frameworks, Geographically Weighted Regression (GWR) and other spatially varying-coefficient approaches represent a promising avenue for explicitly modelling spatial non-stationarity in structure–biodiversity relationships and should be explored in future research (ROBERTS *et al.*, 2017).

Implications for large-scale biodiversity monitoring

Despite these limitations, the results demonstrate the strong potential of spaceborne LiDAR for large-scale forest biodiversity assessment. GEDI provides globally consistent measurements of vertical forest structure, enabling comparative analyses across regions where airborne LiDAR and intensive field inventories are unavailable.

The findings suggest several practical recommendations: prioritizing structural complexity metrics such as PAI and FHD, applying spatially explicit validation strategies, and adopting regionally calibrated or hierarchical modelling frameworks. When combined with complementary sensors and gridded GEDI products, spaceborne LiDAR represents a critical component of next-generation biodiversity monitoring systems capable of supporting conservation planning and ecosystem assessment at continental scales.

Acknowledgements

The author has no acknowledgements to declare.

Competing interests

The author declares no competing interests (financial or non-financial).

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author contributions

The sole author was responsible for the entire study, including conceptualization, methodology, data collection/analysis, writing and editing.

Ethics approval and consent

Not applicable (No human or animal participants were involved in this study).

Data availability

The data generated and/or analyses during this study are available from the corresponding author upon reasonable request.

Supplementary material

Supplementary material for this article (Supplementary Tables S1–S6 and Supplementary Figure S1) is available

References

- ALTMANN, A., TOLOŞI, L., SANDER, O., LENGAUER, T., 2010. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26 (10): 1340–1347. <https://doi.org/10.1093/bioinformatics/btq134>
- ATKINS, J.W., COSTANZA, J., DAHLIN, K.M., DANNENBERG, M.P., ELMORE, A.J., FITZPATRICK, M.C., HAKKENBERG, C.R., HARDIMAN, B.S., KAMOSKE, A., LARUE, E.A., SILVA, C.A., STOVALL, A.E.L., TIELENS, E.K., 2023. Scale dependency of lidar-derived forest structural diversity. *Methods in Ecology and Evolution*, 14 (2): 708–723. <https://doi.org/10.1111/2041-210X.14040>
- BECHTOLD, W.A., PATTERSON, P.L., 2005. *The enhanced Forest Inventory and Analysis program — national sampling design and estimation procedures*. General Technical Report SRS-80. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. 85 p. <https://doi.org/10.2737/SRS-GTR-80>
- BREIMAN, L., 2001. Random forests. *Machine Learning*, 45 (1): 5–32. <https://doi.org/10.1023/A:1010933404324>
- BURNS, P., HAKKENBERG, C.R., GOETZ, S.J., 2024. Multi-resolution gridded maps of vegetation structure from GEDI. *Scientific Data*, 11: 881. <https://doi.org/10.1038/s41597-024-03668-4>
- CHAO, A., CHIU, C.-H., JOST, L., 2014. Unifying species diversity, phylogenetic diversity, functional diversity, and related similarity and differentiation measures. *Ecological Monographs*, 84 (1): 45–67. <https://doi.org/10.1146/annurev-ecolsys-120213-091540>
- COULSTON, J.W., REAMS, G.A., 2004. The effect of blurred plot coordinates on interpolating forest biomass. *Canadian Journal of Forest Research*, 36 (5): 801–807. <https://doi.org/10.1139/X05-265>
- DE CONTO, T., ARMSTON, J., DUBAYAH, R., 2024. Characterizing the structural complexity of the Earth's forests with spaceborne lidar. *Nature Communications*, 15: 8116. <https://doi.org/10.1038/s41467-024-52468-2>
- DHARGAY, S., LYELL, C.S., BROWN, T.P., INBAR, A., SHERIDAN, G.J., LANE, P.N.J., 2022. Performance of GEDI space-borne LiDAR for quantifying structural variation in the temperate forests of south-eastern Australia. *Remote Sensing*, 14 (15): 3615. <https://doi.org/10.3390/rs14153615>
- DHARGAY, S., LYELL, C.S., BROWN, T.P., INBAR, A., SHERIDAN, G.J., LANE, P.N.J., 2020. The Global Ecosystem Dynamics Investigation: high-resolution laser ranging of the Earth's forests and topography. *Science of Remote Sensing*, 1: 100002. <https://doi.org/10.1016/j.srs.2020.100002>
- DUNCANSON, L.I., ARMSTON, J., KELLNER, J.R., DUBAYAH, R., 2022. Aboveground biomass density models for NASA's Global Ecosystem Dynamics Investigation (GEDI) lidar mission. *Remote Sensing of Environment*, 270: 112845. <https://doi.org/10.1016/j.rse.2021.112845>
- FAO, UNEP, 2020. *The State of the World's Forests 2020: Forests, biodiversity and people*. Rome: FAO. 188 p. <https://doi.org/10.4060/ca8642en>
- FRIEDMAN, J.H., 2001. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29 (5): 1189–1232. <https://doi.org/10.1214/aos/1013203451>
- GORELICK, N., HANCHER, M., DIXON, M., ILYUSHCHENKO, S., THAU, D., MOORE, R., 2017. Google Earth Engine: planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202: 18–27. <https://doi.org/10.1016/j.rse.2017.06.031>
- HANCOCK, S., ARMSTON, J., HOFTON, M., SUN, X., TANG, H., DUNCANSON, L.I., KELLNER, J.R., DUBAYAH, R., 2019. The GEDI simulator: a large-footprint wave form lidar simulator for calibration and validation of spaceborne missions. *Earth and Space Science*, 6 (2): 294–310. <https://doi.org/10.1029/2018EA000506>
- HERNÁNDEZ-STEFANONI, J.L., DUPUY, J.M., JOHNSON, K.D., BIRDSEY, R., TUN-DZUL, F., PEDUZZI, A., CAAMAL-SOSA, J.P., SÁNCHEZ-SANTOS, G., LÓPEZ-MERLÍN, D., 2014. Improving species diversity and biomass estimates of tropical dry forests using airborne LiDAR. *Remote Sensing*, 6 (6): 4741–4763. <https://doi.org/10.3390/rs6064741>
- HIJMANS, R.J., CAMERON, S.E., PARRA, J.L., JONES, P.G., JARVIS, A., 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25 (15): 1965–1978. <https://doi.org/10.1002/joc.1276>
- HOERL, A.E., KENNARD, R.W., 1970. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12 (1): 55–67. <https://doi.org/10.1080/00401706.1970.10488634>
- KAMOSKE, A.G., DAHLIN, K.M., READ, Q.D., RECORD, S., STARK, S.C., SERBIN, S.P., ZARNETSKIE, P.L., DORNELAS, M., 2022. Towards mapping biodiversity from above: can fusing lidar and hyperspectral remote sensing predict taxonomic, functional, and phylogenetic tree diversity in temperate forests? *Global Ecology and Biogeography*, 31 (7): 1440–1460. <https://doi.org/10.1111/geb.13516>
- LEUTNER, B.F., REINEKING, B., MÜLLER, J., BACHMANN, M., BEIERKUHNLEIN, C., DECH, S., WEGMANN, M., 2012. Modelling forest α -diversity and floristic composition – on the added value of LiDAR plus hyperspectral remote sensing. *Remote Sensing*, 4 (9): 2818–2845. <https://doi.org/10.3390/rs4092818>
- MACARTHUR, R.H., MACARTHUR, J.W., 1961. On bird species diversity. *Ecology*, 42 (3): 594–598. <https://doi.org/10.2307/1932254>
- MAGURRAN, A.E., 2004. *Measuring biological diversity*. Oxford: Blackwell Publishing. 256 p.
- MANDL, L., STRITIH, A., SEIDL, R., GINZLER, C., SENF, C., 2023. Spaceborne LiDAR for characterizing forest structure across scales in the European Alps. *Remote Sensing in Ecology and Conservation*, 9 (5): 599–614. <https://doi.org/10.1002/rse2.330>

- MOESLUND, J.E., ZLINSZKY, A., EJRNÆS, R., BRUNBJERG, A.K., BØCHER, P.K., SVENNING, J.-C., NORMAND, S., 2019. Light detection and ranging explains diversity of plants, fungi, lichens, and bryophytes across multiple habitats and large geographic extent. *Ecological Applications*, 29 (5): e01907. <https://doi.org/10.1002/eap.1907>
- PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETENTHOFFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU, D., BRUCHER, M., PERROT, M., DUCHESNAY, E., 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830. <https://doi.org/10.5555/1953048.2078195>
- ROBERTS, D.R., BAHN, V., CIUTI, S., BOYCE, M.S., ELITH, J., GUILLERA-ARROITA, G., HAUENSTEIN, S., LAHOZ-MONFORT, J.J., SCHRÖDER, B., THUILLER, W., WARTON, D.I., WINTLE, B.A., HARTIG, F., DORMANN, C.F., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40 (8): 913–929. <https://doi.org/10.1111/ecog.02881>
- SCHNEIDER, F.D., MORSDOFF, F., SCHMID, B., PETCHEY, O. L., HUENI, A., SCHIMEL, D.S., SCHAEPMAN, M.E., 2017. Mapping functional diversity from remotely sensed morphological and physiological forest traits. *Nature Communications*, 8: 1441. <https://doi.org/10.1038/s41467-017-01530-3>
- SIMONSON, W.D., ALLEN, H.D., COOMES, D.A., 2014. Applications of airborne lidar for the assessment of animal species diversity. *Methods in Ecology and Evolution*, 5 (8): 719–728. <https://doi.org/10.1111/2041-210X.12219>
- STANKE, H., FINLEY, A.O., WEED, A.S., WALTERS, B.F., DOMKE, G.M., 2020. rFIA: an R package for estimation of forest attributes with the US Forest Inventory and Analysis database. *Environmental Modelling & Software*, 127: 104664. <https://doi.org/10.1016/j.envsoft.2020.104664>
- TEWS, J., BROSE, U., GRIMM, V., TIELBÖRGER, K., WICHMANN, M.C., SCHWAGER, M., JELTSCH, F., 2004. Animal species diversity driven by habitat heterogeneity/diversity: the importance of keystone structures. *Journal of Biogeography*, 31 (1): 79–92. <https://doi.org/10.1046/j.0305-0270.2003.00994.x>
- TOIVONEN, J., KANGAS, A., MALTAMO, M., KUKKONEN, M., PACKALEN, P., 2023. Assessing biodiversity using forest structure indicators based on airborne laser scanning data. *Forest Ecology and Management*, 546: 121376. <https://doi.org/10.1016/j.foreco.2023.121376>
- TORRESANI, M., ROCCHINI, D., SONNENSCHNEIN, R., ZEBISCH, M., HAUFFE, H.C., HEYM, M., PRETZSCH, H., TONON, G., 2023. LiDAR GEDI derived tree canopy height heterogeneity reveals patterns of biodiversity in forest ecosystems. *Ecological Informatics*, 73: 102082. <https://doi.org/10.1016/j.ecoinf.2023.102082>
- UNEP-WCMC, IUCN, 2020. *Protected Planet Report 2020*. Cambridge, UK and Gland, Switzerland: UNEP-WCMC and IUCN. <https://doi.org/10.34892/jg6t-xn70>
- USDA Forest Service, 2022. *Forest Inventory and Analysis National Core Field Guide*. <https://doi.org/10.2737/FIA-FG-2022>
- WOUDENBERG, S.W., CONKLING, B.L., O'CONNELL, B.M., LAPOINT, E.B., TURNER, J.A., WADDELL, K.L., 2010. *The Forest Inventory and Analysis Database: data-base description and users manual version 4.0 for Phase 2*. General Technical Report RMRS-GTR-245. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. <https://doi.org/10.2737/RMRS-GTR-245>

Received January 26, 2026

Accepted August 5, 2026