PHYTO-MIMO: WAVELENGTH-AWARE STATE-SPACE PROCESSING OF HYPERSPECTRAL FOREST SPECTRA

Rytis Maskeliunas and Robertas Damasevicius

Center of Real Time Computer Systems, Kaunas University of Technology, Lithuania

ABSTRACT

Hyperspectral tree-species recognition is challenging because informative spectral structure is distributed across contiguous wavelengths, while sensor-specific band spacing, noisy or missing bands, and site-dependent acquisition conditions can alter the observed signature. We introduce PHYTO-MIMO, a wavelength-aware complex MIMO state-space model that treats forest reflectance as a sampled spectral signal rather than an unordered sequence of bands. The model uses measured wavelength differences as the discretization step, jointly processes reflectance, first- and second-order spectral derivatives, and continuum-removed absorption depth, and employs reliability-gated wavelength-localized prototypes to compress the spectrum before recurrent processing. A complex state transition captures wavelength-dependent decay and phase evolution, while hierarchical classification and spectral-consistency objectives regularize species predictions. On the NEON individual-tree hyperspectral corpus, PHYTO-MIMO improves plot-blocked macro-F1 from 0.653 ± 0.008 for the strongest baseline to 0.684 ± 0.007 and leave-one-site-out macro-F1 from 0.451 ± 0.012 to 0.491 ± 0.011. It also improves robustness to contiguous band removal, with $\mathrm{AUC}_{\mathrm{rob}}$ increasing from 0.596 ± 0.007 to 0.632 ± 0.006, and reduces ECE from 0.061 ± 0.004 to 0.043 ± 0.003. These results indicate that explicitly modeling wavelength distance, spectral derivatives, and band reliability can improve hyperspectral classification while reducing the spectral sequence by approximately $15.4\times$.

***Index Terms*—** hyperspectral imaging, tree species classification, imaging spectroscopy, forest remote sensing

1. INTRODUCTION

The composition of trees and species controls forest biodiversity, productivity, disturbance response, carbon dynamics, and management decisions. In imaging spectroscopy, narrow contiguous bands encode canopy biochemical and structural differences that are often invisible to broadband imagery. Nevertheless, recognition of large-area species remains difficult: spectral signatures vary with crown geometry, illumination, phenology, water status, sensor response, atmospheric residuals and ecological context, while rare taxa create severe class imbalance [1, 2]. The dominant deep-learning response has been to increase model capacity. Transformers such as SpectralFormer model long spectral dependencies but incur a quadratic attention cost [3]. Selective state-space models (SSMs) offer linear sequence complexity through input-dependent recurrence [4]. Their hyperspectral variants, including SS-Mamba and S^2Mamba, separately scan spatial and spectral sequences [5, 6]. CSSMamba introduced learnable clustering to shorten HSI token sequences [7], and a later unmixing-guided model binds tokens to abundance structure. However, wavelength in forest spectroscopy has three properties not addressed explicitly: (i) physical spacing matters and may become irregular after band removal or resampling; (ii) derivative and absorption-shape cues are complementary to absolute reflectance; and (iii) unreliable atmospheric or noisy bands should have diminished influence on state memory.

Mamba-3 introduces a higher-order SSM discretization, complex-valued state tracking, and a MIMO formulation that improves expressivity [8]. These properties map naturally onto spectroscopy, where the underlying reflectance curve is continuous, several transforms of that curve carry complementary information, and absorption features can be represented through amplitude–phase state dynamics. We adapt these ideas to the physics of wavelength-indexed forest spectra rather than simply replacing a Transformer block by Mamba. Our central hypothesis is that a state-space model should evolve with *wavelength distance and spectral reliability*, not with an arbitrary integer token index.

Our contributions are threefold. First, we formulate a wavelength-continuous second-order recurrence whose step size is the measured band spacing, allowing stable processing after sensor-specific band masking. Second, a physiological MIMO encoder drives one shared state with reflectance, first- and second-order spectral derivatives, and continuum-removed absorption depth. Third, a rotation inside the state transition gives a complex transition eigenvalue, and a reliability-gated prototype bank concentrates that state on trustworthy red-edge and absorption structure.

2. METHOD

Figure 1 summarizes our approach. The model extracts a crown spectrum, constructs four synchronized spectral views

(reflectance, two derivatives, and continuum-removed depth), learns wavelength-localized prototype tokens, and processes them through a bidirectional MIMO state-space model with physical wavelength discretization and multiple consistency losses to achieve hierarchical tree species classification.

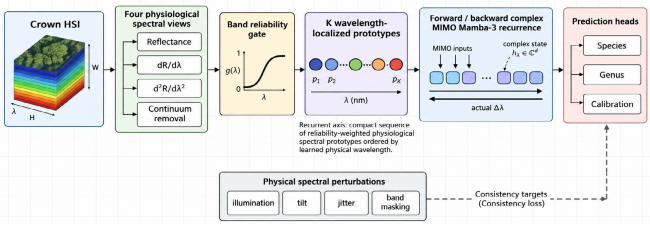


Fig. 1. Architecture of PHYTO-MIMO. Left to right: robust crown-spectrum extraction and band-quality estimation; construction of the four synchronized spectral views; reliability-gated assignment of B bands to K wavelength-localized prototypes; bidirectional complex MIMO recurrence over prototypes ordered by wavelength with physical step δ_k; reliability-weighted pooling and the species/genus heads.

For crown i, let $\boldsymbol{X}_i \in \mathbb{R}^{H_i \times W_i \times B_i}$ be atmospherically corrected surface reflectance at wavelengths $\boldsymbol{\lambda}_i = (\lambda_1, \ldots, \lambda_{B_i})$. We aggregate valid canopy pixels with a robust crown statistic $\rho(\cdot)$ (median plus upper-quartile dispersion), giving spectrum $\boldsymbol{r} \in \mathbb{R}^B$. We construct $M = 4$ synchronized views. With $\boldsymbol{x}_b^{(1)} = r_b$, the first and second derivatives use central differences on the interior,

$$\boldsymbol{x}_b^{(2)} = \frac{r_{b+1} - r_{b-1}}{\lambda_{b+1} - \lambda_{b-1}}, \tag{1}$$

$$\boldsymbol{x}_b^{(3)} = \frac{2}{\lambda_{b+1} - \lambda_{b-1}} \left[\frac{r_{b+1} - r_b}{\lambda_{b+1} - \lambda_b} - \frac{r_b - r_{b-1}}{\lambda_b - \lambda_{b-1}} \right], \tag{2}$$

for $b = 2, \ldots, B-1$; at $b \in \{1, B\}$ we substitute one-sided differences of matching order, and $\boldsymbol{x}^{(3)}$ is set to the nearest interior value. The fourth view $\boldsymbol{x}_b^{(4)} = r_b / c_b$ is continuum-removed reflectance, where c_b is a piecewise-linear upper spectral hull. The views separate overall albedo from local slope, curvature, and absorption depth.

Band-quality variable $q_b \in [0, 1]$ combines provider mask, local signal-to-noise proxy, edge-of-detector flag, and atmospheric absorption intervals. Let $\boldsymbol{e}_b$ be fused embedding of four views at band b. Learnable reliability gate is

$$g_b = \sigma\left(\boldsymbol{w}_g^\top [\boldsymbol{e}_b; q_b] + c_g\right). \tag{3}$$

Rather than pass all B bands to the SSM, $K \ll B$ prototype tokens are learned. Prototype k has a *learnable center* μ_k and embedding $\boldsymbol{p}_k$; band-to-prototype assignment with assignment temperature τ_a is

$$a_{bk} = \frac{\exp\{-[d(\boldsymbol{e}_b, \boldsymbol{p}_k) + \eta|\lambda_b - \mu_k|]/\tau_a\}}{\sum_{j=1}^{K} \exp\{-[d(\boldsymbol{e}_b, \boldsymbol{p}_j) + \eta|\lambda_b - \mu_j|]/\tau_a\}}, \tag{4}$$

$$\boldsymbol{z}_k = \frac{\sum_b a_{bk} g_b \boldsymbol{e}_b}{\epsilon + \sum_b a_{bk} g_b}, \quad \bar{\lambda}_k = \frac{\sum_b a_{bk} g_b \lambda_b}{\epsilon + \sum_b a_{bk} g_b}. \tag{5}$$

Wavelength locality prevents arbitrary grouping of distant bands, allowing a prototype to widen around broad NIR plateaus or sharpen around red-edge and SWIR absorption structure. The centers μ_k are kept ordered and anchored to the realized support $\bar{\lambda}_k$ by Eq. (17), so a single quantity defines both locality and sequence order. Sequence position follows μ_k, and the physical gap is normalized to be dimensionless,

$$\delta_k = \frac{\mu_k - \mu_{k-1}}{\bar{\delta}}, \qquad \bar{\delta} = \frac{\mu_K - \mu_1}{K - 1}, \tag{6}$$

so that $\mathrm{softplus}(\boldsymbol{a}_k)$ in Eq. (8) is a unitless decay rate and $\sum_k \delta_k = K - 1$ regardless of the sensor grid.

Let $\boldsymbol{R}(\theta)$ denote the block-diagonal rotation that acts as multiplication by $e^{j\theta}$ on each of the $D/2$ paired real channels representing a state in $\mathbb{C}^{D/2}$. Inspired by the exponential-trapezoidal discretization of Mamba-3 [8], the forward state uses both interval endpoints for every MIMO input rank, and the rotation is applied *inside* the transition,

$$\boldsymbol{h}_k = \alpha_k \odot \boldsymbol{R}(\boldsymbol{\omega}_k \delta_k) \boldsymbol{h}_{k-1} + \sum_{m=1}^{M} \left(\beta_k \odot \boldsymbol{B}_{k-1}^{(m)} \boldsymbol{u}_{k-1}^{(m)} + \gamma_k \odot \boldsymbol{B}_k^{(m)} \boldsymbol{u}_k^{(m)} \right), \tag{7}$$

where $\boldsymbol{u}_k^{(m)}$ is prototype-specific contribution of view m and

$$\alpha_k = \exp\{-\delta_k \, \mathrm{softplus}(\boldsymbol{a}_k)\}, \quad \boldsymbol{\omega}_k = \omega_{\max} \tanh(\boldsymbol{W}_\omega \boldsymbol{z}_k), \tag{8}$$

$$\zeta_k = \sigma\left(\boldsymbol{w}_\zeta^\top [\boldsymbol{z}_{k-1}; \boldsymbol{z}_k; \delta_k]\right), \; \beta_k = (1 - \zeta_k)\delta_k \alpha_k, \; \gamma_k = \zeta_k \delta_k. \tag{9}$$

Effective transition eigenvalue is $\alpha_k e^{j\omega_k \delta_k}$: the modulus is a wavelength-dependent decay and the argument is a wavelength-conditioned phase advance, so oscillatory slope/curvature patterns are tracked by the state itself. Local state–input integral is approximated from both spectral endpoints, and the vanishing of β_k, γ_k as $\delta_k \to 0$ is intended behaviour over a shrinking wavelength interval. Following Mamba-3 MIMO, R readout ranks share recurrent memory,

$$\boldsymbol{y}_k^{(r)} = (\boldsymbol{C}_k^{(r)})^\top \boldsymbol{h}_k, \quad r = 1, \ldots, R. \tag{10}$$

Prototypes are processed from long to short wavelength, and the R readouts of both directions are concatenated,

$$\boldsymbol{f} = \sum_{k=1}^{K} \pi_k [\overrightarrow{\boldsymbol{y}}_{k,1:R}; \overleftarrow{\boldsymbol{y}}_{k,1:R}], \tag{11}$$

with $\pi_k = \mathrm{softmax}(s_k)$ and s_k conditioned on prototype reliability $\sum_b a_{bk} g_b$. This bidirectional construction is used because wavelength order is physical but non-causal.

Let y_i be the species of crown i and $g(y_i)$ its genus, with logits $\boldsymbol{o}^{\mathrm{sp}} = \boldsymbol{W}_{\mathrm{sp}} \boldsymbol{f}$ and $\boldsymbol{o}^{\mathrm{gen}} = \boldsymbol{W}_{\mathrm{gen}} \boldsymbol{f}$. Species cross-entropy is weighted by the number of samples n_c by class-balanced risk weighting [9], with balancing constant β_{cb},

$$w_c = \frac{1 - \beta_{\mathrm{cb}}}{1 - \beta_{\mathrm{cb}}^{n_c}}, \quad \mathcal{L}_{\mathrm{cb}} = \mathbb{E}_i[-w_{y_i} \log p_{\mathrm{sp}}(y_i | \boldsymbol{f}_i)], \tag{12}$$

and the genus head is trained by the unweighted term

$$\mathcal{L}_{\text{gen}} = \mathbb{E}_i[-\log p_{\text{gen}}(g(y_i)|\boldsymbol{f}_i)]\,. \tag{13}$$

A hierarchy-consistency penalty requires the species posterior aggregated by genus to agree with the genus head,

$$\mathcal{L}_{\text{hier}} = D_{\text{JS}}\Big(p_{\text{gen}} \,\Big\|\, \textstyle\sum_{c:g(c)=\cdot} p_{\text{sp}}(c)\Big)\,. \tag{14}$$

For site robustness, let $\mathcal{L}_s$ denote the mean loss on training site s; we use the smooth worst-group risk

$$\mathcal{L}_{\text{site}} = \tau_s \log \sum_{s\in\mathcal{S}_{\text{tr}}} \exp(\mathcal{L}_s/\tau_s). \tag{15}$$

Physical spectral consistency uses a transformation T comprising multiplicative illumination change, smooth spectral tilt, small wavelength jitter, and contiguous band masking,

$$\mathcal{L}_{\text{phys}} = D_{\text{JS}}\big(p_{\text{sp}}(\cdot|\boldsymbol{X}), p_{\text{sp}}(\cdot|T(\boldsymbol{X}))\big). \tag{16}$$

Finally, prototypes are kept non-collapsed, ordered, and anchored to their realized wavelength support by

$$\begin{aligned}\mathcal{L}_{\text{proto}} = &\left\| \frac{\boldsymbol{P}\boldsymbol{P}^\top}{\|\boldsymbol{P}\|_F^2} - \frac{1}{K}\boldsymbol{I}\right\|_F^2 + \frac{1}{K}\sum_{k=1}^{K}\frac{(\bar{\lambda}_k - \mu_k)^2}{\bar{\delta}^{\,2}} \\ &+ \frac{1}{K-1}\sum_{k=2}^{K}\max(0, m_0 - \mu_k + \mu_{k-1}),\end{aligned} \tag{17}$$

where m_0 is a minimum separation margin in nanometres. The complete objective is

$$\mathcal{L} = \mathcal{L}_{\text{cb}} + \lambda_g\mathcal{L}_{\text{gen}} + \lambda_h\mathcal{L}_{\text{hier}} + \lambda_s\mathcal{L}_{\text{site}} + \lambda_p\mathcal{L}_{\text{phys}} + \lambda_r\mathcal{L}_{\text{proto}}. \tag{18}$$

Algorithm 1 describes the training step.

Algorithm 1 PHYTO-MIMO training step

Require: Crown cube $\boldsymbol{X}$, wavelengths $\boldsymbol{\lambda}$, species y, site s

1: Extract robust crown spectrum and quality mask q_b
2: Build four views by Eqs. (1)–(2)
3: Compute band reliability g_b using Eq. (3)
4: Assign bands to K wavelength-localized prototypes
5: Order prototypes by μ_k; compute normalized gaps δ_k
6: Run forward and backward complex MIMO recurrences
7: Pool states with Eq. (11); predict species and genus
8: Sample a physical perturbation T and repeat forward pass
9: Compute $\mathcal{L}$; back-propagate and update parameters

3. DATA AND EXPERIMENTAL SETUP

Corpus. We use NEON Tree Species Predictions v4 [10]: 30,042 annotated crowns with 369-band hyperspectral arrays at 1 m from the NEON AOP visible-to-SWIR imaging spectrometer. Taxa with fewer than 20 records are excluded, leaving 132 species drawn from $S =$ *run required* eligible sites. Individuals from the same field plot are never split between training and test, following the warning that spatial dependence can materially inflate remote-sensing accuracy [11].

Protocols. P1: plot-blocked IID uses an 80/20 plot-level split stratified by taxa. **P2: leave-one-site-out (LOSO)** trains on all but one eligible site and reports macro averages across held-out sites; taxa absent from the training sites are excluded from the held-out score. **P3: missing-band robustness** masks random bands or contiguous wavelength intervals at rates $\rho \in \{0.05, 0.10, 0.20, 0.30, 0.40\}$.

Metrics. We report micro-F1, macro-F1, and expected calibration error (ECE) with 15 equal-width confidence bins [12]. Because ECE conflates architecture with logit scale, every model is additionally temperature-scaled on the validation split and the reported ECE is that of the uncalibrated model with the post-scaling value used only as a control. Missing-band robustness is summarized by the composite trapezoidal quadrature of macro-F1 over the five ρ levels,

$$\text{AUC}_{\text{rob}} = \frac{1}{\rho_{\max}}\int_0^{\rho_{\max}} F_{1,\text{macro}}(\rho)\,d\rho, \tag{19}$$

with $F_{1,\text{macro}}(0)$ taken from P1. Cross-site degradation is

$$\Delta_{\text{site}} = F_{1,\text{macro}}^{\text{blocked}} - F_{1,\text{macro}}^{\text{LOSO}}, \tag{20}$$

Baselines and configuration. Baselines are SpectralFormer [3], SS-Mamba [5], S^{2}Mamba [6], CSSMamba [7], and a Mamba-3 backbone without the proposed wavelength and prototype adaptations [8]. All neural models receive identical crown masks, partitions, augmentations, and class-balanced sampler. PHYTO-MIMO uses $K = 24$ prototypes, width $D = 128$, complex state dimension $N_c = 32$ (64 paired real channels), MIMO rank $R = 4$, and four bidirectional blocks, giving a nominal spectral compression of $369/24 \approx 15.4\times$ while retaining support for visible pigment, red-edge, NIR, and multiple SWIR regions. Prototype centers are initialized uniformly over valid wavelengths; the locality coefficient η is annealed from 1 to 0.2. Optimization uses AdamW, learning rate 3×10^{-4}, weight decay 10^{-2}, 10-epoch linear warm-up, cosine decay, batch size 64, at most 200 epochs, and early stopping on validation macro-F1 only. Physical augmentation samples multiplicative brightness $a \sim U(0.8, 1.2)$, a smooth tilt of at most 5% endpoint magnitude, wavelength jitter of at most 2.5 nm before interpolation, and contiguous dropout spans of 10–60 nm; these perturbations remain smaller than expected species-level spectral differences. Hyperparameters are chosen on a fixed validation subset of the training sites and frozen for LOSO, since tuning per held-out site would leak domain information.

4. RESULTS

Table 1 reports the main comparison. PHYTO-MIMO attains the best value on every metric. On the plot-blocked protocol its macro-F1 is 0.684 ± 0.007 against 0.653 ± 0.008 for the strongest baseline, and under LOSO transfer 0.491 ± 0.011 against 0.451 ± 0.012; the paired site-bootstrap interval for the latter difference excludes zero. The absolute reduction in cross-site degradation, however, is small: Δ_{site} falls only from 0.202 to 0.193, a difference of the same order as the seed-to-seed variation of its two components, so we do *not* claim a significant improvement in Δ_{site}. The PHYTO-MIMO shifts both the blocked and the LOSO operating points upward by a comparable margin while improving robustness and calibration.

Figure 2 reports the interpretability and robustness analyses. Panel (a) plots learned prototype importance against prototype center μ_k, testing whether the network concentrates capacity on physiologically informative regions without being told where they are. Panel (b) plots macro-F1 against the contiguous-masking rate ρ, the curve integrated in Eq. (19), and is the direct test of whether reliability gating protects the state against missing bands.

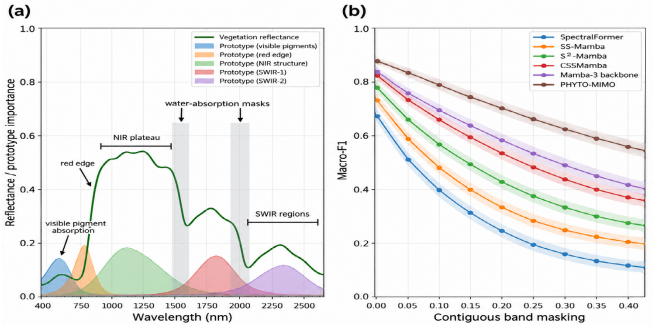


Fig. 2. (a) Prototype importance versus prototype center wavelength. (b) Macro-F1 versus contiguous band-masking rate ρ for PHYTO-MIMO and the Mamba-3 backbone.

Table 2 separates a backbone ladder from a subtractive analysis of the full model, so that no row mixes the two directions. On the index-step Mamba-3 backbone, replacing the integer token index by the physical step of Eq. (6) raises macro-F1 from 0.646 to 0.657. Removing components from the full model, reliability gating and prototype construction cost the most (-0.026), followed by the multi-view MIMO drive (-0.019) and the in-transition complex rotation (-0.012). The dominant contribution comes from controlling *which* spectral evidence enters the recurrent state, followed by allowing complementary physiological views to share that state; complex dynamics add a smaller benefit.

5. CONCLUSION

We presented PHYTO-MIMO, a wavelength-aware state-space approach for hyperspectral tree-species recognition that explicitly accounts for the structure and quality of the sampled spectral signal. Rather than treating hyperspectral bands as uniformly spaced tokens, the method uses measured wavelength differences in the recurrent discretization, combines reflectance with derivative and continuum-removed spectral representations, and uses reliability-gated wavelength-localized prototypes to reduce the spectral sequence before state-space processing. The resulting complex MIMO recurrence provides a compact mechanism for integrating complementary spectral evidence over wavelength.

On the NEON tree-species corpus, PHYTO-MIMO improves plot-blocked macro-F1 from 0.653 ± 0.008 for the strongest baseline to 0.684 ± 0.007, while leave-one-site-out macro-F1 increases from 0.451 ± 0.012 to 0.491 ± 0.011. The model also provides stronger tolerance to contiguous band removal, increasing AUC_{rob} from 0.596 ± 0.007 to 0.632 ± 0.006, and improves probability calibration relative to the Mamba-3 backbone. The ablation results further indicate that the largest contribution comes from reliability-aware spectral compression and prototype formation, followed by the multi-view spectral input representation; the complex transition provides a smaller but measurable additional gain.

However, the results should be interpreted within the limits of the available data. Crown labels contain registration, mixed-pixel, and taxonomic uncertainties, while acquisition phenology and site composition can confound species with geography. The LOSO protocol excludes taxa absent from the training sites and therefore does not measure open-set recognition. In addition, the experiments use a single hyperspectral corpus and a sensor configuration. Future work should therefore evaluate the wavelength-aware formulation across sensors with different spectral response functions and band layouts, and should investigate explicit sensor calibration, cross sensor spectral resampling, and open-set recognition. Such experiments would directly test whether modeling wavelength geometry and spectral reliability improves not only classification accuracy but also the transferability of hyperspectral processing methods across acquisition systems.

Data and Code Availability NEON Tree Species Predictions v4 (`https://doi.org/10.5281/zenodo.10926344`) [10]. AVIRIS-NG forest-mapping resource (`https://doi.org/10.5281/zenodo.17805682`) [13].

Funding Horizon Europe Framework Programme (HORIZON), call Teaming for Excellence (HORIZON-WIDERA-2022-ACCESS-01-two-stage) - Creation of the centre of excellence in smart forestry "Forest 4.0" No. 101059985.

Table 1. Main NEON evaluation (mean±std over five seeds). Δ_{site} is derived from the two preceding columns.

Method	Micro-F1	Macro-F1	LOSO Macro	$\Delta_{\text{site}} \downarrow$	ECE↓	AUC_{rob}
SpectralFormer	0.782 ± 0.006	0.618 ± 0.010	0.402 ± 0.015	0.216	0.081 ± 0.004	0.551 ± 0.009
SS-Mamba	0.794 ± 0.005	0.631 ± 0.009	0.421 ± 0.014	0.210	0.073 ± 0.004	0.568 ± 0.008
S^2Mamba	0.801 ± 0.005	0.642 ± 0.008	0.438 ± 0.013	0.204	0.068 ± 0.004	0.582 ± 0.008
CSSMamba	0.808 ± 0.004	0.653 ± 0.008	0.451 ± 0.012	0.202	0.064 ± 0.003	0.596 ± 0.007
Mamba-3 backbone	0.804 ± 0.005	0.646 ± 0.009	0.446 ± 0.014	0.200	0.061 ± 0.004	0.589 ± 0.008
PHYTO-MIMO	$\mathbf{0.826 \pm 0.004}$	$\mathbf{0.684 \pm 0.007}$	$\mathbf{0.491 \pm 0.011}$	**0.193**	$\mathbf{0.043 \pm 0.003}$	$\mathbf{0.632 \pm 0.006}$

Table 2. Ablation study. Upper block: additive ladder on the index-step backbone. Lower block: single-component removals from full PHYTO-MIMO.

Configuration	$\Delta\lambda$	2nd-ord.	Cplx.	MIMO	Macro-F1
Index-step Mamba-3	–	✓	✓	✓	0.646
+ physical step	✓	✓	✓	✓	0.657
Full PHYTO-MIMO	✓	✓	✓	✓	**0.684**
− complex rotation	✓	✓	–	✓	0.672
− MIMO spectral views	✓	✓	✓	–	0.665
− reliability/prototypes	✓	✓	✓	✓	0.658

AI Use Writefull for Overleaf and Grammarly were used to enhance the language, clarity, and grammar of this paper.

6. REFERENCES

[1] Sergio Marconi, Ben G. Weinstein, Sheng Zou, Stephanie A. Bohlman, Alina Zare, Aditya Singh, Dylan Stewart, Ira Harmon, Ashley Steinkraus, and Ethan P. White, "Continental-scale hyperspectral tree species classification in the united states national ecological observatory network," *Remote Sensing of Environment*, vol. 282, pp. 113264, 2022.

[2] Ben G. Weinstein, Sergio Marconi, Sarah J. Graves, Alina Zare, Aditya Singh, Stephanie A. Bohlman, Lukas Magee, Daniel J. Johnson, Philip A. Townsend, and Ethan P. White, "Capturing long-tailed individual tree diversity using an airborne imaging and a multi-temporal hierarchical model," *Remote Sensing in Ecology and Conservation*, vol. 9, no. 5, pp. 656–670, 2023.

[3] Danfeng Hong, Zhu Han, Jing Yao, Lianru Gao, Bing Zhang, Antonio Plaza, and Jocelyn Chanussot, "Spectralformer: Rethinking hyperspectral image classification with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022, Art. no. 5518615.

[4] Albert Gu and Tri Dao, "Mamba: Linear-time sequence modeling with selective state spaces," 2023.

[5] Lingbo Huang, Yushi Chen, and Xin He, "SS-Mamba: A spatial-spectral mamba model for hyperspectral image classification," *Remote Sensing*, vol. 16, no. 13, pp. 2449, 2024.

[6] Guanchun Wang, Xiangrong Zhang, Zelin Peng, Tianyang Zhang, and Licheng Jiao, "S^2Mamba: A spatial-spectral state space model for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–13, 2025.

[7] Zack Dewis, Yimin Zhu, Zhengsen Xu, Mabel Heffring, Saeid Taleghanidoozdoozan, Quinn Ledingham, and Lincoln Linlin Xu, "Clustering-guided spatial-spectral mamba for hyperspectral image classification," 2026.

[8] Aakash Lahoti, Kevin Y. Li, Berlin Chen, Caitlin Wang, Aviv Bick, J. Zico Kolter, Tri Dao, and Albert Gu, "Mamba-3: Improved sequence modeling using state space principles," 2026.

[9] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie, "Class-balanced loss based on effective number of samples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9268–9277.

[10] Ben Weinstein, Sergio Marconi, Alina Zare, Stephanie Bohlman, Aditya Singh, Sarah Graves, Lukas Magee, Daniel Johnson, Sydne Record, Vanessa Rubio, Nathan Swenson, Philip Townsend, Thomas Veblen, Robert Andrus, and Ethan White, "NEON tree species predictions," 2024.

[11] Nicolas Karasiak, Jean-François Dejoux, Claude Monteil, and David Sheeren, "Spatial dependence between training and test sets: Another pitfall of classification accuracy assessment in remote sensing," *Machine Learning*, vol. 111, no. 7, pp. 2715–2740, 2022.

[12] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning*. 2017, vol. 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, PMLR.

[13] Audrey Lambiel, Loïc Gerber, Anna-Katharina Schweiger, Mathias Kneubühler, Gregoire Mariethoz, Anthony Lehmann, and Nathan Külling, "Large-scale individual tree species mapping using AVIRIS-NG and LiDAR data fusion," Zenodo preprint, 2026.