Article

Not peer-reviewed version

Quantifying Categorical Information Loss in Forest Compositional Mapping: Implications for the Accuracy of Forest Assessment in Lualaba Province (DR Congo)

Médard Mpanda Mukenza *, John Kikuni Tchowa, Felana Nantenaina Ramalason, Heritier Khoji Muteya, Jan Bogaert *, Yannick Useni Sikuzani, Jean-François Bastin

Posted Date: 23 April 2026

doi: 10.20944/preprints202604.1653.v1

Keywords: compositional mapping; subpixel heterogeneity; carbon stock estimation bias; continuous land-cover representation; tropical dry forests

 doi:10.20944/preprints202604.1653.v1

Article

Quantifying Categorical Information Loss in Forest Compositional Mapping: Implications for the Accuracy of Forest Assessment in Lualaba Province (DR Congo)

Médard Mpanda Mukenza [1,2,3]*, John Kikuni Tchowa [2,3], Felana Nantenaina Ramalason [2,4], Heritier Khoji Muteya [1], Jan Bogaert [3,*], Yannick Useni Sikuzani [1] and Jean-François Bastin [3]

[1] Ecology, Ecological Restoration and Landscape Unit, Faculty of Agronomic Sciences, University of Lubumbashi, Lubumbashi P.O. Box 1825, Democratic Republic of the Congo

[2] Department of Plant Science, Faculty of Agronomic Sciences, Université de Kolwezi, Kolwezi, Democratic Republic of the Congo

[3] Teaching and Research Center (TERRA), Gembloux Agro-Bio Tech—Université de Liège, 5030 Gembloux, Belgium

[4] Université d'Antananarivo - École Supérieure des Sciences Agronomiques (ESSA-forêts), Mention Foresterie et Environnement, BP. 175 Ankatso, Antananarivo, 101, Madagascar

* Correspondence: mpandamedard@gmail.com (M.M.M); j.bogaert@uliege.be (J.B)

Abstract

Forests of Lualaba Province (DR Congo) form a compositionally complex mosaic of dry dense forest, gallery forest, and Miombo woodland. Yet, categorical land-cover maps impose discrete boundaries on these inherently continuous vegetation gradients, systematically discarding subpixel compositional information critical for forest monitoring and carbon accounting. The magnitude of this information loss at the landscape scale, however, remains largely unquantified. In this study, we train a Multi-Output Neural Network (MONN) using Sentinel-2 spectral and textural predictors (2025) to estimate the proportional cover of three forest types across the province. Model performance is benchmarked against a normalised Random Forest (RF) using spatial block cross-validation. Categorical information loss is quantified pixel-wise using two complementary metrics, dominant class proportion and Shannon compositional entropy, alongside a derived interpretive quantity, categorical information loss. The MONN slightly outperformed RF (R^2 = 0.648 vs 0.630; RMSE = 0.224 vs 0.229), yet the results reveal a fundamentally heterogeneous landscape structure. The mean dominant-class proportion was only 56.2%, indicating that categorical maps discard, on average, 43.8% of compositional information per pixel. Only 7.9% of forested pixels exceeded the 75% dominance threshold, while Shannon entropy reached 74.1% of its theoretical maximum, indicating that forest types coexist in near-equal proportions across most pixels. This renders categorical attribution structurally inadequate for most of the forested landscape. Across 92.1% of forested pixels, no single forest type achieved clear dominance. These results show that compositional mixing is the dominant structural condition of the landscape, and that compositional mapping is essential for representing tropical forest structure in heterogeneous drylands. By formally quantifying categorical information loss at the landscape scale, this study shows that continuous compositional mapping converts this structural ambiguity into a spatially explicit ecological signal, with direct implications for monitoring vegetation dynamics and biodiversity, highlighting a structural source of error in carbon stock estimation in tropical dry forests.

Keywords: compositional mapping; subpixel heterogeneity; carbon stock estimation bias; continuous land-cover representation; tropical dry forests

 doi:10.20944/preprints202604.1653.v1

1. Introduction

Categorical land-cover maps systematically discard subpixel compositional information: by assigning each pixel to a single discrete class, they impose artificial boundaries on naturally continuous vegetation gradients and obscure the mixed compositions that dominate heterogeneous tropical landscapes. Although widely adopted in global operational products such as Dynamic World and the Copernicus Global Land Service [1,2], categorical approaches systematically fail in transitional landscapes where mixed pixels predominate. In heterogeneous tropical forest mosaics, a single Sentinel-2 pixel routinely encompasses two or more vegetation types in variable proportions. However, the magnitude of the resulting information loss at the landscape scale has never been formally quantified. This information loss has direct consequences for vegetation dynamics monitoring and carbon assessment: when forest-type proportions within a pixel are collapsed into a single label, type-specific biomass equations cannot be applied, compositional shifts preceding structural degradation remain undetected, and carbon stock uncertainty accumulates systematically across the landscape.

Lualaba Province, in south-eastern DR Congo, is a highly relevant case study for this issue. Situated at the biogeographic transition between the Congo Basin's dense forest block and the miombo woodlands of southern Africa (8–12°S, 23–27°E), the province features three structurally distinct forest types, namely dry dense forest, gallery forest, and miombo woodland, that coexist at fine spatial scales, producing a high proportion of mixed pixels at Sentinel-2 resolution. The socio-economic context further complicates this compositional diversity: increasing charcoal production and shifting agriculture are gradually degrading forest cover [3,4], while the recent development of the Lobito economic corridor is expected to accelerate land-use change across the region. Despite these pressures and the significant carbon stocks at risk, Lualaba remains notably under-represented in the remote sensing literature: no study has yet generated continuous compositional maps of its wood vegetation at operational spatial resolution, leaving forest managers and policymakers without the essential spatially explicit information needed to prioritise conservation actions [5].

Despite their operational simplicity, binary forest/non-forest maps impose arbitrary thresholds on continuous canopy-cover gradients, leading to systematic divergences in forest-area estimates across global products [6] and to underestimating degradation by failing to detect sub-threshold cover losses [7]. In dryland systems, this limitation is particularly consequential: Bastin et al. [8] showed that global forest extent in dryland biomes has been underestimated by at least 9% because the categorical approach fails to capture open-canopy tree cover below conventional forest thresholds. However, continuous total woodcover remains insufficient where forest types differ substantially in biomass, carbon stocks, and biodiversity value [9,10], as in Lualaba Province.

Distinguishing proportional cover for each formation is therefore ecologically necessary, not merely a methodological refinement. In heterogeneous Sahelian agropastoral landscapes, Abdou et al. [11] showed that Random Forest regression, which produces continuous fractional cover estimates, better captures within-pixel heterogeneity than categorical classification. Sentinel-2 yielded a slightly but significantly lower RMSE than Landsat 8 ($p < 0.05$). However, area differences between sensors remained small (0.03–3.94%). The theoretical foundations for subpixel composition estimation are well established [12], and the consequences of ignoring subpixel mixing for map uncertainty have been formally documented [13]

Continuous-field methods provide a more suitable framework for mapping landscapes with mixed compositions. While global products such as Vegetation Continuous Fields [14,15] mark significant progress, they do not capture the specific proportions of different tropical forest types at the landscape level [14]. Machine learning regression techniques, especially Random Forest (RF), have been successfully used to estimate continuous woodcover from Sentinel-2 data in African tropical ecosystems [16,17]. Nonetheless, applying separate RF models for each forest type and then adjusting their outputs overlooks interdependencies among coexisting forest formations. Multi-output neural networks (MONN) overcome this by predicting the proportional cover of all forest

 doi:10.20944/preprints202604.1653.v1

types simultaneously using a shared representation, with a softmax layer ensuring that the proportions sum to 1 [18,19].

Wood vegetation cover is a crucial indicator of ecosystem structure and function, supporting carbon storage, water management, biodiversity, and REDD+ reporting [8,20–21]. Two methodological precautions informed the design: stratified random sampling ensured balanced representation across the compositional gradient [22–23], and spatial block cross-validation ensured strict geographic separation of training and test data to prevent inflated accuracy estimates [24–25].

Against this background, this study pursues two objectives: (1) to evaluate whether a Multi-Output Neural Network, by encoding the compositional constraint directly via a softmax output layer, offers a systematic performance advantage over independent, normalised Random Forest models for continuous forest composition mapping in Lualaba Province; and (2) to formally quantify, at the landscape scale, how much categorical information is lost when continuous compositional outputs are reduced to discrete forest-type maps, and to identify which pixels and landscape zones incur the greatest information loss. By establishing the spatial distribution and magnitude of this information loss, the study provides a direct methodological foundation for reducing bias in monitoring vegetation dynamics and carbon stock assessment in this understudied miombo landscape. We hypothesise that the MONN will produce consistent, if modest, performance gains over RF, and that compositional mixing is the dominant structural condition of this landscape, rendering categorical attribution ecologically inadequate across most forested pixels.

2. Materials and Methods

2.1. Study Area

The Lualaba Province (Figure 1), located in the south-east of the Democratic Republic of Congo (approximately 120,000 km²), is situated on the Katangese plateau, a major ecological transition region between the humid forests of the Congo Basin and the drier savannahs of southern Africa. The climate is seasonal, with annual rainfall ranging from 1,200 to 1,600 mm and an average temperature of around 25°C [3,9]. Topographical and edaphic gradients (ferralsols, acrisols, arenosols) structure a complex landscape mosaic that combines miombo woodlands, dry, dense forests, and hydromorphic gallery forests. These formations are organised along continuous ecological and anthropogenic gradients, creating broad transition zones between forest and savannah, and forest and agriculture [10,26,27]. The contrasts in structure and phenology, particularly pronounced in the dry season, induce non-stationary relationships between spectral signatures and woodcover, complicating the application of universal global models [28]. This spatial heterogeneity justifies an ecological stratification and local calibration approach by formation, aiming to represent better the continuous variability of woodcover in transitional landscapes [22,29].

 doi:10.20944/preprints202604.1653.v1

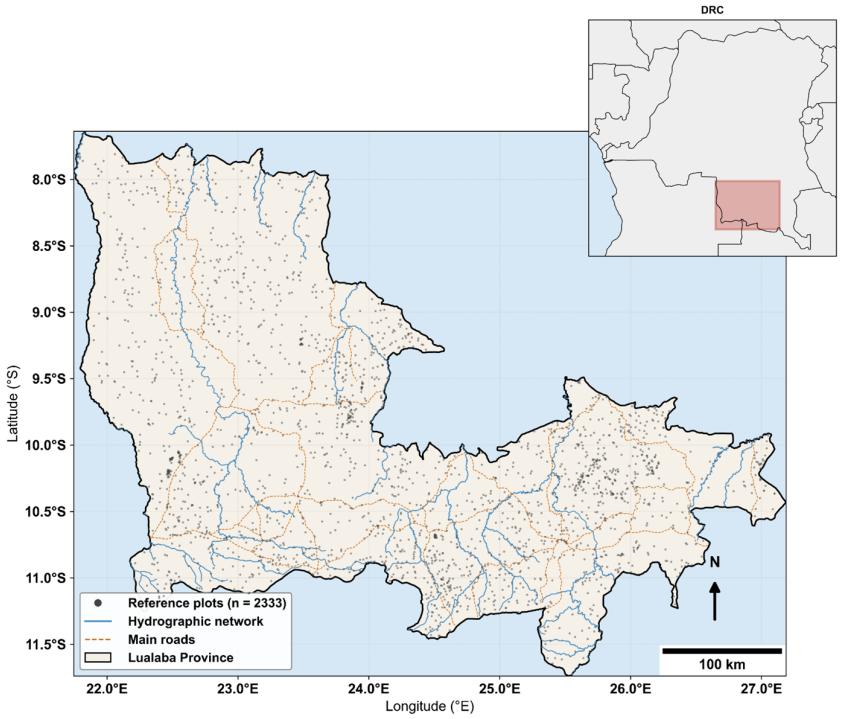

Figure 1. Location of the study area within Lualaba province (DRC), showing administrative boundaries, the main hydrographic network, and the spatial distribution of the 2,333 Collect Earth reference plots.

2.2. Data

2.2.1. Collect Earth Reference Data

Reference data were generated through systematic visual photo-interpretation of 2,333 virtual sample plots (0.25 ha; 50 × 50 m) using Collect Earth Online [29]. The plots were distributed across Lualaba Province using a stratified random sampling design proportional to the spatial extent of each forest type: Dry dense forest, Gallery forest, and Miombo woodland. This approach ensured balanced coverage of the compositional gradient and reduced overestimation bias [22,30]. The final sample comprised 758 plots (32.5%) classified as Dry dense forest, 794 plots (34.0%) as Gallery forest, and 781 plots (33.5%) as Miombo woodland. The frequency distribution of fractional cover values across all plots, showing the compositional structure of the training dataset by forest type, is presented in Appendix E (Figure E1). Within each plot, fractional cover per forest type was computed from systematic point interpretation of very high-resolution imagery, yielding continuous compositional estimates compatible with Sentinel-2 pixel-level modelling [30]: The complete training dataset, comprising spectral predictors and fractional cover values for all 2,333 reference plots, is provided in Table S1 (Supplementary Materials). The reliability of the photo-interpreted data was assessed using an independent field inventory comprising 60 plots (50 × 50 m; 20 per forest type) collected across Lualaba Province between June and August 2025. Strong agreement was observed between photo-interpreted and field-based estimates (mean R^2 = 0.754; mean RMSE = 1.30 percentage points), confirming the robustness of the interpretation protocol. Detailed per-class accuracy metrics and potential biases are provided in Appendix B.

2.2.2. Sentinel-2 Predictors

Satellite data were sourced from the Sentinel-2 MSI Level-2A surface reflectance collection (COPERNICUS/S2_SR_HARMONIZED) and processed using Google Earth Engine [31]. Images were limited to April–November 2025 and filtered to include observations with less than 10% cloud cover: A total of 1,618 images meeting these criteria were retained; full acquisition metadata are provided

 doi:10.20944/preprints202604.1653.v1

in Table S2 (Supplementary Materials). The wet season (December–March) was excluded due to persistent cloud interference and potentially reduced spectral separability among forest types [28]. A pixel-wise median composite was generated from all valid observations, and all 20 m bands were resampled to 10 m using bilinear interpolation for spatial consistency across predictors.

Predictor selection followed a structured, group-based evaluation aligned with configurations detailed in Appendix A (Table A1). Fourteen candidate predictor sets were predefined by combining five thematic groups: (i) Sentinel-2 spectral bands, (ii) vegetation indices, (iii) texture metrics from grey-level co-occurrence matrices (GLCM), (iv) topographic variables, and (v) engineered features. These sets ranged from simple single-family configurations (such as spectral or vegetation-only) to complex composite and engineered models (Table A1). Each set was evaluated through spatial block cross-validation (K = 5 folds; 2,000 m blocks) applied to all 2,333 reference plots, ensuring that predictor selection was based on held-out spatial test performance rather than training fit and therefore remained free of data leakage [23,32]. Performance metrics included R^2, RMSE, and MAE. Predictor sets were ranked based on mean cross-validated R^2, with the eco_medium configuration (13 variables) serving as a composite baseline. The optimal predictor set was identified at the inflexion point of the R^2 convergence curve, defined a priori as the point beyond which additional variables yield marginal improvements of less than $\Delta R^2 = 0.01$ [32,33]. This criterion identified the eco_texture configuration (15 variables) as optimal, achieving the highest cross-validated performance ($R^2 = 0.641$) with a slight but consistent improvement over the eco_medium baseline ($R^2 = 0.636$; $\Delta R^2 = +0.005$).

The chosen eco_texture predictor set comprised 15 variables: eight Sentinel-2 reflectance bands (B3, B4, B5, B6, B7, B8, B11, B12), five vegetation indices (NDVI, EVI, NDMI, NBR, CV-NDVI), and two GLCM texture metrics (NDVI contrast and NDVI entropy). Red-edge and SWIR bands captured chlorophyll content and canopy moisture, respectively; CV-NDVI added phenological contrast; and GLCM texture metrics resolved fine-scale canopy heterogeneity beyond the reach of spectral predictors alone.

2.2.3. Forest Mask

The forest mask serves as a spatial constraint for the MONN: because the softmax output layer distributes predicted cover exclusively among the three forest types, applying the model to non-forest pixels would yield compositionally invalid outputs. The mask therefore restricts predictions to pixels where the compositional framework is ecologically meaningful.

A binary forest mask was derived from a supervised Random Forest classification (500 trees) applied to a Sentinel-2 composite using 10 spectral bands (B2–B12) and NDVI, processed in Google Earth Engine [31]. Dry dense forest, Gallery forest, and Miombo woodland were merged into a single forest stratum [22]. The resulting mask achieved a Producer's Accuracy of 92.4% and a User's Accuracy of 90.9%, indicating reliable spatial delineation with no evidence of systematic bias in forest area estimation. The full accuracy assessment is provided in Appendix D.

2.3. Random Forest Normalised Approach

Three independent Random Forest regression models were trained, one per forest type (RF_Dry dense, RF_Gallery, RF_Miombo), using the ranger package in R [34]. Hyperparameters followed well-established regression defaults: 500 trees, mtry = $\lfloor\sqrt{P}\rfloor$ = 3 candidate variables per split, and a minimum node size of 5 [35]. Hyperparameter tuning was deliberately omitted for three reasons. First, the default mtry = $\lfloor\sqrt{P}\rfloor$ is consistently recommended for regression tasks and rarely yields substantial gains with moderate-sized datasets [36]. Second, nested cross-validation required for unbiased tuning within a spatial block CV framework would have introduced prohibitive computational overhead with five spatial folds and three simultaneous models. Third, and most importantly, fixing hyperparameters to identical defaults across both models ensures that observed performance differences between RF and MONN reflect architectural differences rather than tuning advantages, which is the comparison of scientific interest here. Exploratory tests varying mtry ($\lfloor\sqrt{P}\rfloor$,

 doi:10.20944/preprints202604.1653.v1

⌊P/3⌋, ⌊P/2⌋) and minimum node size (1, 5, 10) on a single spatial fold yielded no consistent R^2 improvement greater than 0.003, corroborating the decision to retain the default parameterisation.

A fundamental structural limitation of this independent RF approach is that each model is trained in isolation, without awareness of the other forest types. This violates the compositional constraint: raw RF outputs are unconstrained continuous values that may fall outside [0, 1] and whose sum across the three types need not equal one [37]. This is precisely the architectural weakness that the MONN addresses intrinsically through its softmax output layer. To make RF outputs compositionally valid for comparison, a pixel-wise post-hoc normalisation was applied: each raw prediction was clipped to [0, 1], then divided by the sum of the three clipped values, yielding proportions that sum to exactly one at each pixel. When the total of clipped predictions was zero, all proportions were set to zero. This normalisation step is a pragmatic correction rather than a principled compositional model.

2.4. Multi-Output Neural Network

The MONN was implemented as a fully connected feedforward neural network in Keras (Chollet, 2015; TensorFlow backend). The architecture comprised four hidden layers with progressively decreasing capacity (512–256–128–64 units), designed to compress the 15-variable spectral and textural input into a compact compositional representation. Each of the first three hidden layers was followed by batch normalisation [38], ReLU activation, and dropout (rates: 0.30, 0.25, and 0.20, respectively) to regularise training and prevent overfitting [39]. The fourth hidden layer used ReLU activation without dropout. The output layer comprised three neurons — one per forest type — activated by a softmax function. This directly encodes the compositional constraint within the network: outputs are strictly positive and sum to exactly 1, without any post hoc correction [40].

The model was trained using the Adam optimiser [41] with an initial learning rate of 0.001 and a batch size of 32. Training ran for a maximum of 300 epochs with two adaptive callbacks: early stopping (patience = 20 epochs, monitoring validation loss, restoring best weights) and learning rate reduction on plateau (factor = 0.5, patience = 8 epochs, minimum learning rate = 10^{-6}). A 15% validation split was held out from the training fold at each cross-validation iteration to monitor generalisation. The loss function was the Kullback–Leibler divergence between observed and predicted compositional distributions. This is more semantically appropriate than mean squared error for probability outputs: it measures the information cost of substituting the true composition with the predicted one and more heavily penalises errors on dominant forest types. Complete architecture specifications and training hyperparameters are provided in Appendix C.

2.5. Spatial Block Cross-Validation

Both models were evaluated within an identical spatial block cross-validation framework using the blockCV package [25], with K = 5 folds and a block size of 2,000 m. The block size was determined as the minimum distance that ensured spatial independence between the training and test sets, as verified by the absence of residual spatial autocorrelation across folds (Moran's I; [24]). Applying the same fold structure to both RF and MONN ensures that performance differences reflect architectural properties alone.

Predictive variability across folds was quantified as the inter-fold coefficient of variation (CV), computed pixel-by-pixel as the ratio of the standard deviation to the mean of the five fold-specific predicted cover values, expressed as a percentage. This yields a spatially explicit map of prediction stability under geographic resampling — a measure of how consistently the model predicts across different spatial partitions of the data, not a probabilistic uncertainty estimate in the Bayesian sense. Zones of high inter-fold CV indicate configurations sensitive to the spatial composition of the training set.

2.6. Performance Metrics

 doi:10.20944/preprints202604.1653.v1

Model performance was quantified using three complementary metrics averaged across the K = 5 spatial folds: the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE) [42]. To assess performance under compositionally demanding conditions, test pixels were stratified into three classes based on the maximum observed cover proportion of any single forest type: pure pixels (dominant proportion > 75%), moderately mixed pixels (50–75%), and highly mixed pixels (< 50%). This stratification captures the gradient from spectrally coherent to compositionally ambiguous pixels and allows evaluation of how both models respond to increasing subpixel heterogeneity.

Following cross-validation, both models were retrained on the complete 2,333 -plot dataset using the same architecture and hyperparameters to produce the final composition maps applied to the Lualaba Province raster stack.

2.7. Continuous Composition and Categorical Classification: Metrics and Framework

Categorical land-cover classification assigns each pixel to a single discrete class, thereby retaining only the dominant cover type and discarding information on minority cover types [43-45]. In contrast, continuous compositional mapping represents each pixel as a vector of proportional cover values that sum to unity, providing a more complete description of within-pixel heterogeneity [12-46]. In this study, we adopt this compositional framework to better capture the structural complexity of forest landscapes, where multiple vegetation types frequently coexist within a single pixel.

Each pixel is described by the proportional cover of three forest types—Dry dense forest, Gallery forest, and Miombo woodland—denoted wc_A, wc_B, and wc_C, respectively. These proportions take values in [0,1], where 1 corresponds to 100% cover, and are constrained to sum to unity:

$$\mathrm{wc\_A} + \mathrm{wc\_B} + \mathrm{wc\_C} = 1. \quad (1)$$

This constant-sum constraint defines the data as compositional in nature, implying that the components are not independent and must be interpreted within a relative framework [37].A minimum presence threshold of 5% was applied to avoid spurious low-proportion contributions inherent to soft classification outputs. The sensitivity of this threshold was evaluated (3–10%) to ensure that it does not drive the observed compositional patterns (Appendix F&G). More generally, let S denote the number of cover types in the compositional vector (S = 3 in this study), allowing the use of general expressions such as 1/S and ln(S) in the formulation of compositional metrics.

To characterise compositional structure and evaluate the implications of reducing continuous information to categorical labels, two complementary metrics are derived: (i) the dominant class proportion and (ii) Shannon compositional entropy. A third quantity, categorical information loss, is derived directly from the proportion of the dominant class and is included as an interpretive aid rather than as an independent metric. These quantities describe complementary aspects of compositional variation, and their relationships are examined in Appendix G.

-The dominant class proportion ($\boldsymbol{P}_{dom}$) is defined as:

$$P_{dom} = \max(\mathrm{wc\_A}, \mathrm{wc\_B}, \mathrm{wc\_C}). \quad (2)$$

This metric represents the only quantitative compositional information retained under categorical classification, ranging from 1/S = 1/3 (33.3%) to 1 (100%), corresponding to complete dominance by a single class. To identify pixels for which categorical attribution remains meaningful, a dominance threshold is required. Because no universal threshold exists, a sensitivity analysis was conducted using three candidate values (0.70, 0.75, and 0.80). For computational efficiency, the analysis was performed on a representative subset of pixels, whereas all summary statistics reported in the main results are derived from the full dataset. Lower thresholds (e.g., 0.50–0.60) were not considered, as they would classify substantially mixed pixels as dominant, thereby reducing interpretability. The sensitivity analysis showed that the proportion of pixels classified as dominant decreases from 12.2% at 0.70 to 4.6% at 0.80, a reduction of 7.6 percentage points. This rapid

 doi:10.20944/preprints202604.1653.v1

contraction confirms that lower thresholds are overly permissive, while thresholds above 0.80 become overly restrictive. A threshold of $P_{dom} > 0.75$ was therefore retained as a pragmatic compromise, ensuring that only clearly dominated pixels are identified while preserving sufficient spatial representation.

-The categorical information loss (E_{cat}) is derived directly from the dominant proportion as its complement:

$$E_{cat} = 1 - P_{dom} \quad (3)$$

This quantity represents the proportion of within-pixel cover assigned to non-dominant forest types, which is discarded when a categorical label is applied. It is explicitly defined as a linear transformation of P_{dom} (its arithmetic complement) and is therefore not an information-theoretic measure in the Shannon or Kullback–Leibler sense: it does not quantify distributional uncertainty or the complexity of the full cover vector. The distributional structure of within-pixel mixing — including how cover is redistributed among non-dominant types — is captured independently by Shannon compositional entropy H′ (Equation 5), which is sensitive to the full probability vector and constitutes the primary metric of compositional complexity in this framework. E_{cat} is retained as a directly interpretable summary of the magnitude of categorical loss at the pixel level, complementary to H′ rather than redundant with it. Together, these two quantities characterise complementary but not independent aspects of compositional structure: E_{cat} reflects the share of composition lost, while H′ reflects how that remaining composition is distributed. The sensitivity analyses presented in Appendices F and G confirm that conclusions are robust to threshold choice and that the metric combination is sufficient to characterise the compositional structure of the landscape without requiring additional divergence measures.

To characterise the internal distribution of cover among all classes, Shannon entropy is computed as [47]:

$$H = -\sum_{i=1}^{s} wc_i \,.\ln(wc_i) \quad (4)$$

with the convention $0 \cdot \ln(0) = 0$. Entropy ranges from 0 (complete dominance) to $\ln(S) = \ln(3) \approx 1.099$, corresponding to maximum compositional mixing. Entropy is normalised following Pielou [48]:

$$H' = \frac{H}{\ln(S)} \in [0,1] \quad (5)$$

Due to the constant-sum constraint of compositional data, H′ and P_{dom} are inherently related but capture different aspects of compositional structure. A strong negative relationship between P_{dom} and H′ is expected and was confirmed empirically (Appendix G). Principal component analysis applied to CLR-transformed proportions revealed that PC1, accounting for 85.0% of compositional variance, primarily contrasts Gallery forest against Dry dense forest and Miombo woodland (loadings: Gallery = +0.777, Dry dense = −0.605, Miombo = −0.172) and shows only weak correlation with P_{dom} ($\varrho = -0.21$). PC2 (15.0% of variance) contrasts Miombo woodland against the other two types (Miombo = +0.798) and is more strongly associated with dominance ($\varrho = 0.41$). These results suggest that the dominance gradient captured by P_{dom} is a secondary rather than primary axis of compositional variation: the dominant structural gradient reflects the redistribution of cover among forest types rather than an overall increase or decrease in mixing. Full details are provided in Appendix G.

3. Results

3.1. Model Performance: Random Forest versus Neural Network

 doi:10.20944/preprints202604.1653.v1

Spatial block cross-validation results are presented in Figure 2. The MONN consistently outperformed RF across all metrics and forest types, though the gains were modest in absolute terms (overall R^2: 0.648 vs 0.630; RMSE: 0.224 vs 0.229; $\Delta R^2 = +0.018$). This modest but consistent advantage reflects the MONN's architecture, which models the three forest types jointly rather than independently, particularly benefiting Dry Dense forest ($\Delta R^2 = +0.025$), where compositional constraints are most informative. Inter-fold variability was similar between models, indicating comparable spatial stability.

Performance stratified by pixel composition class (Table 1) reveals a sharp discontinuity. Both models perform well on pure ($R^2 \approx 0.65$) and moderately mixed pixels ($R^2 \approx 0.63$), with the MONN showing marginal gains in both classes ($\Delta R^2 = +0.014$ and +0.010, respectively). However, both models show near-zero explanatory power for highly mixed pixels (RF $R^2 = 0.008 \pm 0.004$; MONN $R^2 = 0.007 \pm 0.006$; dominant proportion < 50%), despite similar RMSE values across all classes. This sharp R^2 collapse, with no corresponding increase in RMSE, indicates that both models default to predicting near the mean in highly mixed pixels, capturing magnitude but not compositional variance. Critically, highly mixed pixels represent 43.0% of the forested area (Table 1), confirming that this is not a marginal failure mode but the dominant predictive condition.

Table 1. Model performance metrics across pixel composition classes. Values represent the mean ± standard deviation calculated over K = 5 spatial folds. Note: the percentage of forested area per class reflects the composition of the 2,333 reference plots (training sample), not the provincial raster; landscape-level proportions from MONN predictions are reported in Section 3.3 and Table 2. RMSE is expressed as a dimensionless proportion (cover values in [0, 1]). ΔR^2 = MONN R^2 − RF R^2.

Pixel class	% of forested area	RF R^2 (mean ± sd)	RF RMSE (mean ± sd)	MONN R^2 (mean ± sd)	MONN RMSE (mean ± sd)	ΔR^2
Pure (dominant > 75%)	3.0	0.649 ± 0.033	0.249 ± 0.009	0.664 ± 0.047	0.244 ± 0.015	+0.014
Moderately mixed (50–75%)	54.0	0.630 ± 0.065	0.199 ± 0.021	0.640 ± 0.049	0.198 ± 0.018	+0.010
Highly mixed (< 50%)	43.0	0.008 ± 0.004	0.196 ± 0.006	0.007 ± 0.006	0.194 ± 0.009	−0.001

 doi:10.20944/preprints202604.1653.v1

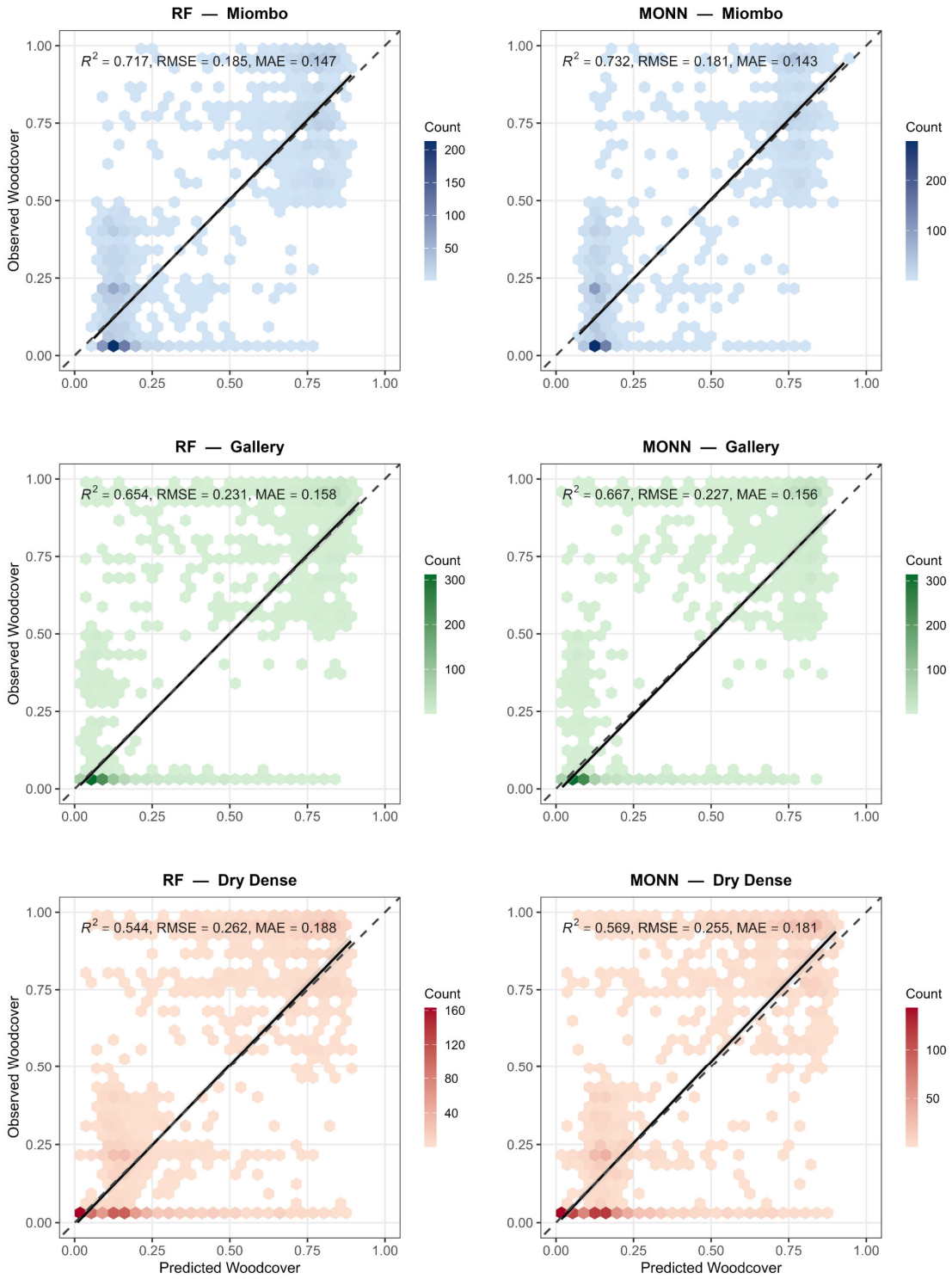

Figure 2. Observed versus predicted woodcover for RF (left) and MONN (right) across Miombo woodland, Gallery forest, and Dry Dense forest. Hexagonal bins represent point density from spatial block cross-validation (k = 5, block = 2,000 m). Solid line = OLS regression (shaded = 95% CI); dashed line = 1:1 reference. Metrics computed on held-out test folds only.

3.2. Spatial Distribution of Predicted Woodcover Composition

The MONN produced spatially consistent woodcover composition maps for the 23.5% of Lualaba Province classified as forested (Figure 3a–c). Miombo woodland dominated the landscape in

 doi:10.20944/preprints202604.1653.v1

mean cover, followed by Dry dense forest and Gallery forest, the latter showing high spatial variability consistent with its linear riparian distribution. The RGB ternary map (Figure 3d) reveals a fine-scale mosaic in which Dry dense forest and Miombo woodland co-occur across the central and southern provinces. In contrast, Gallery forest traces a sparse but persistent network along drainage systems, a spatial structure that no single categorical map can capture without substantial information loss.

Inter-fold predictive variability remained low province-wide (weighted CV = 4.7%), confirming the spatial stability of MONN outputs under geographic resampling. The Spearman correlation between CV and Shannon entropy was negative ($\varrho = -0.298$, $p < 0.001$): inter-fold variability concentrates in the rarest, most spectrally pure pixels, not in the mixed pixels that dominate the landscape. This pattern reflects two compounding mechanisms: the underrepresentation of pure-class pixels in the training dataset and the potentially higher intrinsic spectral variability within spectrally pure types, such as dense forests.

3.3. Compositional Mixing as the Dominant Landscape Condition: Implications for Categorical Mapping

Compositional mixing characterises the forested landscape of Lualaba Province. The mean dominant-class proportion was 56.2%, implying that categorical attribution discards, on average, 43.8% of within-pixel compositional information (Figure 4a–c). Only 7.9% of forested pixels exceeded the 75% purity threshold, while 28.8% fell below 50% — the level at which single-label assignment becomes ecologically arbitrary (Figure 4b). Shannon entropy averaged 74.1% of its theoretical maximum (Figure 4d), confirming that near-complete mixing, rather than dominance, is the prevailing structural state. Across 92.1% of pixels, no single forest type reached the purity level required for defensible categorical attribution.

Virtually no pixel was compositionally pure: only 0.3% of forested pixels showed a single dominant forest type exceeding the 5% presence threshold (Table 2, Panel A). Two-type mixtures accounted for 56.2% of the area and three-type mixtures for 43.5%, yielding 99.7% of pixels with meaningful contributions from at least two forest types. The partition between two- and three-type pixels varied with threshold choice (47.8% to 68.2% for two types across the 3–10% range; Table 2, Panel B), but the near-absence of mono-type pixels was robust across all thresholds tested (0.1% to 1.4%), confirming that the main conclusion does not depend on threshold selection.

Across all three forest types, mean dominant proportions ranged from 0.529 (Miombo woodland) to 0.609 (Dry dense forest), with entropy following the inverse pattern (Table A1). Miombo woodland, which covers 50.3% of the forested area, showed the highest entropy ($H' = 0.792$), reflecting extensive compositional overlap with adjacent types. A Mann-Whitney test confirmed that nominally pure pixels ($P_{dom} > 0.75$) exhibited significantly higher inter-fold predictive variability than mixed pixels (CV: 4.77% vs. 4.39%; $U = 173{,}526{,}728$, $p < 0.001$): prediction is least stable precisely where categorical attribution appears most justified.

All reported proportions should be interpreted in relation to the chosen presence (5%) and dominance (75%) thresholds.

 doi:10.20944/preprints202604.1653.v1

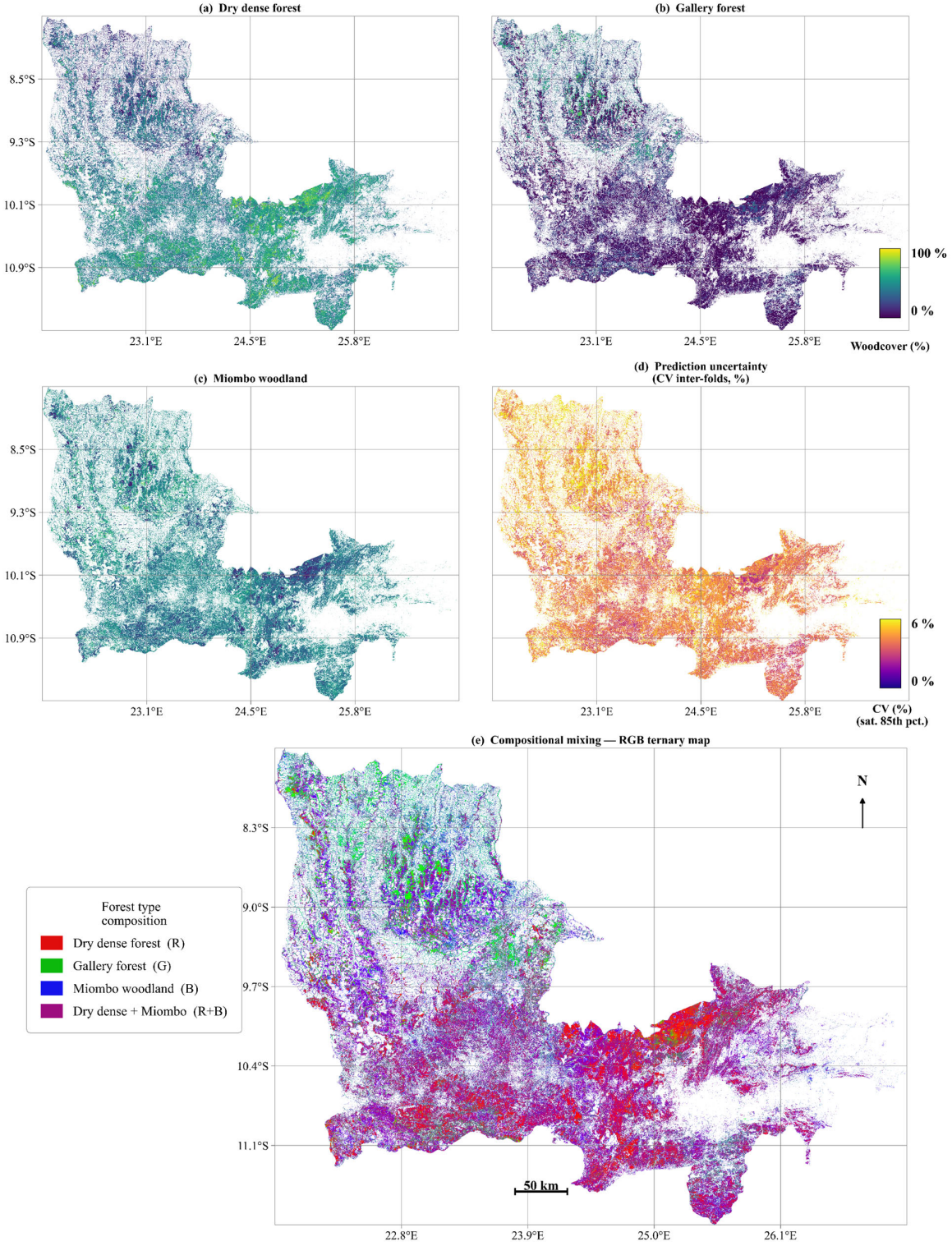

Figure 3. Spatial predictions of woodcover composition and prediction uncertainty across forested pixels of Lualaba Province, derived from the MONN. (a–c) Predicted proportional woodcover (%) per forest type; pixel values sum to 100% at each location. (e) RGB ternary composite. Pure Miombo woodland pixels (blue) are scarce; the predominance of magenta-purple hues indicates that Miombo mostly occurs as a mixture with Dry dense forest. (d) Composition-weighted coefficient of variation (CV = σ_folds / $\bar{p}$ × 100) across five spatially blocked cross-validation folds; colour scale saturated at the 85th percentile to reveal spatial gradients within the low-uncertainty range.

 doi:10.20944/preprints202604.1653.v1

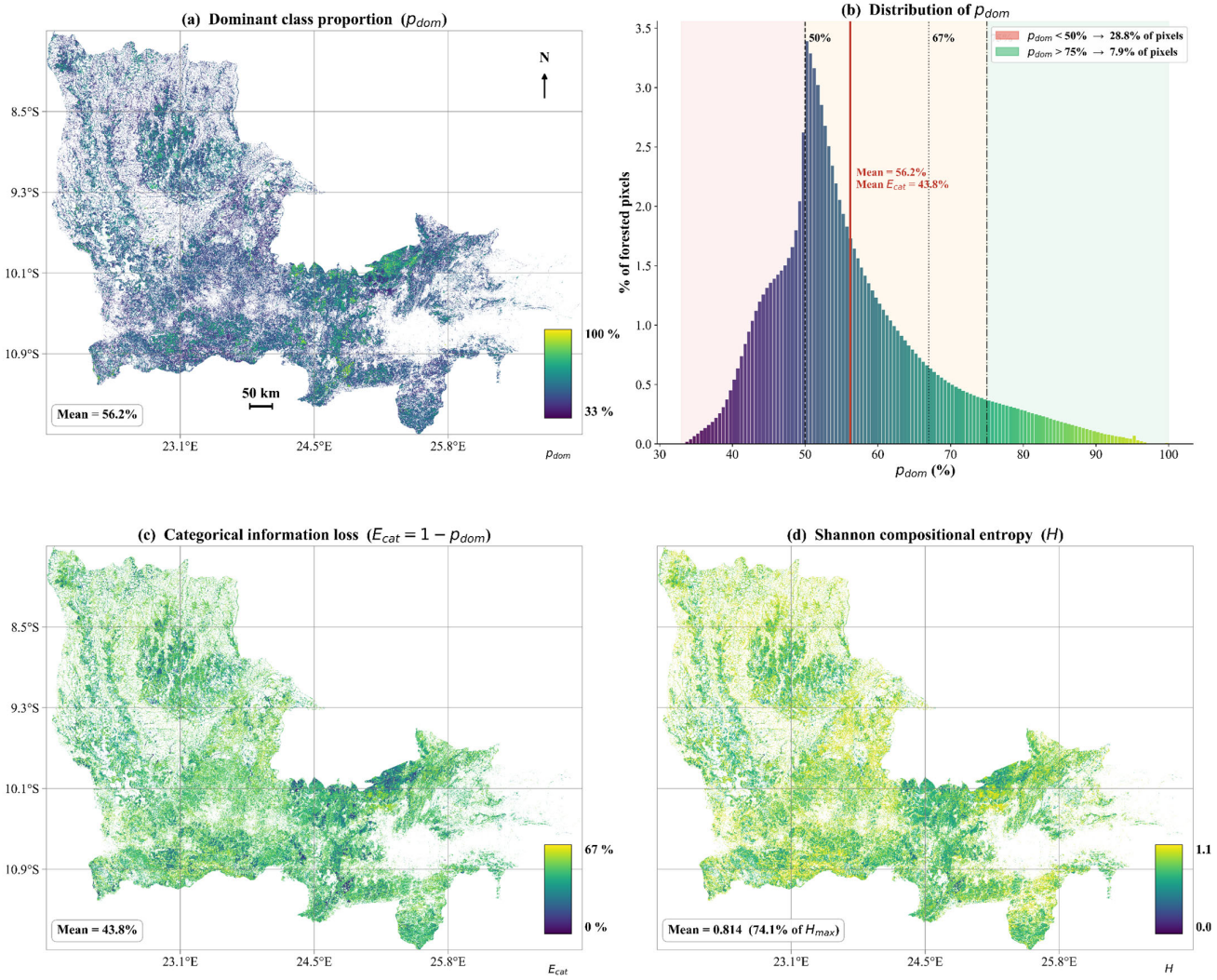

Figure 4. Figure 4. Spatial patterns of compositional dominance and within-pixel heterogeneity across forested pixels in Lualaba Province. (a) Dominant class proportion (P_{dom} = max (wc_A, wc_B, wc_C)), ranging from 1/3 (33%) to 1 (100%), where higher values indicate stronger dominance of a single forest type. (b) Frequency distribution of P_{dom} across all forested pixels; vertical lines mark reference thresholds at 50%, 67%, and 75%, and shaded regions highlight increasing levels of dominance. (c) Categorical residual (E_{cat} = 1 − P_{dom}), representing the proportion of non-dominant cover excluded under categorical classification; higher values indicate greater omission of within-pixel composition. (d) Shannon compositional entropy (H), ranging from 0 to ln(3) ≈ 1.10, reflecting the evenness of cover distribution among forest types, with higher values indicating more heterogeneous compositions. Together, these panels illustrate the structural dominance of compositional mixing and the resulting magnitude of categorical information loss.

Table 2. Pixel mixture structure across the forested area of Lualaba Province. Panel A reports the distribution of pixels by mixture class, defined by the number of forest types exceeding the 5% softmax presence threshold. Panel B reports the sensitivity of this classification to threshold choice.

A. Pixel mixture structure (presence threshold = 5%)	
Mixture class	**Forested pixels (%)**
Single forest type dominant	0.3
Two forest types present	56.2
Three forest types present	43.5
At least two types (total)	99.7

 doi:10.20944/preprints202604.1653.v1

B. Sensitivity to presence threshold			
Threshold	**1 type (%)**	**2 types (%)**	**3 types (%)**
3%	0.1	47.8	52.2
5% (retained)	0.3	56.2	43.5
10%	1.4	68.2	30.5

[2] Values in Panel A are percentages of total forested pixels. Bold row (Panel A: at least two types) is the complement of mono-type pixels. Bold row (Panel B: 5% retained) indicates the threshold used throughout the analysis.

4. Discussion

4.1. MONN Versus RF: consistent Gains In A Compositionally Constrained Regression Task

The MONN consistently outperformed RF across all forest types and folds (R^2: 0.648 vs 0.630; RMSE: 0.224 vs 0.229). The gain, while modest in absolute terms, is architecturally principled: the softmax output layer encodes the unit-sum constraint directly into the forward pass, enabling the model to learn the three forest-type proportions jointly rather than independently [18,40]. This contrasts with the RF approach, where compositional validity is imposed post hoc through normalization, a correction that cannot recover the interdependencies lost during training three separate models. The gain is most pronounced for Dry dense forest ($\Delta R^2 = +0.025$), the type with the sharpest ecological boundaries and thus the most informative compositional constraint. These results align with Lin et al. [49] and Schug et al. [50], who showed that jointly trained architectures outperform independent regression models for fractional cover mapping in heterogeneous landscapes.

The overall R^2 of 0.648 is commensurate with comparable frameworks operating under similar data constraints: Reiner et al. [16] reported R^2 values of 0.55–0.70 for continent-wide African tree cover mapping from a single-date composite, and Abdou et al. [11] obtained comparable accuracy for fractional woodcover in West African agropastoral landscapes using Sentinel-2 and Random Forest regression. Lluïsà et al. [51] achieved R^2 values up to 0.92, but under substantially more favourable conditions: synthetically mixed training data that span the full spectral mixing space by design, and dense multi-date time series providing phenological separability unavailable in a single-season composite. The performance gap relative to Lluïsà et al. [51] therefore reflects structural differences in data availability rather than a model shortcoming. The near-zero R^2 on highly mixed pixels — defined here as pixels where no single forest type exceeds 50% cover (43.0% of the training sample; Table 1), corresponding largely to the 43.5% of the provincial raster exhibiting three-type mixtures (Table 2) — reflects a spectral identifiability limit rather than a model flaw: when multiple forest types co-occur at sub-pixel scale with overlapping spectral signatures, no single-date optical model can reliably partition their contributions [52]. This behaviour is consistent with theoretical limits of linear spectral unmixing in optically similar vegetation types. The MONN is the only framework that makes this structural ambiguity spatially explicit rather than concealing it under a categorical label. For continuous vegetation dynamics monitoring, this transparency is precisely what allows gradual compositional shifts to be tracked over time [53].

4.2. Compositional Mixing as the Dominant Landscape Condition

The pixel-by-pixel quantification of categorical information loss constitutes the central contribution of this study. To our knowledge, no previous work has formally mapped E_{cat} at the provincial scale in a tropical African landscape; the convergence of three independent metrics and the the low province-wide CV (4.7%; Section 3.2) confirms that these patterns are structural rather than artefactual. The sensitivity analysis confirmed that the proportion of pixels classified as dominant decreases sharply from 12.2% at 0.70 to 4.6% at 0.80 (Appendix G), with 0.75 representing the inflection point of this contraction. Similarly, the near-absence of mono-type pixels (0.1–1.4%) was

 doi:10.20944/preprints202604.1653.v1

robust across the full presence threshold range tested (Appendix G), confirming that the compositional mixing conclusion does not depend on either threshold choice.

This result extends a widely acknowledged but rarely formalised limitation in remote sensing. Fernandes [46] identified sub-pixel mixing as the primary source of categorical error in heterogeneous landscapes; Herold et al. [54] showed that global products systematically disagree in transition zones; and Schug et al. [50] confirmed that subpixel fraction mapping overcomes these limitations where patch sizes fall below sensor resolution. Castle et al. [16] recently demonstrated that global forest dataset incongruence translates directly into high uncertainties for conservation and climate policy — an effect our results suggest is structurally unavoidable in landscapes where 92.1% of pixels cannot be unambiguously attributed to a single class. Global categorical map incongruence for forest cover in the DRC region has also been quantified at scale [55], and Sexton et al. [6] showed that even 30 m global products diverge substantially in transitional canopy zones of the kind that dominate Lualaba Province. Picard et al. [56] demonstrated that even the Congo Basin dense forest is a complex structural mosaic; our study extends this conclusion to the drier, miombo-dominated landscapes of southern DRC and, crucially, provides the first pixel-level quantification of the information cost of categorical attribution at full provincial resolution.

The per-type patterns are ecologically coherent. Miombo woodland dominates the landscape yet shows the lowest compositional purity (Table 2), consistent with its open-canopy structure and the co-occurrence of dry, dense forest species along edaphic gradients [9]. Dry dense forest shows the highest purity, reflecting its sharper ecological boundaries on better-drained mid-slope soils [10]. Gallery forest transitions seamlessly into both adjoining formations along hydrological gradients, boundaries that are invisible to categorical maps but spatially explicit in continuous compositional outputs. These ecological gradients match those identified by Picard et al. [56] as primary drivers of forest-type distribution in Central Africa.

The joint behaviour of P_{dom}, E_{cat}, and H clarifies the structure of compositional variation across the landscape. PCA applied to CLR-transformed proportions revealed that the primary axis of variation (PC1, 85.0% of variance) reflects the substitution of Gallery forest by Dry dense and Miombo woodland, and is largely independent of dominance level ($\rho = -0.21$ with P_{dom}). The dominance gradient is instead associated with PC2 (15.0% of variance; $\rho = 0.41$), confirming that overall mixing intensity is a secondary structural axis. Shannon entropy shows a strong negative relationship with P_{dom} ($r = -0.835$), consistent with the constant-sum constraint of compositional data. E_{cat}, as its linear complement, provides no independent structural information but facilitates direct interpretation of compositional loss. Together, these metrics define a hierarchical description of compositional structure where the dominant axis captures forest-type substitution, and the secondary axis captures mixing intensity — a distinction invisible to categorical maps.

4.3. Implications for Vegetation Dynamics Monitoring, Carbon Assessment, and Biodiversity Evaluation

Three limitations bound the scope of this study without undermining its central contribution. First, although the photo-interpretation protocol was validated against independent field plots ($R^2 = 0.754$; Section 2.2.1), the MONN compositional maps have not been validated against independent biomass measurements at the pixel level across the full provincial extent, a prerequisite for operational use in carbon accounting. Second, the softmax output layer structurally prevents compositionally pure pixels: because the exponential function never reaches zero, every forest type receives a strictly positive weight at every pixel, imposing a soft ceiling on P_{dom} and a soft floor on H′. The reported metrics (mean $P_{dom} = 56.2\%$; H′ = 74.1% of H_{max}) therefore reflect both genuine landscape mixing and this architectural constraint, meaning that compositional purity is slightly underestimated in the most spectrally homogeneous zones. Third, Shannon entropy conflates genuine ecological mixing with residual model uncertainty; disentangling these two sources would require a full Bayesian framework, for instance, MC-Dropout inference [57], to separate aleatoric from epistemic uncertainty. The inter-fold CV reported here captures only the latter and should not be interpreted as a proxy for the former.

 doi:10.20944/preprints202604.1653.v1

Two secondary limitations concern threshold choices and temporal scope. The 5% presence threshold used to define meaningful forest-type contributions lacks formal ecological justification. Its sensitivity was evaluated across 3–10%: the qualitative conclusion of near-universal mixing is robust (mono-type pixels: 0.1–1.4% across the range), but the exact partition between two- and three-type pixels varies substantially (Appendix G, Figure G2) and should be treated as an order-of-magnitude indicator rather than a precise ecological quantity. Similarly, the dominance threshold of $P_{dom} > 0.75$ is a pragmatic compromise, confirmed by sensitivity analysis (12.2% to 4.6% across 0.70–0.80; Appendix G, Figure G2), rather than a universal ecological criterion. Models were also trained on 2025 imagery; transferability across years remains untested given Lualaba's pronounced interannual rainfall variability [28], and spatial block cross-validation does not fully capture spectral non-stationarity across a climatically diverse area of 120,000 km² [58,59].

The carbon and biodiversity implications of this compositional mixing are quantifiable. Ryan et al. [9] reported aboveground carbon stocks of 26–34 Mg C ha^{-1} for miombo woodland in Mozambique and substantially higher values for structurally denser dry forest formations — a difference of the same order as the within-pixel compositional uncertainty documented here. When categorical maps assign a pixel to Miombo woodland although it contains 40% Dry dense forest, the applied biomass equation underestimates carbon by a systematic margin that cannot be corrected post hoc without knowing the true fractional composition. Réjou-Méchain et al. [60] showed that propagating such errors across landscape-scale biomass upscaling compounds uncertainties substantially, particularly in heterogeneous tropical systems. Similarly, Higginbottom and Symeonakis [61] demonstrated that vegetation index trajectories underestimate degradation severity in mixed-canopy landscapes because declining contributions from one formation are partially masked by stable or recovering contributions from another — precisely the dynamic that continuous compositional maps would make visible at the pixel level. Grantham et al. [62] quantified the extent of anthropogenic modification in tropical forests and noted that degradation in the miombo belt is chronically underestimated by global categorical products; our results provide a structural explanation for that underestimation. Finally, Wan et al. [63] demonstrated that satellite-derived canopy height and aboveground biomass maps at 10 m resolution are achievable in African tropical forests but require compositionally resolved input layers — an argument for continuous compositional outputs of the kind produced here as precursor data for REDD+ biomass estimation. These implications remain inferential and should be confirmed using biomass-calibrated datasets

Priority perspectives include multi-date extensions to detect gradual compositional changes along the Lobito corridor [3]; integrating GEDI spaceborne LiDAR to improve discrimination between structurally similar formations [64]; and CNN or Vision Transformer architectures that exploit Sentinel-2 spatial context [18,56]. None of these limitations alters the core result: the continuous compositional framework formally quantifies, at the provincial scale and full spatial resolution, what categorical maps appear structurally limited in highly mixed landscapes such as Lualaba, and this quantification is robust to the threshold and architectural choices examined here.

5. Conclusions

This study offers the first large-scale assessment of how much categorical information is lost in mapping tropical forest cover. In Lualaba Province, 92.1% of forest pixels do not meet the 75% purity threshold needed for clear categorical labels. On average, categorical classification omits 43.8% of the within-pixel composition information, and Shannon entropy reaches 74.1% of its maximum potential. These metrics indicate the landscape's structural condition. In Lualaba, compositional mixing is the norm, not an exception, and categorical maps may fail to adequately represent it. The Multi-Output Neural Network, which encodes the unit-sum constraint directly via a softmax layer, outperforms separate Random Forest models and uniquely makes this information loss spatially explicit and measurable. This has practical implications: categorical forest maps in Lualaba and similar miombo landscapes tend to bias carbon stock estimates by mixing different forest types with varying biomass, hide early signs of forest degradation by masking gradual compositional shifts, and reduce

 doi:10.20944/preprints202604.1653.v1

biodiversity transition zones to uniform labels. As land-use pressure along the Lobito corridor grows, the costs of this misrepresentation will increase. Continuous compositional mapping is a better operational approach for landscapes where mixing is the norm. Future work should focus on: first, multi-date mapping to detect gradual compositional changes as early indicators before they become detectable through categories; second, integrating GEDI LiDAR canopy height data to better distinguish similar formations at intermediate cover levels and support direct, type-specific biomass estimates aligned with REDD+ standards; and third, validating the methods with independent biomass and species surveys at the pixel level to develop a practical monitoring tool. This approach can be applied to other complex tropical dry forests where heterogeneity is high and categorical mapping is the only existing method.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/doi/s1, Table S1: Training dataset used for MONN and RF model development, comprising 2,333 reference plots with 15 Sentinel-2 spectral and textural predictors (B3, B4, B5, B6, B7, B8, B11, B12, NDVI, EVI, NDMI, NBR, CV-NDVI, NDVI contrast, NDVI entropy) and fractional cover values for three forest types (Dry dense forest, Gallery forest, Miombo woodland), with plot coordinates (x, y) in geographic decimal degrees; Table S2: Metadata of Sentinel-2 imagery acquired over Lualaba Province used for MONN model training and prediction, including acquisition date, platform, MGRS tile, relative orbit, cloud cover, and spatial coverage of the region of interest.

Author Contributions: Conceptualization, M.M.M., J.-F.B and Y.U.S.; methodology, M.M.M.; software, M.M.M.; validation, M.M.M., J.K.T. and H.K.M.; formal analysis, M.M.M.; investigation, M.M.M.; resources, J.B. (Jan Bogaert) and J.-F.B.; data curation, M.M.M. and F.N.R.; writing -- original draft preparation, M.M.M.; writing -- review and editing, J.-F.B, Y.U.S. and J.B. (Jan Bogaert); visualization, M.M.M.; supervision, J.-F.B and Y.U.S.; funding acquisition, J.-F.B, Y.U.S. and J.B. (Jan Bogaert). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by ARES (Académie de Recherche et d'Enseignement Supérieur) within the framework of the AMORSE project (2024–2026), Arborekol_Kolwezi, which supported fieldwork and data collection.

Data Availability Statement: The authors confirm that all data supporting the findings of this study are available within the article.

Acknowledgments: The authors would like to thank ARES (Academy of Research and Higher Education) for its support in the AMORSE project (2024-2026), Arborekol_Kolwezi. This project has enabled the financing and capacity building of sustainable forest management in Democratic Republic of the Congo.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A

This appendix presents the composition and cross-validated performance of the 14 predictor sets evaluated during feature selection. Table A1 lists the variable families included in each configuration; Table A2 reports the corresponding cross-validated R^2, RMSE, and MAE, and identifies the inflexion point at which adding further variables yields no meaningful performance gain.

Appendix A.1

Table A1. Composition of the 14 candidate predictor sets evaluated for continuous forest cover composition mapping. ★ = selected configuration.

Family	Predictor set	N Sentinel-2 bands	Spectral indices	Texture metrics	Topography

 doi:10.20944/preprints202604.1653.v1

Spectral	spectral_basic	3	B2, B3, B4	—	—	—
Spectral	spectral_swir	3	B8, B11, B12	—	—	—
Spectral	spectral_full	10	B2, B3, B4, B5, B6, B7, B8, B8A, B11, B12	—	—	—
Vegetation	veg_basic	3	—	NDVI, EVI, NDWI	—	—
Vegetation	veg_advanced	7	—	NDVI, EVI, NDMI, NBR, GNDVI, NDRE, SAVI	—	—
Texture	texture	3	—	—	NDVI-contrast, NDVI-entropy, NDVI-variance	—
Topography	topography	3	—	—	—	Elevation, Slope, Aspect
Composite	eco_small	6	B3, B4, B5, B8	NDVI, NDMI	—	—
Composite	eco_medium	13	B3, B4, B5, B6, B7, B8, B11, B12	NDVI, EVI, NDMI, NBR, CV-NDVI	—	—
Composite ★	**eco_texture**	15	B3, B4, B5, B6, B7, B8, B11, B12	NDVI, EVI, NDMI, NBR, CV-NDVI	NDVI-contrast, NDVI-entropy	—
Composite	eco_topography	15	B3, B4, B5, B6, B7, B8, B11, B12	NDVI, EVI, NDMI, NBR, CV-NDVI	—	Elevation, Slope
Complete	full_model	26	B2–B8A, B11, B12 (10 bands)	NDVI, EVI, NDWI, NDMI, NBR, GNDVI, NDRE, SAVI, BSI, CV-NDVI	NDVI-contrast, NDVI-entropy, NDVI-variance	Elevation, Slope, Aspect
Engineered	eco_engineered	21	B3–B8, B11, B12	NDVI, EVI, NDMI, NBR, CV-NDVI, Red/NIR, CIG, NDWI-norm, NIR/SWIR	NDVI×contrast	Elevation, Slope, Elev×Slope
Engineered	full_engineered	34	B2–B8A, B11, B12 (10 bands)	NDVI, EVI, NDWI, NDMI, NBR, GNDVI, NDRE, SAVI, BSI, CV-NDVI + engineered	NDVI-contrast, entropy, variance, NDVI×contrast	Elevation, Slope, Aspect, Elev×Slope

Table A2. Composition of the 14 candidate predictor sets evaluated for continuous forest cover composition mapping. ★ = selected configuration.

Predictor set	Family	N vars	R^2	RMSE	MAE	ΔR^2 vs eco_medium	Notes
spectral_basic	Spectral	3	0.502	0.266	0.196	-0.134	No gain over eco_medium
spectral_swir	Spectral	3	0.549	0.254	0.187	-0.087	No gain over eco_medium
spectral_full	Spectral	10	0.610	0.236	0.169	-0.026	No gain over eco_medium
veg_basic	Vegetation	3	0.595	0.239	0.172	-0.041	No gain over eco_medium

 doi:10.20944/preprints202604.1653.v1

veg_advanced	Vegetation	7	0.629	0.231	0.164	-0.007	No gain over eco_medium
texture	Texture	3	0.102	0.359	0.314	-0.534	Single-family baseline only
topography	Topography	3	0.103	0.359	0.309	-0.533	Single-family baseline only
eco_small	Composite	6	0.590	0.243	0.175	-0.046	No gain over eco_medium
eco_medium	Composite	13	0.636	0.226	0.161	(reference)	
eco_texture ★	**Composite ★**	**15**	**0.641**	**0.226**	**0.160**	**+0.005**	Selected — inflexion point
eco_topography	Composite	15	0.617	0.234	0.166	-0.019	No gain over eco_medium
full_model	Complete	26	0.627	0.231	0.160	-0.009	No gain over eco_medium
eco_engineered	Engineered	21	0.633	0.228	0.163	-0.003	
full_engineered	Engineered	34	0.625	0.232	0.161	-0.011	No gain over eco_medium

[A2] The inflexion point is directly identifiable in Table A2: eco_texture achieves $R^2 = 0.641$ compared to $R^2 = 0.636$ for eco_medium ($\Delta R^2 = +0.005$), while all subsequent sets with additional variables yield equal or lower R^2 ($\Delta R^2 \leq 0$).

Appendix B

This appendix reports per-type validation metrics comparing photo-interpreted fractional cover estimates (Collect Earth Online) against independent field measurements collected across Lualaba Province (June–August 2025). Agreement statistics (R^2, RMSE, MAE, bias) are provided separately for each forest type and averaged across all three types (Table B1).

Table A3. Per-type validation of Collect Earth Online photo-interpreted fractional cover estimates against independent field measurements. n = number of field plots per type (50 × 50 m; 20 per type). RMSE and MAE are expressed in percentage cover points (%). Bias = mean photo-interpreted − mean field-measured cover (%). p-value from two-tailed Pearson correlation test. All types (mean) row reports unweighted mean metrics across the three types.

Forest type	n plots	R^2	RMSE (%)	MAE (%)	Bias (%)	p-value	Interpretation
Dry dense forest	20	0.674	0.961	0.764	+0.004	8.59×10^{-6}	Strong agreement; negligible bias
Gallery forest	20	0.803	1.617	1.068	+0.056	8.71×10^{-8}	Highest R^2; slight positive bias consistent with under-representation in training data
Miombo woodland	20	0.786	1.313	1.057	+0.008	1.68×10^{-7}	Strong agreement; negligible bias
All types (mean)	**60**	**0.754**	**1.297**	**0.963**	**+0.023**	—	Overall strong agreement across the compositional gradient

Appendix C

This appendix provides the complete architecture and training hyperparameter specification of the Multi-Output Neural Network (MONN). All settings reported in Table C1 were held fixed across the five spatial cross-validation folds and for the final model trained on all 2,333 reference plots.

Table A4. Complete specification of the Multi-Output Neural Network (MONN) architecture and training hyperparameters. LR = learning rate; BN = batch normalisation; KL = Kullback–Leibler divergence.

Parameter	Value / Description
Framework	Keras (TensorFlow backend; R interface via keras package)
Architecture	Fully connected feedforward neural network (FCNN)

 doi:10.20944/preprints202604.1653.v1

Input layer	15 neurons (one per predictor variable)
Hidden layer 1	512 neurons — Batch Normalisation — ReLU — Dropout (0.30)
Hidden layer 2	256 neurons — Batch Normalisation — ReLU — Dropout (0.25)
Hidden layer 3	128 neurons — Batch Normalisation — ReLU — Dropout (0.20)
Hidden layer 4	64 neurons — ReLU activation (no dropout)
Output layer	3 neurons — Softmax activation (unit-sum constraint)
Loss function	Kullback–Leibler divergence (KL)
Optimiser	Adam ($\alpha_0 = 0.001$; $\beta1 = 0.9$; $\beta2 = 0.999$; $\varepsilon = 10^{-8}$)
Batch size	32
Maximum epochs	300
Early stopping	Monitor: val_loss; patience = 20 epochs; restore best weights
LR reduction	ReduceLROnPlateau: factor = 0.5; patience = 8; min LR = 10^{-6}
Validation split	15% of training fold (held out for early stopping only)
Input normalisation	Z-score standardisation (mean = 0, sd = 1) per fold, applied to training data; same parameters applied to test data
Random seed	42 (set in R, Python/TensorFlow backends)
Cross-validation	Spatial block CV: K = 5 folds; block size = 2,000 m; systematic fold assignment (blockCV package v2.1)
Final model	Retrained on full 2,333-plot dataset using identical architecture and hyperparameters

Appendix D

This appendix presents the accuracy assessment of the 8-class Random Forest classification used to derive the binary forest mask. Table D1 gives the full confusion matrix (n = 212 reference plots); Table D2 summarises producer's accuracy, user's accuracy, overall accuracy, and Kappa for both the 8-class classification and the derived binary forest/non-forest mask. Only the three natural forest classes (DDF, GF, MW) are merged into the forest stratum; Wooded savannah is classified as non-forest.

Table A5. Confusion matrix of the 8-class Sentinel-2 land-cover classification used to derive the forest mask (n = 212 reference plots). Diagonal cells = correctly classified plots. DDF = Dry dense forest; GF = Gallery forest; MW = Miombo woodland; WS = Wooded savannah; GL = Grassland/Wetland; Ag = Agriculture; Wa = Water; Bu = Built-up/Bare soil. PA = Producer's Accuracy; UA = User's Accuracy. OA = 80.7%; Kappa = 0.776.

Reference \ Predicted	DDF	GF	MW	WS	GL	Ag	Wa	Bu	Total	PA (%)
Dry dense (DDF)	**23**	1	7	1	0	0	0	0	**32**	71.9
Gallery (GF)	1	**34**	0	0	0	0	0	0	**35**	97.1
Miombo (MW)	3	0	**24**	3	0	0	0	0	**30**	80.0
Wooded sav. (WS)	0	0	3	**10**	5	4	0	0	**22**	45.5
Grassland (GL)	0	0	1	4	**21**	0	0	0	**26**	80.8
Agriculture (Ag)	0	0	0	6	0	**22**	0	1	**29**	75.9
Water (Wa)	0	0	0	0	0	0	**6**	0	**6**	100.0
Built-up (Bu)	0	0	0	0	0	1	0	**31**	**32**	96.9
Col. total	**27**	**35**	**35**	**24**	**26**	**27**	**6**	**32**	**212**	—
UA (%)	85.2	97.1	68.6	41.7	80.8	81.5	100.0	96.9	—	—

Table A6. Summary accuracy metrics for the 8-class classification and the derived binary forest mask. Binary mask accuracy is computed by merging the three natural forest classes (DDF + GF + MW) into the forest stratum, and all remaining classes (WS + GL + Ag + Wa + Bu) into the non-forest stratum. Wooded savannah (WS) is classified as non-forest because it represents an anthropogenically degraded wood formation structurally distinct from the three natural forest types targeted by this study.

 doi:10.20944/preprints202604.1653.v1

Metric	8-class classification	Binary forest mask	Interpretation
Overall Accuracy (OA)	80.7%	92.4%	Merging forest classes substantially improves accuracy
Kappa coefficient	0.776	_	Strong agreement across 8 classes
Forest PA	varies by type	92.4%	92.4% of reference forest pixels correctly identified
Forest UA	varies by type	90.9%	90.9% of predicted forest pixels are true forest
Lowest class PA	WS (45.5%)	_	Main confusion: Wooded savannah ↔ other classes
Highest class PA	Water (100%); GF (97.1%)	_	Spectrally distinct classes perfectly or near-perfectly classified

Appendix E

This appendix illustrates the compositional structure of the training dataset. Figure E1 shows the frequency distribution of fractional cover values (0–100%) across the 2,333 Collect Earth reference plots, displayed separately for each forest type and for all plots combined. The bimodal pattern at the tails reflects the stratified sampling design, which deliberately included both spectrally pure and compositionally mixed plots to span the full cover gradient.

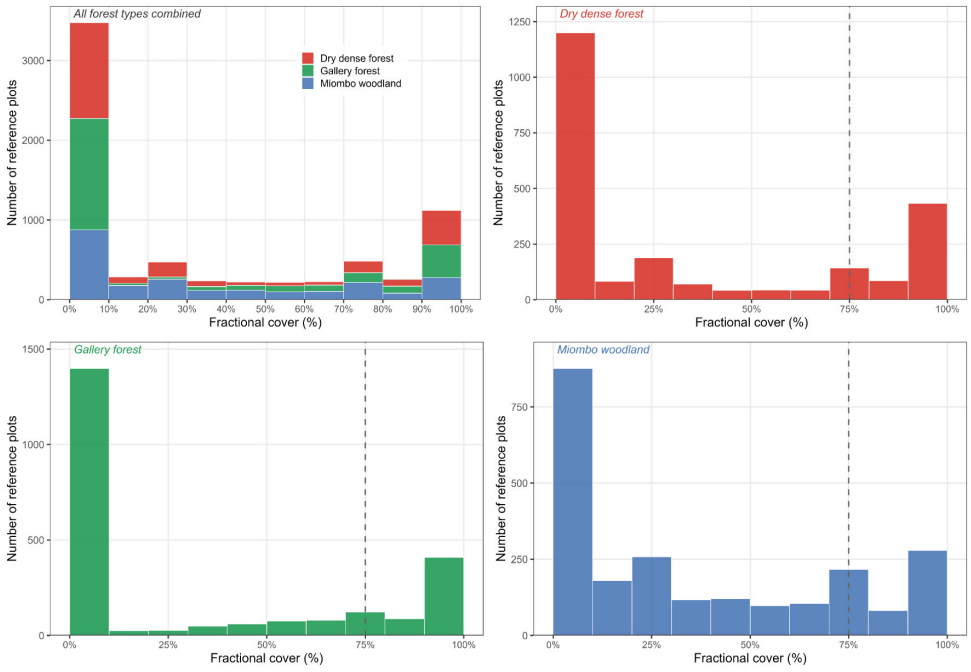

Figure A1. Frequency distribution of fractional cover values (0–100%) across the 2,333 Collect Earth reference plots, shown separately for each forest type (Dry dense forest, Gallery forest, Miombo woodland) and for all plots combined. Bin width = 10 percentage points. Dashed vertical line = 75% purity threshold. Legend for the figure: "The bimodal distribution (high frequency at 0–10% and 75–100%) reflects the stratified sampling design, in which plots were located within spectrally homogeneous zones where one forest type prevails locally. Intermediate cover values (10–75%) represent transition zones between formations and provide the compositional gradient signal used by the model. The gradient from 0% to 100% is represented across all three forest types, confirming that the model was trained on the full compositional range.

Appendix F

 doi:10.20944/preprints202604.1653.v1

This appendix reports per-type compositional statistics derived from MONN predictions at full provincial resolution (162,291,819 forested pixels). Table F1 disaggregates mean P_{dom}, E_{cat}, and H′ by dominant forest type, providing the per-type basis for the landscape-wide averages reported in Section 3.3.

Table A7. Per-type compositional statistics for pixels in which the indicated forest type is dominant (p^{dom} = max among all types). Metrics are computed at full spatial resolution from the MONN continuous predictions. The total row reports global values across all forested pixels. P_{dom} = dominant class proportion; E_{cat} = categorical information loss (1 − P_{dom}); H = normalised Shannon entropy (H / ln (3)).

Dominant forest type	% of forest	Mean P_{dom}	Mean E_{cat} (%)	Mean H
Dry dense forest	36.9	0.609	39.1	0.728
Gallery forest	12.8	0.555	44.5	0.843
Miombo woodland	50.3	0.529	47.1	0.870
All forested pixels	**100.0**	**0.562**	**43.8**	**0.814**

Appendix G

This appendix presents two complementary analyses supporting the compositional framework developed in Section 2.7. Figure G1 shows the PCA biplot of CLR-transformed cover proportions, demonstrating that the primary axis of compositional variation (PC1, 85.0% of variance) reflects forest-type substitution rather than overall mixing intensity, and that P_{dom} is associated with the secondary axis only ($\rho = 0.41$ with PC2). Figure G2 presents the sensitivity analyses for both the dominance threshold (0.70, 0.75, 0.80) and the softmax presence threshold (3%, 5%, 10%), confirming the robustness of the main compositional conclusions to these parameter choices, and the H′– P_{dom} relationship confirming that the two metrics capture complementary, non-redundant dimensions of compositional structure.

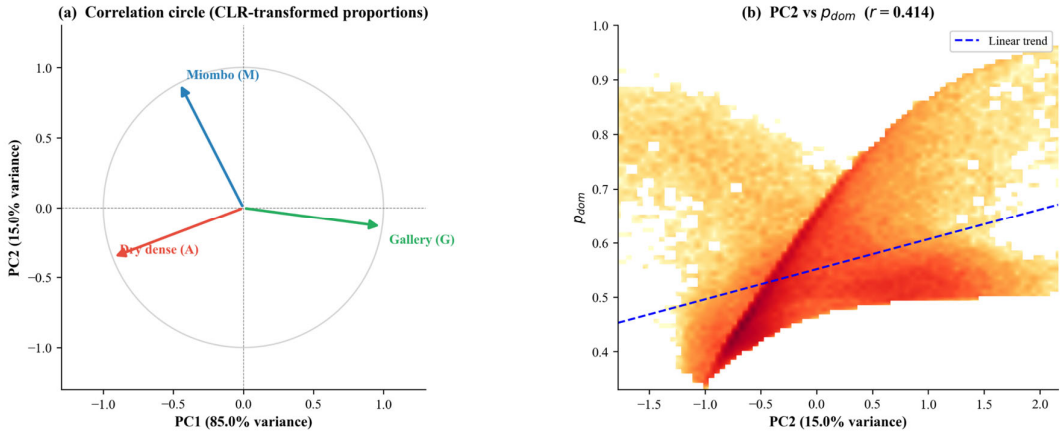

Figure A2. (a) Correlation circle showing the projection of the three forest types onto the first two principal components (PC1: 85.0% of variance; PC2: 15.0%). Gallery forest loads strongly and positively on PC1, while Dry dense forest loads negatively; Miombo woodland loads primarily on PC2. PC1 therefore reflects the compositional gradient opposing Gallery forest against Dry dense and Miombo woodland — a substitution axis structurally independent of overall dominance level (r(PC1, P_{dom}) = −0.21). (b) Density scatter of PC2 scores versus P_{dom} (r = 0.41), suggests that, in this dataset, dominance is not the primary axis of compositional variation. Together, these results indicate that P_{dom} captures an ecologically meaningful but structurally secondary dimension of forest composition: the degree of mixing, rather than the identity of the types involved in that mixing.

 doi:10.20944/preprints202604.1653.v1

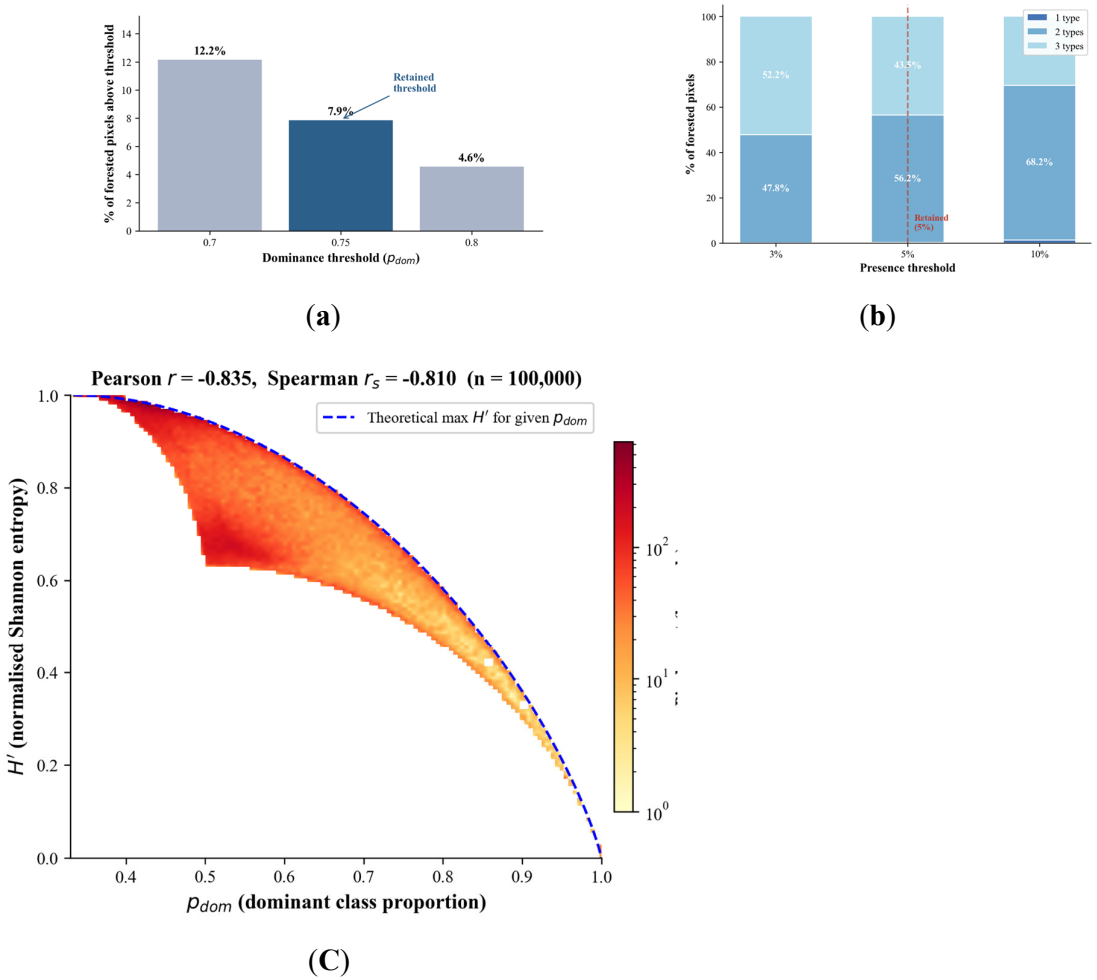

Figure A3. Panel a–c. Sensitivity analyses and compositional metric relationships.

Panel (a) shows the percentage of forested pixels exceeding each dominance threshold (0.70, 0.75, 0.80). The retained threshold ($P_{dom} > 0.75$) yields 7.9% of pixels classified as dominant. The sharp decrease from 12.2% at 0.70 to 4.6% at 0.80 (−7.6 percentage points) identifies 0.75 as an inflection point between overly permissive and overly restrictive classification. Panel (b) shows the distribution of pixels by mixture class (1, 2, or 3 forest types above the softmax presence threshold) across three tested values (3%, 5%, 10%). Two findings hold across the full range: mono-type pixels remain negligible (0.1–1.4%), confirming that near-universal compositional mixing is robust to threshold choice; the two- versus three-type partition varies substantially (47.8–68.2% for two types) and should be read as an order-of-magnitude indicator. Panel (c) shows the density scatter of P_{dom} versus normalised Shannon entropy H′ (n = 100,000; colour scale: pixel count, log). The dashed curve marks the theoretical maximum H′ for a given P_{dom}, assuming an equal distribution of residual cover. The strong negative relationship (Pearson r = 0.835; Spearman r_s = 0.810) confirms that both metrics describe the same primary gradient of overall mixing intensity from complementary angles. The pixel cloud lying below the theoretical maximum indicates that residual cover concentrates in a single minority type rather than being evenly distributed, consistent with Miombo woodland's spatial dominance across the province.

References

1. Buchhorn, M.; Lesiv, M.; Tsendbazar, N.-E.; Herold, M.; Bertels, L.; Smets, B. Copernicus Global Land Cover Layers — Collection 2. *Remote Sens.* **2020**, *12*, 1044. https://doi.org/10.3390/rs12061044

 doi:10.20944/preprints202604.1653.v1

2. Brown, C.F.; Brumby, S.P.; Guzder-Williams, B.; Birch, T.; Hyde, S.B.; Mazzariello, J.; Czerwinski, W.; Pasquarella, V.J.; Haertel, R.; Ilyushchenko, S. Dynamic World, near real-time global 10 m land use land cover mapping. *Sci. Data* **2022**, *9*, 251. https://doi.org/10.1038/s41597-022-01307-4.
3. Useni Sikuzani, Y.; Mpanda Mukenza, M.; Kikuni Tchowa, J.; Kabamb Kanyimb, D.; Malaisse, F.; Bogaert, J. Hierarchical analysis of Miombo woodland spatial dynamics in Lualaba Province (Democratic Republic of the Congo), 1990–2024: Integrating remote sensing and landscape ecology techniques. *Remote Sens*. **2024**, *16*, 3903. https://doi.org/10.3390/rs16203903
4. Tyukavina, A.; Hansen, M.C.; Potapov, P.; Parker, D.; Okpa, C.; Stehman, S.V.; Kommareddy, I.; Turubanova, S. Congo basin forest loss dominated by increasing smallholder clearing. Sci. Adv. 2018, 4, eaat2993. DOI:10.1126/sciadv.aat2993
5. Verhegghen, A.; Eva, H.; Ceccherini, G.; Achard, F.; Gond, V.; Gourlet-Fleury, S.; Cerutti, P.O. The Potential of Sentinel Satellites for Burnt Area Mapping and Monitoring in the Congo Basin Forests. *Remote Sens*. **2016**, *8*, 986. https://doi.org/10.3390/rs8120986.
6. Sexton, J.O.; Song, X.-P.; Feng, M.; Noojipady, P.; Anand, A.; Huang, C.; Kim, D.-H.; Collins, K.M.; Channan, S.; DiMiceli, C. Global, 30-m resolution continuous fields of tree cover. *Int. J. Digit. Earth* **2013**, *6*, 427–448.
7. Castle, S. E., Newton, P., Oldekop, J. A., Baylis, K., & Miller, D. C. (2026). Global forest dataset incongruence creates high uncertainties for conservation, climate, and development policy. One Earth, 9(2).
8. Bastin, J.-F.; Berrahmouni, N.; Grainger, A.; Maniatis, D.; Mollicone, D.; Moore, R.; Patriarca, C.; Picard, N.; Sparrow, B.; Abraham, E.M. The extent of forest in dryland biomes. Science 2017, 356, 635–638. DOI:10.1126/science.aam6527
9. Ryan, C.M.; Williams, M.; Grace, J. Above- and belowground carbon stocks in a miombo woodland landscape of Mozambique. Biotropica 2011, 43, 423–432.
10. White, F. The Vegetation of Africa; UNESCO: Paris, France, 1983.
11. Abdou Amadou, S.; Lawali, D.; Bastin, J.-F.; Bogaert, J.; Michez, A.; Meersmans, J. Effect of Spatial Resolution on Land Cover Mapping in an Agropastoral Area of Niger (Aguié and Mayahi) Using Sentinel-2 and Landsat 8 Imagery Within a Random Forest Regression Framework. *Remote Sens*. **2026**, *18*, 750. https://doi.org/10.3390/rs18050750
12. Atkinson, P.M. Resolution manipulation and sub-pixel mapping. In Remote Sensing Image Analysis: Including the Spatial Domain; De Jong, S.M., Van der Meer, F.D., Eds.; Springer: Dordrecht, The Netherlands, 2004; pp. 51-70.
13. Foody, G.M.; Atkinson, P.M. Uncertainty in Remote Sensing and GIS; Wiley: Chichester, UK, 2002. DOI:10.1002/0470035269
14. Hansen, M.C.; Potapov, P.V.; Moore, R.; Hancher, M.; Turubanova, S.A.; Tyukavina, A.; Thau, D.; Stehman, S.V.; Goetz, S.J.; Loveland, T.R. High-resolution global maps of 21st-century forest cover change. *Science* **2013**, *342*, 850–853. DOI:10.1126/science.1244693
15. Ramalason, F.N.; Rakotondrasoa, O.L.; Vander Linden, A.; Renard, G.; Randriamalala, J.R.; Vereecken, N.J.; Bastin, J.-F. The importance of the sampling design in mapping woody cover in arid ecosystems. *GISci. Remote Sens*. **2026**, *63*, 1. https://doi.org/10.1080/15481603.2026.2658305
16. Belgiu, M.; Drăguț, L. Random forest in remote sensing: A review of applications and future directions. ISPRS J. *Photogramm. Remote Sens*. **2016**, *114*, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011
17. Reiner, F.; Brandt, M.; Tong, X.; Skole, D.; Kariryaa, A.; Ciais, P.; Tucker, C.; Hiernaux, P.; Diouf, A.A.; Wigneron, J.P. More than one quarter of Africa's tree cover is found outside areas previously classified as forest. Nat. Commun. **2023**, *14*, 2258. https://doi.org/10.1038/s41467-023-37880-4
18. Reichstein, M.; Camps-Valls, G.; Stevens, B.; Jung, M.; Denzler, J.; Carvalhais, N.; Prabhat. Deep learning and process understanding for data-driven Earth system science. *Nature* **2019**, *566*, 195–204. https://doi.org/10.1038/s41586-019-0912-1
19. Kattenborn, T.; Leitloff, J.; Schiefer, F.; Hinz, S. Review on convolutional neural networks (CNN) in vegetation remote sensing. ISPRS J. Photogramm. Remote Sens. **2021**, *173*, 24–49. https://doi.org/10.1016/j.isprsjprs.2020.12.010

 doi:10.20944/preprints202604.1653.v1

20. Pan, Y.; Birdsey, R.A.; Fang, J.; Houghton, R.; Kauppi, P.E.; Kurz, W.A.; Phillips, O.L.; Shvidenko, A.; Lewis, S.L.; Canadell, J.G. A large and persistent carbon sink in the world's forests. *Science* **2011**, *333*, 988–993. DOI: 10.1126/science.1201609
21. Olander, L.P.; Gibbs, H.K.; Steininger, M.; Swenson, J.J.; Murray, B.C. Reference scenarios for deforestation and forest degradation in support of REDD. Environ. *Res. Lett*. **2008**, *3*, 025011. **DOI** 10.1088/1748-9326/3/2/025011
22. Olofsson, P.; Foody, G.M.; Herold, M.; Stehman, S.V.; Woodcock, C.E.; Wulder, M.A. Good practices for estimating area and assessing accuracy of land change. Remote *Sens. Environ*. **2014**, *148*, 42–57. https://doi.org/10.1016/j.rse.2014.02.015
23. Colditz, R.R. An evaluation of different training sample allocation schemes for discrete and continuous land cover classification using decision tree-based algorithms. *Remote Sens*. **2015**, *7*, 9655–9681. https://doi.org/10.3390/rs70809655
24. 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. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **2017**, *40*, 913–929. doi: 10.1111/ecog.02881
25. Valavi, R.; Elith, J.; Lahoz-Monfort, J.J.; Guillera-Arroita, G. blockCV: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods Ecol. Evol*. **2019**, *10*, 225–232. https://doi.org/10.1111/2041-210X.13107
26. FAO. Global Forest Resources Assessment 2020: Main Report; Food and Agriculture Organization of the United Nations: Rome, Italy, **2020**.
27. Timberlake, J.; Chidumayo, E.; Sawadogo, L. Distribution and characteristics of African dry forests and woodlands. In The Dry Forests and Woodlands of Africa: Managing for Products and Services; Chidumayo, E.N., Gumbo, D.J., Eds.; Earthscan: London, UK, **2010**; pp. 11–42.
28. Vrieling, A.; Meroni, M.; Darvishzadeh, R.; Skidmore, A.K.; Wang, T.; Zurita-Milla, R.; Oosterbeek, K.; O'Connor, B.; Paganini, M. Vegetation phenology from Sentinel-2 and field cameras for a Dutch barrier island. *Remote Sens. Environ*. **2018**, *215*, 517–529. https://doi.org/10.1016/j.rse.2018.03.014
29. Bey, A.; Sánchez-Paus Díaz, A.; Maniatis, D.; Marchi, G.; Mollicone, D.; Ricci, S.; Bastin, J.-F.; Moore, R.; Federici, S.; Rezende, M.; et al. Collect Earth: Land Use and Land Cover Assessment through Augmented Visual Interpretation. *Remote Sens*. **2016**, *8*, 807. https://doi.org/10.3390/rs8100807.
30. Colditz, R.R.; Schmidt, M.; Conrad, C.; Hansen, M.C.; Dech, S. Land cover classification with coarse spatial resolution data to derive continuous and discrete maps for complex regions. *Remote Sens. Environ*. **2011**, *115*, 3264–3275. https://doi.org/10.1016/j.rse.2011.07.010
31. Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ*. **2017**, *202*, 18–27. https://doi.org/10.1016/j.rse.2017.06.031
32. Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed.; Springer: New York, NY, USA, **2009**.
33. Georganos, S.; Grippa, T.; Vanhuysse, S.; Lennert, M.; Shimoni, M.; Kalogirou, S.; Wolff, E. Less is more: Optimizing classification performance through feature selection in a very-high-resolution remote sensing object-based urban application. GIScience *Remote Sens*. **2018**, *55*, 221–242. https://doi.org/10.1080/15481603.2017.1408892
34. Wright, M.N.; Ziegler, A. ranger: A fast implementation of random forests for high dimensional data in C++ and R. J. Stat. Softw. **2017**, *77*, 1–17.
35. Breiman, L. Random forests. Mach. Learn. **2001**, *45*, 5–32. https://doi.org/10.1023/A:1010933404324
36. Probst, P.; Wright, M.N.; Boulesteix, A.-L. Hyperparameters and tuning strategies for random forest. WIREs Data Mining Knowl. Discov. **2019**, *9*, e1301.
37. Aitchison, J. The Statistical Analysis of Compositional Data; Chapman & Hall: London, UK, **1986**.
38. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the 32nd International Conference on Machine Learning (ICML), Lille, France, 6–11 July 2015; pp. 448–456.

 doi:10.20944/preprints202604.1653.v1

39. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 2014, 15, 1929–1958.
40. Goodfellow, I.; Bengio, Y.; Courville, A. Deep Learning; MIT Press: Cambridge, MA, USA, 2016.
41. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 2015.
42. Willmott, C.J.; Matsuura, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. 2005, 30, 79–82. https://www.jstor.org/stable/24869236
43. Foody, G.M.; Atkinson, P.M. Uncertainty in Remote Sensing and GIS; Wiley: Chichester, UK, 2002. DOI:10.1002/0470035269
44. Fisher, P.F. The pixel: a snare and a delusion. *Int. J. Remote Sens.* **1997**, *18*, 679–685. https://doi.org/10.1080/014311697219015
45. Cracknell, A.P. Synergy in remote sensing — what's in a pixel? *Int. J. Remote Sens.* **1998**, *19*, 2025–2047. https://doi.org/10.1080/014311698214848
46. Fernandes, R.; Fraser, R.; Latifovic, R.; Cihlar, J.; Beaubien, J.; Du, Y. Approaches to fractional land cover and continuous field mapping: A comparative assessment over the BOREAS study region. Remote Sens. Environ. 2004, 89, 234–251. https://doi.org/10.1016/j.rse.2002.06.006
47. Shannon, C.E. A mathematical theory of communication. Bell Syst. Tech. J. **1948**, *27*, 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
48. Pielou, E.C. The measurement of diversity in different types of biological collections. *J. Theor. Biol*. **1966**, *13*, 131–144. https://doi.org/10.1016/0022-5193(66)90013-0
49. Yu, R.; Li, S.; Zhang, B.; Zhang, H. A deep transfer learning method for estimating fractional vegetation cover of Sentinel-2 multispectral images. IEEE Geosci. *Remote Sens. Lett*. **2022**, *19*, 6005605. https://doi.org/10.1109/LGRS.2021.3125429.
50. Schug, F.; Pfoch, K.A.; Pham, V.-D.; van der Linden, S.; Okujeni, A.; Frantz, D.; Radeloff, V.C. Land cover fraction mapping across global biomes with Landsat data, spatially generalized regression models and spectral-temporal metrics. *Remote Sens. Environ*. **2024**. https://doi.org/10.1016/j.rse.2024.114260
51. Lluïsà, J.; Ogaya, R.; Sardans, J.; Peñuelas, J. Mapping tree species fractions in temperate mixed forests using Sentinel-2 time series and synthetically mixed training data. *Remote Sens. Environ*. **2025**. https://doi.org/10.1016/j.rse.2025.114740
52. Somers, B.; Asner, G.P.; Tits, L.; Coppin, P. Endmember variability in spectral mixture analysis: A review. *Remote Sens. Environ*. **2011**, *115*, 1603–1616. https://doi.org/10.1016/j.rse.2011.03.003
53. Zhu, Z.; Zhang, J.; Yang, Z.; Aljaddani, A.H.; Cohen, W.B.; Qiu, S.; Zhou, C. Continuous monitoring of land disturbance based on Landsat time series. *Remote Sens. Environ*. **2020**, *238*, 111116. https://doi.org/10.1016/j.rse.2019.03.009
54. Herold, D.M.; Fedor, D.B.; Caldwell, S.; Liu, Y. The effects of transformational and change leadership on employees' commitment to a change: A multilevel study. J. Appl. Psychol. 2008, 93, 346–357. https://doi.org/10.1037/0021-9010.93.2.346
55. Potapov, P.; Hansen, M.C.; Pickens, A.; Hernandez-Serna, A.; Tyukavina, A.; Turubanova, S.; Zalles, V.; Li, X.; Khan, A.; Stolle, F. The global 2000-2020 land cover and land use change dataset derived from the Landsat archive: The GLC-SHARE dataset. Front. *Remote Sens*. **2022**, *3*, 856903. https://doi.org/10.3389/frsen.2022.856903
56. Picard, J.; Nungi-Pambu Dembi, M.M.; Barbier, N.; Cornu, G.; Couteron, P.; et al. Combining satellite and field data reveals Congo's forest types structure, functioning and composition. Remote Sens. Ecol. Conserv. 2025, 11, 200–220. https://doi.org/10.1002/rse2.419.
57. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the 33rd International Conference on Machine Learning (ICML), New York, NY, USA, **2016**; pp. 1050–1059.
58. Meyer, H.; Reudenbach, C.; Wollauer, S.; Nauss, T. Importance of spatial predictor variable selection in machine learning applications — Moving from data reproduction to spatial prediction. *Ecol. Model.* **2018**, *411*, 108815. https://doi.org/10.1016/j.ecolmodel.2019.108815

 doi:10.20944/preprints202604.1653.v1

59. Wadoux, A.M.J.-C.; Heuvelink, G.B.M.; De Bruin, S.; Brus, D.J. Spatial cross-validation is not the right way to evaluate map accuracy. *Ecol. Model.* **2021**, *457*, 109692. https://doi.org/10.1016/j.ecolmodel.2021.109692
60. Réjou-Méchain, M.; Barbier, N.; Couteron, P.; Ploton, P.; Vincent, G.; Herold, M.; Mermoz, S.; Saatchi, S.; Chave, J.; de Boissieu, F. Upscaling forest biomass from field to satellite measurements: Sources of errors and ways to reduce them. Surv. Geophys. 2019, 40, 881–911. https://doi.org/10.1007/s10712-019-09532-0
61. Higginbottom, T.P.; Symeonakis, E. Assessing land degradation and desertification using vegetation index data: Current frameworks and future directions. *Remote Sens.* **2014**, *6*, 9552–9575. https://doi.org/10.3390/rs6109552
62. Grantham, H.S.; Duncan, A.; Evans, T.D.; Jones, K.R.; Beyer, H.L.; Schuster, R.; Walston, J.; Ray, J.C.; Robinson, J.G.; Callow, M. Anthropogenic modification of forests means only 40% of remaining forests have high ecosystem integrity. Nat. Commun. 2020, 11, 5978. https://doi.org/10.1038/s41467-020-19493-3
63. Wan, L.; Ciais, P.; de Truchis, A.; Sean, E.; Fischer, F.J.; Purnell, D.; Ploton, P. Satellite-based mapping of annual canopy height and aboveground biomass in African dense forests. Front. Remote Sens. 2025, 6. https://doi.org/10.3389/frsen.2025.1724950. https://doi.org/10.3389/frsen.2025.1724950
64. Duncanson, L.; Kellner, J.R.; Armston, J.; Dubayah, R.; Minor, D.M.; Hancock, S.; Healey, S.P.; Patterson, P.L.; Saarela, S.; Marselis, S. Aboveground biomass density models for NASA's Global Ecosystem Dynamics Investigation (GEDI) lidar mission. Remote Sens. Environ. 2022, 270, 112845. https://doi.org/10.1016/j.rse.2021.112845