Ecological Informatics 95 (2026) 103726

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Unsupervised machine learning of LiDAR-derived forest structure relevant to ladder and canopy fuels in Minnesota's Superior National Forest

Nasrin Salehnia [a], Peter T. Wolter [a,*], Brian R. Sturtevant [b], Dalia Abbas Iossifov [c]

[a] Department of Natural Resource Ecology and Management, Iowa State University, Ames, IA, USA
[b] Northern Research Station, USDA Forest Service, Rhinelander, WI, USA
[c] USDA Forest Service, Forest Products Laboratory, Washington, D.C., USA

ARTICLE INFO

Keywords:
Self-organizing maps
Canopy height model
Principal component analysis
Wildfire
Clustering
Kawishiwi Ranger District

ABSTRACT

Understanding forest canopy structure is essential for monitoring ecosystem function, carbon dynamics, and wildfire risk. In this study, we developed an unsupervised classification framework that integrates 21 LiDAR-derived vertical structure metrics with Principal Component Analysis (PCA) and Self-Organizing Maps (SOM) to characterize canopy heterogeneity across the Kawishiwi Ranger District of the Superior National Forest in Minnesota, USA. The resulting structural classes support ladder- and canopy-fuel–relevant assessment and management. The PCA-based dimensionality reduction preserved key structural variation, while the SOM algorithm maintained topological relationships among feature vectors, enabling ecologically meaningful clustering. We applied K-means clustering to the trained SOM and selected the optimal number of clusters using the Silhouette score, identifying four distinct structural types—ranging from tall, multilayered canopies to sparse early-successional stands—whose spatial patterns support resilience-oriented planning, fuel-relevant structural stratification, and a district-wide baseline for monitoring and prioritizing areas for further evaluation under local objectives and constraints. Cluster-based canopy structure mapping revealed strong alignment with ecological conditions such as successional stage and disturbance history. This approach demonstrates the capacity of LiDAR and unsupervised machine learning to support scalable forest structure assessment. While promising, future integration with field-based validation and supervised learning could enhance ecological interpretability and operational applicability.

1. Introduction

Forest canopy structure—including tree height, canopy fractional cover, stem density, and vertical arrangement—affects key ecosystem processes such as carbon storage and wildlife habitat, while also controlling the spatial distribution of fuels and vertical continuity (ladder fuels) that drive wildland fire behavior (Asner et al., 2012; Cazzolla Gatti et al., 2017). Across North American boreal and mixed-hardwood forests, variation in canopy height distributions and gap dynamics have been shown to explain much of the spatial heterogeneity in biomass accumulation and species richness (Jucker et al., 2018; Sun et al., 2023). Meanwhile, in fire-prone landscapes, the vertical continuity of live and dead foliage and the complexity of under- and mid-story fuels govern flame spread rates and severity, directly influencing landscape-scale fire behavior and post-fire recovery trajectories (Collins et al., 2021; Verheyen et al., 2024).

Airborne LiDAR (Light Detection and Ranging) has revolutionized forest structural mapping by delivering dense, three-dimensional point clouds over hundreds to thousands of square kilometers (Guo et al., 2022; Hudak et al., 2016; Swatantran et al., 2011). From these point clouds one can derive dozens to hundreds of structural predictors—canopy height percentiles (H10, H50, H75, etc.), strata-specific mean heights, measures of vertical evenness (standard deviation, variance), and higher-order moments quantifying profile skew or kurtosis (Stavros et al., 2018; Torresan et al., 2016). These rich metrics underpin key applications: aboveground biomass estimation (Graves et al., 2018), species classification (Dalponte and Coomes, 2016), habitat suitability modeling (Balestra et al., 2024; Marselis et al., 2020), and parameterization of advanced fire-behavior models (Parsons et al., 2011). However, the explosion of LiDAR-derived predictors creates a high-dimensional, highly correlated data space that challenges traditional mapping workflows, risking overfitting or masking ecologically

* Corresponding author.
E-mail address: ptwolter@iastate.edu (P.T. Wolter).

https://doi.org/10.1016/j.ecoinf.2026.103726
Received 11 August 2025; Received in revised form 5 February 2026; Accepted 17 March 2026
Available online 18 March 2026

important gradients if variables are naively combined (Coops et al., 2021). Recent advancements, particularly with high-resolution airborne LiDAR surveys, now enable wall-to-wall characterization of these canopy attributes at fine, sub-10-m resolutions, offering unprecedented detail for ecological assessments (Shao et al., 2024; Skowronski et al., 2020).

Despite these advances, a persistent analytical challenge is translating dozens of LiDAR-derived, highly correlated vertical-structure metrics into a small number of interpretable, spatially coherent structural classes at management-relevant extents. Many LiDAR applications focus on predicting individual endpoints (e.g., biomass, species, habitat metrics, or fire-model inputs), but such workflows do not necessarily yield a transferable structural typology for district-wide stratification and monitoring, and uninformed aggregation of many correlated predictors can obscure key gradients or inflate model complexity. As a result, there is a need for a transparent, scalable approach that reduces redundancy, preserves continuous structural gradients, and produces a wall-to-wall structural classification that is useful for fuel-relevant stratification and broader forest monitoring.

Dimensionality-reduction techniques such as Principal Component Analysis (PCA) provide a straightforward remedy by transforming correlated predictors into a smaller set of orthogonal axes that capture the dominant patterns of variation (Chen et al., 2024). Yet PCA alone does not directly yield spatially coherent classes. Self-Organizing Maps (SOMs) address this by projecting high-dimensional data onto a two-dimensional grid of neurons, preserving topological relationships so that similar input vectors map to nearby neurons (Kohonen, 2001; Mingoti and Lima, 2006). SOMs have been employed in a variety of ecological and remote-sensing contexts—identifying tree species assemblages from LiDAR and hyperspectral data (Zhang and Qiu, 2012), mapping habitat types (Hopton and Mayer, 2006), and forest disturbance detection (Park et al., 2018).

In a related study, Fujino and Yoshida (2006) combined PCA with SOMs to delineate forestry zones, illustrating how these networks can integrate dimensionality reduction with topology-preserving clustering. Klobucar and Subasic (2012) demonstrated that SOMs effectively disentangle complex forest inventory data, uncovering stand-level structural and compositional gradients that traditional statistical techniques often overlook. Their work underscored the strength of SOMs for merging heterogeneous inputs—from remote sensing metrics to GIS layers—to inform more nuanced forest management strategies. Sulkava and Hollmén (2003) applied SOMs to group and interpret foliar nutrient measurements, demonstrating the method's ability to reveal ecological gradients in coniferous foliar chemistry. Together, these examples highlight SOMs' aptitude for capturing complex, nonlinear patterns and maintaining neighborhood relationships, making them particularly well suited for ecological analysis and forest planning.

Across numerous disciplines, SOMs have been widely recognized for their ability to condense high-dimensional data into lower-dimensional maps, uncover clustering, and provide intuitive visualizations—tools that aid in revealing hidden structures and relationships during exploratory analyses (Forest et al., 2021; Ponmalai and Kamath, 2019).

In this study, we develop and apply a unified PCA–SOM–K-means workflow for the Kawishiwi Ranger District (KRD) of the Superior National Forest in northern Minnesota. We first derive 21 LiDAR-based structural predictors—selected for their demonstrated relevance to canopy fuel continuity and vertical architecture (Engelstad et al., 2019)—at 10 m resolution across the entire district. After min–max scaling, we use PCA to reduce the 21 derivative layers to the fewest orthogonal components explaining $\geq$ 95% of the total variance, balancing information retention against complexity. Next, we perform a systematic SOM grid search over 3×3 to 15×15 neuron grids, training each map for 2000 iterations and evaluating the sum of quantization and topographic errors to identify the optimal map size. Finally, we apply K-means clustering to the SOM neuron weight vectors, selecting the cluster count (k) that maximizes average silhouette width, and map the resulting clusters back to each 10 m pixel.

Accordingly, our objective was to develop a LiDAR-only, unsupervised classification framework for mapping forest vertical structure across the KRD. Specifically, we aimed to: (1) reduce redundancy among LiDAR-derived structural metrics while retaining dominant gradients in canopy height and layering; (2) delineate a parsimonious set of structurally distinct classes and map their spatial distributions at 10 m resolution; and (3) interpret the resulting classes in terms of canopy height distribution, vertical stratification, and structural heterogeneity relevant to ladder- and canopy-fuel stratification and post-disturbance structural trajectories.

We exported the resulting structural clusters as continuous GeoTIFF raster layers and summarized each cluster with violin-plot diagnostics, yielding an ecologically interpretable partitioning of canopy architecture that captures both the dominant height gradient and subtler patterns of mid-story density and upper-canopy heterogeneity. By combining PCA's dimensionality reduction, SOM's topology-preserving mapping, and K-means clustering, the pipeline delivers a scalable, transparent framework for high-resolution forest-structure mapping—-providing inputs that can inform wildfire-risk modeling, habitat-suitability assessments, and long-term ecological monitoring across fire-sensitive hemiboreal forests (the boreal–temperate transition zone) such as those of northern Minnesota (Brandt, 2009).

Beyond producing an interpretable canopy-structure map, unsupervised structural clustering provides a transferable stratification layer that can support multiple downstream applications. Examples include (i) wildfire fuels and fire-behavior stratification by identifying structure classes consistent with ladder-fuel continuity and high vertical heterogeneity, (ii) habitat-structure assessment by mapping gradients in canopy height, layering, and complexity relevant to species distributions, and (iii) monitoring and stratified sampling by providing consistent structure classes for targeted field checks and repeated LiDAR-based change assessments.

2. Methodology

2.1. Study area and dataset

The Kawishiwi Ranger District (KRD) is located in the Superior National Forest in northeastern Minnesota, USA (47.7° N Lat., 91.3° W Lon.). The landscape is characterized by hemiboreal mixed forests, interspersed with lakes and wetlands, making it a complex and ecologically significant area for forest monitoring (Fig. 1). See Sturtevant et al. (2012) for a comprehensive description of this study landscape.

Airborne LiDAR sensor data collected between 2018 and 2021 as part of the USGS 3D Elevation Program (3DEP) served as the data source for this study. Across the KRD boundary (2,907,181,565 m^2), the LAS dataset contains ~82.27 billion returns, corresponding to an average density of ~28.3 points m^{-2} (and ~ 16.4 first returns m^{-2}). We imported approximately 3470 LAZ tiles (Fig. A1) into a LAS dataset using ArcGIS Pro (Esri, 2024) and processed all products in a consistent projected coordinate reference system (NAD83 / UTM Zone 15 N) to preserve metric distances and areas during aggregation and mapping. All raster processing was performed in NAD83(2011) / UTM Zone 15 N with elevations referenced to NAVD88 (Geoid12B). From this dataset, we generated three raster surfaces: a Digital Terrain Model (DTM) (interpolated from ground returns), a Digital Surface Model (DSM) (derived from the highest returns), and a Canopy Height Model (CHM). We processed all raster files at 1 m spatial resolution to align with the scale of analysis (Eq. 1):

$$CHM_{1m} = DSM_{1m} - DTM_{1m} \tag{1}$$

Prior to metric computation, CHM heights were filtered to the physically plausible range used throughout the analysis (0–40 m), and cells dominated by invalid/no-data values were excluded. We then aggregated the CHM into 10 m × 10 m grid cells. The 10 m grain size was

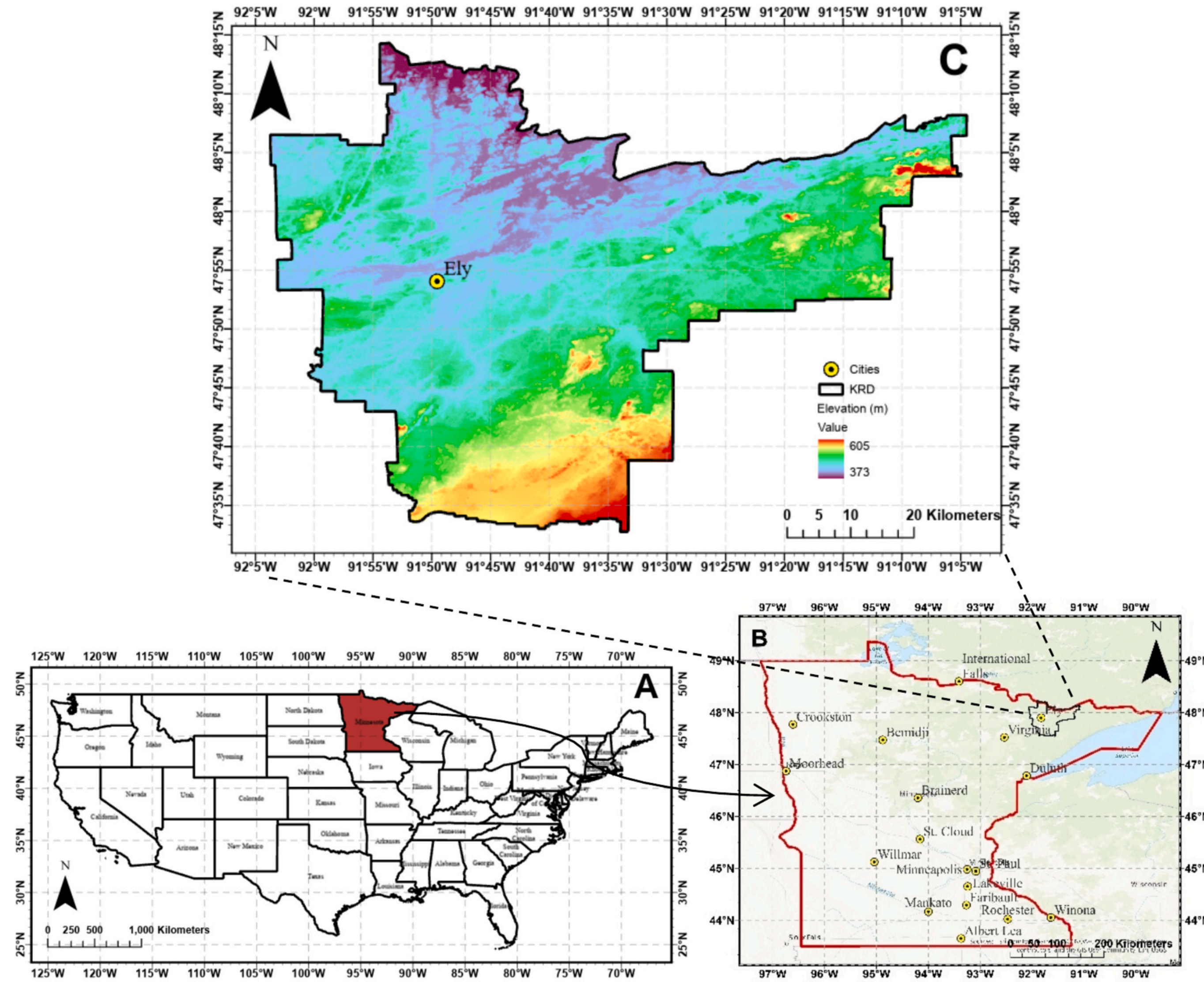

Fig. 1. (A) The conterminous United States with Minnesota highlighted in red. (B) Minnesota inset showing the Kawishiwi Ranger District (KRD, black outline) in the northeastern portion of the state; principal towns (e.g., Ely and Duluth) are labelled for reference. (C) Hill-shaded 1/3-arc-second (≈ 10 m) digital-elevation model of the KRD. Major lakes and rivers are delineated, and the city of Ely is marked. Light tones correspond to higher ridges, darker tones to lower terrain. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

chosen because (i) it approximates the average crown footprint of mature conifers and mixed hardwoods (Wolter et al., 2009) in the study area, and (ii) it reduces random height noise without obscuring stand-level structural heterogeneity. As such, we calculated the 21 LiDAR-derived Forest structure predictors at this 10 m resolution.

2.2. Derivation of structural predictors

We extracted 21 structural metrics (Table 1) used in forest structure analysis from the CHM and point cloud data using Python libraries. The complete processing scripts and dataset are available in the associated reproducibility package (Salehnia et al., 2026; doi:10.17632/26pw833vxb.1). The 21 LiDAR-based structure metrics consist of five functional categories—Central Tendency, Variability, L-Moments, Shape & Distribution, and Canopy Stratification & Vegetation Structure—to better represent their statistical or ecological role. These predictors are adapted and extended from Engelstad et al. (2019), which demonstrated their relevance for canopy fuel estimation and wildfire modeling in this region.

2.3. Data preprocessing and PCA

We consolidated all 21 canopy-structure predictors into a single, multi-layer dataset, and we excluded all records with missing values to prevent bias in subsequent analyses (Han et al., 2012). To ensure equal weighting across variables, we scaled each feature to the interval [0,1] via min–max normalization as Eq. 2 (Pedregosa et al., 2011).

$$\widetilde{x}_{ij} = \frac{x_{ij} - min_i(x_{ij})}{max_i(x_{ij}) - min_i(x_{ij})} \quad (2)$$

where x_{ij} is the original value of feature j for sample i, $min_i(x_{ij})$ and $max_i(x_{ij})$ are the minimum and maximum of that feature over all samples, and $\widetilde{x}_{ij} \in [0, 1]$ is the normalized value. This approach preserves the original distribution shape while preventing variables with large numeric ranges from dominating the analysis. Once scaled, we conducted Principal Component Analysis (PCA) (Abdi and Williams, 2010; Jolliffe, 2002). PCA identifies orthogonal axes (principal components, PCs) that sequentially explain the greatest remaining variance in the data.

Table 1
Grouped LiDAR-derived canopy metrics used to describe forest structure and fuel- relevant characteristics, adapted from Engelstad et al. (2019).

Central Tendency Metrics	H50PCT	Average height 50th percentile (median)
	H60PCT	Average height 60th percentile
	H70PCT	Average height 70th percentile
	H75PCT	Average height 75th percentile
	H10PCT	Average height 10th percentile
	HQUAD	Quadratic mean height
Variability Metrics	HSTD	Standard deviation of all return heights
	HVAR	Variance of heights
	HCV	Coefficient of variation of heights
	MEDMAD	Median absolute deviation from median height
	MEDMODE	Median absolute deviation from mode height
L-Moments	LMOM2	Second L-moment (scale/ dispersion)
	LMOM3	Third L-moment (skewness)
	LMOM4	Fourth L-moment (kurtosis)
Shape & Distribution Metrics	LSKEW	L-moment skewness
	HSKEW	Kurtosis of heights
	CRR	Canopy relief ratio (relative vertical profile)
Canopy Stratification & Vegetation Structure	STRATUM5_MEAN	Mean height of vegetation >5 m and ≤ 10 m
	STRATUM7_MEAN	Mean height of vegetation >20 m and ≤ 30 m
	STRATUM7_SD	Std. deviation of vegetation >20 m and ≤ 30 m
	STRATUM7	Percentage of vegetation >20 m and ≤ 30 m

In this regard, let X be the $n \times p$ matrix of min–max–normalized predictors. PCA proceeds by the following steps (Eqs. 3–7):

First, we computed the covariance matrix as shown in Eq. 3:

$$\sum = \frac{1}{n-1} X^T X \in \mathbb{R}^{p \times p} \tag{3}$$

Next, we perform the eigen-decomposition, obtaining the ordered eigenvalues ($\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$) and their corresponding orthonormal eigenvectors e_j by solving Eq. 4:

$$\sum e_j = \lambda_j e_j, j = 1, \ldots, p \tag{4}$$

In this step, principal-component scores are computed as the score of sample i on component j is (Eq. 5):

$$z_{ij} = X_i^T e_j \tag{5}$$

In this turn, the proportion of variance explained by PC_j is (Eq. 6):

$$\eta_j = \frac{\lambda_j}{\sum_{k=1}^{p} \lambda_k}, \sum_{j=1}^{p} \eta_j = 1 \tag{6}$$

Finally, we used cumulative explained variance, as Eq. 7:

$$C(m) = \sum_{j=1}^{m} \eta_j \tag{7}$$

and chose the smallest m such that $C(m) \geq 0.95$ (Bro and Smilde, 2003). Indeed, we plotted the cumulative explained variance ratio (the “scree plot”) to determine how many PCs were needed to capture at least 95% of the total variance. In our analysis, this occurred at $m = 6$.

For interpretability and diagnostics, we saved several outputs, including scree plot (showing the cumulative variance as a function of component count (Fig. 6)), component loading matrix (detailing the contribution of each original predictor to the first several PCs (Fig. 7)), and pairwise scatterplots (Fig. A2) of PC1–PC6 (visualized with kernel-density diagonals, to inspect data clustering and identify potential outliers).

2.4. Self_Organizing map (SOM)

To segment the PCA-reduced canopy metrics into coherent groups, we employed a Self-Organizing Map (SOM), an unsupervised neural-network method that projects high-dimensional data onto a two-dimensional neuron grid while preserving topological relationships (Kohonen, 2001). We used the MiniSom v2.2 implementation (Vettigli, 2018) to evaluate square SOM grids of size $m \times m$, with m ranging from 3 to 15 in steps of 2. For every candidate grid we applied the following procedure:

For every candidate SOM grid, we first seeded neuron weights with random values drawn from the full range of each principal-component axis, ensuring that the initial map covered the data space comprehensively (Kohonen, 2013). The map was then trained for 2000 iterations presented in random order; at each iteration the current input vector $x(t)$ was paired with its best-matching unit (BMU), defined as the neuron whose weight vector minimized the Euclidean distance to $x(t)$. After training, two quality metrics were computed: the quantization error (QE), i.e., the mean Euclidean distance between all inputs and their respective BMUs, which gauges representation accuracy; and the topographic error (TE), i.e., the proportion of inputs for which the first- and second-closest neurons are not adjacent, which gauges how faithfully neighborhood relations are preserved. We summed the two metrics for each grid size, and we selected the grid with the smallest value of QE + TE as the optimal compromise between accuracy and topological smoothness (Vesanto and Alhoniemi, 2000). For example, Fig. 2 shows a simple scheme of the SOM concept.

For selecting the final SOM training, by using the selected grid size from the previous procedure, we performed a refined training pass of 4000 iterations. At each iteration for input x(t) three main steps is run:

I) to identify BMU, find neuron c by minimizing $\|x(t) - w_i(t)\|$
II) adjust the BMU and its neighbors according to a Gaussian kernel:

$$\Delta w_i(t) = \alpha(t) \exp\left(-\frac{\|r_c - r_i\|^2}{2\sigma(t)^2}\right)[x(t) - w_i(t)] \tag{8}$$

where r_i is the grid position of neuron i, and α(t) and σ(t) decay linearly from initial values ($\alpha_0 = 0.3$, $\sigma_0 = 1.5$) to zero (Kohonen, 2001).

III) Early iterations encourage global organization, while later steps fine-tune local clusters (convergence)

Upon completion, we visualized the SOM's U-matrix (unified distance matrix) to inspect inter-neuron distances and emergent cluster boundaries.

2.5. Clustering on SOM Vectors

After training, each neuron in the $m \times m$ SOM carries a reference vector in the reduced PC space. To partition these neurons into discrete groups, we applied the K-Means algorithm (Niknam et al., 2011) to the set of m^2 reference vectors. We explored values of k from 2 to 14, evaluating cluster quality via the average Silhouette score (Arbelaitz et al., 2013; Salehnia et al., 2019). The optimal k chosen was the one that maximized this metric, balancing intra-cluster cohesion against inter-cluster separation. Finally, each original data pixel inherited the label of its neuron's cluster, thereby mapping every 10 m observation to one of the k discrete classes.

In the proposed workflow, PCA reduces redundancy among LiDAR

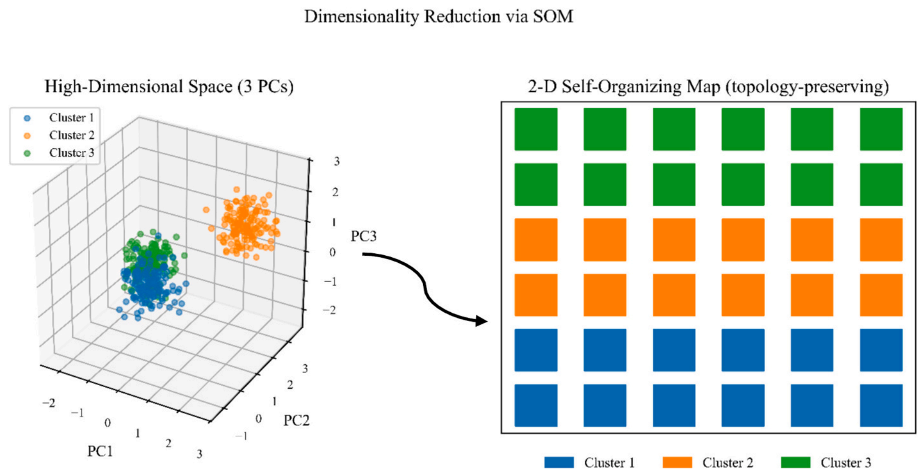

Fig. 2. A simple scheme of the SOM performance regarding dimensionality reduction, with three PCs and three clusters.

metrics, the SOM learns a topology-preserving set of representative canopy-structure prototypes (neuron weight) that help reduce sensitivity to noise and improve the stability and interpretability of the final mapped classes, and k-means clusters these prototypes into a small number of discrete structure classes for wall-to-wall mapping. To quantify the contribution of each step, we performed a baseline method comparison of k-means, PCA → k-means, and PCA → SOM → k-means (all with k = 4) using standard internal clustering validity indices; results are provided in Appendix Table A3.

3. Results

3.1. LiDAR point cloud and CHM characterization

To document the realism of the LiDAR inputs used to derive the CHM and all subsequent structural predictors, we include a representative subset of the raw point cloud from the study area (Fig. 3). This subset corresponds to a ~ 700 m × 700 m analysis window and, in this example, intersects two USGS 3DEP LiDAR tiles; returns from both tiles were combined for the visualization in Fig. 3. The plan-view point cloud (Fig. 3A) shows spatially coherent patterns in LiDAR return height above local ground across this window, including contiguous areas dominated by taller vegetation returns interspersed with lower vegetation and open/low-return conditions. The absence of obvious striping, tiling seams, or implausible clusters of extreme heights supports consistent LiDAR acquisition/processing across the mapped extent. The transect (P0–P1) highlights a profile location crossing multiple canopy conditions, providing a direct qualitative check of the vertical structure observed by the sensor.

The vertical cross-section (Fig. 3B; ±25 m slice around the transect) reveals a physically plausible vertical distribution of returns, with a continuous near-ground envelope and abundant vegetation returns extending upward through the canopy profile. The point cloud shows clear structural stratification, including a dense layer of lower returns and distinct mid-to-upper canopy returns, with local canopy tops reaching on the order of ~30–35 m in this example window. Importantly, taller returns occur in spatially coherent patches along the transect rather than as isolated spikes, supporting that the LiDAR observations capture meaningful structural gradients. Together, the plan view and profile provide transparent "raw-data" context demonstrating coherent horizontal and vertical forest-structure patterns, which are then summarized by the CHM and aggregated to 10 m for the predictor derivation and clustering workflow described in Sections 2.2–2.5.

Because all LiDAR-derived predictors in this study originate from the

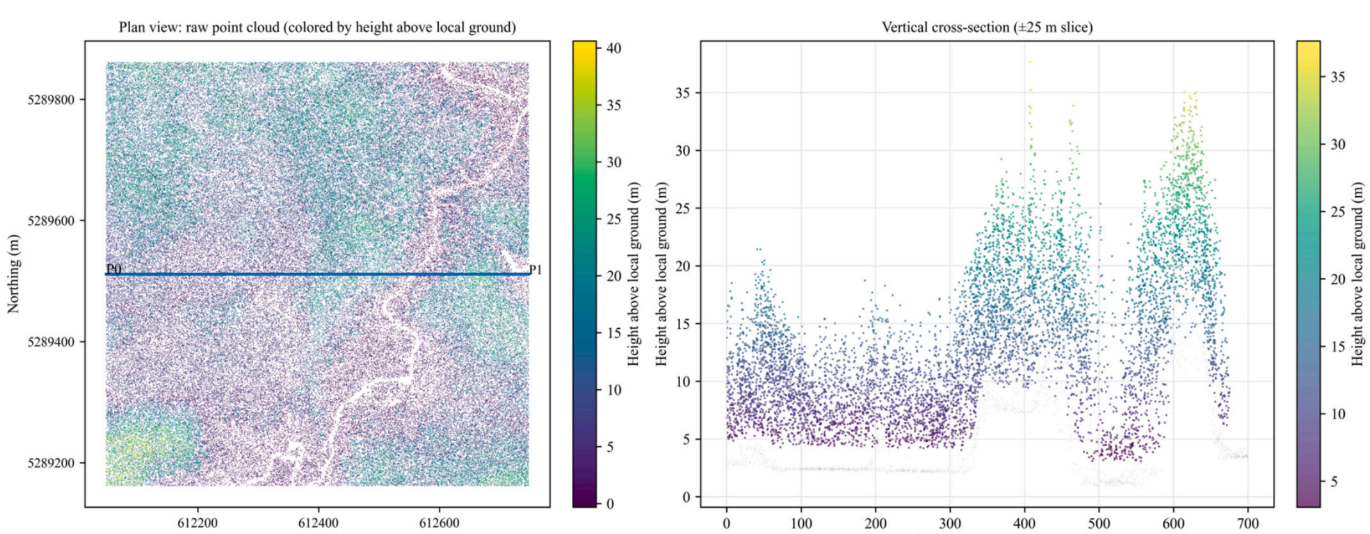

Fig. 3. Raw LiDAR point-cloud quality-control visualization for a representative analysis window in the Kawishiwi Ranger District study area. (A) Plan view of LiDAR returns colored by height above local ground; the P0–P1 line marks the transect. (B) Vertical cross-section of returns within a ± 25 m slice around the transect. The window intersects two USGS 3DEP tiles (combined for visualization). Axes are in projected meters; colors show height above local ground (m).

cleaned 1 m CHM described in Section 2.1, we first summarize the CHM distribution and provide representative spatial examples to document product realism and the effectiveness of the cleaning steps (0–40 m). The CHM height distribution (Fig. 4) is strongly right-skewed, with most 1 m cells concentrated at low canopy heights and a long tail toward taller canopy. This pattern is consistent with the expected structure of a heterogeneous forest mosaic containing open/short-canopy conditions, intermediate canopy, and localized tall-canopy patches.

To complement the distributional summary, we present representative 1 m CHM subsets after cleaning (Fig. 5) that illustrate spatial patterns at fine grain. Each panel shows a 700 m × 700 m window (∼49 ha; 0.49 km^2) extracted from the 1 m CHM, providing a stand-to-patch–scale view of canopy structure. The subsets capture three contrasting structural conditions: (A) a tall-canopy example (high mean CHM), (B) a short/open-canopy example (low mean CHM), and (C) a structurally complex canopy example characterized by high within-window variability. Across these panels, canopy-height patterns are spatially coherent (e.g., contiguous tall-canopy patches, gaps, and heterogeneous mixtures), supporting that the cleaned CHM retains meaningful structural gradients while minimizing artifacts associated with invalid/no-data cells and implausible heights. Together, the histogram and representative subsets confirm that the CHM is physically plausible and suitable for aggregation to 10 m and for computing the LiDAR-derived forest-structure predictors used in the subsequent analyses.

Whereas Fig. 5 presents the cleaned 1 m CHM alone, Appendix Fig. A4 overlays the derived 10 m structural classes on the same windows to show that class boundaries align with coherent differences in canopy height and within-window structural variability. It is discussed later in the cluster-characterization results.

3.2. PCA diagnostics

The cumulative explained-variance scree plot (Fig. 6) demonstrates that the first six PCs collectively capture approximately 96% of the total variance in the 21 canopy-structure predictors. Individually, PC1 and PC2 account for the largest proportions—58.5% and 21.0%, respectively—reflecting gradients in mean canopy height and vertical heterogeneity, while PC 3 (9.8%) further distinguishes canopy density patterns. PCs 4–6 contribute incremental gains of 3.7%, 1.9%, and 1.5%, respectively, bringing the cumulative total above the 95% threshold and justifying their retention for downstream clustering. This dimensionality reduction from 21 to 6 axes streamlined the dataset by over 70% while preserving almost all its informative variability.

Accordingly, by considering the loading-matrix heatmap (Fig. 7), one can draw several key insights into the relevance of our LiDAR-derived canopy structural predictors:

PC1 (≈ 58.5% of variance) is overwhelmingly driven by canopy height percentiles (CRR, H50PCT, H60PCT, H70PCT, H75PCT all load ∼ + 0.30 to +0.35) and, to a lesser extent, the mean height of the lower stratum (STRATUM5_MEAN, +0.11). In combination, these variables capture overall height of the forest canopy. PC2 (≈ 21.0% of variance) contrasts height-distribution spread and skew (HSTD, +0.37; LMOM2, +0.36; HVAR, +0.27) with moderate contributions from upper-stratum mean and variability (STRATUM7_MEAN, +0.30; STRATUM7_SD, +0.34). This axis represents vertical heterogeneity—how even or clumped foliage is through the canopy profile. PC3 (≈ 9.8% of variance) loads strongly on upper-stratum metrics (STRATUM7, +0.39; STRATUM7_MEAN, +0.40; STRATUM7_SD, +0.41) and on the fourth-moment metric LMOM4 (+0.31), distinguishing upper-canopy density and complexity from lower-canopy structure. PC4 (≈ 3.7% of variance) is dominated by STRATUM5_MEAN (+0.84), the mean height of the mid-story (5–10 m layer). This axis isolates mid-story height variation independent of overall canopy stature. PC5 (≈ 1.9% of variance) loads primarily on the upper-stratum standard deviation (STRATUM7_SD, +0.63), capturing upper-canopy heterogeneity (crown clumpiness or variability above ∼20 m). PC6 (≈ 1.5% of variance) shows moderate loadings on lower-tail and mode metrics (H10PCT, +0.41; MEDMODE, +0.44), reflecting the lower-canopy leaf-height distribution tails.

Although PC4–PC6 together account for only ∼6% of the total variance, by feeding all six PCs into the SOM we ensure the map is sensitive not just to the primary height axis (PC1) but also to mid-story structure (PC4) and upper-canopy heterogeneity (PC5). In practice, this suggests that two stands with identical median heights (i.e., similar PC1 scores)—but differing in mid-story development or crown-clumpiness—can occupy distinct regions on the SOM and end up in separate clusters.

Hence, while PCs 1–3 capture the primary canopy gradients and are easiest to interpret (≈ 87% of variance), we retained all six components (≥ 95% variance) for the subsequent SOM clustering to ensure no meaningful structural signal was discarded. Moreover, the pairwise plots of the first six PCs (Appendix, Fig. A2) confirm that no unexpected clustering or outliers remain in the reduced space, and that the PCs are approximately uncorrelated (as expected by construction). In particular, PC1 vs. PC2 shows a broad triangular cloud with no obvious skew, while PC3 vs. PC1 and PC3 vs. PC2 reveal slight curvature but no distinct subgroups beyond those later recovered by the SOM. The smooth, cloud-

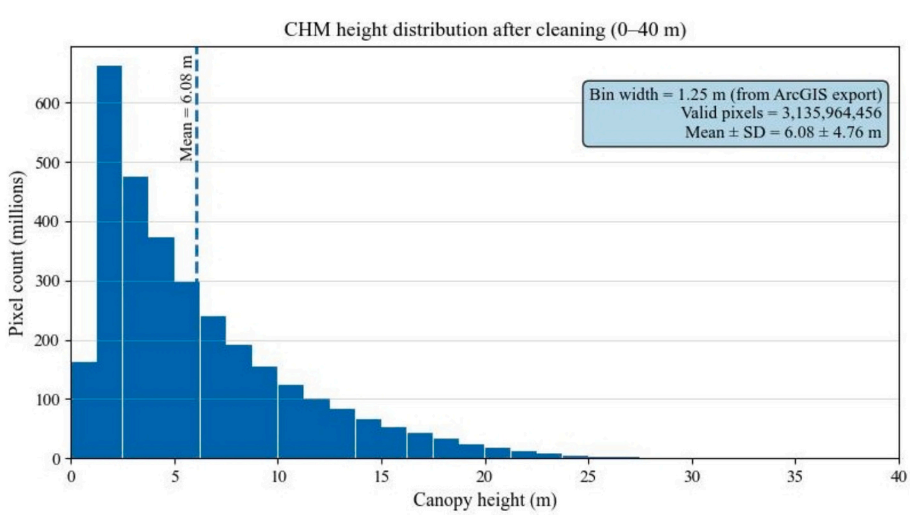

Fig. 4. Histogram of 1 m canopy height model (CHM) values after cleaning to the physically plausible range (0–40 m). Bars show pixel counts (millions) in 1.25 m height bins, and the dashed vertical line indicates the mean canopy height (6.08 m). The inset reports the bin width, number of valid pixels, and mean ± standard deviation for the cleaned CHM.

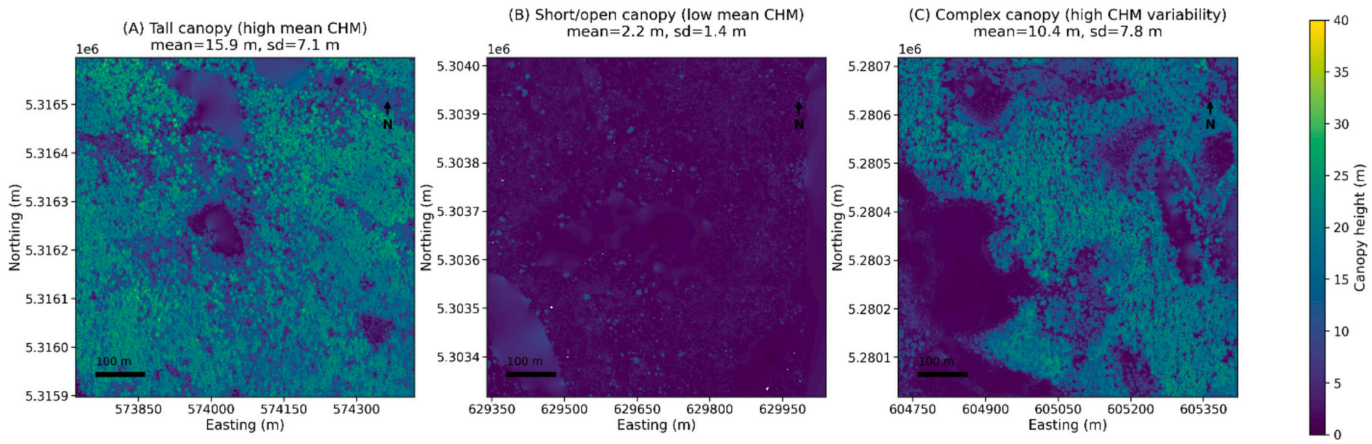

Fig. 5. Representative 1 m canopy height model (CHM) subsets after cleaning to the physically plausible range (0–40 m), illustrating contrasting canopy structural conditions across the study area: (A) tall canopy (high mean CHM), (B) short/open canopy (low mean CHM), and (C) complex canopy (high CHM variability). Panel titles report the mean and standard deviation of CHM within each subset. Maps are shown in projected coordinates (Easting/Northing, m); north arrows and 100 m scale bars are included. The colour bar indicates canopy height (m).

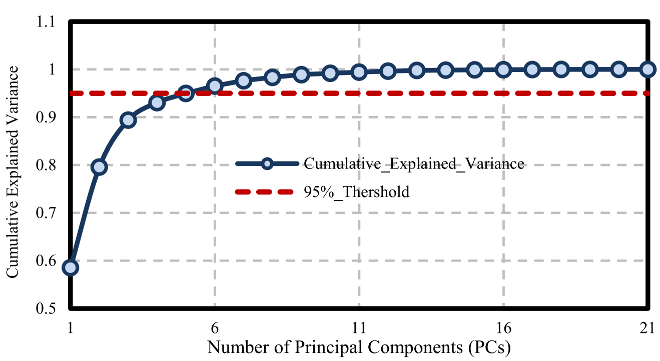

Fig. 6. Cumulative explained-variance scree plot for the PCA on 21 canopy-structure predictors. Circles mark the cumulative variance captured by each successive principal component; the red dotted line shows the 95% variance threshold. Six components are required to exceed this cutoff. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

like appearance across all off-diagonal panels underscores that the retained PCs capture continuous variation rather than discrete artifacts, supporting their use as inputs for topology-preserving clustering.

3.3. SOM grid-size selection and topology evaluation

We evaluated square SOM grids from 3 × 3 to 15 × 15 neurons. We trained each candidate map for 2000 iterations, after which we recorded respective QE, TE, and sum errors (Table 2). The QE declined steadily with increasing grid size—dropping from 0.281 (3 × 3) to 0.150 (15 × 15)—because additional neurons better approximated the six-dimensional PC space. Conversely, TE rose sharply beyond the 5 × 5 map, indicating growing difficulty in preserving neighborhood relationships on larger grids. The 5 × 5 grid produced the lowest combined error (QE + TE = 0.226), offering the best compromise between representation fidelity (low QE) and topology preservation (low TE). After selecting this optimal grid, we re-initialized the 5 × 5 SOM and performed a longer, 4000-iteration training pass to ensure full weight convergence before applying K-means. The resulting 25 prototype vectors capture the dominant structural gradients in the PCA-reduced feature space while preserving contiguous neighborhoods, enabling ecologically interpretable clustering and smooth spatial boundaries in the final GeoTIFF outputs.

After finalizing the 5 × 5 SOM structure and completing the extended training phase, we visualized the Unified Distance Matrix (U-Matrix) to interpret the internal topology of the trained map and to identify emergent cluster boundaries. The U-Matrix (Fig. 8) encodes the average Euclidean distance between each neuron and its immediate neighbors. This visualization is particularly valuable because it reveals not only neuron positions relative to one another in high-dimensional space but also where transitions between structurally distinct forest types are likely to occur. In this representation, darker cells correspond to higher inter-neuron distances, indicating abrupt shifts in canopy structure between adjacent neurons. These dark zones typically emerge as ridge lines separating distinct regions and are suggestive of potential cluster boundaries. Notably, cells at positions such as (2,1), (2,2), (2,3), (3,3), and (4,3) exhibit high distance values—highlighting well-defined separations across the map. In contrast, lighter cells (e.g., lower-right and upper-left corners) exhibit low distances, reflecting internally coherent areas where neurons (and their associated input data) are highly similar—indicative of stable and compact clusters.

The spatial coherence of dark bands within the U-Matrix affirms the SOM's ability to preserve topological relationships, even after reducing the original 21-dimensional canopy feature space to just six principal components. Furthermore, these ridge-like transitions guide the subsequent K-means clustering step by revealing where cluster divisions should ideally occur. Overall, the U-Matrix provides an intuitive visual confirmation that the trained SOM has organized the forest structure data into meaningful and separable regions, each representing a distinct ecological composition or successional stage.

3.4. Cluster number determination

To identify the optimal number of canopy-structure clusters, we applied K-means clustering to the 25 SOM prototype vectors across a candidate range of k = 2–14. Two diagnostic metrics were used to evaluate clustering performance: total within-cluster variance (inertia) and the Silhouette Score, which measures the balance between cluster compactness and separation (Fig. 9A–B). The inertia curve (Fig. 9A) exhibits a pronounced elbow at k = 4, beyond which additional clusters yield diminishing returns in variance reduction. Simultaneously, the Silhouette Score peaks at k = 4 (≈ 0.34; Fig. 9B), indicating the highest degree of internal cohesion and inter-cluster separability. The convergence of both metrics on the same value strongly supports the selection of k = 4 as the optimal number of clusters. This choice balances parsimony with structural detail and ensures that downstream interpretations reflect meaningful differentiation in canopy structure. Accordingly, all subsequent analyses—including U-matrix

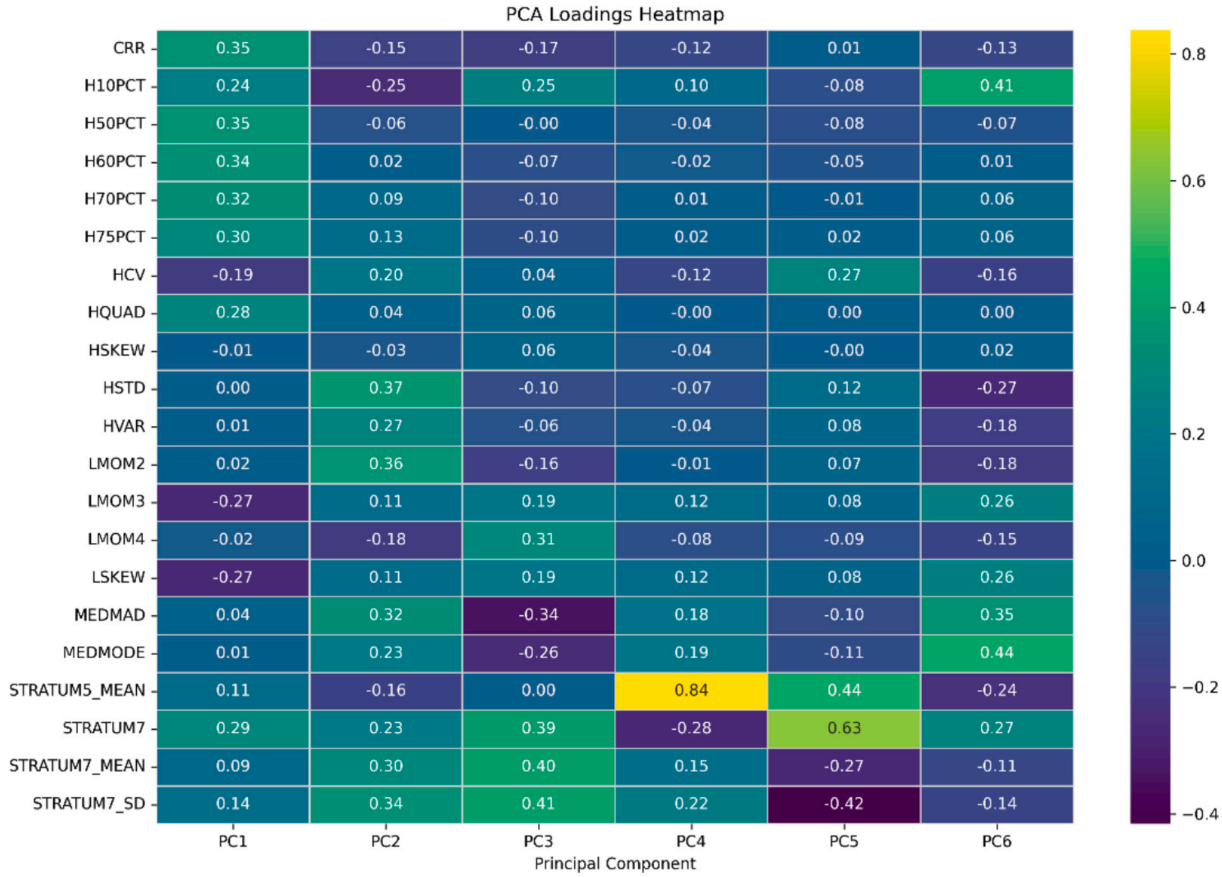

Fig. 7. PCA loadings heatmap showing the contribution of each original canopy-structure predictor to the first six principal components. Cell colors indicate loading magnitude and sign (purple = strong negative loading; yellow = strong positive loading), with annotated values for clarity. High loadings of canopy height percentiles (e.g., H50PCT, H60PCT, H70PCT, H75PCT) on PC1 reflect overall vertical height, while metrics of lower-stratum mean height (STRATUM5_MEAN) and upper-stratum heterogeneity (STRATUM7, STRATUM7_MEAN) dominate PCs 4 and 5, respectively, summarizing major structural gradients in the forest canopy. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Quantization error (QE), topographic error (TE), and their sum for candidate SOM grid sizes (3 × 3 to 15 × 15). The 5 × 5 grid (bold) yields the minimum combined error (0.226), and was the grid selected for the final 4000-iteration training phase.

Grid_Size	Quantization_Error	Topographic_Error	Sum_Error	Grid_Dim
3	0.281	0.001	0.281	3 × 3
5	**0.212**	**0.014**	**0.226**	**5 × 5**
7	0.184	0.053	0.238	7 × 7
9	0.171	0.127	0.298	9 × 9
11	0.157	0.163	0.320	11 × 11
13	0.150	0.140	0.290	13 × 13
15	0.151	0.189	0.340	15 × 15

interpretation, spatial mapping, and ecological assessment—were conducted based on this four-cluster configuration.

3.5. Cluster characterization

To better understand the structural differences among the four identified clusters, six key canopy metrics were selected for visualization using violin plots (Fig. 10). These metrics—CRR, HSTD, STRATUM7_SD, STRATUM5_MEAN, STRATUM7, and MEDMODE—were chosen based on their dominant contributions to the first six principal components, as illustrated in the PCA loadings heatmap (Fig. 7). Together, they represent the most influential axes of structural variation in the dataset.

The violin plots reveal distinct forest structural characteristics across clusters. Cluster 1 shows the highest values for CRR, STRATUM5_MEAN, and STRATUM7, indicating a structurally complex canopy with well-

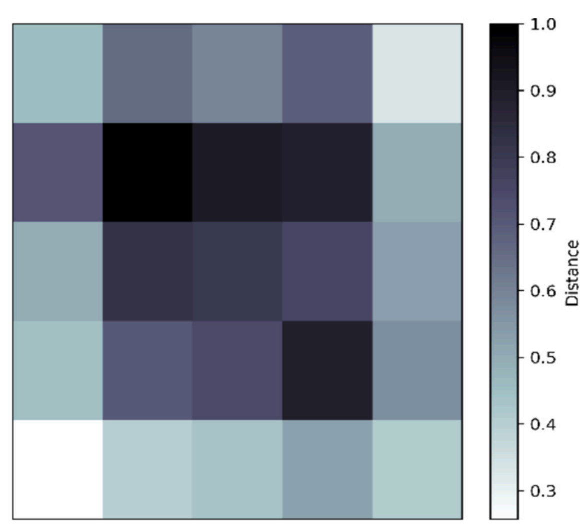

Fig. 8. Self-Organizing Map (SOM) U-Matrix for the optimal 5 × 5 topology trained on PCA-reduced features. Each cell represents a SOM neuron, and the colour indicates the average distance to its immediate neighbors. Darker shades highlight greater dissimilarities—suggesting potential cluster boundaries—-while lighter shades represent more homogeneous regions.

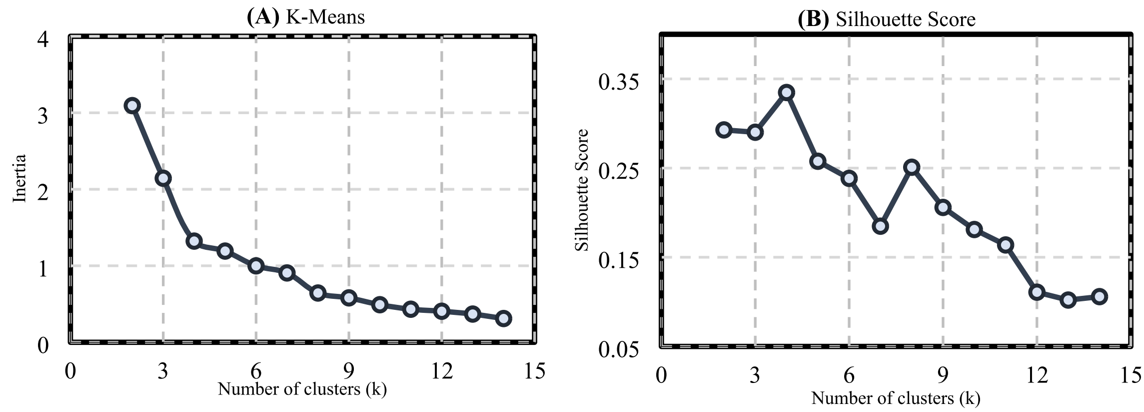

Fig. 9. Cluster number evaluation using K-means on SOM codebook vectors: (A) Elbow plot showing inertia (total within-cluster variance) across tested cluster numbers (k = 2–14). A clear elbow is visible at k = 4. (B) Silhouette Score plot indicating peak clustering performance at k = 4, supporting the selection of four topologically consistent clusters.

developed vertical layering and dense upper strata. These patterns suggest mature forest stands with multiple canopy layers and substantial foliage at both mid and upper heights. The moderate variation in HSTD and STRATUM7_SD further supports the presence of vertical complexity without extreme heterogeneity. Cluster 2 exhibits high vertical heterogeneity, with the highest standard deviation of height (HSTD) and STRATUM7_SD, reflecting a mixture of canopy heights likely caused by regenerating or mixed-age forest layers. Broader distributions in MEDMODE and CRR also point to diverse vertical structures and canopy gaps, characteristic of transitional or recovering forest areas. In contrast, Cluster 3 demonstrates relatively uniform vertical structure, with the lowest variability in height (HSTD) and small values for STRATUM7_SD. It also has moderate canopy heights and lower STRATUM7 coverage, indicative of even-aged stands dominated by mid-sized trees with sparse upper strata and minimal understory. While both clusters (Clusters 2 and 3) exhibit intermediate canopy structures, a closer comparison of the height-related metrics reveals that Cluster 3 generally ranks second in overall canopy height (after cluster 1). Specifically, Cluster 3 exhibits higher values in median-based metrics such as H50 and H60, suggesting more uniformly tall canopies. In contrast, Cluster 2 shows slightly higher values only in the upper percentiles (e.g., H70 and H75), likely indicating the presence of isolated tall trees amid a more variable canopy structure. This pattern reinforces the characterization of Cluster 3 as more structurally uniform and vertically continuous, compared to the heterogeneous and regenerating nature of Cluster 2.

Finally, Cluster 4 is characterized by consistently low values across most structural metrics, particularly STRATUM7, CRR, and STRATUM7_SD. This suggests sparse, vertically simple forests with limited canopy height and weak stratification—likely representing understocked or early successional stages. The narrow distributions in MEDMODE support a homogeneous vertical profile with minimal structural differentiation. Altogether, the selected variables provide an informative summary of the structural contrasts among the clusters, aligning well with the full-metric statistical profiles and offering valuable ecological interpretations. These findings lay the foundation for targeted forest management and restoration strategies (Olszewski and Bailey, 2022).

3.6. Spatial patterns

The spatial distribution of the four SOM-derived forest structural clusters across the KRD is presented in Fig. 11. The map reveals clear geographic organization and ecological coherence in the spatial placement of clusters.

Cluster 1 (mature forests) exhibits a scattered yet discernible presence across the Kawishiwi landscape, with notable concentrations in the northern and northwestern zones—corresponding to remnant high-canopy stands indicative of older forest structures. These areas likely correspond to older-growth or minimally disturbed forest remnants, where canopy layering has had time to develop.

Regenerating and vertically heterogeneous forests (Cluster 2) are scattered across the landscape and commonly occur in structural transition zones where a mix of low saplings and residual mid-canopy trees produces the greatest within-pixel height dispersion (highest HSTD ≈ 7 m and HVAR ≈ 50 m^2). The marked negative skew (HSKEW ≈ -1.2) indicates an abundance of short returns relative to tall ones, consistent with active regrowth following recent canopy opening.

Mid-story–dominated, structurally uniform stands (Cluster 3) show a more compact and clustered spatial distribution, particularly in areas characterized by relatively even-aged structure or past management. These patterns are consistent with structurally homogeneous stands that can arise under uniform stand development pathways. In contrast, sparse forests with low canopy complexity (Cluster 4) dominate much of the landscape matrix, especially in central and eastern sectors and in low-elevation or hydrologically constrained zones. These areas are consistent with early-successional conditions, recent disturbance, or locations where regeneration may be limited by environmental constraints. Altogether, the spatial arrangement of clusters supports the ecological plausibility of the SOM–PCA framework and highlights its potential for forest monitoring, structural classification, and landscape-scale planning.

4. Discussion

This study presents a scalable, data-driven framework to characterize forest vertical and horizontal structural heterogeneity using unsupervised clustering of LiDAR-derived metrics. By integrating PCA with SOM and K-means, we identified four distinct structural types across the KRD of the Superior National Forest in northern Minnesota, each with unique ecological implications and management priorities (Fig. 12). The clustering captures both vertical complexity and successional gradients, offering a spatially explicit foundation for adaptive forest monitoring, restoration planning, and valuable predicators of forest species composition, as tree size metrics (e.g., bole diameter, height, and canopy characteristics) are often species-specific (Wolter and Townsend, 2011). Previous efforts have leveraged multi-resolution satellite sensor data (Wolter et al., 2009) and combinations of multi-sensor and multi-resolution satellite sensor data (Wolter and Townsend, 2011) to model and map forest structure and composition in northern Minnesota. Nevertheless, such efforts to accurately map species-level forest composition in the upper Midwest remains a challenge, as some tree species are spectrally confused (e.g., *Pinus banksiana* and *Picea mariana*)

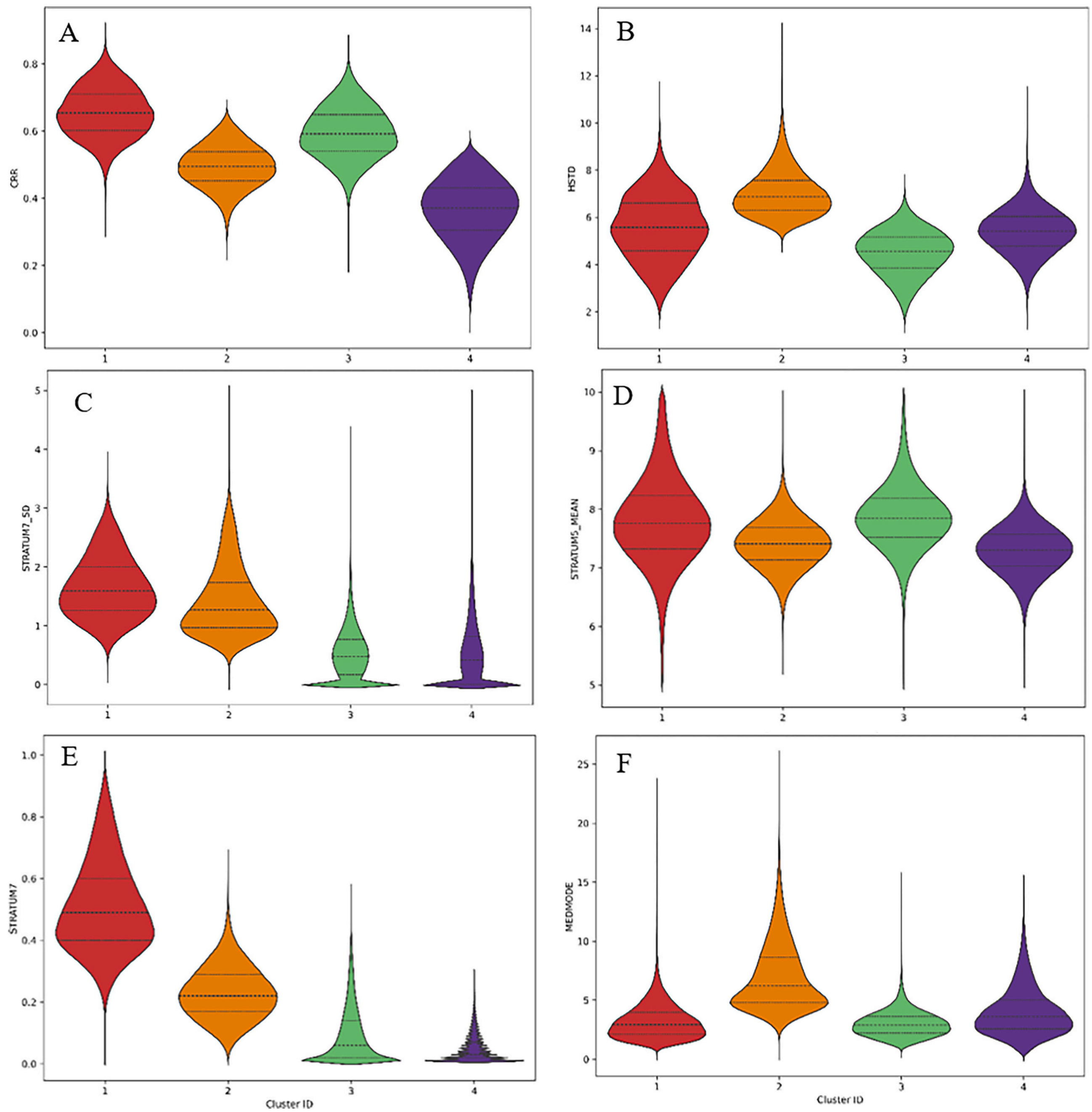

Fig. 10. Violin plots of six key LiDAR-derived canopy structure metrics across the four PCA-based clusters: (A) Canopy Relief Ratio (CRR), (B) Standard Deviation of Height (HSTD), (C) Standard deviation of STRATUM7 (STRATUM7_SD), (D) Mean height of vegetation of STRATUM5 (STRATUM5_MEAN), (E) STRATUM7, and (F) Median absolute deviations from mode height (MEDMODE). These metrics were selected based on their strong contributions to the first six principal components (see Fig. 4) and summarize dominant patterns in vertical forest structure.

or occupy understory positions that make space-borne detection difficult (Wolter et al., 2021; Wolter and Townsend, 2011). However, inclusion of vertical forest structure information (i.e., height, length of live crown, bole diameter, and canopy diameter) has improved forest composition modeling, since many tree species have unique 3D structural characteristics (Wolter and Townsend, 2011). Hence, our LiDAR-based methods and derivatives may serve to overcome the limitations described above, by providing robust, high-resolution, 3D clustering that is crucial for understanding forest ecosystem dynamics.

To clarify the practical value of the derived canopy-structure classes, we outline three common downstream applications of the cluster map (supported by the cluster signatures in Fig. 11):

1. Wildland fire fuels / fire behavior stratification: the cluster map provides spatial strata that summarize vertical structure relevant to fire behavior (e.g., stands with dense mid-canopy and high vertical heterogeneity that are consistent with ladder-fuel-like conditions). We summarize per-cluster distributions of key structural descriptors (e.g., height percentiles and variability/stratification metrics) to show how clusters distinguish fuel-relevant structural regimes and how they could be used as inputs or constraints for canopy-fuel parameterization and risk screening.
2. Habitat structure assessment: many species respond more strongly to structural complexity (height, layering, heterogeneity) than to dominant species alone. The cluster map provides canopy-structure classes that can be used as covariates or strata for habitat

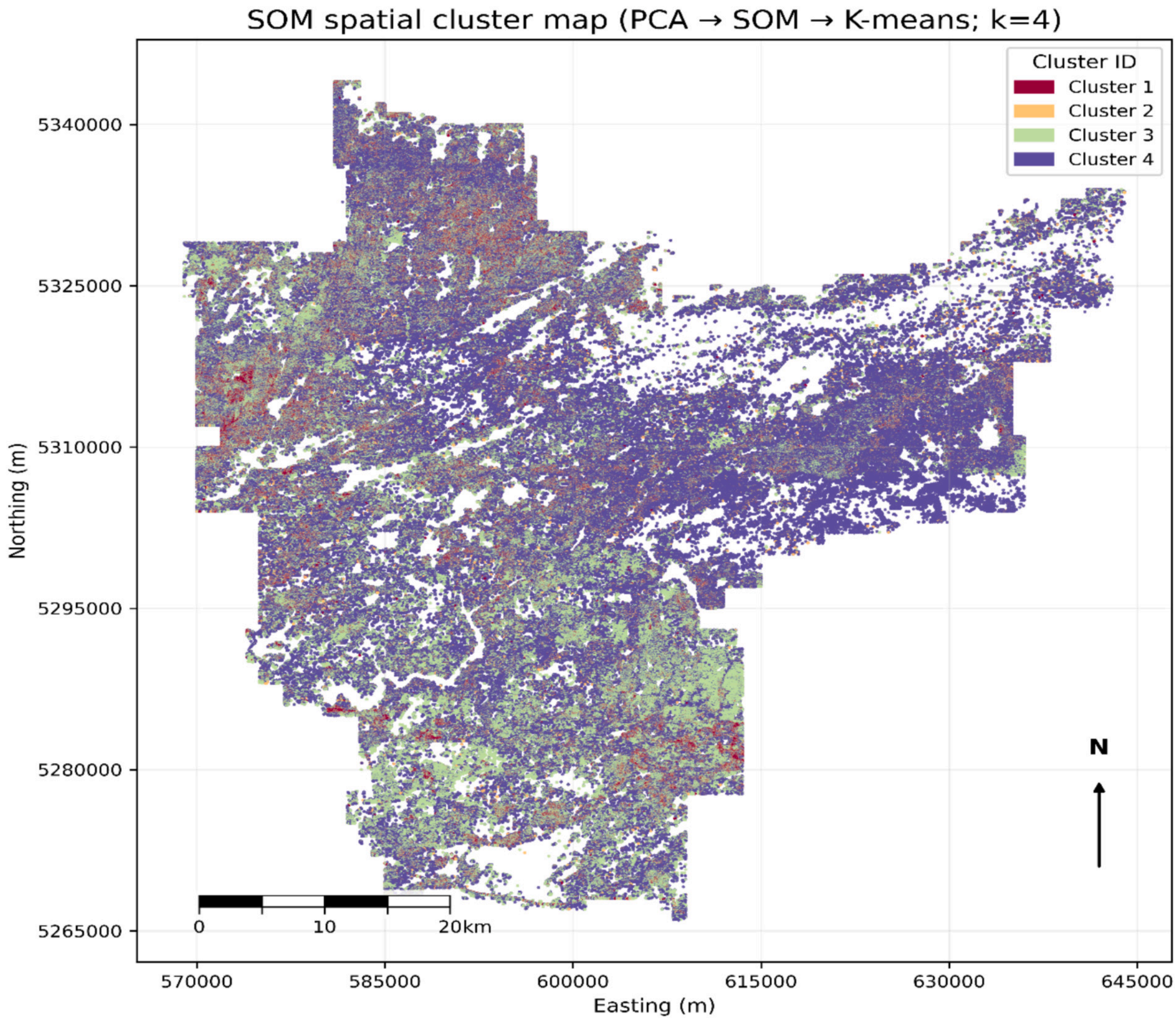

Fig. 11. Spatial distribution of SOM-PCA forest structural clusters across the KRD, illustrating distinct geographic patterns associated with canopy complexity and successional stage across four clusters.

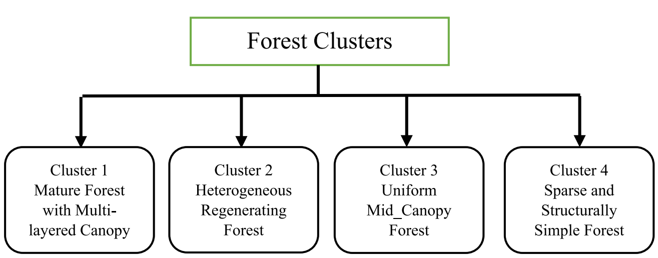

Fig. 12. Flowchart summarizing the four forest structural clusters derived from PCA–SOM–K-means workflow and their corresponding potential planning applications (fuel-relevant stratification, habitat-structure assessment, and monitoring/stratified sampling). Each cluster reflects distinct canopy characteristics—from complex, mature forests (Cluster 1) to sparse, regenerating stands (Cluster 4).

suitability modeling (e.g., identifying patches with tall, multi-layer structure versus short, simple canopies). We report the cluster-wise structural signatures to demonstrate that clusters represent ecologically interpretable habitat structure gradients.

3. Monitoring and stratified sampling: the cluster map offers a consistent structural stratification to guide plot allocation (stratified sampling) and to track change through time as new LiDAR becomes available. Because the classes are defined in a standardized feature space, repeated mapping enables targeted field checks and change detection within comparable structure classes.

Against this applied backdrop, we interpret Clusters 1–4 derived from unsupervised clustering using the structural signatures in Fig. 12, highlighting how each class relates to wildfire-relevant fuel structure, habitat complexity, and monitoring value. Cluster 1 comprises the tallest, most vertically complex forests in the KRD and therefore represents the greatest pools of live biomass and carbon, and role in biodiversity conservation. However, a high, closed overstory in this hemiboreal mixedwood setting does not necessarily translate into reduced ladder-fuel hazards. Shade-tolerant balsam fir (*Abies balsamea*) and other mid-story conifers commonly persist beneath dense canopies throughout northern Minnesota, maintaining continuous live fuels from the forest floor into the lower crowns (Légaré et al., 2002; Messier et al., 1998). Wolter et al. (2021) found no clear relationship between overstory cover type/closure and the abundance of balsam-fir ladder fuels, and earlier LiDAR studies in the region (e.g., Engelstad et al., 2019) showed that moderate-density point clouds could not reliably resolve these shaded strata. While Cluster 1 signifies stands with high crowns and layered structure—attributes often associated with mature, high-value habitat—this structure does not necessarily imply that ladder fuels are absent; understory and midstory vegetation may still provide a continuous vertical pathway for fire to move from surface fuels into the canopy. Field verification or denser LiDAR acquisitions are recommended before prescribing fire-hazard reductions or fuel-treatment priorities in these tall, structurally complex stands.

Cluster 2 reflects regenerating stands with pronounced vertical variability. Such heterogeneity is indicative of early- to mid-successional dynamics following disturbance, aligning with transitional forest conditions described by Franklin et al. (2002) and Ritter et al. (2022). From a fuels and resilience perspective, these stands may warrant closer attention because vertical layering can increase ladder-fuel connectivity. The mapped Cluster 2 areas can support planning by indicating where targeted field checks of vertical structure and fuel connectivity may be most informative, consistent with local management objectives and site constraints.

Cluster 3 is composed of structurally uniform mid-story canopies and moderate biomass. Although less complex than Cluster 1, this group, if predominantly coniferous, may present elevated fire risk due to vertical fuel continuity. These areas provide a spatial basis for screening and prioritizing locations for further assessment of fuel-structure conditions (e.g., vertical continuity) and for scenario-based planning; any actions will ultimately be determined by local objectives, constraints, and site conditions (Finkral and Evans, 2008; Thomas and Gale, 2015).

Cluster 4 comprises sparse, structurally simple patches that most likely correspond to very early-successional or recently disturbed sites. Their open canopy could be desirable—e.g., as wildlife foraging habitat or fuel breaks—or it may signal a need for accelerated recovery, depending on management goals for the Superior National Forest. These areas may warrant prioritized field evaluation to distinguish early-successional conditions from environmentally constrained regeneration and to assess fuel-structure characteristics; they can support scenario-based planning, but any actions will ultimately be determined by local objectives, constraints, and site conditions (Drake et al., 2015; Greenberg et al., 2011; Rodriguez Franco et al., 2024). In short, Cluster 4 represents structurally flexible sites; the optimal intervention, if any, depends on whether the desired future condition is rapid canopy closure, sustained openness, or an intermediate habitat mosaic.

Together, these clusters highlight the potential of machine learning-based classification to support landscape-scale forest monitoring. The alignment of cluster characteristics with known ecological trajectories reinforces the validity of our method.

From a methodological perspective, our PCA–SOM–K-means workflow addresses the challenge of high-dimensional LiDAR data by combining dimensionality reduction and topologically aware clustering. This hybrid approach offers advantages over traditional unsupervised methods by preserving nonlinear structural gradients and improving cluster interpretability. Notably, our use of the U-matrix in SOM supports visual validation of structural transition zones, while the discrete spatial map allows geographic assessment of forest patterns.

Comparison with previous studies further supports our results. For instance, similar structural typologies have been identified in temperate forests using field inventory or airborne LiDAR (e.g., Silva et al., 2016; White et al., 2013), though few have applied machine learning to delineate structural clusters at this resolution and extent. Our approach complements plot-based surveys by offering continuous, wall-to-wall clustering without requiring prior labeling.

However, limitations exist. While the resulting clusters align well with known structural gradients, their ecological functions and temporal stability remain to be validated through field-based measurements. Incorporating ground truth data and supervised classification approaches in future work would likely yield more precise and ecologically interpretable outcomes. Additionally, although our clusters capture how canopy material is distributed vertically, the algorithm is agnostic to the composition of that material. For fuel-mapping and fire-behavior applications, the distinction between live conifer needles, live deciduous foliage, and dead woody debris is critical because each burns under different moisture and wind conditions. Hence, structural clustering should be paired with complementary information—e.g., species-probability layers from multispectral or hyperspectral imagery, LiDAR-intensity–based hardwood/softwood separation, or ground-plot inventories—that identifies the composition and live/dead status of returns within each height band. The choice of which ancillary layers to fuse, and which LiDAR metrics to prioritize, will depend on the management question at hand, underscoring the need for expert judgment when selecting predictor sets, tuning clustering parameters, and interpreting the resulting groups.

Overall, this study illustrates the potential of combining unsupervised machine learning with LiDAR-derived metrics to classify forest structure at high spatial resolution. Such approaches can enhance adaptive forest management strategies and contribute to the resilience of forested landscapes in the face of ongoing environmental change.

5. Conclusion

This study presents a data-driven framework for unsupervised classification of forest canopy structure by integrating 21 LiDAR-derived vertical structure metrics with dimensionality reduction via Principal Component Analysis (PCA) and clustering through Self-Organizing Maps (SOM) over the KRD of the Superior National Forest in northern Minnesota. The resulting structural clusters—assessed through wall-to-wall spatial mapping and interpretation of their LiDAR metric profiles—capture key gradients in canopy height distribution, vertical stratification, and internal heterogeneity. The SOM-based approach preserves topological continuity within the reduced feature space, enabling delineation of ecologically meaningful structural types that reflect forest condition and successional dynamics. Our findings highlight the utility of combining LiDAR and unsupervised learning techniques to support landscape-scale forest structure assessment, particularly for post-disturbance evaluation, restoration planning, and canopy fuel stratification. Moreover, these structure clusters could serve as powerful co-predictors of species-specific forest composition when combined with multi-temporal optical satellite sensor data. While promising, future integration with field-based inventory data and supervised classification methods will be essential to strengthen ecological interpretation and enhance applicability for operational forest monitoring.

CRediT authorship contribution statement

Nasrin Salehnia: Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Peter T. Wolter:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition. **Brian R. Sturtevant:** Writing – review & editing, Project administration, Funding acquisition. **Dalia Abbas Iossifov:** Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by the U.S. Department of Agriculture, Forest Service, under Grant No. 24-JV-11111137-108, and the North Central Research Station, under Grant No. GR-029426-00001. The authors would like to thank Dr. Christel C. Kern (Forest Service, Northern Research Station) for her expert review and constructive feedback on the manuscript. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and should not be construed to represent any official USDA or U.S. Government

determination or policy.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2026.103726.

Data availability

The data and code supporting the findings of this study are openly available in the Mendeley Data repository at doi:10.17632/26pw833vxb.1 under the title "Reproducibility package: PCA–SOM–KMeans clustering of LiDAR-derived structural predictors".

References

Abdi, H., Williams, L.J., 2010. Principal component analysis. Wiley Interdiscipl. Rev.: Comput. Statist. 2 (4), 433–459. https://doi.org/10.1002/wics.101.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J.M., Perona, I., 2013. An extensive comparative study of cluster validity indices. Pattern Recogn. 46 (1), 243–256. https://doi.org/10.1016/j.patcog.2012.07.021.

Asner, G.P., Clark, J.K., Mascaro, J., Galindo García, G.A., Chadwick, K.D., Navarrete Encinales, D.A., Paez-Acosta, G., Cabrera Montenegro, E., Kennedy-Bowdoin, T., Duque, Á., Balaji, A., von Hildebrand, P., Maatoug, L., Phillips Bernal, J.F., Yepes Quintero, A.P., Knapp, D.E., García Dávila, M.C., Jacobson, J., Ordóñez, M.F., 2012. High-resolution mapping of forest carbon stocks in the Colombian Amazon. Biogeosciences 9, 2683–2696. https://doi.org/10.5194/bg-9-2683-2012.

Balestra, M., Marselis, S., Sankey, T.T., et al., 2024. LiDAR data fusion to improve forest attribute estimates: a review. Curr. For. Rep. 10, 281–297. https://doi.org/10.1007/s40725-024-00223-7.

Brandt, J.P., 2009. The extent of the north American boreal zone. Environ. Rev. 17, 101–161. https://doi.org/10.1139/A09-004.

Bro, R., Smilde, A.K., 2003. Principal component analysis. Anal. Methods 6 (9), 2812–2831. https://doi.org/10.1039/C3AY41907J.

Cazzolla Gatti, R., Di Paola, A., Bombelli, A., Noce, S., Valentini, R., 2017. Exploring the relationship between canopy height and terrestrial plant diversity. Plant Ecol. 218 (7), 899–908. https://doi.org/10.1007/s11258-017-0738-6.

Chen, C., Yuan, X., Gan, S., et al., 2024. A new strategy based on multi-source remote sensing data for improving the accuracy of land use/cover change classification. Sci. Rep. 14, 26855. https://doi.org/10.1038/s41598-024-75329-w.

Collins, L., Bradstock, R.A., Clarke, H., Clarke, M.F., Nolan, R.H., Penman, T.D., 2021. The 2019/2020 mega-fires exposed Australian ecosystems to an unprecedented extent of high-severity fire. Environ. Res. Lett. 16, 044029. https://doi.org/10.1088/1748-9326/abeb9e.

Coops, N.C., Tompalski, P., Goodbody, T.R.H., Queinnec, M., Luther, J.E., et al., 2021. Modelling lidar-derived estimates of forest attributes over space and time: a review of approaches and future trends. Remote Sens. Environ. 260. https://doi.org/10.1016/j.rse.2021.112477.

Dalponte, M., Coomes, D.A., 2016. Tree-centric mapping of forest carbon density from airborne laser scanning and hyperspectral data. Methods Ecol. Evol. 7 (10), 1236–1245. https://doi.org/10.1111/2041-210X.12575.

Drake, J.A., Carrucan, A., Jackson, W.R., Cavagnaro, T.R., Patti, A.F., 2015. Biochar application during reforestation alters species present and soil chemistry. Sci. Total Environ. 514, 359–365. https://doi.org/10.1016/j.scitotenv.2015.02.012.

Engelstad, P.S., Falkowski, M., Wolter, P., Poznanovic, A., Johnson, P., 2019. Estimating canopy fuel attributes from low-density LiDAR. Fire 2 (3), 38. https://doi.org/10.3390/fire2030038.

Esri, 2024. ArcGIS Pro, Version 3.5.2. Environmental Systems Research Institute, Redlands, CA, USA. Available at: https://www.esri.com/en-us/arcgis/products/arcgis-pro/overview.

Finkral, A.J., Evans, A.M., 2008. The effects of a thinning treatment on carbon stocks in a northern Arizona ponderosa pine forest. For. Ecol. Manag. 255 (7), 2743–2750. https://doi.org/10.1016/j.foreco.2008.01.041.

Forest, F., Lebbah, M., Azzag, H., et al., 2021. Deep embedded self-organizing maps for joint representation learning and topology-preserving clustering. Neural Comput. & Applic. 33, 17439–17469. https://doi.org/10.1007/s00521-021-06331-w.

Franklin, J.F., Spies, T.A., Pelt, R.V., Carey, A.B., Thornburgh, D.A., et al., 2002. Disturbances and structural development of natural forest ecosystems with silvicultural implications, using Douglas-fir forests as an example. For. Ecol. Manag. 155, 399–423. https://doi.org/10.1016/s0378-1127(01)00575-8.

Fujino, M., Yoshida, M., 2006. Development and validation of a method of forestry region classification using PCA and cluster analysis together with SOM algorithm. J. Jpn. Forest Soc. 88 (4), 221–230. https://doi.org/10.4005/jjfs.88.221.

Graves, S.J., Caughlin, T.T., Asner, G.P., Bohlman, S.A., 2018. A tree-based approach to biomass estimation from remote sensing data in a tropical agricultural landscape. Remote Sens. Environ. 218, 32–43. https://doi.org/10.1016/j.rse.2018.09.009.

Greenberg, C.H., Collins, B., Thompson, F.R., 2011. Sustaining young forest communities. Chapter 1 introduction: what are early successional habitats, why are they important, and how can they be sustained? Manag. Forest Ecosyst. 21, 1–10. https://link.springer.com/content/pdf/10.1007/978-94-007-1620-9.pdf.

Guo, Q., Liang, X., Li, W., Jin, S., Guan, H., et al., 2022. LiDAR remote sensing of Forest ecosystems: Applications and prospects. In: Li, B., Shi, X., Zhu, A.X., Wang, C., Lin, H. (Eds.), New Thinking in GIScience. Springer, Singapore. https://doi.org/10.1007/978-981-19-3816-0_24.

Han, J., Kamber, M., Pei, J., 2012. Data Mining: Concepts and Techniques, 3rd ed. Morgan Kaufmann.

Hopton, M.E., Mayer, A.L., 2006. Using self-organizing maps to explore patterns in species richness and protection. Biodivers. Conserv. 15, 4477–4494. https://doi.org/10.1007/s10531-005-5099-0.

Hudak, A.T., Bright, B.C., Pokswinski, S.M., Loudermilk, E.L., O'Brien, J.J., et al., 2016. Mapping forest structure and composition from low-density LiDAR for informed forest, fuel, and fire management at Eglin air Force Base, Florida, USA. Can. J. Remote. Sens. 42 (5), 411–427. https://doi.org/10.1080/07038992.2016.1217482.

Jolliffe, I.T., 2002. Principal Component Analysis, 2nd ed. Springer.

Jucker, T., Bongalov, B., Burslem, D.F.R.P., Nilus, R., Dalponte, M., Lewis, S.L., Phillips, O.L., Qie, L., Coomes, D.A., 2018. Topography shapes the structure, composition and function of tropical forest landscapes. Ecol. Lett. 21 (7), 989–1000. https://doi.org/10.1111/ele.12964.

Klobucar, D., Subasic, M., 2012. Using self-organizing maps in the visualization and analysis of forest inventory. iForest 5, 216–223. https://doi.org/10.3832/ifor0629-005.

Kohonen, T., 2001. Self-Organizing Maps, 3rd edition. Springer, New York. https://doi.org/10.1007/978-3-642-56927-2.

Kohonen, T., 2013. Essentials of the self-organizing map. Neural Netw. 37, 52–65. https://doi.org/10.1016/j.neunet.2012.09.018.

Légaré, S., Bergeron, Y., Paré, D., 2002. Influence of forest composition on understory cover in boreal mixedwood forests of western Quebec. Silva Fennica 36 (1), 353–366. https://doi.org/10.14214/sf.567.

Marselis, S.M., Abernethy, K., Alonso, A., Armston, J., Baker, T.R., et al., 2020. Evaluating the potential of full-waveform lidar for mapping pan-tropical tree species richness. Glob. Ecol. Biogeogr. 29 (10), 1799–1816. https://doi.org/10.1111/geb.13158.

Messier, C., Parent, S., Bergeron, Y., 1998. Effects of overstory and understory vegetation on the understory light environment in mixed boreal forests. J. Veg. Sci. 9 (4), 511–520. https://doi.org/10.2307/3237266.

Mingoti, S.A., Lima, J.O., 2006. Comparing SOM neural network with fuzzy c-means, K-means and traditional hierarchical clustering algorithms. Eur. J. Oper. Res. 174 (3), 1742–1759. https://doi.org/10.1016/j.ejor.2005.03.039.

Niknam, T., Taherian Fard, E., Pourjafarian, N., Rousta, A., 2011. An efficient hybrid algorithm based on modified imperialist competitive algorithm and K-means for data clustering. Eng. Appl. Artif. Intell. 24, 306–317. https://doi.org/10.1016/j.engappai.2010.10.001.

Olszewski, J.H., Bailey, J.D., 2022. LiDAR as a tool for assessing change in vertical fuel continuity following restoration. Forests 13 (4), 503. https://doi.org/10.3390/f13040503.

Park, Y.S., Chon, T.S., Bae, M.J., Kim, D.H., Lek, S., 2018. Multivariate data analysis by means of self-organizing maps. In: Recknagel, F., Michener, W. (Eds.), Ecological Informatics. Springer, Cham. https://doi.org/10.1007/978-3-319-59928-1_12.

Parsons, R.A., Mell, W.E., McCauley, P., 2011. Linking 3D spatial models of fuels and fire: effects of spatial heterogeneity on fire behavior. Ecol. Model. 222 (3), 679–691. https://doi.org/10.1016/j.ecolmodel.2010.10.023.

Pedregosa, F., Varoquaux, G., Gramfort, A., et al., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Ponmalai, R., Kamath, C., 2019. Self-Organizing Maps and Their Applications to Data Analysis. https://doi.org/10.2172/1566795.

Ritter, S.M., Hoffman, C.M., Battaglia, M.A., Jain, T.B., 2022. Restoration and fuel hazard reduction result in equivalent reductions in crown fire behavior in dry conifer forests. Ecol. Appl. 32 (7), e2682. https://doi.org/10.1002/eap.2682.

Rodriguez Franco, C., Page-Dumroese, D.S., Pierson, D., Nicosia, T., 2024. Biochar utilization as a forestry climate-smart tool. Sustainability 16 (5), 1714. https://doi.org/10.3390/su16051714.

Salehnia, N., Salehnia, N., Ansari, H., et al., 2019. Climate data clustering effects on arid and semi-arid rainfed wheat yield: a comparison of artificial intelligence and K-means approaches. Int. J. Biometeorol. 63, 861–872. https://doi.org/10.1007/s00484-019-01699-w.

Salehnia, N., Wolter, P., Sturtevant, B., Abbas Iossifov, D., 2026. Reproducibility package: PCA–SOM–KMeans clustering of LiDAR-derived structural predictors. In: Mendeley Data. https://doi.org/10.17632/26pw833vxb.1. V1.

Shao, J., Lin, Y.C., Wingren, C., Shin, S.Y., et al., 2024. Large-scale inventory in natural forests with mobile LiDAR point clouds. Sci. Remote Sens. 10, 100168. https://doi.org/10.1016/j.srs.2024.100168.

Silva, C.A., Hudak, A.T., Vierling, L.A., Loudermilk, E.L., O'Brien, J.J., Hiers, et al., 2016. Imputing individual longleaf pine (*Pinus palustris* Mill.) tree attributes from field and LiDAR data. Can. J. Remote. Sens. 42 (5), 554–573. https://doi.org/10.1080/07038992.2016.1196582.

Skowronski, N.S., Gallagher, M.R., Warner, T.A., 2020. Decomposing the interactions between fire severity and canopy fuel structure using multi-temporal, active, and passive remote sensing approaches. Fire 3 (1), 7. https://doi.org/10.3390/fire3010007.

Stavros, E.N., Coen, J., Peterson, B., Singh, H., et al., 2018. Use of imaging spectroscopy and LIDAR to characterize fuels for fire behavior prediction. Remote Sens. Appl.: Soc. Environ. 11, 41–50. https://doi.org/10.1016/j.rsase.2018.04.010.

Sturtevant, B.R., Miranda, B.R., Shinneman, D.J., Gustafson, E.J., Wolter, P.T., 2012. Comparing modern and presettlement forest dynamics of a subboreal wilderness: does spruce budworm enhance fire risk? Ecol. Appl. 22 (4), 1278–1296. https://research.fs.usda.gov/treesearch/40930.

Sulkava, M., Hollmén, J., 2003. Finding profiles of forest nutrition by clustering of the SOM. In: Proceedings of the Workshop on "Self-organizing maps". Kitakyushu (Japan) 11-14 September 2003, pp. 243–248.

Sun, Z., Sonsuthi, A., Jucker, T., Ali, A., Cao, M., Liu, F., Cao, G., Hu, T., Ma, Q., Guo, Q., Lin, L., 2023. Top canopy height and stem size variation enhance aboveground biomass across spatial scales in seasonal tropical forests. Plants 12 (6), 1343. https://doi.org/10.3390/plants12061343.

Swatantran, A., Dubayah, R., Roberts, D., Hofton, M., Blair, B., 2011. Mapping biomass and stress in the Sierra Nevada using lidar and hyperspectral data fusion. Remote Sens. Environ. 115 (11), 2917–2930. https://doi.org/10.1016/j.rse.2010.08.027.

Thomas, S.C., Gale, N., 2015. Biochar and forest restoration: a review and meta-analysis of tree growth responses. New For. 46, 931–946. https://doi.org/10.1007/s11056-015-9491-7.

Torresan, C., Corona, P., Scrinzi, G., Marsal, J.V., 2016. Using classification trees to predict forest structure types from LiDAR data. Ann. For. Res. 59 (2). https://doi.org/10.15287/afr.2016.423.

Verheyen, K., Gillerot, L., Blondeel, H., De Frenne, P., De Pauw, K., Depauw, L., et al., 2024. Forest canopies as nature-based solutions to mitigate global change effects on people and nature. J. Ecol. 112 (11), 2451–2461. https://doi.org/10.1111/1365-2745.14345.

Vesanto, J., Alhoniemi, E., 2000. Clustering of the self-organizing map. IEEE Trans. Neural Netw. 11 (3), 586–600. https://doi.org/10.1109/72.846731.

Vettigli, G., 2018. MiniSom: Minimalistic and NumPy-Based Implementation of the Self Organizing Map. GitHub repository.

White, J.C., Wulder, M.A., Vastaranta, M., Coops, N.C., Pitt, D., Woods, M., 2013. The utility of image-based point clouds for Forest inventory: a comparison with airborne laser scanning. Forests 4 (3), 518–536. https://doi.org/10.3390/f4030518.

Wolter, P.T., Townsend, P.A., 2011. Multi-sensor data fusion for estimating forest species composition and abundance in northern Minnesota. Remote Sens. Environ. 115 (2), 671–691. https://doi.org/10.1016/j.rse.2010.10.010.

Wolter, P.T., Townsend, P.A., Sturtevant, B.R., 2009. Estimation of forest structural parameters using 5 and 10 meter SPOT-5 satellite data. Remote Sens. Environ. 113 (9), 2019–2036. https://doi.org/10.1016/j.rse.2009.05.009.

Wolter, P.T., Olbrich, J.J., Johnson, P.J., 2021. Modeling sub-boreal forest canopy bulk density in Minnesota, USA, using synthetic aperture radar and optical satellite sensor data. Fire Ecol. 17, 26. https://doi.org/10.1186/s42408-021-00112-5.

Zhang, C., Qiu, F., 2012. Mapping individual tree species in an urban forest using airborne Lidar data and hyperspectral imagery. Photogramm. Eng. Remote. Sens. 78 (10), 1079–1087. https://doi.org/10.14358/PERS.78.10.1079.