Science of Remote Sensing 14 (2026) 100466

Contents lists available at ScienceDirect

Science of Remote Sensing

journal homepage: www.sciencedirect.com/journal/science-of-remote-sensing

Full Length Article

Geospatial foundation models enable data-efficient tree species mapping in temperate mountain forests

James G.C. Ball [a,*], Jana Annika Wicklein [b,c], Zhengpeng Feng [d], Jovana Knezevic [d], Sadiq Jaffer [d], Anil Madhavapeddy [d], Clement Atzberger [e], Michele Dalponte [b], David A. Coomes [a,*]

[a] Conservation Research Institute, Department of Plant Sciences, University of Cambridge, Cambridge, UK
[b] Research and Innovation Centre, Fondazione Edmund Mach, 38098 San Michele all'Adige, Italy
[c] Department of Information Engineering and Computer Science, University of Trento, Italy
[d] Department of Computer Science and Technology, University of Cambridge, Cambridge, UK
[e] CYCLOPS MRV, United States

ARTICLE INFO

Dataset link: https://zenodo.org/records/18924120, https://github.com/PatBall1/trentino-trees

Keywords:
Geospatial foundation models
Remote sensing
Tree species classification
Mountain forests
Biodiversity monitoring

ABSTRACT

Accurate mapping of tree species from satellite data remains challenging in heterogeneous mountain forests due to environmental gradients, mixed stands, and limited training labels. Geospatial foundation models (GFMs) learn rich representations from large multi-sensor archives, but their utility for species-level mapping remains unclear. Here, we evaluate two GFM embeddings, AlphaEarth and Tessera, for tree species classification in a demanding mountain landscape (Trentino, Italy; 18 species and groups), versus conventional Sentinel-1+2 composites, across experiments spanning classification accuracy, label efficiency, classifier complexity, environmental covariates, label impurity, and temporal transferability. GFM embeddings consistently outperform conventional baselines (weighted F1 = 0.83 vs. 0.80; macro F1 = 0.55 vs. 0.50), approaching saturation with only 5% of training parcels while organising species into ecologically meaningful taxonomic and functional groupings. Realising this advantage requires a nonlinear classifier: a compact neural network yields the largest single gain, whereas a linear classifier on GFM embeddings underperforms a neural network on conventional composites. Classification is robust to moderate label impurity, and training with parcel-level species proportions as soft labels improves minority-species discrimination (macro F1 = 0.586 –0.589) without purity filtering; adding terrain-derived covariates provides no further benefit. Temporal transfer across years degrades performance: weighted F1 falls from 0.84 to 0.77 (−9%) for Tessera and from 0.85 to 0.73 (−14%) for AlphaEarth, with disproportionate losses for rare species. These results show that GFMs shift the primary bottleneck in species mapping from feature engineering toward the availability, quality, and temporal alignment of reference data.

1. Introduction

Accurate, up-to-date regional maps of tree species are essential for biodiversity assessment, carbon estimation, and evidence-based forest management, because species composition regulates productivity, disturbance susceptibility, and the delivery of key ecosystem services (Tilman et al., 2013; Brockerhoff et al., 2017; Bai et al., 2024) . Field inventories remain the most reliable source of species-level data, yet their sparse coverage and high per-unit cost limit their utility at landscape and regional scales (McRoberts et al., 2010; Bergseng et al., 2015). National forest inventory programmes therefore increasingly integrate plot networks with satellite observations to expand spatial coverage cost-effectively (McRoberts and Tomppo, 2007; White et al., 2016; Immitzer and Atzberger, 2023) . The need for scalable species mapping is especially acute in mountain regions, where climate warming is driving elevational shifts of forest belts and amplifying disturbance dynamics – drought, windthrow, and bark-beetle-linked mortality – across the European Alps (Seidl et al., 2017; Senf and Seidl, 2021; Frei et al., 2023), underscoring the urgency of robust, repeatable approaches to monitor forest trajectories and support adaptive management (Directorate-General for Environment, 2023; De Keersmaecker et al., 2024).

Traditional satellite-based approaches remain constrained in their ability to discriminate tree species reliably over large, heterogeneous

* Corresponding authors.
E-mail addresses: jgcb3@cam.ac.uk (J.G.C. Ball), dac18@cam.ac.uk (D.A. Coomes).

https://doi.org/10.1016/j.srs.2026.100466
Received 7 April 2026; Received in revised form 9 June 2026; Accepted 6 July 2026
Available online 10 July 2026

List of Abbreviations

AEF	AlphaEarth Foundations	NDVI	Normalised Difference Vegetation Index
BRDF	Bidirectional Reflectance Distrib. Function	NDWI	Normalised Difference Water Index
CLC	CORINE Land Cover	PropL1	Proportion L1 error
CLMS	Copernicus Land Monitoring Service	RF	Random Forest
DEM	Digital Elevation Model	SAR	Synthetic Aperture Radar
F1	F1 score	SCL	Scene Classification Layer
GFM	Geospatial Foundation Model	SRTM	Shuttle Radar Topography Mission
GRD	Ground Range Detected	TPI	Topographic Position Index
IW	Interferometric Wide swath	UMAP	Uniform Manifold Approx. & Projection
KL	Kullback–Leibler divergence	UTM	Universal Transverse Mercator
KNN	K-Nearest Neighbours	VV/VH	SAR polarisations
LightGBM	Light Gradient Boosting Machine	WF1	Weighted F1 score
LR	Logistic Regression	XGBoost	eXtreme Gradient Boosting
MF1	Macro-averaged F1 score		
MLP	Multi-Layer Perceptron		
NBR	Normalised Burn Ratio		

landscapes (Fassnacht et al., 2016). Multispectral sensors such as Sentinel-2 provide broad spectral bands often insufficient to detect subtle canopy variations associated with species-specific differences in leaf biochemistry and pigment concentrations (Dalponte et al., 2012; Immitzer et al., 2019; Zhong et al., 2024) , while deca-metric spatial resolution (10–30 m) leads to mixed pixels that obscure individual crowns or fine-scale species mosaics characteristic of mountain forests (Xu et al., 2021). Species identification from individual images is further confounded by phenological variation (Cambridge Centre for Earth Observation, 2026; Dalponte et al., 2025), canopy understory variability (Rautiainen et al., 2011), and snow seasonality (Wang et al., 2023), meaning single-date imagery rarely captures optimal discriminative conditions. Multi-season time series can address temporal confounds but are cumbersome to process, involving vast data volumes often impacted by irregular cloud cover (Wang et al., 2022). Fusing complementary modalities – such as synthetic aperture radar and optical reflectance – can mitigate ambiguities (Blickensdörfer et al., 2024), but multimodal integration has historically required bespoke pre-processing, cross-sensor alignment, and task-specific pipelines. Topographic effects introduce additional complexity: Bidirectional Reflectance Distribution Function (BRDF) effects from variable illumination and slope-induced reflectance changes reduce the transferability of spectral signatures across terrain (Dong et al., 2020; Kellndorfer et al., 1998), so that models trained for one site or season often fail to generalise (Tuia et al., 2016) . Together, these limitations have historically confined detailed tree species mapping to small study areas or narrow sets of species groups.

Recent advances in geospatial foundation models (GFMs) offer a potential solution. Trained on massive multi-sensor archives using self-supervised objectives, GFMs learn information-rich embeddings that capture spectral and temporal features of the Earth's surface without requiring labelled data (Xiao et al., 2025) . By distilling geospatially relevant patterns directly from unlabelled imagery, they enable downstream classifiers to be trained on far smaller sets of local labels — addressing one of the main bottlenecks in species-level remote sensing (Ball et al., 2024) . Models such as AlphaEarth (Brown et al., 2025) and Tessera (Feng et al., 2025) provide latent representations derived from petabyte-scale, globally distributed satellite time series that can potentially be transferred to regional tasks where labelled data are sparse or heterogeneous. In principle, such models address several long-standing obstacles simultaneously: they ingest multi-temporal and multi-modal inputs during pre-training, learning to handle cloud gaps, phenological variation, and sensor differences without site-specific calibration.

Despite this promise, several questions about how GFM representations relate to species-level ground reality remain unresolved. First, although GFMs have been applied to vegetation monitoring, their use for tree species mapping from satellite imagery remains limited (Braham et al., 2025), and it is unclear whether their representations reliably encode species-discriminative information in structurally complex mountain forests (Brandt et al., 2025; Ishikawa et al., 2025; Jin and Wu, 2025). Second, a central claim of foundation models is improved label efficiency, yet few studies quantify how performance scales with labelled data under a fixed evaluation design; in forest inventories – where labels are costly and spatially uneven – understanding this relationship is critical (Brus et al., 2012; Ramezan et al., 2021). Third, the downstream classifier capacity needed to exploit GFM embeddings has not been systematically assessed: if embeddings capture species-discriminative structure, simple classifiers may suffice; if not, additional complexity may be necessary and could change conclusions about representation quality. Fourth, label noise and mixed-species stands are common in real inventories, yet it remains unclear whether GFM representations can leverage noisy or aggregated labels to improve discrimination (Bolyn et al., 2022). Finally, temporal stability of learned representations is largely untested — it is unclear whether embeddings retain species-discriminative power under interannual variation in phenology and acquisition conditions (Lisaius et al., 2026).

In this study, we evaluate whether globally pre-trained GFMs can serve – without fine-tuning – as effective and data-efficient representations for regional-scale tree species mapping. Using forest cover in the Autonomous Province of Trento (Italy) as a case study – characterised by steep environmental gradients, mixed stands, strong phenological variability, and detailed parcel-level inventories – we compare embeddings from AlphaEarth and Tessera with conventional Sentinel-1+2 composites. Our analysis focuses on assessing the ecological information encoded in each representation. To our knowledge, this is the first study to map detailed tree species distributions from space using geospatial foundation models. Specifically, we ask:

1. Do GFM embeddings encode species-discriminative information more effectively than conventional satellite composites?
2. Does increasing downstream classifier capacity lead to consistent performance gains when applied to geospatial foundation-model embeddings?
3. How sensitive is classification performance to training-data availability and label purity?
4. Do supplementary environmental variables provide additional predictive power beyond satellite-derived representations?
5. How does classification performance change under cross-year transfer?

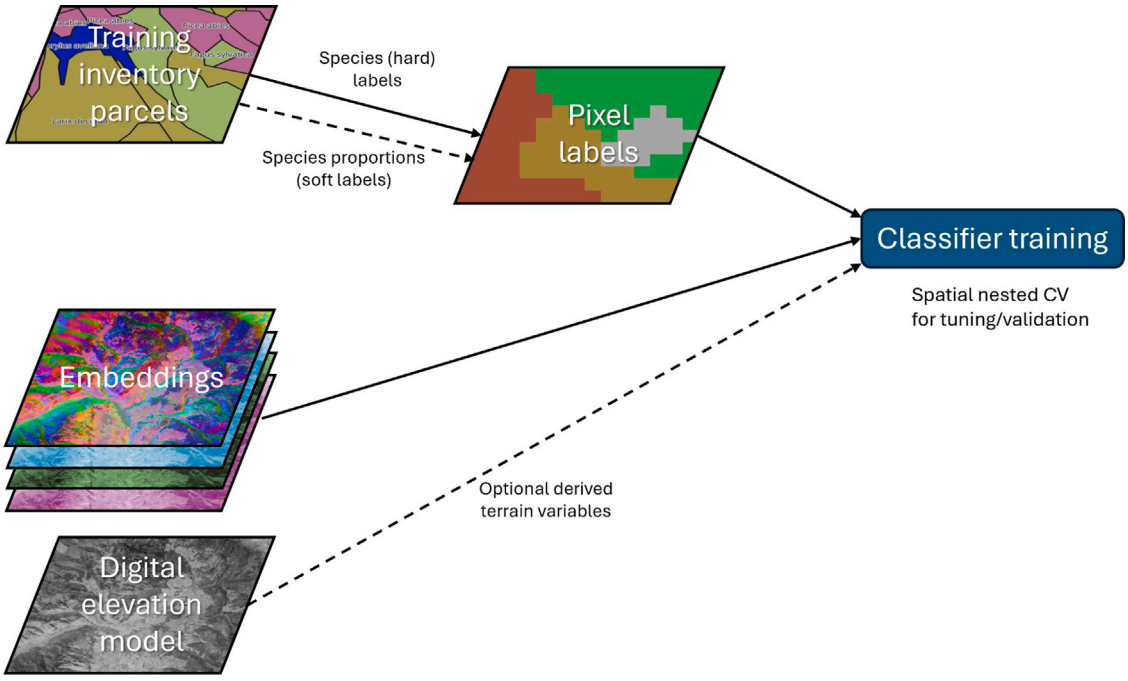

Fig. 1. Conceptual overview of the species classification workflow. Parcel-level forest inventory data are rasterised to pixel-level labels or as mixed composition soft labels. Geospatial foundation-model embeddings provide pixel-level feature representations for classifier training. Optional terrain variables are included in ablation experiments.

2. Methods

2.1. Conceptual overview

This study evaluates whether geospatial foundation-model (GFM) embeddings provide ecologically meaningful and data-efficient representations for tree-species classification. We compare two globally pretrained GFM embeddings (AlphaEarth; Brown et al. 2025 and Tessera; Feng et al. 2025) with conventional Sentinel-1/2 composites under controlled modelling conditions.

All training and prediction was conducted at the pixel level (with evaluation also at the parcel level), producing discrete species predictions compatible with existing wall-to-wall forest maps (e.g. De Keersmaecker et al. 2024). Although fractional approaches are increasingly used for mixed stands (e.g. Bolyn et al. 2022), we adopt pixel-wise classification to keep the focus on representation choice and downstream model behaviour.

Classifier capacity is treated as a controlled experimental axis: we evaluate classifiers spanning from linear models to neural networks, enabling us to disentangle whether performance is primarily driven by representation quality or downstream model capacity.

Model behaviour is assessed along five dimensions: (i) overall classification performance; (ii) label efficiency under reduced training data; (iii) sensitivity to training-label purity; (iv) the contribution of ancillary environmental covariates; and (v) temporal generalisation under cross-year transfer. Results follow a reference configuration, then targeted ablation analyses.

A schematic overview of the modelling pipeline is provided in Fig. 1 and evaluation process in Fig. 2. Detailed descriptions of individual components are presented in the following sections. Unless otherwise stated, results refer to an MLP classifier trained on 2018 embeddings using parcel-level cross-validation, with training labels filtered to ≥60% dominant species cover.

2.2. Study region and forest inventory

2.2.1. Study region: Trentino, Italy

The Autonomous Province of Trento (Trentino), Italy, provides a challenging test case for species-level forest mapping. The province covers 6207 km^2, of which approximately 60% is forested. The region spans steep elevational, climatic, and biogeographic gradients, producing sharp transitions between distinct forest communities (Agnoletti and Biasi, 2013; Pedrotti, 2018). This heterogeneity, coupled with phenological variation and complex illumination from rugged terrain, challenges satellite-based species classification (Weiss and Walsh, 2009; Hu et al., 2025), with mountainous topography also introducing SAR distortions such as layover and shadowing. Trentino therefore represents a stringent test of whether satellite-derived representations can encode species-discriminative information robust to terrain and mixed stand composition.

These gradients produce pronounced elevational zonation: broadleaved stands (dominated by European beech) at lower montane elevations give way to Norway spruce–silver fir forests in the montane belt, and to European larch–stone pine stands at the subalpine level (Gasparini et al., 2022). Valley floors below ~600 m are largely non-forested, and forests transition into alpine meadows above ~2200 m. Full species composition and elevation distributions are provided in Table 1.

2.2.2. Inventory data

As reference data, we used the provincial forest assessment plan (*piani di assestamento*) provided by the Forest Service of the Province of Trento, collected from 2010–2021 (see Fig. S.1 for temporal distribution). The full dataset is described and publicly released in Dalponte and Andreatta (2026). The inventory divides the regional forest area into 83,000 management units (mean area 2.89 ha; maximum 519 ha) covering 260,000 ha in total. For each unit, trained foresters recorded tree species present and their proportional canopy cover through in-field visual inspection, following provincial technical guidelines (Dalponte and Andreatta, 2026). Species are recorded when their

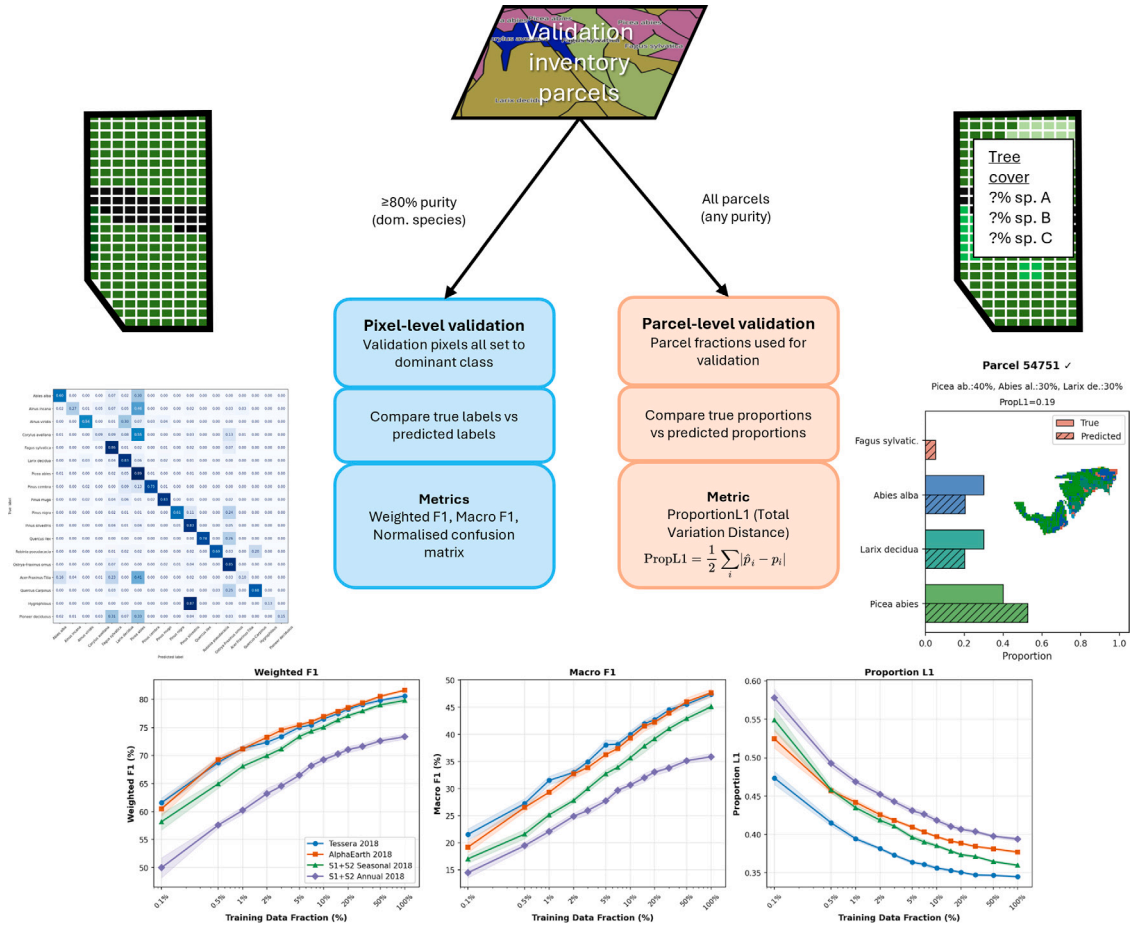

Fig. 2. Schematic of the dual validation framework. Validation inventory parcels (top) are evaluated at two complementary levels. Left branch (pixel-level): parcels with dominant-species purity ≥80% are selected and all pixels are assigned the dominant species label; per-pixel predictions are compared against these labels to compute weighted F1, macro F1, and a normalised confusion matrix. Right branch (parcel-level): all parcels, including mixed-species stands, are used; pixel-level predictions are aggregated into predicted species proportions and compared against true inventory fractions using the Proportion L1 metric (Total Variation Distance). Black pixels represent filtered non-forest classes such as roads.

estimated cover exceeds approximately 5%; below this threshold only presence is noted (mean 3.55 species per unit; range 1–15). Parcels under 0.2 ha were excluded as uninformative given pixel size and co-registration uncertainty.

Using coexistence analysis (Jaccard similarity and hierarchical clustering) in consultation with local forestry experts, we grouped less common species into ecologically meaningful community classes, yielding 13 dominant species and five community groups (Table 1; Section S.2) (see Fig. 3).

2.3. Remote sensing inputs

2.3.1. Geospatial foundation model embeddings

Geospatial foundation models produce embeddings, which are compact numerical representations of each pixel that summarise high-dimensional remote-sensing information – such as spectral signatures, temporal dynamics, and cross-sensor relationships – into a fixed-length vector (Xiao et al., 2025). These vectors inhabit a latent space, where proximity between points often corresponds to similarity in underlying surface properties according to the representation learned during self-supervised training. In this space, pixels with similar land-cover composition, phenology, or structure tend to cluster together, while dissimilar surfaces are separated by larger distances. Importantly, the latent space provides a continuous geometry in which operations such as measuring distances, identifying neighbourhoods, or anchoring representations to reference examples become meaningful.

AlphaEarth Foundations (AEF) is a global geospatial foundation model that produces annual 64-dimensional embeddings for every 10 m land pixel. It encodes multi-source remote sensing data from Sentinel-1 SAR, Sentinel-2 optical, and Landsat optical and thermal imagery. In addition to these AEF input sources, the approach uses additional target variables spanning a broader set of modalities, including LiDAR structure (GEDI), climate variables (ERA5), gravity fields (GRACE), and topography (DEM), implemented via implicit decoder objectives. Instead of using single images, representations are derived from satellite "video sequences" using a Space–Time–Precision (STP) encoder that supports continuous time, enabling the integration of irregularly sampled observations into a consistent temporal representation. During training, a student network is optimised to reproduce the embeddings of a teacher network via a consistency loss, encouraging robustness to missing or perturbed inputs, while a text-contrastive objective introduces a weak alignment with geolocated text descriptions. Together, these design choices aim to produce surface representations that are transferable across downstream tasks. The resulting embeddings are accessible via Google Earth Engine. Further details are provided in Brown et al. (2025).

Tessera is an open geospatial foundation model that generates annual 128-dimensional embeddings for every 10 m Sentinel pixel by jointly modelling optical and SAR satellite time series (Feng et al., 2025). Building on the concept of the spectral–temporal Barlow Twins (Lisaius et al., 2024), the model adopts a dual-encoder architecture in which annual time series of Sentinel-2 optical reflectance and Sentinel-1 SAR backscatter sequences are processed through separate encoders

Table 1
Target classes with training sample sizes and elevation distributions. The 18 classes comprise 13 dominant species retained as individual classes and 5 community groups formed by merging ecologically co-occurring minor species. Note that the pixel count is derived from pixels contained in parcels in which the species is dominant and area is calculated from the proportional cover of the forestry parcels, so correspondence is not exact.

Class	Common name	Parcels	Pixels	Area (ha)	Elev. range (m)	Elev. mean (m)
Dominant species						
Picea abies	Norway spruce	33,885	9,555,996	91,009	151–2849	1482
Fagus sylvatica	European beech	11,991	3,540,217	33,688	130–2008	1160
Larix decidua	European larch	11,274	3,515,723	43,715	200–2611	1700
Pinus sylvestris	Scots pine	6255	1,559,357	16,080	76–2794	1011
Abies alba	Silver fir	4464	1,406,263	15,052	245–2192	1297
Pinus mugo	Mountain pine	1887	841,517	9443	318–2366	1715
Pinus nigra	Black pine	1832	404,008	4025	73–1802	705
Alnus alnobetula	Green alder	1680	381,005	6413	815–2561	1910
Pinus cembra	Swiss stone pine	772	279,831	4419	1439–2455	1980
Corylus avellana	Common hazel	662	65,432	1245	211–2205	1087
Robinia pseudacacia	Black locust	447	52,813	512	201–1322	690
Alnus incana	Grey alder	299	37,140	434	468–2160	1327
Quercus ilex	Holm oak	120	57,819	576	63–1303	388
Community/species groups						
Ostrya–Fraxinus ornus[a]	Thermophilous deciduous	6304	2,095,415	21,346	72–1914	790
Quercus-Carpinus[b]	Oak-hornbeam woodland	692	114,640	1641	152–1459	795
Pioneer deciduous[c]	Early successional	610	81,989	2592	223–2131	1374
Acer–Fraxinus–Tilia[d]	Mixed mesophilous	444	59,162	1811	139–1864	1092
Hygrophilous[e]	Riparian/wetland	66	6060	193	165–2068	1097
Total	(18 classes)	**83,684**	**24,054,387**	**254,192**		

[a] *Ostrya carpinifolia*, *Fraxinus ornus*, *Quercus pubescens*, *Sorbus aria*, *Acer campestre*, *Ulmus minor*, *Celtis australis*.
[b] *Carpinus betulus*, *Castanea sativa*, *Prunus avium*, *Quercus cerris*, *Q. petraea*, *Q. robur*.
[c] *Betula pendula*, *Populus tremula*, *Salix caprea*, *Sorbus aucuparia*, *Ailanthus altissima*, *Laburnum alpinum*, *L. anagyroides*.
[d] *Acer platanoides*, *A. pseudoplatanus*, *Fraxinus excelsior*, *Taxus baccata*, *Tilia cordata*, *T. platyphyllos*, *Ulmus glabra*.
[e] *Alnus glutinosa*, *Populus alba*, *Salix alba*, *S. cinerea*, *S. eleagnos*.

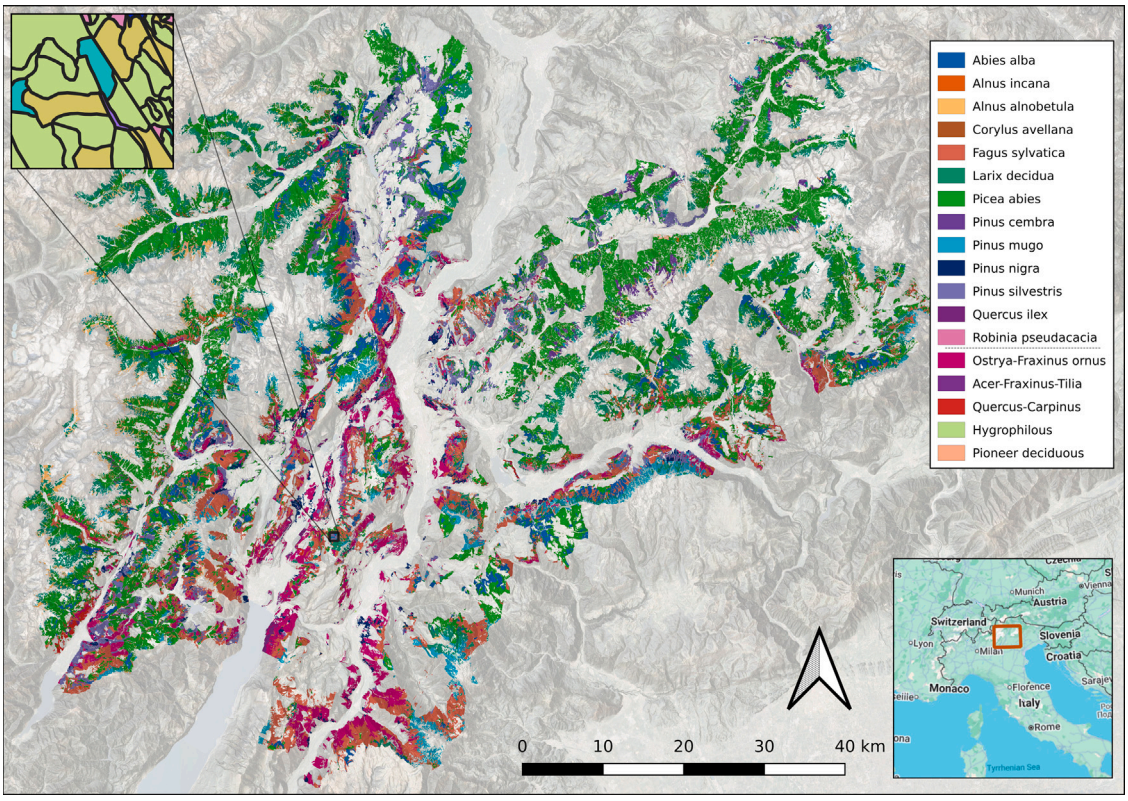

Fig. 3. Trentino with the colours showing 18 labelled tree species and species group classes of the surveyed forest plots. The top-left inset shows a zoom on a dense section of the irregular forest inventory parcels which define the pixel-level species labels within them. The bottom-right inset shows the study region within its context of northern Italy.

and fused into a unified representation. Tessera is trained on a global dataset using a self-supervised Barlow Twins objective (Zbontar et al., 2021; Lisaius et al., 2024) combined with sparse random temporal sampling, encouraging invariance to missing observations and irregular acquisition patterns. Robustness is further promoted through global shuffling, which reduces spatial autocorrelation, and mix-up regularisation. The model learns characteristic spectral–temporal patterns of land surfaces without labelled or ancillary data. The resulting embeddings

capture the temporal evolution of 10-band spectral signatures and SAR backscattering in VV and VH, and are precomputed as annual global rasters, accessible via the `geotessera` Python package.[1] For details, the reader is referred to Feng et al. (2025).

Although both foundation models were *trained* on multi-year archives, the embeddings used in this study are annual products. For Tessera, the 2018 (or 2019) embeddings encode exclusively the Sentinel observations from that calendar year, ensuring strict temporal equivalence with the Sentinel-1/2 composites. For AlphaEarth, the published embeddings are likewise annual products, although its continuous-time STP encoder may in principle draw on observations near the year boundaries. We therefore treat both GFMs as annual representations that are directly comparable with the year-matched Sentinel-1/2 composites, and return in the Discussion to the question of how AlphaEarth's patch-level, multi-modal pre-training shapes the embeddings.

2.3.2. Satellite composite baselines

Sentinel-2 (optical). We used Sentinel-2 Level-2 A surface reflectance imagery to generate annual and seasonal composite feature stacks for 2018. All scenes intersecting the study region were filtered to retain acquisitions with <60% scene-level cloud cover. The 60% threshold applies at scene level; all retained scenes undergo pixel-level cloud, shadow, and cirrus masking via the Sen2Cor SCL layer and QA60 bitmask, so no cloudy pixels enter any composite. Reflectance values were scaled to unitless surface reflectance (0–1) (see Gao et al., 2024).

To quantify the contribution of seasonal information in a conventional multi-sensor baseline, we produced two alternative Sentinel-2 representations: (i) an annual per-pixel median composite over all valid observations in the year, and (ii) seasonal composites computed as per-pixel medians within three phenologically meaningful periods – spring (March–May), summer (June–August), and autumn (September–November) – with winter excluded due to pervasive snow cover at higher elevations. Consistent with common practice in tree-species and forest type mapping, retaining seasonal structure aims to preserve phenological variation that is attenuated by annual aggregation.

All Sentinel-2 spectral bands (10 m, 20 m, 60 m; excluding atmospheric band B10) were retained. For each seasonal composite (and for the annual composite), we additionally computed NDVI, NDWI (McFeeters), and NBR.

Sentinel-1 (SAR). To incorporate cloud-independent structural and moisture-sensitive information, we also derived Sentinel-1 GRD (IW mode) seasonal and annual features for 2018 using dual polarisation VV and VH, along with the incidence angle band provided in the collection. The Earth Engine Sentinel-1 GRD collection is radiometrically calibrated and terrain corrected, with backscatter provided in decibels (dB). We formed seasonal (spring/summer/autumn) and annual Sentinel-1 composites by aggregating observations within the same date ranges as Sentinel-2. To reduce bias when compositing on a log scale, we computed medians in linear backscatter space and converted back to dB after aggregation (details in Section S.3). In addition to VV and VH backscatter, we derived simple polarisation features (VV–VH in dB and VV/VH ratio) commonly used to capture vegetation structural differences.

Gridding and alignment. All composites were exported on a fixed 10 m UTM grid aligned exactly to the study region. Sentinel-2 bands at 20 m and 60 m were resampled to 10 m; Sentinel-1 features were exported on the same target grid. Full implementation details (mask classes, bitmasks, resampling choices, and export parameters) are provided in Section S.3 (see Table 2).

[1] https://geotessera.org

2.4. Label construction

Pixels within a forestry parcel were assigned the parcel's dominant species label (determined from proportional inventory cover; see Fig. 1) for model training and testing. To improve the quality and quantity of the labelled pixels we applied some additional pre-processing steps (described below).

2.4.1. Constraints of parcel-level labels

Forest inventory data in Trentino provide species composition at the level of management parcels rather than at the level of individual trees or pixels. Each parcel reports the proportional canopy cover of multiple species, often representing structurally and compositionally heterogeneous stands. As a result, direct use of parcel labels for pixel-level classification introduces several sources of ambiguity. Parcels frequently contain mixtures of species, parcel boundaries are irregular, and non-forest elements (e.g., roads, clearings, rocky outcrops) are not explicitly represented in the inventory.

Critically, parcel labels are not pixel-accurate ground truths. Treating them as such risks inflating apparent classification performance for dominant species while systematically penalising predictions of secondary species that are present but spatially intermixed. Conversely, discarding mixed parcels entirely would reduce the volume of useable training data and disproportionately exclude ecologically important stand types. Addressing this trade-off between label quantity and label contamination is central to enabling data-efficient species mapping from satellite imagery using forest inventories as reference data. To address this, we also explore a soft-label training formulation in which pixel-level targets encode the full parcel-level species composition vector rather than a single dominant species (Section 2.8.2).

2.4.2. Land-cover masking

To ensure that species training and inference were restricted to ecologically plausible locations, we filtered non-forest areas during training label construction. Because inventory-derived species proportions refer exclusively to tree canopy cover (and exclude other ground cover such as roads or grassland), pixels corresponding to non-forest land-cover types (e.g. built surfaces, water bodies, bare ground, grassland, or persistent snow) were filtered to prevent implausible species assignments.

We used an independent, freely available, global land-cover product – the 10 m Global Land Cover 2020 (Copernicus Land Monitoring Service / EU CLMS, 2020) – to identify forested pixels and gate species label assignment accordingly. Pixels classified as forest were retained as eligible for species inference, while non-forest pixels were assigned fixed non-tree classes to ensure consistent handling across all representations (see Section S.1.3 for the full filtering workflow).

2.5. Model selection

2.5.1. Classifier architectures

We selected classifiers spanning a gradient of expressive capacity to assess how downstream model complexity influences the utility of different input representations. Logistic Regression (LR; Hastie et al. 2009) provides a linear baseline. K-Nearest Neighbours (kNN; Cover and Hart, 1967) offers a non-parametric alternative relying on local similarity in feature space. Random Forests (RFs; Breiman, 2001) serve as a widely adopted nonlinear baseline in ecological remote sensing, robust to noisy labels and class imbalance (Belgiu and Drăguţ, 2016) . We additionally include two gradient-boosted decision-tree ensembles, XGBoost (Chen and Guestrin, 2016) and LightGBM (Ke et al., 2017) , which are among the strongest general-purpose tabular classifiers and complement the bagged Random Forest with a boosting-based ensemble strategy. Multilayer perceptrons (MLPs; Rumelhart et al. 1986) can learn smooth nonlinear transformations, potentially exploiting structure that simpler methods cannot capture.

Table 2
Comparison of remote sensing predictors used to train the species classification models. For all layers, years 2018 (for training/validation) and 2019 (for transferability) were acquired which fall within the inventory collection period (2010–2021).

Input	Features	Data type	Resolution	Modality
AlphaEarth	64	GFM embeddings	10 m	Optical + SAR
Tessera	128	GFM embeddings	10 m	Optical + SAR
S1+S2 seasonal	60	Seasonal composite	10 m	Optical + SAR
S1+S2 annual	20	Annual composite	10 m	Optical + SAR

2.5.2. Training and validation design

We employed nested parcel-level cross-validation with five outer folds for test estimation and Bayesian optimisation (Optuna; Akiba et al. 2019) for inner-loop hyperparameter tuning. Pixels belonging to the same parcel were never split across folds, preventing spatial leakage from within-parcel autocorrelation. This parcel-aware design reflects the intended gap-filling application, in which unsurveyed parcels are inferred from a partially surveyed landscape; we additionally assess robustness to between-parcel spatial autocorrelation using a stricter spatial-block cross-validation (5 km grid) in Section S.8.5 (Table S.19). Preprocessing was computed on training data only, and models were retrained on the full outer-fold training set before evaluation. Performance is reported as the mean ± standard deviation across folds.

Training parcels were restricted to ≥ 60% dominant-species cover (except in Section 2.8.2), while pixel-level evaluation used a higher-purity subset (≥80%). PropL1 was computed across all validation parcels regardless of purity (Section 2.6). This partitioning, purity criteria, and evaluation protocol were held constant across all analyses. For reproducibility in downstream experiments (Sections 2.8–2.9), hyperparameters were fixed using a representative outer-fold configuration.

2.5.3. Model training and hyperparameter tuning

Hyperparameters were selected by minimising a composite validation metric balancing pixel-level accuracy with parcel-level compositional realism:

$$\text{Combined} = \text{PropL1} - 0.5 \times \text{Macro F1} \tag{1}$$

where PropL1 is the parcel-level species proportion error (Section 2.6.2) and Macro F1 the pixel-level macro F1 score (Section 2.6.1). The weighting ensures that a 1% improvement in Macro F1 offsets a 0.5% increase in proportion error, encouraging models that detect less common species while maintaining realistic aggregate composition. This objective was used exclusively for hyperparameter selection.

2.5.4. Class imbalance handling

To mitigate class imbalance (Table 1), the three most abundant species were downsampled within each training fold to match the pixel count of the fourth most abundant class (Section S.5 .2). Downsampling was parcel-stratified to preserve spatial and environmental variability, and was applied during training only; validation and test data were left untouched. We tested the sensitivity of our results to this choice through a balanced-versus-unbalanced ablation (Supplementary Table S.17): removing downsampling slightly raised weighted F1 but did not improve macro F1, our primary tuning metric (Section 2.6.1), confirming that the conclusions are not an artefact of the balancing strategy.

2.6. Performance metrics and interpretation

2.6.1. Pixel-level classification metrics

Pixel-level performance was assessed using F1 scores on validation parcels with ≥80% dominant-species purity. The F1 score for a class is the harmonic mean of its precision (the proportion of pixels predicted as that class that are correct) and its recall (the proportion of that class's pixels that are correctly identified). Because the harmonic mean is dominated by the smaller of the two quantities, F1 is high only when a class is both detected reliably and predicted precisely, making it a more informative summary than overall accuracy when classes are highly imbalanced. Weighted F1 (WF1) averages per-class F1 scores weighted by class prevalence, reflecting overall accuracy dominated by common species. Macro F1 (MF1) gives equal weight to all species regardless of prevalence, emphasising performance on rare classes; MF1 was therefore prioritised in model tuning. Normalised confusion matrices reveal systematic misclassification patterns across functional and phylogenetic groups.

2.6.2. Parcel-level species proportion error

To evaluate performance at a management-relevant scale across all mixed parcels, we used the Proportion L1 metric (PropL1), defined as the Total Variation Distance between predicted and reference species-proportion vectors, equivalent to Bray–Curtis dissimilarity for proportional data (Gibbs and Su, 2002; Bray and Curtis, 1957):

$$\text{PropL1} = \frac{1}{2}\sum_{i=1}^{C} |\hat{p}_i - p_i| \tag{2}$$

where $\hat{p}_i$ and p_i denote predicted and reference proportions for species i, respectively. PropL1 ranges from 0 (perfect agreement) to 1 (complete mismatch) and was computed across all validation parcels regardless of purity. Unlike pixel-level metrics, PropL1 makes no assumptions about stand composition, providing a robust assessment in heterogeneous landscapes.

2.7. Effect of downstream classifier capacity

We assess how classification performance varies across classifiers of increasing capacity, holding training protocol, cross-validation design, and evaluation metrics fixed (Section 2.5). For each representation, we compare linear (logistic regression), instance-based (kNN), tree-ensemble (Random Forest, XGBoost, and LightGBM), and neural-network (MLP) classifiers to distinguish representation-limited from classifier-limited regimes.

Unless otherwise stated, all subsequent results use the best-performing MLP configuration from this comparison.

2.8. Analyses of representation quality and supervision sensitivity

2.8.1. Label availability and fraction-curve analysis

To quantify label efficiency, models were trained on increasing fractions of available training parcels (0.1% to 100%), selected by stratified random sampling at the parcel level. Hyperparameters were fixed at values from full-data tuning (Section 2.5), isolating the effect of label availability. At each fraction, performance was assessed via 5-fold parcel-aware cross-validation repeated across three random seeds. Class-capped downsampling (Section 2.5.4) was applied at every fraction, and stratified parcel sampling ensured that all classes remained represented even at the smallest fractions.

2.8.2. Sensitivity to label purity and weak supervision

We varied the minimum dominant-species cover threshold required for training parcels (30% dominance, ~83k parcels, to 100% dominance, ~4.9k parcels), capturing the trade-off between label noise and training data volume. Pixel-level validation used the fixed ≥80% purity subset; PropL1 was computed across all parcels.

Table 3
Classification performance of MLP models trained on different input representations, evaluated using 5-fold nested cross-validation. Best value per column in bold.

Input representation	Accuracy	Weighted F1	Macro F1	Proportion L1
Tessera	0.817 ± 0.009	0.827 ± 0.009	**0.551 ± 0.019**	**0.349 ± 0.004**
AlphaEarth	**0.824 ± 0.006**	**0.833 ± 0.006**	0.537 ± 0.015	0.384 ± 0.002
S1+2 seasonal	0.788 ± 0.011	0.803 ± 0.010	0.503 ± 0.017	0.368 ± 0.004
S1+2 annual	0.729 ± 0.007	0.745 ± 0.005	0.414 ± 0.022	0.401 ± 0.002

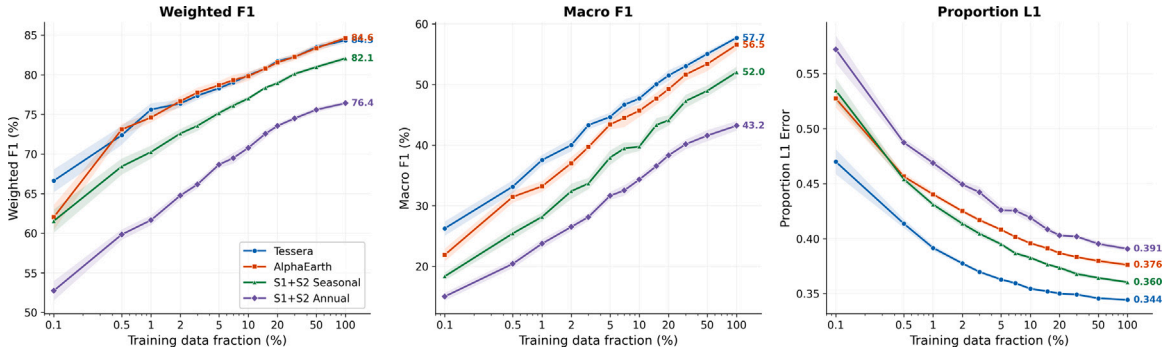

Fig. 4. Label efficiency curves comparing Tessera 2018, AlphaEarth 2018 and Sentinel-1+2 seasonal and annual composites feature representations for multi-label forest species classification. Performance metrics are shown as functions of training data fraction for an MLP classifier trained on forest parcel data from Trentino. (Left) Weighted F1-score. (Middle) Macro F1-score. (Right) Proportion L1 error. Shaded regions indicate 95% confidence intervals. Here, 100% of training data is ~43,834 training parcels (~5.9M training pixels).

We also evaluated a soft-label formulation in which pixel-level targets encode parcel-level species proportions rather than hard dominant-species assignments. Neural networks were trained with cross-entropy loss generalised to soft targets, so that the training objective accommodates mixed-species parcels rather than forcing every pixel toward a single dominant label. Performance was compared to standard hard-label training across the same purity thresholds. We additionally tested an alternative label-refinement strategy – iterative label distillation, which reassigns within-parcel labels using embedding-space prototypes (Section S.4) – but it did not yield consistent improvements.

2.8.3. Ancillary environmental covariates

We augmented input features with terrain covariates from the SRTM DEM (30 m; Farr et al., 2007): elevation, slope, aspect (sine/cosine encoded), and topographic position index (TPI; Wilson and Gallant 2000), yielding five additional channels. All features were z-scored (standardised) using training-data statistics. Models with terrain covariates were tuned using the same procedure described in Section 2.5.

2.9. Temporal transfer experiments

To assess temporal generalisability, models trained on 2018 data were applied to 2019 without retraining. Parcel assignments to training and evaluation folds were fixed across years to prevent spatial overlap, and the same purity thresholds and evaluation protocol described in Section 2.5 were applied.

No fine-tuning or domain adaptation was performed, providing a conservative assessment of how well classification boundaries transfer across years. This experiment serves as a stress test of representation robustness under realistic interannual variation in phenology and acquisition conditions, rather than a demonstration of multi-year modelling strategies.

3. Results

3.1. Classification accuracy, label efficiency, and ecological structure

Across experiments, geospatial foundation-model embeddings outperformed conventional Sentinel-1+2 spectral composites for species-level tree classification. Using the best-performing MLP configuration for each representation, Tessera and AlphaEarth achieved similar overall accuracy (weighted F1 = 0.827 ± 0.009 and 0.833 ± 0.006, respectively) and higher macro F1 than Sentinel-1+2 baselines, indicating improved discrimination of less dominant species (Table 3; per-species precision, recall, and F1 in Table S.16). Sentinel-1+2 annual composites performed substantially worse across all metrics, consistent with temporally aggregated spectral features being insufficient for fine-grained species discrimination in heterogeneous mountain forests.

Label efficiency. Fig. 4 shows how performance scales with the fraction of labelled training parcels (100% corresponds to ~43,834 parcels on average across folds). For weighted F1, both foundation-model embeddings exhibit strong label efficiency, approaching saturation with ≤5% of training parcels and converging rapidly as supervision increases. However, this comes at the cost of approximately 10 percentage points of macro F1 relative to the full training set, disproportionately affecting rare species; practitioners must weigh this trade-off between labelling effort and per-species reliability. Tessera shows a small but consistent advantage at the lowest training fractions, while differences between Tessera and AlphaEarth diminish with additional training data. Sentinel-1+2 baselines lag behind the foundation models across all training fractions, with annual composites consistently underperforming seasonal composites.

Macro F1 increases more gradually with training fraction, with Tessera achieving the highest values across most fractions; both Sentinel-1+2 variants show substantially weaker and more slowly saturating performance.

Parcel-level composition fidelity. Tessera achieves the lowest proportion error (PropL1) across all training fractions. Sentinel-1+2 seasonal composites consistently outperform AlphaEarth on PropL1 despite lower pixel-level F1, indicating that representations can differ in how well they preserve aggregate class frequencies. Fig. 5 illustrates pixel-level species predictions for three mixed validation parcels.

Embedding-space structure. To examine the ecological structure underlying these performance gains, we visualised Tessera embeddings in three dimensions using UMAP (Fig. 6). The low-dimensional manifold appears to be organised primarily along broad environmental gradients. Elevation emerges as a dominant organising axis spanning both coniferous and broadleaf taxa, while separation by functional type is clearly evident. At finer taxonomic scales, genera and species

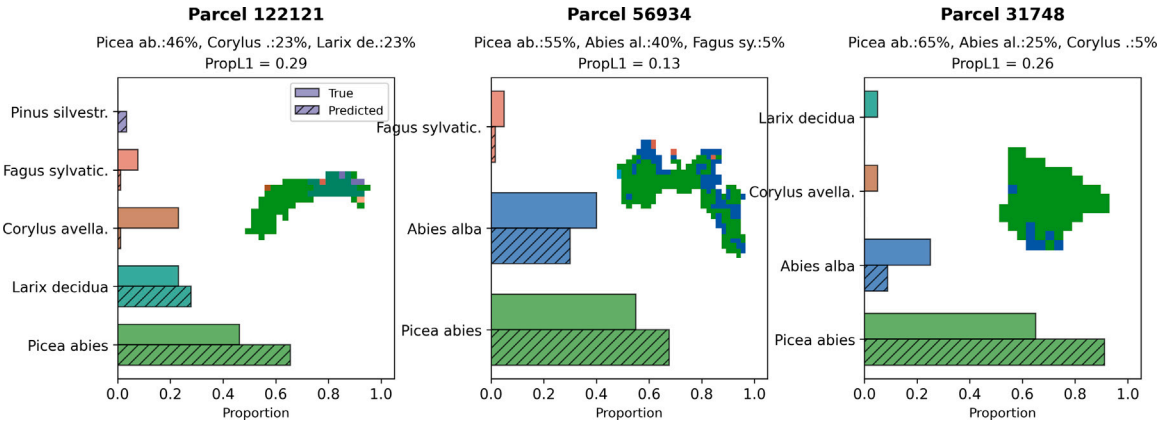

Fig. 5. Pixel-level species predictions for three mixed validation parcels using the MLP classifier with Tessera embeddings. Bar charts compare true species proportions (solid bars) against predicted proportions aggregated from pixel classifications (hatched bars). Spatial prediction maps (insets) show individual pixel classifications coloured by species. The PropL1 score quantifies the total variation distance between true and predicted proportions (0 = perfect match, 1 = complete mismatch). Species colours are consistent across maps and bars.

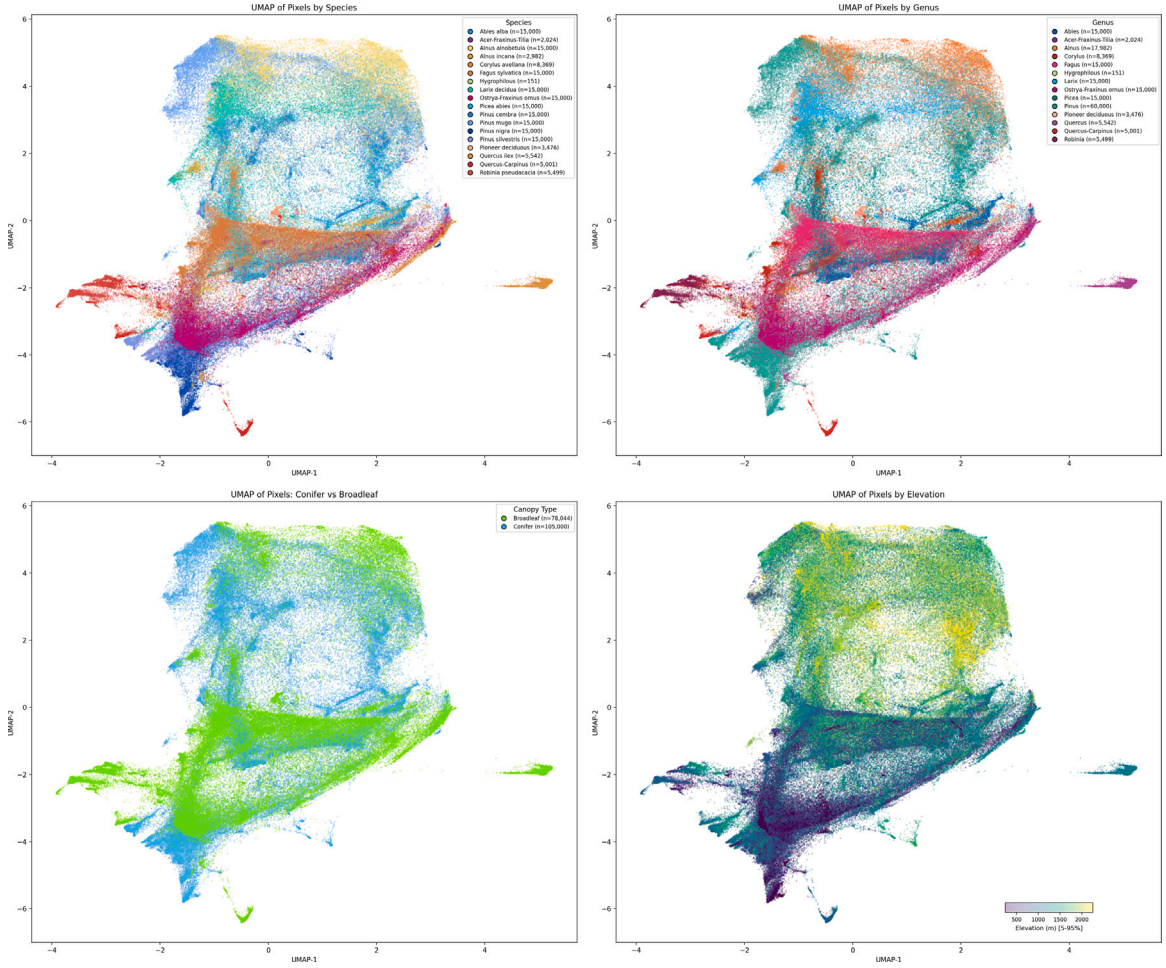

Fig. 6. UMAP representations of Tessera embeddings coloured by species (or species groups), genus, broadleaf/conifer functional types and elevation. Point ordering by the 3rd UMAP dimension gives the effect of a z-axis and demonstrates separation in this 3rd dimension. Individual species group facets are available in Fig. S.20; UMAP projections coloured by individual species (pre-merge) in Figs S.13–S.15.

occupy consistent regions of the manifold, with within-genus separation aligned with known ecological and elevational niches. Ecologically broad genera (e.g. *Pinus*) show broader dispersion, consistent with wide environmental amplitude.

Error structure. The row-normalised confusion matrix (Fig. 7) exhibits a coherent block-diagonal structure. Several taxa show high recall, including Norway spruce (*Picea abies*; 0.86), European beech (*Fagus sylvatica*; 0.85), mountain pine (*Pinus mugo*; 0.84), European

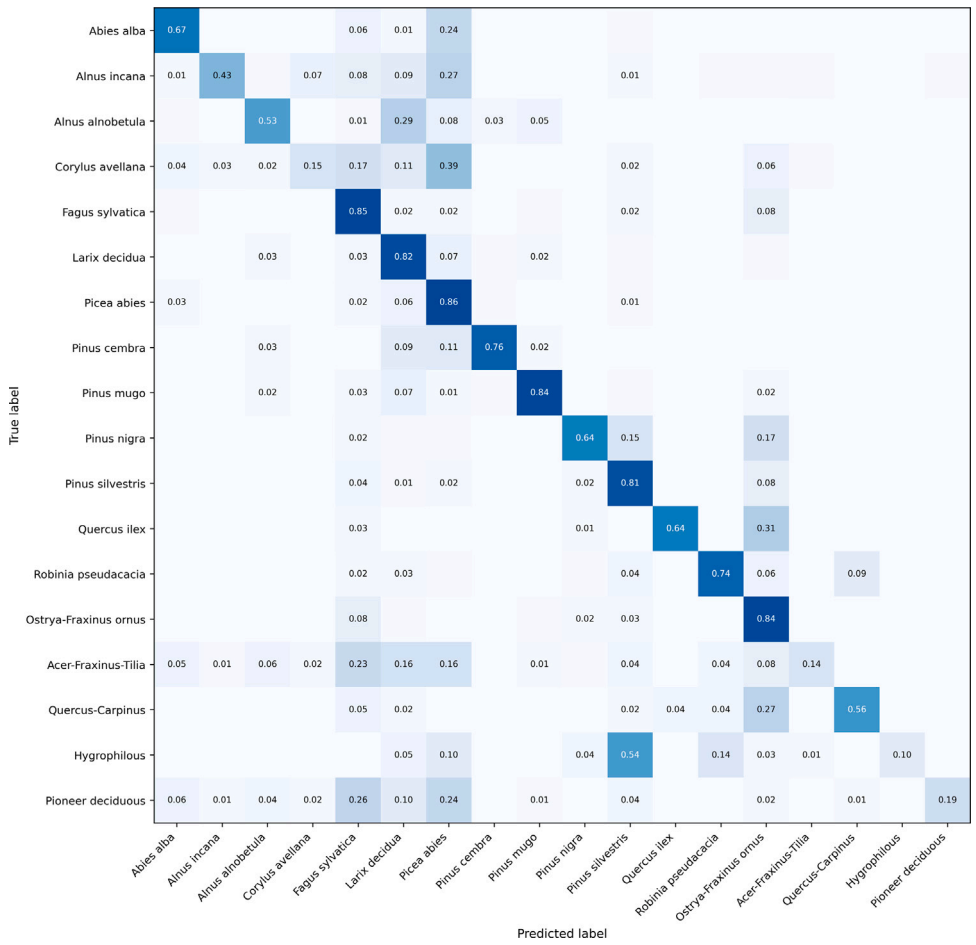

Fig. 7. Row-normalised confusion matrix for pixel-level species classification on validation parcels using MLP trained on Tessera embeddings. Eighteen species and merged species groups are shown.

larch (*Larix decidua*; 0.82), and Scots pine (*Pinus sylvestris*; 0.81). Misclassifications are concentrated within functional groups and genera: silver fir (*Abies alba*) is most often confused with Norway spruce (0.24), and moderate confusion occurs among pine species. Broadleaf taxa show more heterogeneous performance; mixed broadleaf assemblages (e.g. oak–hornbeam, hop-hornbeam–manna ash) are frequently confused with one another, partly reflecting ecological and spectral overlap but also the internal heterogeneity of merged composite classes. Overall, coniferous species are classified with notably higher mean per-class F1 (0.69) than broadleaf taxa (0.46), and errors are strongly structured by phylogenetic relatedness, functional type, and stand composition. Excluding boundary pixels via 1-pixel erosion improves weighted F1 by ~3 percentage points, confirming modest mixed-pixel effects at parcel edges (Table S.20). Results are robust to spatial blocking: relative to a matched parcel-aware control (both arms using fixed hyperparameters and a common training subsample to isolate fold geometry), spatial-block cross-validation (5 km grid) reduces weighted F1 by approximately 1 percentage point while preserving representation rankings (Table S.19).

3.2. Effect of downstream classifier capacity

To assess the sensitivity of classification performance to downstream model capacity, we evaluated the classifiers described in Section 2.7, spanning increasing expressive power (Fig. 9; full numerical results with per-fold standard deviations in Table S.8 and Fig. S.10). Both classifier choice and input representation affected performance, with their relative importance depending on the metric considered.

For all input representations, MLPs achieved the highest performance, followed by Random Forest and k-nearest neighbour classifiers, while linear models performed considerably worse. Using Tessera embeddings, macro F1 increased from 0.38 (LR) to 0.55 (MLP), a gain of +0.17; weighted F1 increased from 0.72 to 0.83 (+0.11). AlphaEarth embeddings showed comparable gains. The effect of switching input representation while holding the classifier fixed was smaller for macro F1: with an MLP, moving from S1+2 seasonal composites to Tessera improved macro F1 by +0.05 and weighted F1 by +0.02. Thus, for macro F1 – the metric most sensitive to minority-species discrimination – classifier choice had a larger effect than representation choice. A linear classifier applied to foundation-model embeddings performed below a seasonal S1+2 composite paired with an MLP.

Beyond this nonlinear threshold, additional capacity yielded diminishing returns: the gain from Random Forest to MLP was modest (+0.08 macro F1 on Tessera), and hyperparameter optimisation favoured shallow MLPs (one or two hidden layers) with strong regularisation, with the improvements concentrated in macro F1 (i.e. in the discrimination of rarer species). We discuss the interpretation of these capacity effects in Section 4.2.

3.3. Do environmental covariates help?

We tested whether supplementing foundation-model embeddings with terrain-derived covariates improves species classification performance. Across all evaluated metrics, the inclusion of terrain-derived covariates produced no meaningful performance gains relative to using embeddings alone (Table 4). For both Tessera and AlphaEarth

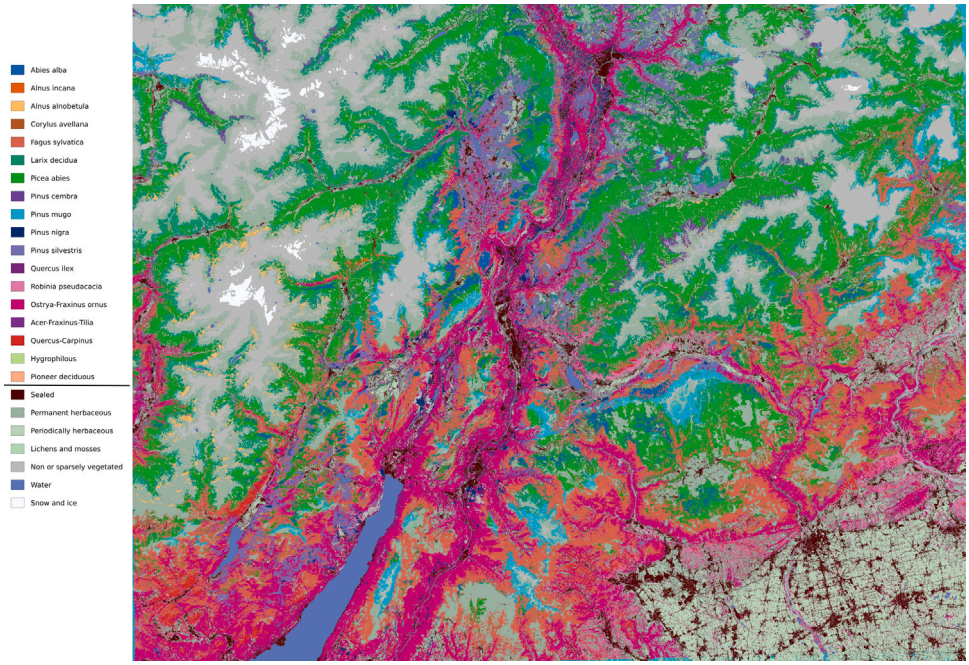

Fig. 8. Wall-to-wall species composition map of Trentino at 10 m resolution. Forest pixels are classified into 18 tree species and species groups by an MLP trained with soft labels (KL-divergence loss) on Tessera foundation-model embeddings, using provincial forest-inventory polygons as reference data. Non-forest land cover shown as the map background is taken from the Copernicus Land Monitoring Service CLC+ Backbone 2018 product, used here purely for cartographic display; the forest mask applied during model training and inference instead derives from the 10 m Global Land Cover 2020 product (see Methods).

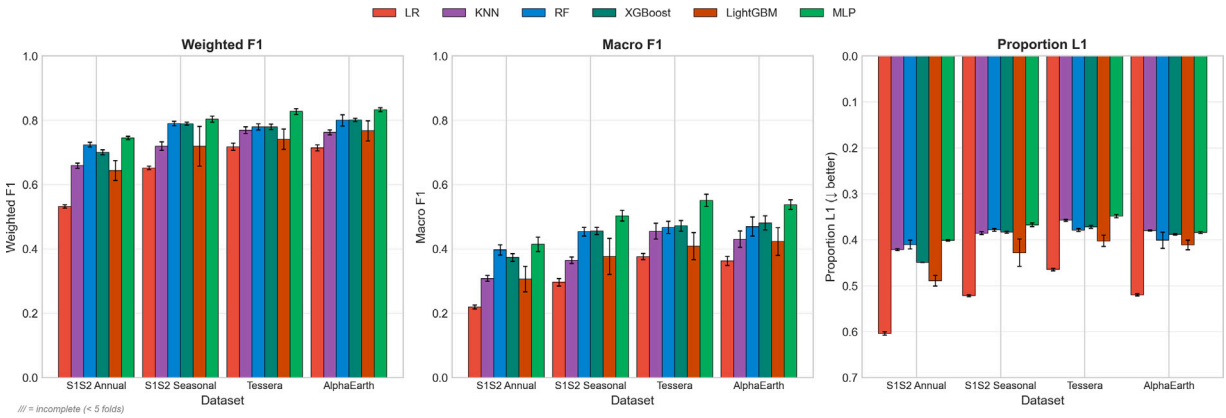

Fig. 9. Comparison of classification performance across six model architectures and input feature sets using 5-fold nested cross-validation with class-balanced training data. Metrics shown are (a) Weighted F1-score, (b) Macro F1-score, and (c) Proportion L1 error (inverted y-axis). Models compared: Logistic Regression (LR), K-Nearest Neighbours (KNN), Random Forest (RF), XGBoost, LightGBM, and Multi-Layer Perceptron (MLP). XGBoost and LightGBM are the gradient-boosting classifiers recommended by Reviewer 2. Input features: Sentinel-1/2 annual composite, Sentinel-1/2 seasonal composite, Tessera, and AlphaEarth embeddings. Error bars indicate standard deviation across folds.

embeddings trained on 2018 data, weighted F1, macro F1, and parcel-level proportion error (PropL1) remained effectively unchanged, with differences consistently < 0.005 across metrics.

These differences were small relative to cross-validation variability, confirming that terrain covariates did not benefit minority species any more than overall accuracy.

3.4. Sensitivity to label purity and weak supervision

We evaluated sensitivity to training-label purity by varying the minimum dominant-species threshold required for parcels included in training, while holding the overall training fraction constant (Fig. 10; Tables S.10–S.11). Pixel-level classification performance was comparatively robust across a broad range of purity thresholds. Weighted F1 increased slightly from low to intermediate purities (peaking around ~70%–80%) and then declined sharply when training was restricted to near-monocultures (≥90%–100%), consistent with a reduction in available training data. Macro F1 followed a similar pattern but showed greater sensitivity to high-purity filtering, indicating that aggressively excluding mixed parcels disproportionately degrades performance for less common species.

Table 4
Effect of adding terrain-derived covariates (DEM) to foundation-model embeddings for the MLP classifier (2018 data, 5-fold nested cross-validation). Differences between embeddings-only and +geofeatures are negligible (<0.005) for both representations.

Metric	Tessera		AlphaEarth	
	Embeddings only	+Geofeatures (DEM)	Embeddings only	+Geofeatures (DEM)
Accuracy	0.817	0.817	0.824	0.821
Weighted F1	0.827	0.827	0.833	0.831
Macro F1	0.551	0.548	0.537	0.540
Prop. L1	0.349	0.350	0.384	0.384

Table 5
Temporal transfer performance. MLP classifiers trained on 2018 embeddings evaluated on held-out 2018 parcels (within-year) and 2019 parcels (cross-year transfer). Values are mean ± s.d. across 5-fold cross-validation.

Input features	Weighted F1		Macro F1		PropL1	
	2018	2019	2018	2019	2018	2019
Tessera	0.842 ± 0.006	0.767 ± 0.006	0.573 ± 0.011	0.462 ± 0.004	0.346 ± 0.001	0.400 ± 0.005
AlphaEarth	0.848 ± 0.006	0.728 ± 0.011	0.568 ± 0.011	0.405 ± 0.030	0.375 ± 0.003	0.459 ± 0.009

Parcel-level composition fidelity exhibited the opposite dependence. Proportion error (PropL1) was lowest when mixed parcels were retained, remained close to this minimum across low-to-intermediate thresholds, and then rose as stricter purity filters were applied — markedly so at ≥90%. Thus, while high-purity filtering may marginally improve dominant-class assignment, it degrades the ability to recover realistic within-parcel species composition.

To test whether aggregate parcel-level species proportions could be leveraged, we trained MLP classifiers using soft labels in which each pixel received the full parcel-level species composition vector as its training target, optimised with KL-divergence loss (see Section S.6). Evaluated at 30% purity (the lowest threshold, maximising training volume), soft-label training achieved peak macro F1 of 0.586 for Tessera and 0.589 for AlphaEarth, exceeding the respective hard-label peaks of 0.581 and 0.565 obtained at 80% and 60% purity. In both cases, soft labels reached their best performance at low purity thresholds (30%–50%) using all available training data, while hard labels required filtering to high-purity parcels. Proportion L1 error – measuring the fidelity of predicted species proportions at the landscape scale – improved consistently under soft-label training, decreasing by 0.006–0.008 across the stable purity range for both embeddings (e.g. 0.335 vs. 0.343 for Tessera and 0.368 vs. 0.377 for AlphaEarth at 30% purity). These gains were disproportionately concentrated in less common species.

Across the purity curve, the soft-label advantage was largest at low purity thresholds where mixed parcels dominate, and narrowed at intermediate purity (70%–80%) where dominant-species labels are already largely correct (Fig. 10). The pattern was consistent across both foundation models, with AlphaEarth showing a larger macro F1 advantage (+0.03 at 30% purity) than Tessera (+0.02), suggesting that soft labels may partially compensate for weaker embedding-level species discrimination. Proportion L1 error showed a consistent advantage for soft labels across the full purity range for both Tessera and AlphaEarth, reflecting improved recovery of landscape-scale species proportions.

Taken together, these results show that foundation-model embeddings are tolerant to moderate label impurity in dominant-species training labels and that retaining mixed parcels does not substantially penalise pixel-level classification performance. Soft-label training provides targeted benefits for improving discrimination of less common species and, to a lesser extent, for recovering parcel-level species composition in heterogeneous stands.

3.5. Temporal transferability

To assess temporal generalisation, we trained MLP classifiers on 2018 embeddings and evaluated performance on both held-out 2018 parcels (within-year validation) and parcels from 2019 (cross-year transfer), using identical parcel-level partitions and high-purity validation criteria in both cases (Table 5; Fig. 11; see also Figs S.21–S.22). This design isolates temporal domain shift while preventing spatial leakage between years. To isolate the temporal effect, both years share a fixed-hyperparameter configuration; the within-year values reported here are therefore not directly comparable to the fully tuned results in Table 3 (and the parcel-aware control in Table S.19), and are intended for comparison across years rather than against the fully tuned configuration.

When transferred to 2019 embeddings without retraining, performance declined for both models. Tessera embeddings demonstrated greater temporal stability: weighted F1 decreased by 8.9% relative to within-year performance, compared to 14.1% for AlphaEarth. Differences were more pronounced for macro F1 (Tessera – 19.4%; AlphaEarth – 28.7%), indicating that minority species are disproportionately affected by temporal domain shift. Parcel-level composition error (PropL1) increased by 15.6% for Tessera and 22.3% for AlphaEarth.

Overall, while both embeddings support strong within-year classification, interannual domain shift remains a significant constraint. Tessera exhibits comparatively greater stability under cross-year transfer, though neither model achieves fully temporally invariant performance.

4. Discussion

This study shows that geospatial foundation-model embeddings enable accurate and label-efficient tree species mapping in a highly heterogeneous mountain landscape, consistently outperforming conventional Sentinel-1+2 composites. Below, we discuss each experimental finding in turn: the nature and magnitude of the performance advantage (Section 4.1), the limited role of classifier complexity (4.2) and environmental covariates (4.3), the implications of soft-label supervision (4.4), temporal transferability as a remaining challenge (4.5), the broader operational context (4.6), limitations (4.7) and future directions (4.8).

4.1. Why foundation-model embeddings outperform multispectral baselines

The superior performance of geospatial foundation-model embeddings relative to conventional, seasonal multispectral Sentinel-1+2 composites reflects a difference in how ecological information is represented. Traditional species-mapping approaches rely on hand-crafted spectral features derived from single dates or seasonal summaries, which are only weak proxies for the structural, phenological, biochemical and functional traits that differentiate tree species in heterogeneous forests. In contrast, foundation models are trained to integrate spectral information across time and sensor modalities, producing representations that integrate over cloud gaps, atmospheric variability, and

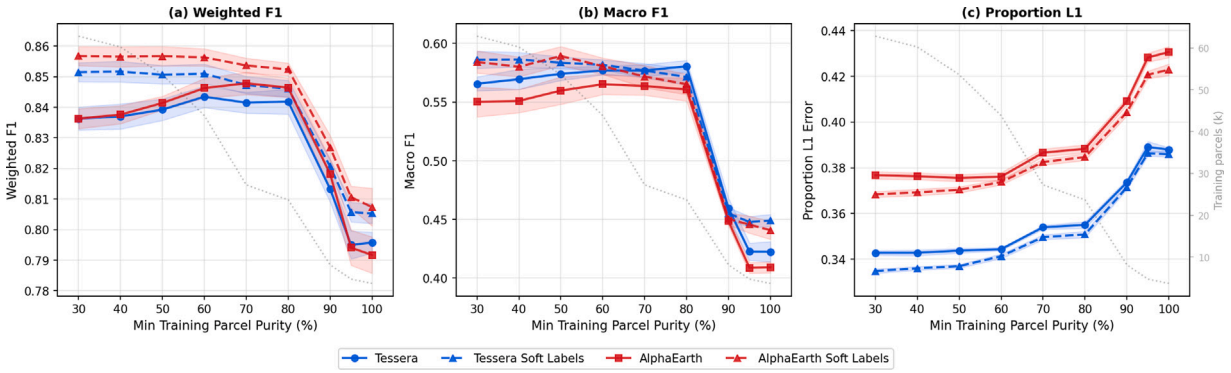

Fig. 10. Effect of training parcel purity threshold on MLP classification performance for Tessera and AlphaEarth. Solid lines: hard-label training (argmax); dashed lines: soft-label training (proportion targets). (a) Weighted F1. (b) Macro F1. (c) Proportion L1 error. Shaded ribbons: 95% confidence intervals from 5-fold CV repeated with 3 random seeds. Pixel-level validation uses parcels with ≥80% purity; parcel-level validation (PropL1) uses all parcels. Dotted grey line: number of available training parcels (right axis).

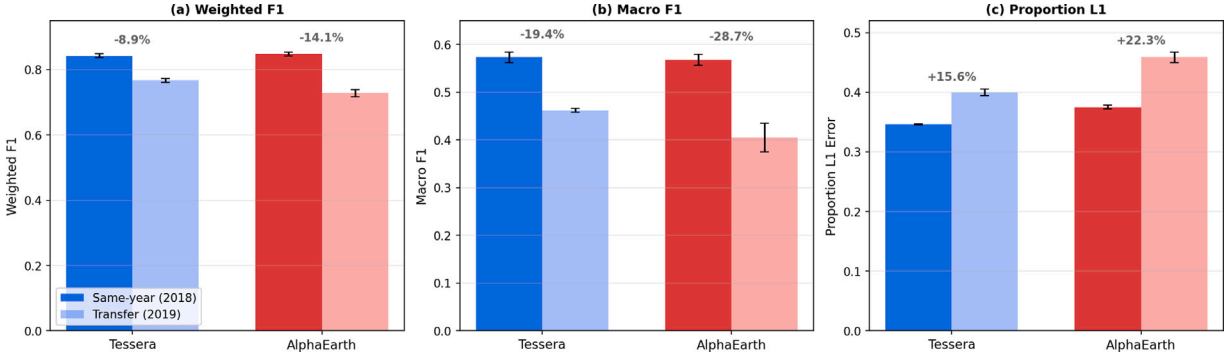

Fig. 11. Temporal transfer performance for Tessera and AlphaEarth embeddings. MLP classifiers trained on 2018 data evaluated on same-year holdout parcels (dark bars) and 2019 transfer data (light bars). (a) Weighted F1 score. (b) Macro F1 score. (c) Proportion L1 error. Percentages above bars indicate relative performance degradation from 2018 to 2019. Error bars show ±1 standard deviation across 5 CV folds.

irregular acquisition patterns while preserving ecologically meaningful signals (Xiao et al., 2025). Although recent standardised benchmarks show that GFMs do not consistently outperform supervised baselines across all Earth observation tasks (Marsocci et al., 2025), our results demonstrate that species-level forest classification in complex terrain is a domain where their advantages are pronounced.

The label efficiency observed here – with embeddings approaching saturation using ≤5% of available training parcels – suggests that discriminative structure closely aligned with species-relevant traits is already present in the representation rather than being learned during downstream training. That macro F1 continues to climb more steadily than weighted F1 confirms that rare-species discrimination remains the limiting constraint. The UMAP projections reinforce this interpretation, revealing coherent organisation along environmental gradients and taxonomic groupings without task-specific supervision.

These results compare favourably with the broader tree species classification literature. Reviews of satellite-based species mapping have noted that most studies use fewer than ten classes and often achieve overall accuracies in the range of 70%–85% with conventional features (Fassnacht et al., 2016; Zhong et al., 2024). Previous work in the same study region using fused multispectral and hyperspectral airborne data achieved strong discrimination among dominant conifers but required data sources with limited spatial and temporal coverage (Dalponte et al., 2012), while national-scale Sentinel time-series approaches in comparable European forests have reported lower accuracies at coarser taxonomic resolution (Blickensdörfer et al., 2024). Here, we classify 18 species and species groups – a finer scheme – and achieve a weighted F1 of 0.83 using only 10 m resolution Sentinel-derived embeddings, without airborne or hyperspectral data. This level of performance in a topographically complex mountain landscape with strong illumination and BRDF effects suggests that foundation-model representations overcome several of the key limitations historically associated with satellite-based species discrimination, including the need for carefully engineered multi-temporal feature sets and sensor-specific tuning (Immitzer et al., 2019). By contrast, multispectral composites collapse temporal variation into summary statistics, obscuring phenological dynamics critical for species discrimination (Xu et al., 2021). Richer multi-temporal feature sets can partially compensate, but at considerable processing overhead and sensitivity to cloud cover (Wang et al., 2022) — challenges that foundation models sidestep by learning directly from complete, multi-year image collections.

An interesting nuance is the pattern of relative advantage between the two foundation models investigated in our study. Recent comparative evaluations of GFM embeddings for tree species tasks have generally treated them as interchangeable feature extractors (Ishikawa et al., 2025), but our results reveal a more differentiated picture. Tessera outperforms AlphaEarth on macro F1 (0.551 vs. 0.537) and parcel-level proportion error (PropL1 = 0.349 vs. 0.384), while AlphaEarth achieves marginally higher weighted F1 (0.833 vs. 0.827; weighted F1 was not prioritised in classification tuning). This divergence is not straightforwardly explained by embedding dimensionality, since Tessera's advantage is concentrated in minority-species discrimination and compositional fidelity rather than in overall pixel accuracy. Instead, the two models' architectural differences offer a more compelling explanation. AlphaEarth operates at the patch level and is trained using auxiliary targets spanning multiple Earth-system modalities, including lidar structure, climate, and topography. This design intends to prioritise robustness and generality across tasks and spatial

contexts, but may over-smooth fine-grained spectral–phenological variation within parcels, explaining its weaker recovery of within-parcel species proportions. A complementary parcel-level aggregation test (Table S.18) is consistent with this interpretation: averaging embeddings within parcels before classification slightly improves Tessera, for which it suppresses within-parcel pixel noise, but degrades AlphaEarth, whose patch-level embeddings are already spatially smooth so that further averaging mainly removes discriminative variation. Tessera, by contrast, operates at the pixel level and is trained purely from remote-sensing time series using a self-supervised Barlow Twins objective (Lisaius et al., 2024). Its representations may therefore preserve canopy-trait variation that is particularly informative for differentiating co-occurring species and recovering realistic species mixtures at the parcel scale. This trade-off highlights that no single foundation-model representation is currently universally optimal: embeddings tuned for broad environmental generalisation may sacrifice local discriminative power, while models focused on sensor-derived signals may excel in region-specific ecological tasks and, as shown in our temporal transfer results, can retain greater stability across years.

The ecological coherence of the error structure provides further evidence that foundation-model representations encode biologically meaningful variation. Misclassifications are concentrated along taxonomic and functional axes – the silver fir–Norway spruce confusion pair, a persistent challenge in this region (Dalponte et al., 2012, 2025), and pine-species clusters – rather than randomly distributed. The consistently higher classification of conifers (mean per-class F1 = 0.69) compared with broadleaf taxa (0.46) reflects several compounding factors: broadleaf seasonal trajectories are broadly shared among co-occurring deciduous taxa, providing strong functional-type contrast but limited between-species signal at 10 m resolution; broadleaf species more frequently co-occur in mixed stands, introducing label ambiguity; and five of the eleven broadleaf classes are merged community groups. These patterns suggest that residual errors largely reflect genuine biochemical–structural similarity among co-occurring taxa rather than arbitrary noise, and for forest management the most frequent misclassifications occur between species that share similar silvicultural treatment.

Per-species comparison across representations (Table S.16) reveals that GFM embeddings provide the largest relative benefit for phenologically dynamic broadleaf species. Among deciduous broadleaves, *Fagus sylvatica* achieves F1 = 0.80 with both Tessera and AlphaEarth, compared with 0.73 for annual and 0.79 for seasonal S1+2 composites, while the thermophilous *Ostrya–Fraxinus ornus* group reaches F1 = 0.82 with embeddings versus 0.76 for annual composites. For rarer broadleaf taxa, the advantage is more pronounced: *Corylus avellana* reaches F1 = 0.18–0.21 with embeddings versus 0.03–0.16 with composites, and the hardest mixed-broadleaf classes (*Acer–Fraxinus–Tilia*, pioneer deciduous formations) remain challenging across all representations (F1 $\leq$ 0.23) yet still derive their largest gains from the embeddings. *Robinia pseudacacia*, a phenologically distinctive invasive, achieves F1 = 0.75 with Tessera versus 0.66 with annual composites, consistent with embeddings better capturing species-specific phenological trajectories. These patterns support the interpretation that GFM representations encode fine-grained temporal signatures that are most beneficial where species differ in leaf flush timing, senescence, or canopy structure.

Taken together, these findings indicate that the performance advantage of foundation models stems primarily from a more expressive and ecologically aligned representation of the landscape itself, rather than from downstream modelling choices. To probe the influence of spatial autocorrelation, we additionally evaluated a stricter spatial-block cross-validation (5 km grid; Table S.19). The representation rankings, overall accuracy, and parcel-level proportion estimates were preserved, and weighted F1 fell only slightly (by ~1–1.5 percentage points). Macro F1, however, declined more substantially (by ~6–7 percentage points), concentrated in the rarest, spatially clustered species. We do not wish to downplay this reduction: discrimination of rare species is genuinely sensitive to whether geographically nearby reference parcels are available during training. It also reflects a deliberately conservative test, since spatial blocking emulates extrapolation to entirely unsampled regions, whereas the parcel-aware design matches the intended gap-filling application of inferring unsurveyed parcels interspersed among surveyed ones. As a complementary check on mixed-pixel and boundary effects, parcel-erosion tests that exclude the outer pixel ring of each parcel produced only minor changes in accuracy (Table S.20), indicating that boundary contamination is not a major driver of the reported performance.

4.2. Classifier capacity: necessary but quickly saturating

A natural question arising from the strong performance of foundation-model embeddings is whether additional gains can be achieved through increased classifier capacity. Our results show that classifier choice matters – a linear classifier applied to foundation-model embeddings underperforms an MLP applied to conventional S1+2 seasonal composites – but that returns diminish rapidly once a moderately nonlinear model is used.

The step from linear to nonlinear classifiers produced the largest single-factor improvement observed in the study: macro F1 increased by +0.17 on Tessera embeddings, exceeding the gain from switching input representations with a fixed MLP (+0.05 from S1+2 seasonal to Tessera). This indicates that the discriminative information in foundation-model embeddings is not linearly separable, and that a nonlinear classifier is required to exploit it. However, once this threshold is crossed, further increases in complexity yield bounded gains. Shallow MLPs provided consistent improvements over Random Forest baselines – the dominant classifier in operational remote-sensing classification (Belgiu and Drăguţ, 2016) – primarily reflected in macro-averaged performance (+0.08 on Tessera), while deeper or more flexible architectures failed to yield further improvements. The concentration of performance gains in macro F1 indicates that nonlinear classifiers mainly benefit minority species, likely by smoothing decision boundaries in regions of the embedding space where rare classes form small but coherent clusters. The relative stability of weighted F1 across nonlinear classifiers implies that, for dominant species, the representation is already sufficiently structured for tree-based or instance-based methods to achieve strong performance.

The nature of the discriminative signal offers a mechanistic explanation for this pattern. Species identity is encoded in a pixel's spectral–temporal trajectory – phenology, leaf biochemistry – which is captured by per-pixel embeddings. Tessera's training explicitly decorrelates spatial neighbourhoods through global shuffling, so the resulting embeddings encode spectral–temporal properties of individual pixels rather than spatial structure. For species mapping this is precisely the relevant signal, and a per-pixel MLP is sufficient to exploit it. By contrast, structural attributes such as canopy height or biomass may benefit from spatial context – shadow geometry, textural arrangement, gap structure – explaining why the same embeddings benefit from spatial decoders such as UNets when applied to regression tasks (Feng et al., 2025).

From a practical perspective, these results indicate that at least a shallow nonlinear classifier is needed to exploit foundation-model embeddings for species mapping, but that pursuing increasingly complex architectures beyond this point offers little additional benefit.

4.3. The limited role of ancillary environmental covariates

Environmental variables such as elevation, slope, and topographic position are well known to structure forest composition in mountain landscapes (Weiss and Walsh, 2009). These covariates are therefore commonly incorporated into species distribution and classification models to compensate for topography-induced spectral distortions

(Dong et al., 2020) and to exploit the strong species–site relationships in mountain forests (Pu, 2021). It is therefore reasonable to ask whether explicitly providing such covariates can improve species discrimination beyond what is achievable with foundation-model embeddings alone. In this study, however, we find no evidence that ancillary environmental data provide meaningful additional predictive power.

Augmenting foundation-model embeddings with terrain-derived variables yielded no improvement in classification performance for either representation. For AlphaEarth, this outcome is perhaps unsurprising: its training explicitly incorporates DEM-derived topographic targets among its auxiliary objectives, so elevation and slope information is likely already directly encoded in the embedding. For Tessera, which is trained purely from optical and SAR time series without any explicit environmental inputs, the null result is more noteworthy. One interpretation is that key abiotic gradients are already indirectly captured through their ecological effects on canopy phenology, structure, and biochemical composition – a hypothesis supported by the UMAP visualisation showing clear elevation-related structure in Tessera embedding space – but alternative explanations should be considered. Simple feature concatenation may be a suboptimal integration strategy, the 30 m DEM resampled to 10 m may be too coarse to add meaningful local information, and the terrain-derived variables tested (elevation, slope, aspect, and topographic position index) may not capture the most relevant environmental gradients for species discrimination. The null result therefore demonstrates that terrain covariates are redundant under the specific experimental design employed here, but does not establish that all environmental information is already encoded in the embeddings.

For species mapping, the practical implication is that incorporating terrain-derived covariates via simple concatenation adds complexity, data dependencies, and preprocessing overhead without delivering commensurate gains in accuracy under the conditions tested. Whether alternative integration strategies (e.g. attention-based fusion, multi-scale environmental features, or higher-resolution terrain data) could yield improvements remains an open question. While explicit environmental data may still play an important role in other ecological modelling tasks – particularly those involving extrapolation beyond observed conditions – the present results suggest that in this study region they are not a prerequisite for effective species discrimination when using foundation-model representations.

4.4. Label quality, soft supervision, and the limits of hard classification

A common assumption in species-mapping workflows is that label noise represents a primary constraint on model performance, particularly when training data are derived from coarse or mixed inventory units (Bolyn et al., 2022). Our results paint a more nuanced picture: classification from foundation-model embeddings is surprisingly robust to moderate label impurity, hard-label refinement strategies did not help, but reframing supervision as soft labels unlocks gains that neither approach achieves alone.

The robustness to label impurity is notable, particularly given the wider literature showing that training-set size and composition can substantially affect supervised classification accuracy (Ramezan et al., 2021). Under hard-label training, performance remains nearly flat from 30% to 80% dominant-species purity, with aggressive filtering to near-monocultures (≥90%) degrading performance as data volume collapses. We hypothesise that foundation-model embeddings provide sufficient spectral–temporal structure to tolerate moderate label uncertainty: at lower purity thresholds, increased noise is offset by greater sample availability, while at very high thresholds, data scarcity degrades performance, particularly for minority species. The failure of label distillation (Section S.4) – iteratively reassigning hard labels within parcels using embedding-space prototypes – to improve pixel-level accuracy reinforces this interpretation: at 10 m resolution, where individual pixels integrate reflectance from multiple tree crowns, re-sorting discrete labels within a parcel cannot recover information that hard labels are structurally unable to encode.

The success of soft-label training, by contrast, reveals a deeper issue. The distinction between hard-label distillation and soft-label training is not merely technical but epistemic. The species composition of a parcel is known from forest inventory data, but the spatial arrangement of species within that parcel is not; assigning each pixel a single species therefore fabricates pixel-level certainty from parcel-level knowledge. Simultaneously, a 10 m Sentinel-2 pixel integrates reflectance from multiple tree crowns and is itself a spectral mixture. Hard labels thus impose a single-species target on a mixed-species input, using information that was never spatially resolved. Soft labels resolve both mismatches, encoding exactly the compositional knowledge available from the inventory while presenting the model with targets that are consistent with the mixed nature of the spectral input.

The consequences are visible across the purity curve (Section 3.4; Fig. 10). Where hard-label training produces an optimum at ∼80% purity, with performance falling on either side, the soft-label curve continues to improve from 30% to 50% purity. Mixed-species stands, which are noise sources under hard labelling, become informatively labelled training data under soft supervision, eliminating the trade-off between label quality and data volume. This pattern is consistent across both Tessera and AlphaEarth embeddings, with the latter showing a larger macro F1 advantage (+0.024 at peak), suggesting that soft labels can partially compensate for weaker embedding-level species discrimination. The improvement in Proportion L1 error (0.006–0.008 lower across the stable purity range for both embeddings) confirms that soft labels also improve recovery of landscape-scale species proportions. The macro F1 gains indicate that the primary beneficiaries are less common species that often occur as secondary components in mixed stands rather than forming monospecific parcels; under hard labelling their training signal is systematically suppressed whenever they are not the parcel majority, and soft labels recover it. Given that parcel-level species composition is already routinely recorded in national forest inventories (Brus et al., 2012), and that the majority of real forest parcels contain some degree of species mixing, these results suggest that the widespread practice of converting fractional composition data into hard per-pixel labels discards substantial useable information — and that the limiting factor is not label noise per se, but the assumption that supervision must be discrete. Our approach is complementary to recent work on sub-pixel species-fraction mapping using synthetically mixed training spectra (Klehr et al., 2025), which tackles the same fundamental problem – mixed pixels and discrete labels — from the feature-engineering side. Together, these lines of evidence suggest that moving from hard classification to fractional or soft-label formulations represents a broad methodological shift rather than an incremental refinement.

4.5. Temporal transferability as the remaining frontier

Despite the strong within-year performance, temporal generalisation under single-year training remains a challenge. Domain shift between training and deployment periods is well known in remote-sensing classification (Tuia et al., 2016); a key question is whether large-scale pre-training confers temporal robustness. Our results show that it does not fully do so: applying 2018-trained classifiers to 2019 data produced substantial accuracy declines concentrated in macro-averaged metrics, indicating that rare species are particularly sensitive to interannual domain shift.

Several factors contribute. Interannual phenological variation alters spectral signals encoded in embeddings (Cambridge Centre for Earth Observation, 2026), and differences in snow persistence, cloud availability, and acquisition geometry exacerbate domain shift. Critically, the 2018–2019 year-pair coincides with Storm Vaia (Seidl et al., 2017; Senf and Seidl, 2021), which caused widespread windthrow across Trentino, making it impossible to fully separate representational drift

from genuine landscape change. Because these results are based on a single year-pair in one study region, the magnitude of temporal degradation should not be generalised without broader evaluation.

Under cross-year transfer, Tessera retains a greater proportion of its within-year performance than AlphaEarth, suggesting that its purely remote-sensing-driven training captures more temporally invariant features. This finding is consistent with early evidence from embedding-based crop classification (Lisaius et al., 2026) and independent evaluation of AlphaEarth (Ma et al., 2025), confirming that interannual variability in frozen embeddings is a systematic consideration rather than a study-specific artefact.

At the same time, temporal variation in embedding space may also be an opportunity: temporal trajectories could be exploited to study phenological shifts, stress responses, or disturbance dynamics, bridging species mapping with functional monitoring. Operationally, these results caution that naïvely applying a static, single-year classifier across years risks conflating phenological variation with genuine compositional change. Overall, temporal generalisation remains a critical frontier; addressing it through multi-year training, temporal domain adaptation, or phenology-aware representation learning will be essential for operational deployment.

4.6. Implications for operational forest monitoring

These findings have direct implications for national forest inventory programmes and conservation agencies. Foundation-model embeddings achieve strong classification with minimal downstream complexity and as few as 5% of available training parcels, substantially reducing labelling effort compared to conventional approaches (McRoberts and Tomppo, 2007; White et al., 2016; McRoberts et al., 2010). For agencies that already collect parcel-level species proportions, the soft-label finding is particularly actionable: existing compositional data routinely discarded when converted to hard labels can instead be used directly as training targets, improving minority-species discrimination without additional field effort.

In the context of emerging policy needs such as the EU Biodiversity Strategy for 2030 and associated Forest Strategy (Directorate-General for Environment, 2023), wall-to-wall species maps at 10 m resolution (Fig. 8) represent a step beyond the genus-level products currently available at European scale (De Keersmaecker et al., 2024) — provided that temporal transfer challenges can be addressed. Realising this potential at continental scales will require attention to cross-regional transferability, inventory harmonisation, and temporally robust training strategies.

4.7. Limitations

Several limitations should be noted when interpreting these results. First, all experiments are conducted in a single study region (Trentino), which, while ecologically complex and demanding, represents only one biogeographic setting; the degree to which findings generalise to other forest types, climatic zones, or inventory systems remains to be tested. Second, temporal transfer is evaluated over a single year-pair (2018–2019) that coincides with Storm Vaia (Section 4.5), making it difficult to fully disentangle representational drift from genuine landscape change. Third, we adopt a pixel-level hard classification approach rather than sub-pixel fractional prediction. At Sentinel-2's 10 m resolution, individual pixels typically capture one or a few tree crowns, limiting the degree of sub-pixel species mixing; moreover, pixel-level classification is broadly compatible with existing operational mapping workflows and inventory systems. Our soft-label experiments further demonstrate that compositional reference data can be exploited effectively within this framework, though a fully fractional formulation could offer additional benefits in genuinely mixed stands. Fourth, foundation-model embeddings are evaluated in a frozen, off-the-shelf configuration without fine-tuning; task-specific adaptation has shown promise for vegetation mapping (Ishikawa et al., 2025) and could potentially improve performance, particularly for rare species, but would complicate the assessment of representation quality per se. Similarly, hierarchical classification schemes that exploit the taxonomic structure of species labels (Jin and Wu, 2025) or hyperspectral foundation models with richer spectral resolution (Braham et al., 2025) remain unexplored directions that could further improve discrimination of spectrally similar taxa. Finally, the reference data are parcel-level inventory proportions rather than individual tree-level labels; while this is standard for forest inventory studies, it means that pixel-level predictions are validated against spatially aggregated proportions rather than at the scale of individual stems.

4.8. Future directions

The strong label efficiency demonstrated here – approaching saturation performance from ≤5% of parcels – raises the question of whether even sparser supervision could suffice if training samples are selected more strategically. In preliminary experiments comparing alternative sampling strategies (Section S.7.4; Table S.14), coreset selection guided by K-means clustering in the embedding space consistently outperformed random stratified sampling across both foundation models, improving weighted F1 by up to 1.6 percentage points at 5% training fraction and by 3.3 points at 0.1% (~45 parcels). These results point toward few-shot species mapping in data-sparse regions, where a small number of strategically selected reference parcels – guided by unsupervised structure in the embedding space – may be sufficient for operational classification without region-specific campaigns. More broadly, extending this approach to multi-year training with explicit temporal domain adaptation, sub-pixel fractional prediction building on the soft-label formulation explored here, explicit spatial-context modelling of these per-pixel representations through patch-based architectures or spatial decoders, and task-specific fine-tuning of embeddings would address several of the limitations noted above and move the approach closer to wall-to-wall operational deployment.

4.9. Conclusions

Geospatial foundation-model embeddings enable accurate and label-efficient tree species mapping in complex mountain forests, consistently outperforming conventional Sentinel-1+2 composites when paired with a nonlinear classifier. Embeddings encode species-discriminative information more effectively than conventional composites, with the advantage most pronounced for minority species; all representations approach performance saturation with limited training data, but embeddings do so at a consistently higher ceiling. Realising this advantage requires at least a moderately nonlinear classifier – the step from linear to nonlinear models produces the single largest improvement – but further increases in model capacity yield diminishing returns. Classification is robust to moderate label impurity, and soft species-composition labels substantially improve minority-species discrimination, suggesting that converting fractional inventory data into discrete labels discards useable information. Supplementary environmental covariates provide no additional predictive power when concatenated with embeddings. Temporal transfer across years reveals marked performance degradation, particularly for rare species, with embedding choice influencing temporal stability.

These findings are based on a single study region and require multi-site validation before generalisation. Nonetheless, they suggest that geospatial foundation models shift a primary bottleneck in species mapping from feature engineering toward the availability, quality, and temporal alignment of ecological reference data. Realising their full potential for operational biodiversity monitoring will require temporally consistent training strategies and reference datasets that embrace the compositional complexity of real forest stands.

CRediT authorship contribution statement

James G.C. Ball: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jana Annika Wicklein:** Writing – review & editing, Investigation, Data curation. **Zhengpeng Feng:** Writing – review & editing, Software. **Jovana Knezevic:** Writing – review & editing. **Sadiq Jaffer:** Writing – review & editing, Software. **Anil Madhavapeddy:** Writing – review & editing, Supervision, Resources, Funding acquisition. **Clement Atzberger:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Michele Dalponte:** Writing – review & editing, Resources, Data curation. **David A. Coomes:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to edit and proofread text drafted by the authors. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

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.

Acknowledgements

JGCB was supported by the Medical Research Council (MR/Z505456/1). SJ was funded by a charitable donation from John Bernstein. ZF was supported by the Robert Sansom Studentship. We thank Forestmap.ai Research for providing computational resources used in this study.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.srs.2026.100466.

Data availability

The forest inventory data used in this study were provided by the Autonomous Province of Trento and are available upon reasonable request to the corresponding author. Processed inventory parcels, foundation-model embeddings, other intermediate data and landscape predictions are available on Zenodo: https://zenodo.org/records/18924120 . Code is available at https://github.com/PatBall1/trentino-trees. All experiments used Python 3.13, scikit-learn 1.7.2, and rasterio 1.4.3, with random seed 42 for reproducibility. Run scripts and configuration files are included in the repository.

References

Agnoletti, M., Biasi, R., 2013. Trentino alto adige. In: Agnoletti, M. (Ed.), Italian Historical Rural Landscapes: Cultural Values for the Environment and Rural Development. Springer Netherlands, Dordrecht, ISBN: 978-94-007-5354-9, pp. 247–262. http://dx.doi.org/10.1007/978-94-007-5354-9_10.

Akiba, T., et al., 2019. Optuna: A next-generation hyperparameter optimization framework. arXiv:1907.10902 [cs].

Bai, J., et al., 2024. Tree species diversity impacts on ecosystem services of temperate forests. Ecol. Indic. (ISSN: 1470-160X) 167, 112639. http://dx.doi.org/10.1016/j.ecolind.2024.112639, URL https://www.sciencedirect.com/science/article/pii/S1470160X24010963.

Ball, J.G.C., et al., 2024. Harnessing temporal and spectral dimensionality to map and identify species of individual trees in diverse tropical forests. bioRxiv, 2024.06.24.600405.

Belgiu, M., Drăguţ, L., 2016. Random forest in remote sensing: A review of applications and future directions. ISPRS J. Photogramm. Remote Sens. (ISSN: 0924-2716) 114, 24–31. http://dx.doi.org/10.1016/j.isprsjprs.2016.01.011, URL https://www.sciencedirect.com/science/article/pii/S0924271616000265.

Bergseng, E., et al., 2015. Assessing forest inventory information obtained from different inventory approaches and remote sensing data sources. Ann. For. Sci. (ISSN: 1297-966X) 72 (1), 33–45. http://dx.doi.org/10.1007/s13595-014-0389-x.

Blickensdörfer, L., et al., 2024. National tree species mapping using sentinel-1/2 time series and german national forest inventory data. Remote Sens. Environ. (ISSN: 0034-4257) 304, 114069. http://dx.doi.org/10.1016/j.rse.2024.114069, URL https://www.sciencedirect.com/science/article/pii/S0034425724000804.

Bolyn, C., et al., 2022. Mapping tree species proportions from satellite imagery using spectral–spatial deep learning. Remote Sens. Environ. (ISSN: 0034-4257) 280, 113205. http://dx.doi.org/10.1016/j.rse.2022.113205, URL https://www.sciencedirect.com/science/article/pii/S0034425722003145.

Braham, N.A.A., et al., 2025. SpectralEarth: Training hyperspectral foundation models at scale. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. (ISSN: 2151-1535) 18, 16780–16797. http://dx.doi.org/10.1109/JSTARS.2025.3581451, URL https://ieeexplore.ieee.org/document/11045062.

Brandt, M., et al., 2025. High-resolution sensors and deep learning models for tree resource monitoring. Nat. Rev. Electr. Eng. (ISSN: 2948-1201) 2 (1), 13–26. http://dx.doi.org/10.1038/s44287-024-00116-8, URL https://www.nature.com/articles/s44287-024-00116-8.

Bray, J.R., Curtis, J.T., 1957. An ordination of the upland forest communities of southern wisconsin. Ecol. Monograph. (ISSN: 0012-9615) 27 (4), 326–349. http://dx.doi.org/10.2307/1942268, URL https://www.jstor.org/stable/1942268.

Breiman, L., 2001. Random forests. Mach. Learn. (ISSN: 1573-0565) 45 (1), 5–32. http://dx.doi.org/10.1023/A:1010933404324.

Brockerhoff, E.G., et al., 2017. Forest biodiversity, ecosystem functioning and the provision of ecosystem services. Biodivers. Conserv. 26 (13), 3005–3035.

Brown, C.F., et al., 2025. AlphaEarth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. arXiv:2507.22291 [cs].

Brus, D.J., et al., 2012. Statistical mapping of tree species over Europe. Eur. J. For. Res. (ISSN: 1612-4677) 131 (1), 145–157. http://dx.doi.org/10.1007/s10342-011-0513-5.

Cambridge Centre for Earth Observation, 2026. URL https://github.com/ucam-eo/geotessera. original-date: 2025-07-02T10:24:37Z.

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD'16, Association for Computing Machinery, New York, NY, USA, pp. 785–794. http://dx.doi.org/10.1145/2939672.2939785.

Copernicus Land Monitoring Service / EU CLMS, 2020. Land cover 2020 (raster 10 m), global, annual - version 1. URL https://land.copernicus.eu/en/products/global-dynamic-land-cover/land-cover-2020-raster-10-m-global-annual.

Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE Trans. Inform. Theory (ISSN: 1557-9654) 13 (1), 21–27. http://dx.doi.org/10.1109/TIT.1967.1053964, URL https://ieeexplore.ieee.org/document/1053964.

Dalponte, M., Andreatta, D., 2026. A spatially explicit dataset of upper canopy tree species composition of public forests of the Autonomous Province of Trento, Italy. Data Brief 66, 112736. http://dx.doi.org/10.1016/j.dib.2026.112736.

Dalponte, M., Bruzzone, L., Gianelle, D., 2012. Tree species classification in the Southern Alps based on the fusion of very high geometrical resolution multispectral/hyperspectral images and LiDAR data. Remote Sens. Environ. 123, 258–270.

Dalponte, M., et al., 2025. Canopy spectral responses of temperate forests to late spring frost and hot drought events assessed with Sentinel-2 NDVI time series. Remote. Sens. Appl.: Soc. Environ. (ISSN: 2352-9385) 40, 101737. http://dx.doi.org/10.1016/j.rsase.2025.101737, URL https://www.sciencedirect.com/science/article/pii/S2352938525002903.

De Keersmaecker, W., et al., 2024. ForestPaths: European tree genus map. URL https://zenodo.org/records/13341104.

Directorate-General for Environment, 2023. Proposal for a Regulation on a Forest Monitoring Framework - Environment, URL https://environment.ec.europa.eu/publications/proposal-regulation-forest-monitoring-framework_en.

Dong, C., et al., 2020. The effect of topographic correction on forest tree species classification accuracy. Remote. Sens. (ISSN: 2072-4292) 12 (5), 787. http://dx.doi.org/10.3390/rs12050787, URL https://www.mdpi.com/2072-4292/12/5/787.

Farr, T.G., et al., 2007. The shuttle radar topography mission. Rev. Geophys. (ISSN: 1944-9208) 45 (2), http://dx.doi.org/10.1029/2005RG000183, URL https://onlinelibrary.wiley.com/doi/abs/10.1029/2005RG000183, _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2005RG000183.

Fassnacht, F.E., et al., 2016. Review of studies on tree species classification from remotely sensed data. Remote Sens. Environ. 186, 64–87.

Feng, Z., et al., 2025. TESSERA: Temporal embeddings of surface spectra for earth representation and analysis. version: 4. URL arXiv:2506.20380 [cs].

Frei, E.R., et al., 2023. Evidence for 40 years of treeline shift in a central Alpine Valley. Forests (ISSN: 1999-4907) 14 (2), 412. http://dx.doi.org/10.3390/f14020412, URL https://www.mdpi.com/1999-4907/14/2/412.

Gao, X., et al., 2024. Comparison of cloud-mask algorithms and machine-learning methods using sentinel-2 imagery for mapping paddy rice in Jianghan plain. Remote. Sens. (ISSN: 2072-4292) 16 (7), http://dx.doi.org/10.3390/rs16071305, Company: Multidisciplinary Digital Publishing Institute Distributor: Multidisciplinary Digital Publishing Institute Institution: Multidisciplinary Digital Publishing Institute Label: Multidisciplinary Digital Publishing Institute, URL https://www.mdpi.com/2072-4292/16/7/1305.

Gasparini, P., et al. (Eds.), 2022. Italian national forest inventory—methods and results of the third survey: Inventario nazionale delle foreste e dei serbatoi forestali di carbonio—metodi e risultati della terza indagine. In: Springer Tracts in Civil Engineering. Springer International Publishing, Cham, http://dx.doi.org/10.1007/978-3-030-98678-0, ISBN: 978-3-030-98677-3 978-3-030-98678-0. URL https://link.springer.com/10.1007/978-3-030-98678-0.

Gibbs, A.L., Su, F.E., 2002. On choosing and bounding probability metrics. Int. Stat. Rev. (ISSN: 1751-5823) 70 (3), 419–435. http://dx.doi.org/10.1111/j.1751-5823.2002.tb00178.x, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-5823.2002.tb00178.x, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1751-5823.2002.tb00178.x.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning. In: Springer Series in Statistics. Springer, New York, NY, http://dx.doi.org/10.1007/978-0-387-84858-7, ISBN: 978-0-387-84857-0 978-0-387-84858-7. URL http://link.springer.com/10.1007/978-0-387-84858-7.

Hu, Z., et al., 2025. Application of landsat high spatial resolution phenological synthesized data in mountainous land cover classification. Remote. Sens. (ISSN: 2072-4292) 17 (15), 2603. http://dx.doi.org/10.3390/rs17152603, URL https://www.mdpi.com/2072-4292/17/15/2603.

Immitzer, M., Atzberger, C., 2023. Tree species diversity mapping—Success stories and possible ways forward. Remote. Sens. (ISSN: 2072-4292) 15 (12), 3074. http://dx.doi.org/10.3390/rs15123074, URL https://www.mdpi.com/2072-4292/15/12/3074.

Immitzer, M., et al., 2019. Optimal input features for tree species classification in central europe based on multi-temporal sentinel-2 data. Remote. Sens. (ISSN: 2072-4292) 11 (22), 2599. http://dx.doi.org/10.3390/rs11222599, URL https://www.mdpi.com/2072-4292/11/22/2599.

Ishikawa, T., et al., 2025. Deep pre-trained time series features for tree species classification in the dutch forest inventory. arXiv:2508.18829 [cs].

Jin, Y., Wu, C., 2025. Plant taxonomy meets deep learning: Hierarchical classification for fine-grained tree species identification with hyperspectral foundation models. URL https://papers.ssrn.com/abstract=5573998.

Ke, G., et al., 2017. LightGBM: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc., pp. 3146–3154, URL https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html.

Kellndorfer, J., et al., 1998. Toward consistent regional-to-global-scale vegetation characterization using orbital SAR systems. IEEE Trans. Geosci. Remote Sens. (ISSN: 1558-0644) 36 (5), 1396–1411. http://dx.doi.org/10.1109/36.718844, URL https://ieeexplore.ieee.org/document/718844.

Klehr, D., et al., 2025. Mapping tree species fractions in temperate mixed forests using sentinel-2 time series and synthetically mixed training data. Remote Sens. Environ. (ISSN: 0034-4257) 323, 114740. http://dx.doi.org/10.1016/j.rse.2025.114740, URL https://www.sciencedirect.com/science/article/pii/S0034425725001440.

Lisaius, M.C., et al., 2024. Using barlow twins to create representations from cloud-corrupted remote sensing time series. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. (ISSN: 2151-1535) 17, 13162–13168. http://dx.doi.org/10.1109/JSTARS.2024.3426044, URL https://ieeexplore.ieee.org/document/10592304.

Lisaius, M.C., et al., 2026. Embedding -based crop type classification in the groundnut basin of Senegal. arXiv:2601.16900 [cs].

Ma, Y., et al., 2025. Harvesting AlphaEarth: Benchmarking the geospatial foundation model for agricultural downstream tasks. arXiv:2601.00857 [cs].

Marsocci, V., et al., 2025. PANGAEA: A global and inclusive benchmark for geospatial foundation models. arXiv:2412.04204 [cs].

McRoberts, R.E., Tomppo, E.O., 2007. Remote sensing support for national forest inventories. Remote Sens. Environ. (ISSN: 0034-4257) 110 (4), 412–419. http://dx.doi.org/10.1016/j.rse.2006.09.034, URL https://www.sciencedirect.com/science/article/pii/S0034425707001721.

McRoberts, R.E., Tomppo, E.O., Næsset, E., 2010. Advances and emerging issues in national forest inventories. Scand. J. For. Res. (ISSN: 0282-7581) 25 (4), 368–381. http://dx.doi.org/10.1080/02827581.2010.496739.

Pedrotti, F., 2018. Vegetation series along climatic gradients in the central southern alps (trentino- Alto Adige Region). In: Pedrotti, F. (Ed.), Climate Gradients and Biodiversity in Mountains of Italy. Springer International Publishing, Cham, ISBN: 978-3-319-67967-9, pp. 51–81. http://dx.doi.org/10.1007/978-3-319-67967-9_3.

Pu, R., 2021. Mapping tree species using advanced remote sensing technologies: A state-of-the-art review and perspective. J. Remote. Sens. 2021. http://dx.doi.org/10.34133/2021/9812624, URL https://spj.science.org/doi/full/10.34133/2021/9812624.

Ramezan, C.A., et al., 2021. Effects of training set size on supervised machine-learning land-cover classification of large-area high-resolution remotely sensed data. Remote. Sens. (ISSN: 2072-4292) 13 (3), http://dx.doi.org/10.3390/rs13030368, URL https://www.mdpi.com/2072-4292/13/3/368.

Rautiainen, M., et al., 2011. Seasonal reflectance dynamics of common understory types in a northern European boreal forest. Remote Sens. Environ. (ISSN: 0034-4257) 115 (12), 3020–3028. http://dx.doi.org/10.1016/j.rse.2011.06.005, URL https://www.sciencedirect.com/science/article/pii/S003442571100229X.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature (ISSN: 1476-4687) 323 (6088), 533–536. http://dx.doi.org/10.1038/323533a0, URL https://www.nature.com/articles/323533a0.

Seidl, R., et al., 2017. Forest disturbances under climate change. Nat. Clim. Chang. (ISSN: 1758-6798) 7 (6), 395–402. http://dx.doi.org/10.1038/nclimate3303, URL https://www.nature.com/articles/nclimate3303.

Senf, C., Seidl, R., 2021. Persistent impacts of the 2018 drought on forest disturbance regimes in Europe. Biogeosciences (ISSN: 1726-4170) 18 (18), 5223–5230. http://dx.doi.org/10.5194/bg-18-5223-2021, URL https://bg.copernicus.org/articles/18/5223/2021/.

Tilman, D., Kinzig, A.P., Pacala, S., 2013. The Functional Consequences of Biodiversity: Empirical Progress and Theoretical Extensions. Princeton University Press, ISBN: 978-1-4008-4730-3, http://dx.doi.org/10.1515/9781400847303, URL https://www.degruyterbrill.com/document/doi/10.1515/9781400847303/html.

Tuia, D., Persello, C., Bruzzone, L., 2016. Domain adaptation for the classification of remote sensing data: An overview of recent advances. IEEE Geosci. Remote. Sens. Mag. (ISSN: 2168-6831) 4 (2), 41–57. http://dx.doi.org/10.1109/MGRS.2016.2548504, URL https://ieeexplore.ieee.org/abstract/document/7486184.

Wang, D., et al., 2022. An empirical study of remote sensing pretraining. IEEE Trans. Geosci. Remote Sens. 61, 1–20.

Wang, H., et al., 2023. Leaf economics fundamentals explained by optimality principles. Sci. Adv. 9 (3), eadd5667.

Weiss, D.J., Walsh, S.J., 2009. Remote sensing of mountain environments. Geogr. Compass (ISSN: 1749-8198) 3 (1), 1–21. http://dx.doi.org/10.1111/j.1749-8198.2008.00200.x, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-8198.2008.00200.x, _eprint: https://compass.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1749-8198.2008.00200.x.

White, J.C., et al., 2016. Remote sensing technologies for enhancing forest inventories: A review. Can. J. Remote Sens. (ISSN: 0703-8992) 42 (5), 619–641. http://dx.doi.org/10.1080/07038992.2016.1207484.

Wilson, J.P., Gallant, J.C., 2000. Terrain Analysis: Principles and Applications. John Wiley & Sons, ISBN: 978-0-471-32188-0.

Xiao, A., et al., 2025. Foundation models for remote sensing and earth observation: A survey. IEEE Geosci. Remote. Sens. Mag. (ISSN: 2168-6831) 2–29. http://dx.doi.org/10.1109/MGRS.2025.3576766, URL https://ieeexplore.ieee.org/abstract/document/11097335.

Xu, K., et al., 2021. How spatial resolution affects forest phenology and tree-species classification based on satellite and up-scaled time-series images. Remote. Sens. (ISSN: 2072-4292) 13 (14), 2716. http://dx.doi.org/10.3390/rs13142716, URL https://www.mdpi.com/2072-4292/13/14/2716.

Zbontar, J., et al., 2021. Barlow twins: Self-supervised learning via redundancy reduction. arXiv:2103.03230 [cs].

Zhong, L., et al., 2024. A review: Tree species classification based on remote sensing data and classic deep learning-based methods. Forests (ISSN: 1999-4907) 15 (5), 852. http://dx.doi.org/10.3390/f15050852, URL https://www.mdpi.com/1999-4907/15/5/852.