AlphaEarth Embeddings Deliver France-Wide Deciduous–Evergreen Maps Without Local GPU Training

Arthur Calvi *

Independent work, Paris, France

Correspondence*:
Arthur Calvi
arthur.calvi@icloud.com

Unpublished working manuscript – not peer reviewed. August 2026.

ABSTRACT

Foundation models provide precomputed embeddings that could replace user-defined remote-sensing features, but their benefit over user-defined baseline features remains unclear under controlled evaluation. We benchmark Google's AlphaEarth annual 10 m embeddings against harmonic descriptors fitted to Sentinel-2 vegetation-index time series for the task of classifying deciduous and evergreen forests across metropolitan France. Using 14.1 million labeled forest pixels and tile-based spatial cross-validation for large forest ecological regions in France, we train the same lightweight classifiers on 14-dimensional feature sets derived either from embeddings or from harmonic features. Embeddings deliver a consistent classification advantage, increasing mean macro-F1 by 3.7 percentage points across Logistic Regression, Linear SVM, and Random Forest classifiers. Beyond their higher accuracy, embedding-based predictions are better calibrated and produce more spatially coherent maps without post-processing. A label-source diagnostic (BD Forêt polygons versus in-situ/expert labels) reveals measurable source shift, motivating pooled training for national robustness, while a simple test using both embeddings and harmonics features suggests limited but non-zero complementarity of adding harmonics to the embeddings. Overall, the embeddings can act as a drop-in replacement for harmonic features for our classification purpose, thus reducing reliance on GPU training.

Keywords: foundation models, AlphaEarth embeddings, feature engineering, harmonic analysis, Random Forest, phenology, deciduous-evergreen

1 INTRODUCTION

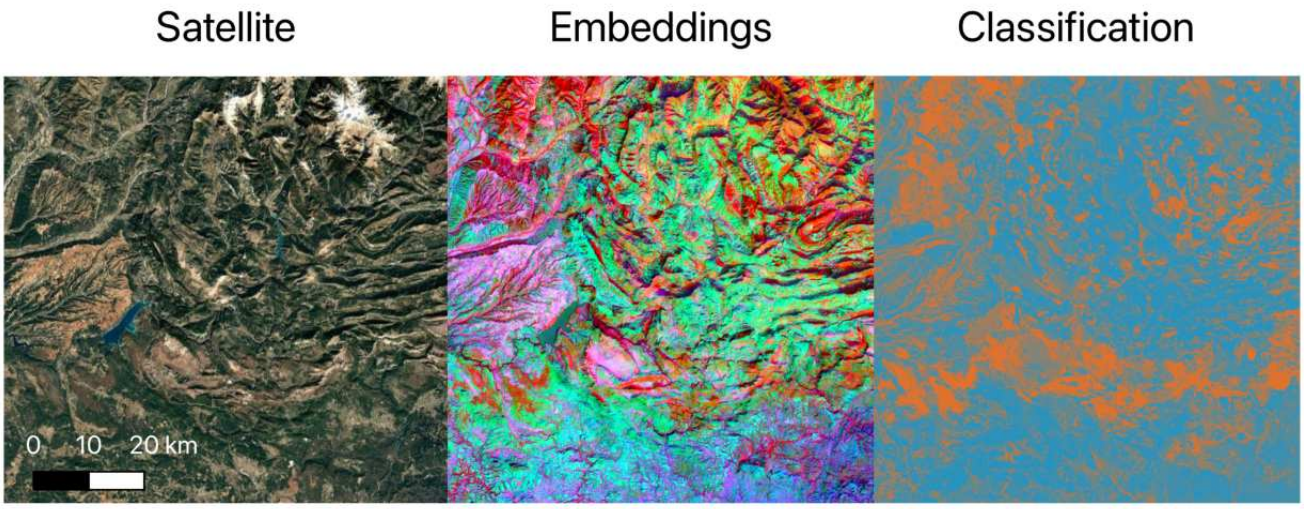

Figure 1. Southeast France preview: true-color Sentinel-2 (left), AlphaEarth embedding visualization (middle), and classification map (right). The classification is not masked with the national forest mask, all pixel are classified.

Remote sensing is providing imagery at unprecedented volumes for mapping forest attributes, but creating accurate maps from sparse and heterogeneous labels remains challenging (Gorelick et al., 2017; Zhu and Woodcock, 2014). In practice, national mapping pipelines often face a compute-versus-craft dilemma. One option to perform forest types classification is to create user-defined interpretable features based on spectral indices, seasonal metrics, etc... and train lightweight classifiers. The other is to learn features end-to-end with deep neural networks at the cost of heavier training and deployment.

The craft strategy has deep roots in Earth observation. Fitting vegetation-index time series with simple functions yields compact phenology descriptors such as the amplitude and timing of greenness cycles (Jönsson and Eklundh, 2002). Time-series methods such as BFAST (Verbesselt et al., 2010), harmonic regression (Wilson et al., 2018), and break-point detection algorithms (Kennedy et al., 2010; Zhu and Woodcock, 2014) remain widely used for land cover classification and forest types mapping based on phenological signals because they capture the seasonal behavior related to forest types with a few parameters and can be deployed with modest computation. Many national-scale forest classification products build on user-defined features. The Copernicus Dominant Leaf Type layer, for instance, uses multi-year Sentinel-2 indices to classify broadleaf versus conifer forest at 10–20 m resolution (European Union, 2024). By using indices such as NDVI, NBR, or EVI and summarising their seasonal cycles, user-defined features such as harmonics provide intuitive descriptors for forest-type mapping that work well with simple classifiers (e.g. Random Forests (Belgiu and Drăguţ, 2016)), including in operational toolchains such as iota2 (Inglada et al., 2017).

In contrast, the compute-intensive strategy uses deep learning with task-specific features directly derived from multi-date imagery. Architectures from UNet and vision transformers have been applied to land cover and phenology mapping with notable gains in detail and accuracy; for example, continental-scale phenology maps have been produced by integrating Landsat and Sentinel time series into deep networks (Bolton et al., 2020). Emerging large-scale models such as billion-parameter vision transformers (Cha et al.,

2023) and generative approaches such as DiffusionSat (Khanna et al., 2023) can capture complex spatial and temporal patterns beyond the reach of simple index-based descriptors. However, these approaches typically require substantial GPU resources and enough local labels to adapt the model to each region or task, limiting their applicability for routine national mapping.

Foundation models for Earth observation have recently emerged as a third path (Alp, 2025). These models are pretrained (often self-supervised) on large multi-sensor archives and provide generic embeddings that can be used as inputs to lightweight downstream classifiers. The key promise is representation reuse: feature quality is obtained once during pretraining, then transferred to new mapping tasks with minimal additional computation (Xie et al., 2024). Early examples of foundation models include SatMAE (Cong et al., 2022), which learns from unlabeled satellite data with masked-autoencoder objectives, and Prithvi (Singh et al., 2023), a multi-temporal Earth-observation model designed for downstream geospatial tasks. The Presto model further illustrates the trend toward learned temporal representations that remain usable under missing observations and heterogeneous inputs (Ishikawa et al., 2025). Recently, Google's AlphaEarth model produced a 64-dimensional annual embedding field at 10 m resolution for 2017-2024 (Alp, 2025) with global coverage, making embeddings available as analysis-ready features for multiple applications.

Given the rapid rise of foundation embeddings, a question is how much practical benefit they provide over the previous domain-informed and user-defined features under the same transparent evaluation. Recent work has started to benchmark foundation representations in remote sensing, but controlled comparisons against classical baselines remain limited (Dionelis et al., 2024; Xie et al., 2024). Two practical points are particularly important for operational users. First, "information content" is best assessed by downstream performance: in addition to testing whether one feature family can be predicted from another, one can directly test whether combining them improves classification. Second, training supervision is rarely homogeneous at national scale. Reference datasets often mix photo-interpreted polygons and in-situ or expert-verified observations; if these supervision sources differ in noise, class definitions, or sampling focus, then model fit and generalisation can depend strongly on which source is used for training and evaluation.

In this work, we provide a controlled benchmark of AlphaEarth embeddings versus harmonics of multipectral Sentinel-2 time-series descriptors for a concrete national mapping task: the classification of deciduous versus evergreen forest types across metropolitan France. This binary classification task is ecologically meaningful and operationally relevant, yet remains challenging in ecotones and Mediterranean landscapes where evergreen and deciduous trees co-exist (Li et al., 2023; Francini et al., 2024). We compare (i) harmonic descriptors fitted to a single annual cycle of Sentinel-2 index observations and (ii) AlphaEarth embeddings, while holding the training data, spatial splits, and classifier models identical. We use three classifier models: logistic regression, linear Support Vector Machine, and Random Forest. To enforce spatial independence, our evaluation uses tile-based cross-validation rather than pixel-wise splits. Classification performance is reported nationally and by the large forest ecological regions known as GRECO (Institut national de l'information géographique et forestière, 2013).

Beyond the benchmark of our binary forest type classification between embeddings and harmonics features, we include two targeted diagnostics motivated by the questions raised above. First, we test the complementarity of harmonics with embeddings by training models based on both input data and evaluating whether the fused input data improves performance relative to embeddings alone. Second, we quantify label-source shift by training classification models on one type of labels (BD Forêt photo-interpreted polygons versus in-situ/expert-derived labels) and evaluating their results against the two others. We also relate eco-region deltas to the regional density of in-situ labels to test whether spatial sampling drives

the shift, motivating pooled training as the most robust national strategy. Finally, we assess the temporal portability of the embeddings by training a model trained on one year of embeddings to classify forest types during other years, as a practical test for making annual updates without having to retrain models. Together, these analyses illustrate the performances of models using foundation embeddings in a representation-focused, data-aware evaluation framework for national-scale mapping of evergreen and deciduous forest types.

2 DATA AND METHODS

Our experimental design isolates representation quality under controlled conditions. We compare the use of user-defined widely used harmonic features with foundation-model embeddings as input data for the same classification models, while holding the training data and feature dimensionality identical. This keeps the evaluation focused on the features themselves rather than on downstream model choices.

2.1 Reference data (labels) and sampling units

France spans diverse climates and forest types, including continental, alpine forests with a marked seasonality, and Mediterranean, Atlantic pine forests where seasonality is less obvious. We structure both sampling and reporting using the eleven large forest ecoregions *Grandes Régions Ecologiques* (GRECO) that capture major differences in climate, relief, and dominant forest formations (Institut national de l'information géographique et forestière, 2013). Table 1 summarises the GRECO eco-regions and the number of labelled pixels sampled in each.

2.1.0.1 Tiles for training and test.

To control for spatial autocorrelation and define the train/test separation explicit, we partition the forest domain into non-overlapping $2.5\,$km tiles. Tiles play two roles. First, they define the unit of spatial independence in cross-validation: a tile belongs entirely to either the training set or the validation set. Second, they provide a practical unit to balance the geographic coverage across ecological contexts. The number of tiles per GRECO is computed according to the effective area of forest. We selected 639 tiles distributed across all GRECO regions (Figure 2). Each tile is assigned to a GRECO region using its spatial overlay (e.g., the dominant GRECO area within the tile), and the tile set is balanced so that all ecoregions are represented in each fold.

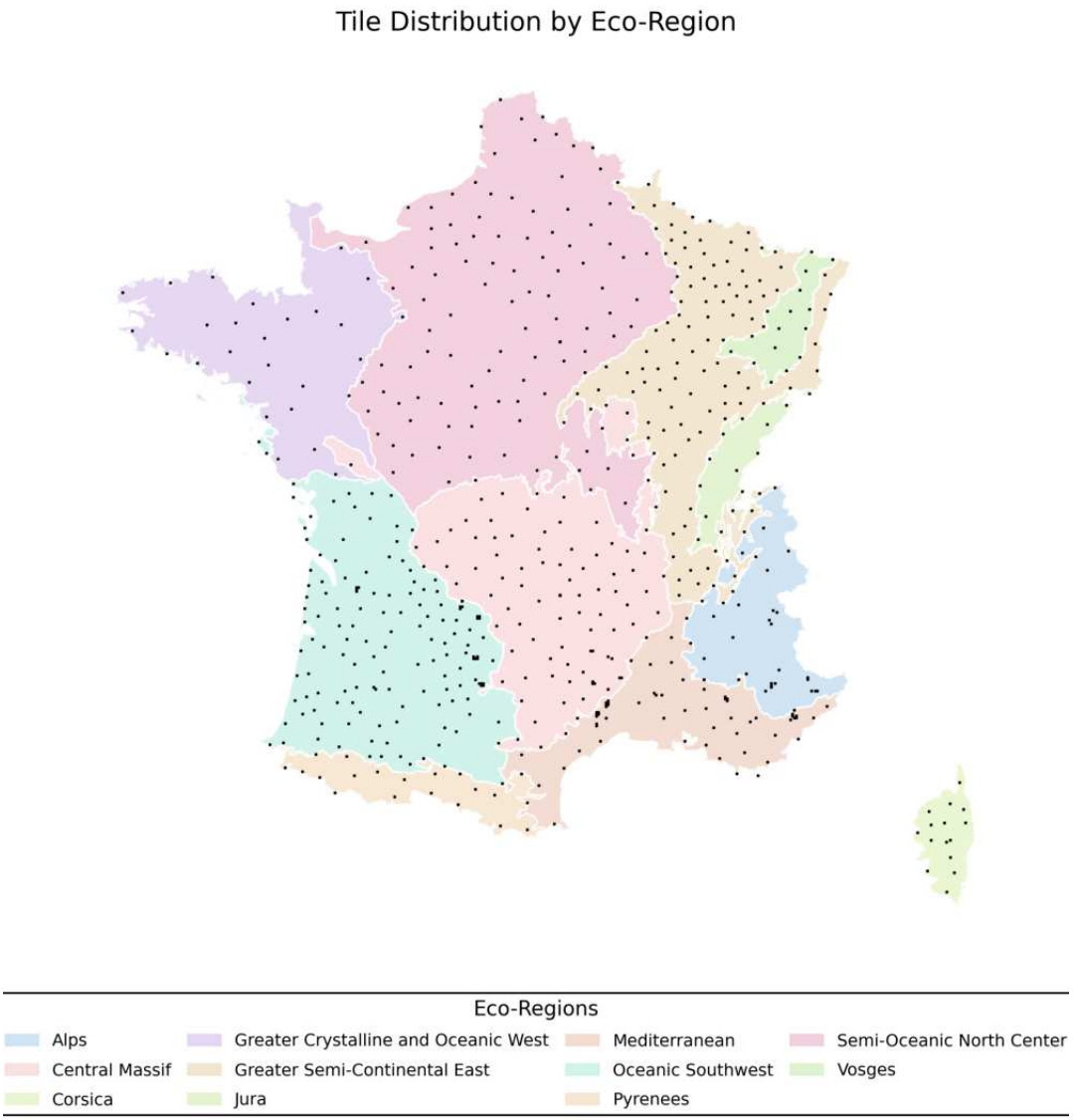

Figure 2. Distribution of the 639 2.5 km tiles used for sampling and spatial cross-validation. Background colours show GRECO ecoregions. Label-source composition varies across tiles and is summarised in Section 3.1.

2.1.0.2 Two label families.

- **In-situ / expert-derived labels .** These labels have a high precision but a limited spatial extent. They include (i) PureForest LiDAR patches in the south (50 m mono-specific forest plots pairing 0.2 m aerial imagery with $\sim$40 pts m^{-2} airborne LiDAR with 18 tree species grouped into 13 classes. (Gaydon et al., 2024), (ii) RENECOFORe inventory plots (Ulrich et al., 2021), and (iii) LiDAR-corrected GPS points from the Tree Position Calibration campaign (Office National des Forêts, 2024; Institut national de l'information géographique et forestière, 2024b). These sources provide the most reliable anchors in heterogeneous or spectrally ambiguous landscapes.
- **Photo-interpreted polygons (wall-to-wall coverage).** BD Forêt V2 provides national coverage derived from aerial photo interpretation for an epoch varying by region between 2007 and 2018, with polygons reporting forest types and mixed conifer / mixed broadleaved classes (Institut national de

l'information géographique et forestière, 2024a). Because the polygon labels can include mixed pixels along boundaries; to reduce edge contamination we apply a $100\,\mathrm{m}$ negative buffer before sampling pixels.

Across the dataset, in-situ/expert labels represent 11.4% of the samples and BD Forêt polygons 88.6%. The different species or genus were pooled into an evergreen or deciduous category, with 75.5% deciduous and 24.5% evergreen, forming a single training set of 14.1 million labeled pixel, each attached to an identifier of its original data source. Formally, the training data can be written as

$$\mathcal{D} = \{(\mathbf{x}_i,\ y_i,\ s_i,\ g_i,\ t_i)\}_{i=1}^{N}, \tag{1}$$

where $\mathbf{x}_i$ is the feature vector (harmonics, embeddings, or both), $y_i \in \{\text{deciduous}, \text{evergreen}\}$ is the target label, s_i indicates the label source (in-situ/expert vs photo-interpretation), g_i is the GRECO region, and t_i is the tile id used for spatial splitting. The source tag s_i is not used as a model input; it is retained to quantify potential domain shift between label families (Section 3.1).

Code	Region (English name)	Climate and terrain	Dominant forest structure	Training pixels
A	Greater Crystalline and Oceanic West	Humid Atlantic plains and low hills, dense bocage networks	Oak–chestnut coppice with maritime pine and Sitka spruce plantations	0.49 M
B	Semi-Oceanic North Centre	Loess plateaus and chalk cuestas of the Paris Basin	Pedunculate/sessile oak with hornbeam or beech; Scots/Corsican pine on sandy soils	2.39 M
C	Greater Semi-Continental East	Ardennes and Lorraine uplands with colder winters	Beech–fir and spruce on mesic plateaus, mixed oak lowlands	2.80 M
D	Vosges	Steep crystalline range with high precipitation	Silver fir–beech high forests with spruce and Douglas-fir plantations	0.39 M
E	Jura	Limestone plateaus under cool montane climate	Calcareous beech–fir mosaics and mixed spruce on karst slopes	0.20 M
F	Oceanic Southwest	Landes coastal plain and Atlantic piedmont	Maritime pine estates interleaved with humid oak, alder, and chestnut stands	2.91 M
G	Central Massif	Volcanic plateaus and valleys with montane climate	Beech–fir belts, chestnut groves, Scots/Douglas pine on poorer soils	1.92 M
H	Alps	Sharp elevational gradients and deep glacial valleys	Deciduous foothills grading to spruce–larch–fir subalpine belts	0.65 M
I	Pyrenees	Atlantic–Mediterranean transition, steep valleys	Oak/beech foothills with fir–spruce upper slopes and Mediterranean pine on south-facing flanks	0.56 M
J	Mediterranean	Coastal ranges and plateaus with intense summer drought	Evergreen holm and cork oak maquis, Aleppo and maritime pine stands	1.43 M
K	Corsica	Crystalline massif with rugged relief	Lowland evergreen maquis and extensive Laricio pine forests above 900 m	0.35 M

Table 1. GRECO ecoregions used for sampling and evaluation; counts sum to 14.1 million labelled pixels.

2.2 Label-source shift diagnostic

Because our labels are a mix of photo-interpreted polygons (BD Forêt) with in-situ pixels, we quantify cross-source generalisation in a controlled diagnostic. We construct two source-specific training sets:

$$\mathcal{D}^{(1)} = \{(\mathbf{x}_i, y_i) : s_i = \text{BD Forêt}\}, \qquad \mathcal{D}^{(2)} = \{(\mathbf{x}_i, y_i) : s_i = \text{in-situ/expert}\}. \tag{2}$$

Using the same tile-based folds as in the main benchmark, we train models on one source at a time and evaluate them on (i) BD Forêt pixels, (ii) in-situ pixels, and (iii) the pooled validation set. We compute macro-F1 by GRECO region for each source-trained model and compare the regional deltas with the regional share of in-situ labels. We report this diagnostic using harmonic features to isolate label/source effects from representation effects.

2.3 Feature representations: harmonic descriptors and foundation embeddings

We compare two families of feature computed on the same forest pixels. The first family (HARM) encodes annual phenology explicitly via a harmonics parametric model fitted to Sentinel-2 index time series

2.3.0.1 Harmonic descriptors (HARM).

From Sentinel-2 Level-2A surface reflectances, we compute four vegetation indices (NDVI, EVI, NBR, and CRSWIR; CRSWIR definition in Supplement) after cloud masking with `s2cloudless` (75% probability threshold). For each index trajectory $z(t)$, we fit a two-harmonic regression over one annual cycle:

$$z(t) = c + \sum_{k=1}^{2} \left[a_k \cos\left(\frac{2\pi k t}{T}\right) + b_k \sin\left(\frac{2\pi k t}{T}\right) \right] + \varepsilon(t), \tag{3}$$

where T is one year and $\varepsilon(t)$ captures departures from periodicity. We derive compact descriptors from the fitted coefficients: offset c, amplitudes $A_k = \sqrt{a_k^2 + b_k^2}$, and phase angle $\phi_k = \operatorname{atan2}(b_k, a_k)$ encoded as $(\cos\phi_k, \sin\phi_k) = (a_k/A_k,\ b_k/A_k)$ to avoid discontinuities at 2π, plus the residual variance $\sigma^2 = \mathrm{Var}(\varepsilon)$. This yields a pool of 32 candidate harmonic descriptors (4 indices $\times$ [c, A_1, $\cos\phi_1$, $\sin\phi_1$, A_2, $\cos\phi_2$, $\sin\phi_2$, σ^2]). Such harmonic summaries are widely used as interpretable phenology features in forest types classifcations mapping (Wilson et al., 2018; Bolton et al., 2020).

To match the downstream model budget and keep the benchmark compact, we retain a 14-dimensional subset (Table 2) using recursive feature elimination with cross-validation (RFECV), performed within the training tiles of each fold. This subset defines the **HARM-14** feature set used in the main comparison.

2.3.0.2 AlphaEarth embeddings (EMB).

AlphaEarth provides annual 10 m embedding rasters where each pixel is represented by a 64-dimensional vector learned by a large self-supervised model trained on petabyte-scale multi-sensor archives (Alp, 2025). The embedding channels are not designed to be individually interpretable; instead, they act as generic descriptors that summarize spectral-temporal trajectories and local spatial context in a way that can be exploited by lightweight classifiers.

We define an **EMB-14** benchmark by selecting 14 embedding dimensions from the 64-channel stack using the same RFECV procedure as for HARM. This parity setting enables an apples-to-apples comparison at fixed feature dimensionality. Because embedding performance can improve when using more than 14

channels, we additionally report (in Supplementary Table S1) the performance of higher-dimensional embedding variants (e.g. EMB-64 or the RFECV-optimal k), so the main results can be interpreted as a conservative estimate of the embedding advantage.

HARM descriptor	**Ecological signal**
NDVI first-harmonic amplitude	Seasonal strength of broadleaf greenness pulses
NDVI first-harmonic phase (cosine/sine)	Timing of green-up and senescence transitions
NDVI second-harmonic phase (sine)	Asymmetry between rapid spring and gradual autumn trajectories
NDVI offset	Mean canopy greenness throughout the year
NBR first-harmonic amplitude	Annual moisture and structural contrast between canopies and soil
NBR first-harmonic phase (cosine)	Calendar timing of minimum fuel moisture
NBR second-harmonic phase (cosine)	Secondary moisture cycle in bimodal climates
NBR offset	Average woody biomass signal
NBR residual variance	Short-term disturbance departures from harmonic behaviour
NBR first-harmonic phase (sine)	Complementary timing cue for fuel moisture (sine component)
CRSWIR first-harmonic phase (cosine)	Timing of peak shortwave-infrared water stress
CRSWIR second-harmonic phase (cosine)	Recovery pattern of evergreen water content
CRSWIR offset	Mean canopy water and lignin content
CRSWIR residual variance	Fine-scale heterogeneity in SWIR response

Table 2. Fourteen harmonic descriptors retained by RFECV from 32 candidates (4 indices × [offset, two harmonic amplitudes, sine/cosine phase terms, residual variance]). Phase is represented as sine/cosine components and each component counts as a separate feature. Together these define the **HARM-14** set.

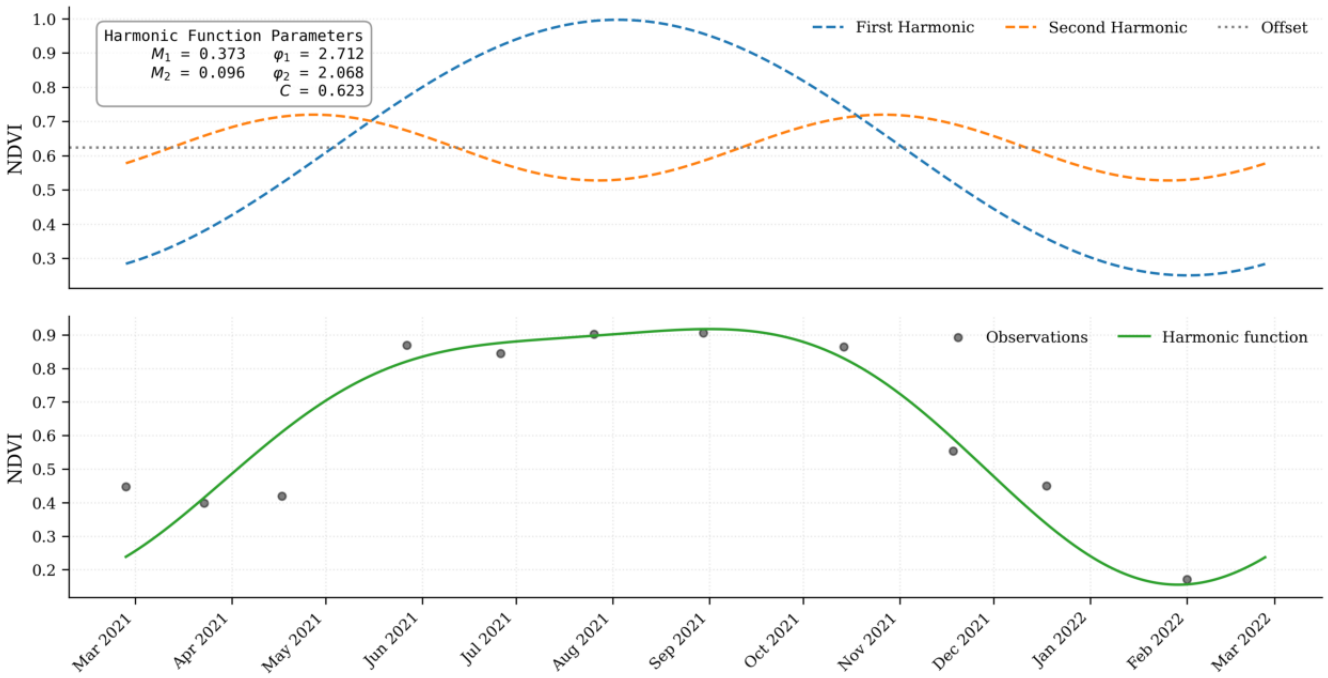

Figure 3. Harmonic decomposition of an NDVI annual curve; offsets, amplitudes, phases, and residual variance form the harmonic features. The 12 and 6-month components stack to reproduce asymmetric trajectories (e.g., rapid spring green-up, slower autumn decline, or a secondary pulse).

2.4 Classifiers and evaluation protocol

We use three simple classifiers : Logistic Regression (elastic-net), Linear SVM, and Random Forest. All comparisons use the same labelled pixels, the same sample-weighting scheme, and the same tile-based spatial splits; only the feature representation changes (HARM-14, EMB-14, or their concatenation).

Linear models are trained on standardised features (zero mean, unit variance) computed on the training folds only. Random Forests use the raw feature values. Class imbalance is handled via balanced class weights (and equivalent sample weighting when applicable), so the loss gives comparable importance to deciduous and evergreen pixels.

2.4.0.1 Tile-based spatial cross-validation.

To enforce spatial independence, we split data at the $2.5\,\mathrm{km}$ tile level. Let $\mathcal{T}$ be the set of tiles and let $\mathcal{T}_k$ denote the tiles assigned to fold k. For fold k, the validation set is

$$\mathcal{D}_{\mathrm{val}}^{(k)} = \{(\mathbf{x}_i, y_i) : t_i \in \mathcal{T}_k\}, \tag{4}$$

and the training set is defined by the complementary tiles. This prevents leakage through local spatial autocorrelation because no tile contributes pixels to both training and validation in the same fold.

2.4.0.2 GRECO-stratified folds.

Each tile is assigned a GRECO label. We then build folds by stratified assignment within GRECO groups: tiles are shuffled inside each GRECO and distributed across the $K = 5$ folds in a round-robin fashion. This ensures that every fold contains tiles from all GRECO regions, while keeping the GRECO composition approximately balanced between folds.

2.4.0.3 Hyperparameters.

To keep the benchmark representation-focused and avoid representation-specific tuning as an extra degree of freedom, hyperparameters are selected once on the harmonic baseline and kept fixed across feature sets.

2.5 Feature fusion: embeddings plus harmonics

If the HARM features contain information not present in the embeddings (or vice versa), then a model trained on both features should improve predictive performance under the same evaluation protocol.

For each pixel i, we define a 14-dimensional embedding vector $\mathbf{e}_i \in \mathbb{R}^{14}$ and a 14-dimensional harmonic vector $\mathbf{h}_i \in \mathbb{R}^{14}$ (selected as described above). We then form the fused representation by simple concatenation:

$$\mathbf{x}_i^{(\mathrm{FUSE})} = \begin{bmatrix} \mathbf{e}_i & \mathbf{h}_i \end{bmatrix} \in \mathbb{R}^{28}. \tag{5}$$

The experiment with HARM + EMB is simple, with no interaction terms, and no spatial post-processing, so any gain can be attributed to information added by the extra feature family. We trained the same classifiers as in the main benchmark (Logistic Regression and Random Forest) using the same spatial splits and weighting scheme. Because fusion doubles the feature dimension (28 vs 14), it is not a complexity-controlled comparison. An AIC-style penalty is not defined for the RF head, so we treat fusion as a complementarity diagnostic and keep hyperparameters fixed. Because embeddings and harmonic features were computed in separate streams, the fusion dataset was built by an inner join on pixel identifiers (tile id and pixel coordinates; year when applicable), yielding the intersection of valid pixels across both feature

stacks. This detail does not change the logic of the test (*does HARM add information beyond EMB?*), but it should be kept in mind when comparing absolute scores to single-feature baselines computed on slightly different sample counts. Performance is reported as overall accuracy and macro-F1 averaged over 5 cross-validation folds.

2.6 Temporal Portability of embeddings

We trained our model with a RF classifier on the embedding data for the year 2023 and applied it with embedding data from 2018, 2020, and 2022 to classify forest types during those years. The overall accuracy (OA) was recomputed on the same reference pixels (mixing BD Forêt polygons with in-situ labels for 2022) to isolate the effect of temporal drift in the embeddings with the train-test split being the same. Table 8 reports these metrics; Figure 5 shows representative tiles.

2.7 Comparison with previous datasets

We assessed the consistency of our deciduous/evergreen classification against two existing datsets, the Copernicus Dominant Leaf Type (DLT; broadleaf vs. conifer) and the BD Forêt V2. Comparisons are restricted to forest pixels using a national forest mask (Institut National de l'Information Géographique et Forestière, 2024; national de l'information géographique et forestière , IGN). We computed per-tile confusion matrices between our predictions and each existing dataset, then aggregated them nationally and by GRECO. Because BD Forêt contributes to the training supervision, agreement with BD Forêt should be interpreted as *map-to-map consistency*, not as an independent validation. Because DLT uses a different definition of forest classes (broadleaf vs. conifer rather than deciduous vs. evergreen), part of the disagreement—especially in holm oak and larch areas—reflects class-definition differences rather than model error.

2.8 Calibration and spatial coherence metrics

2.8.0.1 Calibration.

We computed the expected calibration error (ECE) from $M = 10$ equally spaced bins over predicted evergreen probability $p \in [0, 1]$. Let B_m be the set of samples in bin m, with empirical accuracy $\mathrm{acc}(B_m)$ and mean confidence $\mathrm{conf}(B_m)$. We define

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|, \qquad \mathrm{MCE} = \max_{m\in\{1,\ldots,M\}} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|. \tag{6}$$

2.8.0.2 Spatial coherence.

We quantify map fragmentation on the raw 10 m prediction grid (no morphological smoothing), after masking non-forest pixels. Patch density is the number of 8-connected components of the predicted classes per 100 km^2 of forest:

$$\mathrm{PatchDensity} = \frac{N_{\mathrm{cc}}}{A_{\mathrm{forest}}/100}, \tag{7}$$

where N_{cc} is the total number of connected components and A_{forest} is forest area in km^2. Edge density measures boundary length per km^2 of forest:

$$\mathrm{EdgeDensity} = \frac{L_{\mathrm{edge}}}{A_{\mathrm{forest}}}, \tag{8}$$

where L_{edge} is computed from 4-neighbour class transitions, converted to km using the 10 m pixel size. Lower values indicate less "salt-and-pepper" fragmentation.

3 RESULTS

3.1 Label-source shift

We quantify cross-source generalisation between BD Forêt polygon labels and in-situ labels using the diagnostic described in Section 2.2. Two consistent patterns emerge (Table 3; full metrics in Supplementary Table S5. First, evaluation on in-situ labels yields higher macro-F1 than evaluation on BD Forêt polygons, regardless of the training source. Second, models trained only on in-situ labels transfer less well to BD Forêt than the reverse, indicating a measurable source shift and/or higher effective label noise in photo-interpreted polygons. At the eco-region level, the sign and magnitude of the train-source delta track the spatial distribution of in-situ labels: regions with minimal in-situ labels coverage (Semi-Oceanic North Center, Vosges, Greater Semi-Continental East) show the largest penalties when training only on in-situ labels, whereas the Mediterranean region (40% of the in-situ labels) is the only region where the in-situ-trained Random Forest exceeds the BD Forêt-trained Random Forest. Across the 11 regions, the percentage of in-situ labels correlates with the macro-F1 delta ($r \approx 0.61$ for Random Forest, $r \approx 0.39$ for Logistic Regression), suggesting that spatial sampling explains much of the cross-source shift, with exceptions such as the Alps and Corsica.

For national mapping, training on the pooled dataset gives the best performances. the Random Forest trained on the full harmonic dataset reaches a cross-validated macro-F1 of 0.874, exceeding both single-features training on the pooled evaluation.

Classifier	Train source	F1 (BD Forêt)	F1 (in-situ)
Random Forest	BD Forêt	0.861	0.884
	In-situ	0.835	0.922
Logistic Regression	BD Forêt	0.834	0.845
	In-situ	0.818	0.874

Table 3. Condensed label-source shift summary (macro-F1). Models are trained on one source and evaluated on both label sources; full metrics are in Supplementary Table S5

3.2 Classification Performance

We found that the embeddings consistently outperform harmonic features regardless of the classifier used. When averaged across three classifier algorithms (Logistic Regression, Linear SVM, and Random Forest), the embedding give a mean accuracy of 91.9% (+2.8 pp) and a macro-F1 of 89.7% (+3.7 pp) compared to the harmonics (Table 4); macro-F1 is the unweighted mean of class-wise F1 scores. This confirms that the embeddings superiority lies in their intrinsic representation, not in the complexity of the downstream classifier. Even with a simple Logistic Regression, the embeddings deliver performances comparable to a Random Forest trained on harmonics.

Classifier	Harmonic (HARM) OA	Harmonic (HARM) F1	Embedding (EMB) OA	Embedding (EMB) F1	Gain (EMB - HARM) ΔOA (pp)	Gain (EMB - HARM) ΔF1 (pp)
Logistic Regression	0.883	0.851	0.915	0.893	+3.2	+4.2
Linear SVM	0.886	0.854	0.915	0.893	+2.9	+3.9
Random Forest	0.904	0.874	0.926	0.905	+2.3	+3.1
Average	**0.891**	**0.860**	**0.919**	**0.897**	**+2.8**	**+3.7**

Table 4. Cross-validated performance across three classifiers using matched 14-dimensional feature sets. Embeddings provide a consistent uplift independent of model complexity.

3.3 Regional performances

The performance improvements from the embedding vary across the GRECO regions (Table 5), with pixel counts indicating that several regions are supported by fewer than one million reference samples. The largest improvements occur in heterogeneous regions with ecotones and mixed stands (e.g. Oceanic Southwest, Pyrenees, Corsica), where the macro-F1 score of embeddings is several percent higher than the harmonics. These regons are known to have mixed pixels and sharp transitions at 10 m resolution; the embedding representation likely helps because it encodes spatial context beyond per-pixel seasonal summaries. The Mediterranean region (J) remains challenging for both feature sets, with only marginal gains for embeddings (+0.7 pp in macro-F1). This suggests that the bottleneck in this regime is less about representation capacity and more about intrinsic separability and label ambiguity at 10 m. For instance, evergreen broadleaved sclerophylls (e.g. *Quercus ilex*) and Mediterranean pines (e.g. *Pinus halepensis*) can have similar seasonal trajectories, and understory green-up after autumn rains can partially decouple canopy and ground signals (Asner, 1998; Helman et al., 2015).

GRECO region	N (M)	OA_H	OA_E	ΔOA	$F1_H$	$F1_E$	ΔF1
National (Weighted)	**14.09**	**0.903**	**0.926**	**+2.3**	**0.875**	**0.905**	**+3.1**
A: Crystalline/Oceanic West	0.49	0.873	0.923	+5.0	0.815	0.883	+6.8
B: Semi-Oceanic North	2.46	0.937	0.956	+1.9	0.856	0.899	+4.3
C: Semi-Continental East	2.87	0.952	0.960	+0.8	0.811	0.851	+4.1
D: Vosges	0.40	0.880	0.907	+2.7	0.840	0.875	+3.5
E: Jura	0.20	0.883	0.921	+3.8	0.775	0.794	+1.9
F: Oceanic Southwest	2.66	0.925	0.951	+2.6	0.867	0.933	+6.5
G: Central Massif	1.93	0.895	0.928	+3.3	0.887	0.923	+3.6
H: Alps	0.72	0.869	0.895	+2.6	0.854	0.877	+2.3
I: Pyrenees	0.56	0.930	0.957	+2.7	0.775	0.883	+10.8
J: Mediterranean	1.46	0.811	0.818	+0.7	0.791	0.798	+0.7
K: Corsica	0.35	0.654	0.675	+2.1	0.554	0.642	+8.8

Table 5. Cross-validated overall accuracy (OA) and macro-F1 for HARM vs EMB (Random Forest models) by GRECO region. N (M) denotes the number of reference pixels in millions from the harmonic training set. Gains are substantial in mixed Atlantic and montane zones (A, F, I, K) but limited in the Mediterranean (J).

In Mediterranean forests, separability between evergreen broadleaved oaks (e.g., *Quercus ilex*) and conifers (e.g., *Pinus halepensis*) can be weak at 10 m resolution. Differences often hinge on subtle near-infrared plateau behaviour driven by leaf/needle structure, internal scattering, and crown clumping (Asner, 1998). In addition, understory green-up following autumn rains can introduce a second phenological pulse beneath an evergreen canopy, creating mixed temporal signals that challenge simple seasonal descriptors (Helman et al., 2015).

3.4 Spatial coherence and calibration

We found that the embeddings improve not only the classification accuracy but also the *quality* of the maps. Calibration errors drop substantially (ECE 0.033 vs 0.059; MCE 0.114 vs 0.184; Table 6), meaning that predicted probabilities better match empirical frequencies. This matters operationally: probability thresholds can be interpreted as confidence levels when prioritising human review or downstream analyses. Spatial coherence also improves without post-processing. Relative to harmonics, patch density decreases by 65% and edge density by 55% (Table 6; definitions in Section 2.8), indicating a strong reduction in "salt-and-pepper" fragmentation. Figure 4 illustrates this effect: embedding-based maps form contiguous stand-like patches with cleaner boundaries, whereas harmonic maps exhibit isolated flips and ragged edges.

Metric	HARM	EMB	Change	Interpretation
Expected calibration error	0.059	**0.033**	**-44%**	Probabilities better aligned to outcomes
Maximum calibration error	0.184	**0.114**	**-38%**	Worst-case bin disparity halves
Edge density (km km^{-2})	8.79	**3.92**	**-55%**	Fewer fragmented boundaries
Patch density (components / 100 km^2)	8462	**2925**	**-65%**	Less salt-and-pepper noise

Table 6. Calibration and spatial coherence metrics for harmonic vs embedding maps. Lower is better; edge density is edge length per km^2 of forest, patch density counts components per 100 km^2.

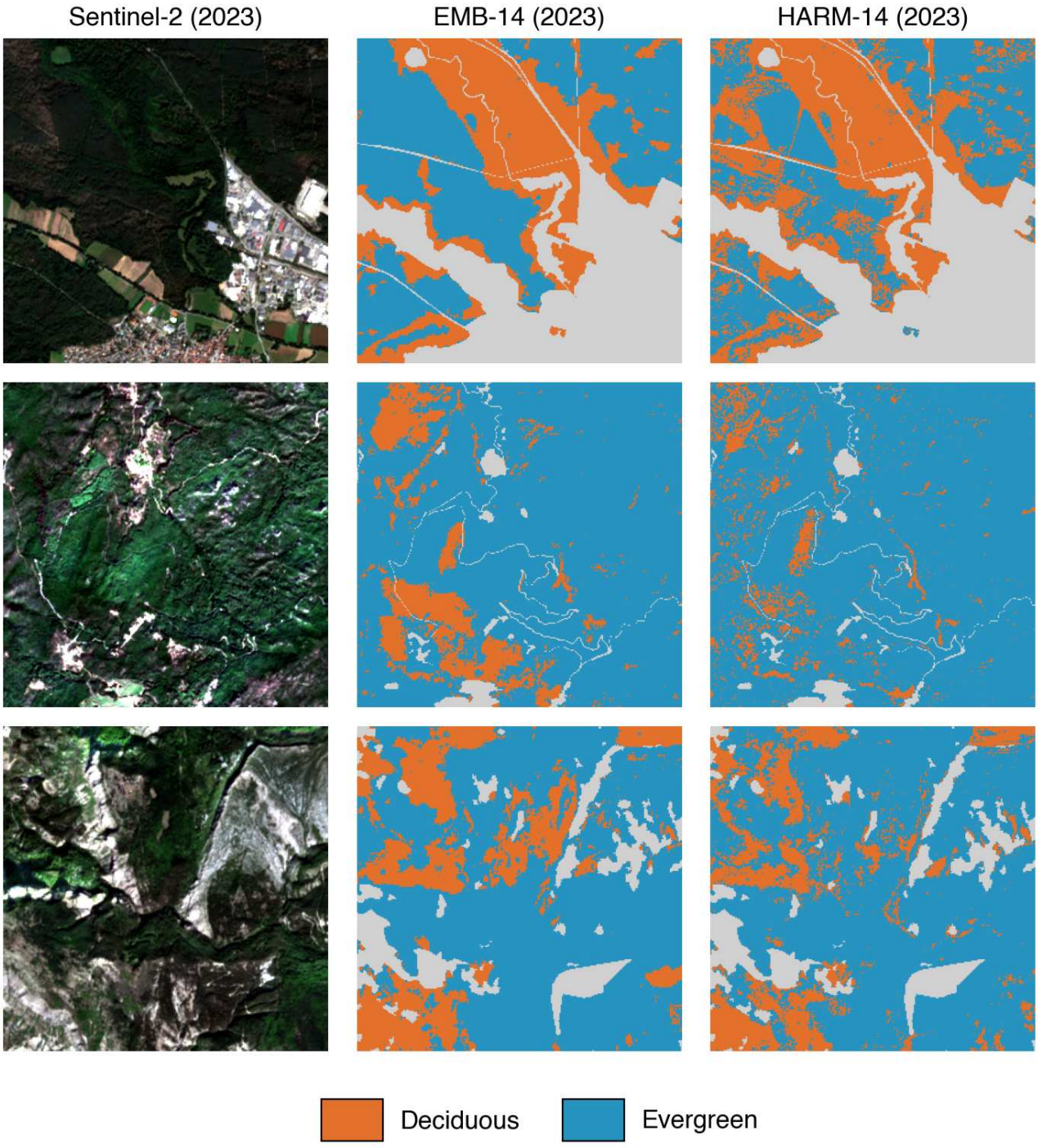

Figure 4. Representative tiles comparing harmonics and embeddings. Embeddings suppress salt-and-pepper artefacts while keeping sharp boundaries in mixed stands.

3.5 Models including both embeddings and harmonics

Adding harmonic to the embeddings yields a small but consistent improvement for the non-linear model, and a marginal improvement in overall accuracy for the linear model (Table 7). For the Logistic Regression, the OA increases slightly (0.915 $\rightarrow$ 0.916) while macro-F1 remains essentially unchanged (0.893 vs 0.891), suggesting that any additional signal is small and not robustly exploitable under a linear decision rule. For

Random Forest, the fused feature set produces a consistent uplift in both accuracy and macro-F1 relative to EMB alone, indicating that harmonic timing/amplitude cues can complement the embedding representation when the classifier can model non-linear interactions.

	HARM-14		EMB-14		EMB-14 + HARM-14	
Classifier	**OA**	**F1**	**OA**	**F1**	**OA**	**F1**
Logistic Regression	0.883	0.851	0.915	0.893	0.916	0.891
Random Forest	0.904	0.874	0.926	0.905	0.931	0.908

Table 7. Complementarity experiment. Models are trained on harmonics (HARM-14), embeddings (EMB-14), or their concatenation (28 features). The fusion yields no clear gain for the linear classifier, but a small and consistent improvement for Random Forest, consistent with partial complementarity between feature families.

The gain of performance for the models using both features is modest compared to that obtained with embeddings relative to harmonics, but it is informative: it suggests that the selected 14 embedding dimensions do not fully subsume the phenology-oriented harmonic descriptors, and that combining both can yield incremental benefits in difficult regimes (e.g., southern or montane zones) .

3.6 Temporal stability

We applied the Random Forest trained on the 2023 embeddings to embeddings from earlier years (2018, 2020, 2022) and re-evaluated it on the same reference pixels (Table 8). Accuracy is 2.9–3.4 percentage points lower for 2018–2022 than for 2023, indicating moderate temporal drift. Visual examples (Figure 5) show broadly consistent spatial patterns; the largest year-to-year differences occur in areas affected by disturbances in the displayed tiles.

Year	**OA**	**ΔOA vs 2023 (pp)**
2018	0.940	−2.9
2020	0.938	−3.1
2022	0.935	−3.4
2023 (baseline)	**0.969**	**0.0**

Table 8. Temporal evaluation of the 2023 embedding Random Forest on earlier embedding vintages. Higher is better for overall accuracy (OA); all rows use the same reference pixels, so the 2023 baseline reflects training-set performance rather than cross-validated metrics.

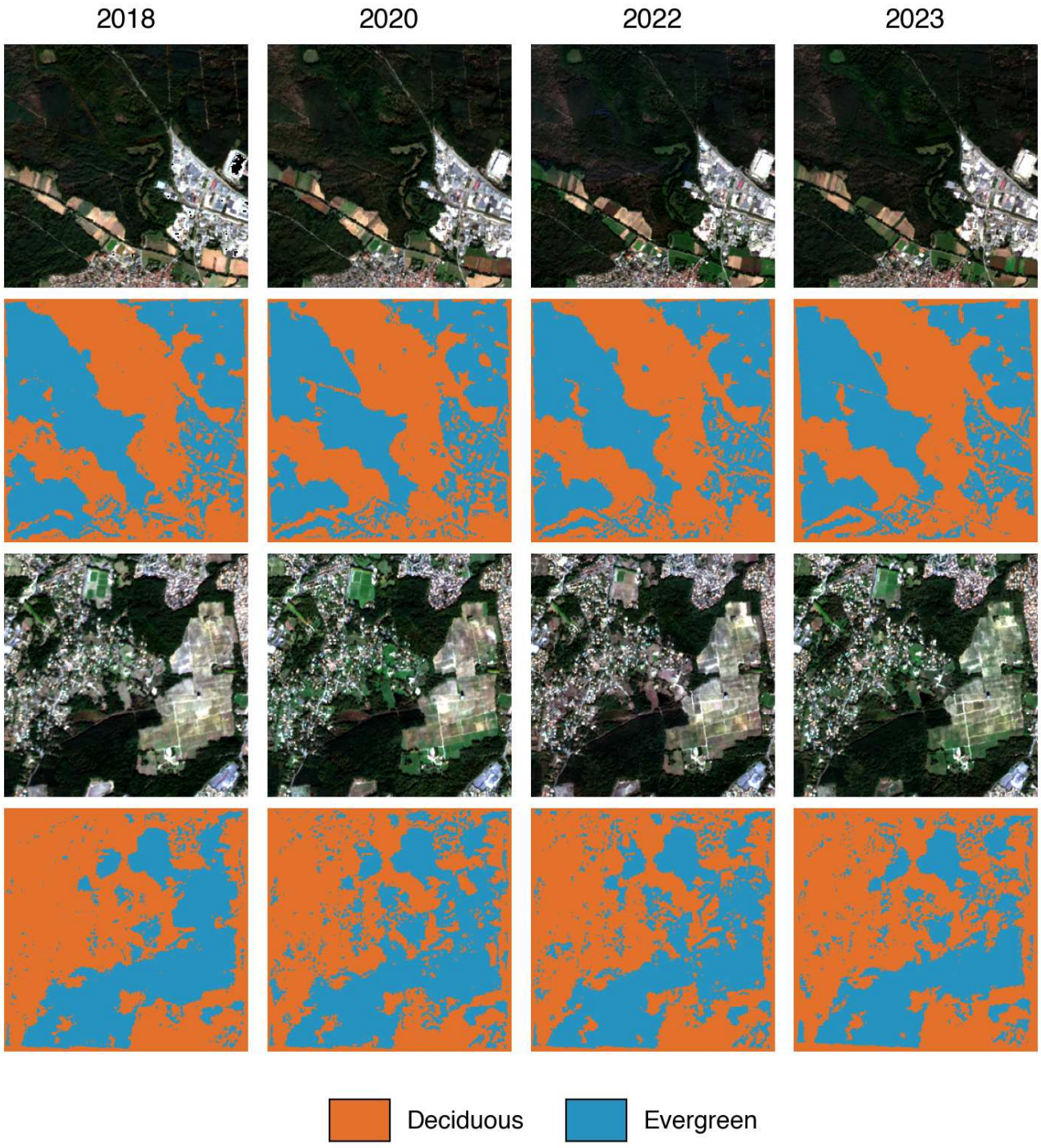

Figure 5. Temporal behaviour on three GRECO tiles across 2018, 2020, 2022, and 2023. Embedding predictions remain stable.

3.7 National map

Figure 6 presents the deciduous–evergreen classification across the entire France produced with the Random Forest classifier. The map exhibits coherent stand patterns in the Atlantic and mountain regions and sharp transitions along major ecological boundaries (e.g. foothills and plateau-to-mountain gradients).

The output is the direct classifier prediction on the 10 m grid (no morphological smoothing), so the observed coherence reflects the representation and the model rather than post-processing.

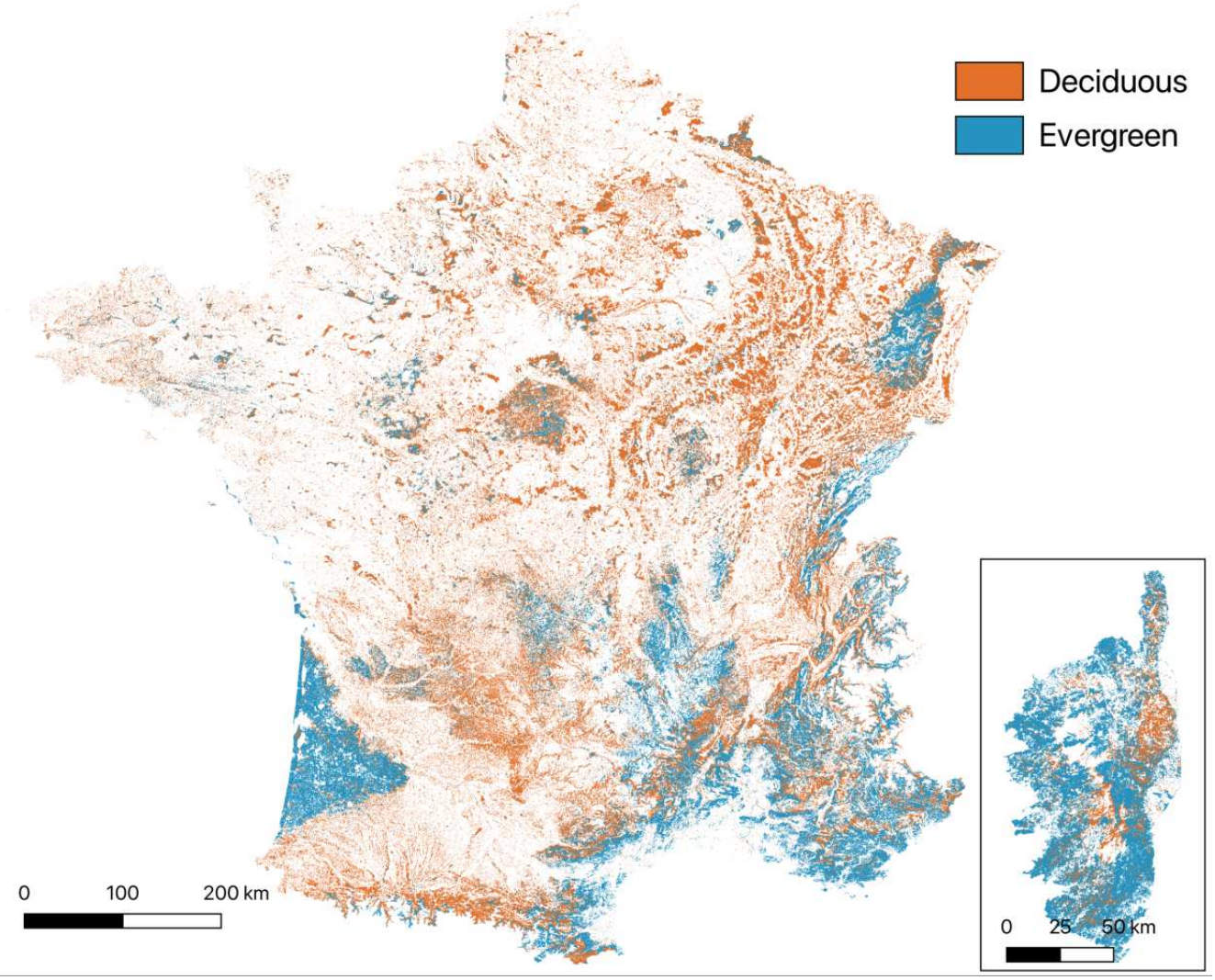

Figure 6. France 2023 deciduous–evergreen map from the embedding model. Orange = deciduous, cyan = evergreen. Embeddings yield smoother patterns while preserving ecological gradients.

3.8 Comparison with other datasets

We compare our maps to Copernicus DLT and BD Forêt V2 to assess consistency on forest pixels (Section 2.7). National agreement is moderate (overall accuracy ≈0.63; Table 9) and very similar for embedding- and harmonic-based predictions, indicating that the embedding improvement of performances over harmonics does not come at the expense of consistency with existing products. Overall agreement near 0.63 indicates substantial differences between products, consistent with definition mismatch (broadleaf/conifer vs deciduous/evergreen) and differing mapping epochs.

Interpretation must account for label and definition differences. BD Forêt contributes to training supervision, so agreement with BD Forêt partly reflects consistency with the reference used for learning. DLT targets broadleaf vs. conifer rather than deciduous vs. evergreen; disagreements in holm oak and larch

regions therefore mix taxonomy mismatch and classification error. Figure 7 provides site-level examples of these effects. To reduce this mismatch in Mediterranean evergreen broadleaved forests, we also report DLT agreement excluding the Mediterranean GRECO region (J) (Table 9).

Comparison	Overall accuracy	ΔOA (pp)	Kappa	F1 (macro)	ΔF1 (pp)	F1 (decid./ever.)
Embedding vs DLT	0.628	**+0.1**	0.159	0.574	0.4	0.725 / 0.423
Harmonic vs DLT	0.627		0.169	0.578		0.722 / 0.434
Embedding vs DLT (excl. Med.)	0.648	0.1	0.177	0.585	0.5	0.747 / 0.422
Harmonic vs DLT (excl. Med.)	0.648		0.189	0.589		0.744 / 0.435
Embedding vs BD Forêt	0.638	0.1	0.164	0.582	0.3	0.735 / 0.428
Harmonic vs BD Forêt	0.639		0.170	0.585		0.734 / 0.436

Table 9. National agreement between our maps and legacy products (forest pixels only). Higher is better; deltas show absolute differences between embeddings and harmonics in percentage points. Rows labelled "excl. Med." exclude the Mediterranean GRECO region (J). DLT maps broadleaf vs conifer rather than deciduous vs evergreen, so disagreements in holm oak or larch areas partly reflect class-definition differences.

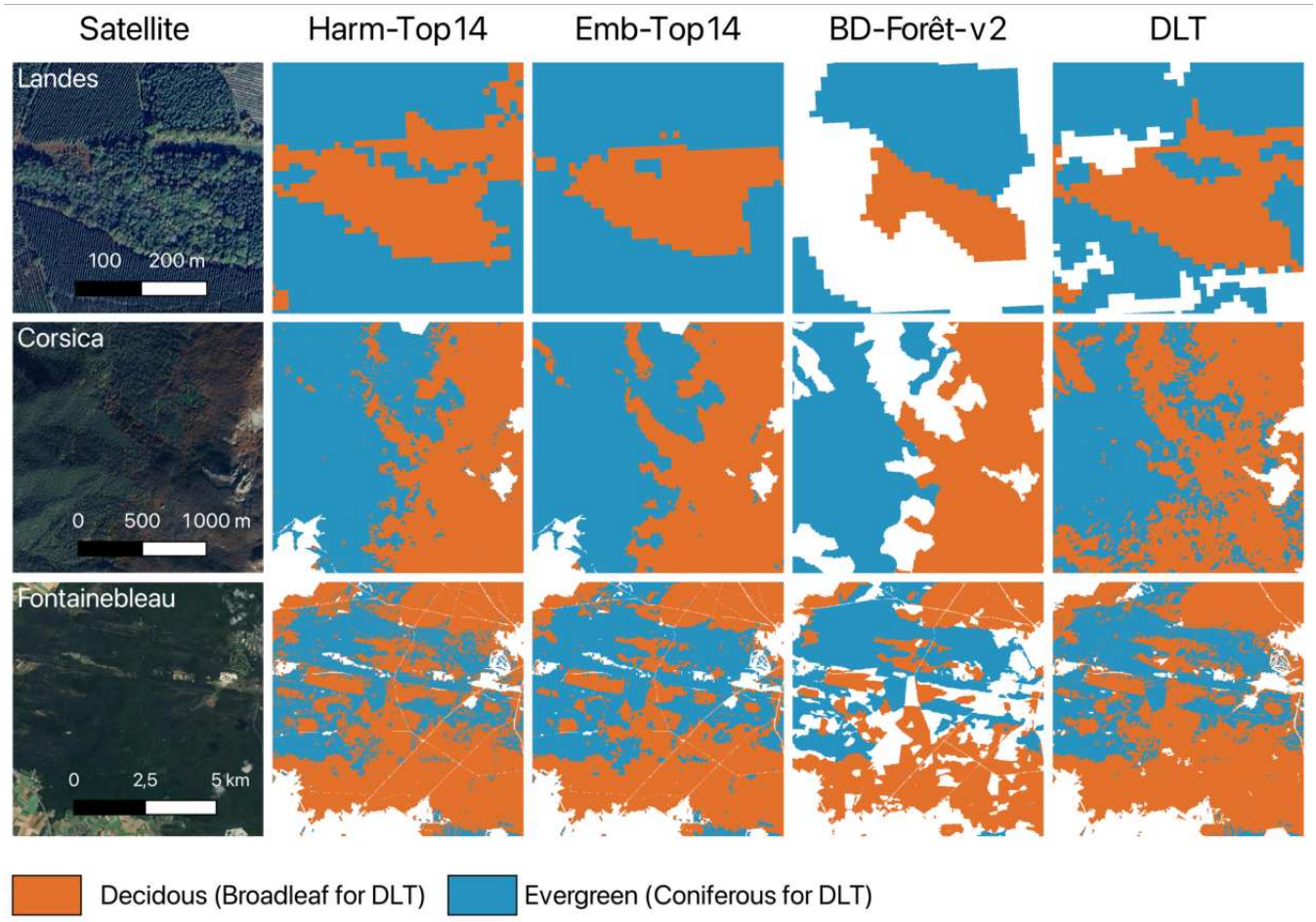

Figure 7. Three sites comparing embedding map with BD Forêt and Copernicus DLT. Embeddings reduce speckle and highlight DLT evergreen-oak inconsistencies.

4 DISCUSSION

4.1 Representation advantage

Our experimental design isolates the contribution of the input features holding training data, classifier models, and feature dimensionality constant. Under these controlled conditions, the AlphaEarth embeddings provide a systematic and algorithm-agnostic advantage over harmonic descriptors. Across the three classifiers, embeddings increase the national mean accuracy by 2.8 percentage points and macro-F1 by 3.7 points (Table 4), despite hyperparameters being tuned on the harmonic baseline and transferred unchanged to the embedding features. Moreover, the dimensionality-matched EMB-14 setting is conservative: Random Forest performance continues to improve when using more embedding channels (Supplementary Material, Section S1), so the reported gains should be interpreted as a lower bound.

The better metrics across Logistic Regression, Linear SVM, and Random Forest indicates that the improvement does not arise from interactions with a particular model class, but from a better separation of deciduous and evergreen classes in feature space. Even linear models trained on embeddings achieve performance similar to or exceeding that of a non-linear Random Forest trained on harmonics. This is a central result: for this national-scale phenology task, the critical step is not increasing downstream model complexity but upgrading the input representation. This result aligns with recent work showing that generic remote-sensing foundation models can yield strong transfer performance with shallow heads (e.g. Xie et al., 2024; Dionelis et al., 2024).

To assess information overlap between handcrafted and learned representations, we test complementarity directly through feature fusion rather than by attempting to regress one feature family onto the other. Concatenating EMB-14 and HARM-14 yields no clear gain for the linear model, but produces a small and consistent improvement for Random Forest (macro-F1 0.905 $\rightarrow$ 0.908; Table 7). Because the fusion experiment is computed on the intersection of valid pixels across both pipelines, the absolute delta is modest and should be read cautiously; nevertheless, the fold-consistent uplift supports the conclusion that harmonic timing/amplitude cues add limited but non-zero task-relevant signal beyond the restricted embedding subset.

4.2 Regional differences

The regional breakdown highlights where embeddings provide the largest accuracy improvement. Gains in macro-F1 are most pronounced in heterogeneous Atlantic and montane regions such as the Oceanic Southwest, Pyrenees, and Corsica (Table 5), where mixed stands and sharp transitions are common at 10 m resolution. In such settings, harmonic descriptors treat each pixel independently and impose a smooth periodic form on annual trajectories, which can be brittle under mixed signals and missing observations. AlphaEarth embeddings, by construction, encode local spatial context in addition to spectral–temporal cues, which likely helps stabilise decisions near boundaries. A tile-level analysis of environmental covariates supports this interpretation: embedding advantages increase with landscape heterogeneity (Supplementary Material, Section S10).

In contrast, the Mediterranean ecoregion shows only marginal improvement (+0.7 percentage points in macro-F1). Here, the main limitation appears to be intrinsic separability and label ambiguity rather than representation capacity. Evergreen broadleaved oaks and pines can have similar spectral and seasonal signatures at 10 m resolution, and understory dynamics (e.g. autumn greening following the first rains) can partially decouple canopy and ground signals. Under these conditions, richer representations help only up

to a point; remaining gains are more likely to come from refined class definitions and higher-confidence supervision than from additional modelling complexity.

The gains in the Alps (+2.3 pp macro-F1) and Jura (+1.9 pp) illustrate another pattern. In mountain regions, snow cover, topographic shadows, and variable illumination generate departures from the smooth annual curves assumed by harmonic models. Embeddings, trained to be robust to missing or corrupted inputs, appear less sensitive to these artefacts, which likely reduces non-phenological variation leaking into fitted harmonic parameters.

4.3 Calibration and map coherence

From an operational perspective, the most salient advantages between harmonic and embedding maps are not only in accuracy but in calibration error and spatial coherence. The expected calibration error is reduced by 44% (from 0.059 to 0.033), and maximum calibration error by 38% (Table 6). For national mapping agencies, well-calibrated probabilities are essential: they allow practitioners to define confidence thresholds for automatic labelling, prioritize human review, and propagate uncertainty into downstream analyses (e.g. biomass estimation or habitat modelling).

Spatial metrics show equally substantial improvements. Edge density and patch density both drop strongly (by 55% and 65%, respectively), indicating a marked reduction in "salt-and-pepper" noise without any explicit spatial smoothing. Visual inspection confirms that embedding-based maps maintain sharp stand boundaries while suppressing isolated misclassified pixels (Figure 4). This behaviour is consistent with the spatial inductive bias of the AlphaEarth model: by encoding local neighbourhood context, the embeddings implicitly enforce that nearby pixels with similar imagery map to similar feature vectors. The result is a more coherent map that better matches how human interpreters perceive forest patterns.

At the product level, both embedding and harmonics maps show comparable agreement with BD Forêt V2 and Copernicus DLT (Table 9). This indicates that the embedding model does not simply overfit to the training data at the expense of alignment with existing products. Rather, it delivers the expected refinement: moderate to high agreement with legacy maps while producing cleaner, more up-to-date spatial patterns, especially in ecotones and mixed Atlantic/montane mosaics (Figures 6 and 7).

4.4 Label and supervision shift

The Mediterranean plateau and the moderate national agreement with legacy products underscore the central role of training data quality and definition mismatch. The label-source diagnostic further shows that performance is systematically higher on in-situ/expert labels, and that models trained only on in-situ supervision transfer less well to BD Forêt polygons, consistent with a measurable source shift and/or higher effective noise in photo-interpreted labels. These results motivate pooled training for national robustness: even though in-situ labels represent only 11.4% of the samples, mixing both supervision families yields the most reliable overall performance.

These observations support a shift toward data-centric workflows for remote sensing applications that employ foundation models. For tasks such as deciduous–evergreen mapping, the incremental benefit of designing new handcrafted indices or more elaborate classifier architectures is likely smaller than the benefit of improving the reference data: harmonizing class definitions, explicitly modelling mixed pixels, refining boundaries with high-resolution imagery or LiDAR, and systematically characterizing label uncertainty. Foundation embeddings help by removing the need to engineer features for every new map, allowing domain experts to invest their time in curating, validating, and extending high-confidence training sets.

In this sense, foundation models do not eliminate the need for ecological expertise; they rebalance it. Expert knowledge shifts from crafting features that approximate phenological processes (e.g. harmonic phases and amplitudes) to designing sampling schemes, quality-control protocols, and uncertainty analyses that ensure that the supervision signal is as informative as possible. In regions like the Mediterranean where authoritative labels exist but separability is intrinsically low, both richer labels and clearer class definitions remain bottlenecks, and legacy products (DLT, BD Forêt) also show lower agreement. The reliance on buffered BD Forêt polygons for 88.6% of training pixels also underscores how much progress depends on expanding high-confidence ground labels beyond the current 11.4% anchor set.

4.5 Operational implications and outlook

A practical advantage of AlphaEarth embeddings is that they can be used without local GPU training. In our workflow, feature extraction is delegated to the pre-computed annual embedding field, and only lightweight classifiers are trained on standard CPU infrastructure. Once fitted, the embedding-based Random Forest can generate national maps at 10 m resolution with conventional geospatial tooling. Because the embeddings are pre-computed, increasing embedding dimensionality primarily affects I/O and the downstream model rather than requiring additional GPU inference; in our feature-selection analysis, performance continues to improve when using more than 14 channels (Supplementary Material, Section S1).

The temporal stability analysis further supports operational use. Applying the 2023 model to 2018, 2020, and 2022 embedding vintages yields only modest declines in accuracy and macro-F1 (Table 8), and visual inspection shows limited year-to-year flicker outside disturbance zones (Figure 5). This stability is crucial for long-term monitoring: changes in the map can more confidently be attributed to real forest dynamics rather than model drift.

Several limitations point to clear avenues for future work. First, we evaluate a single foundation model and a single binary task; extending the analysis to multiple foundation backbones and to multi-class forest typologies (Supplementary Material, Section S9) would clarify how general the observed advantages are.Second, we treat embeddings as fixed features; task-specific fine-tuning of the embedding model or of a shallow adaptation layer may further improve performance, but would forgo the convenience of precomputed global layers. Third, we operate at the pixel level; simple neighbourhood features or stand-level aggregation could further stabilize boundaries, though the current embeddings already encode local context. Finally, the current evaluation is restricted to France; applying the same experimental design in other biomes would test the portability of both the embeddings and the proposed training strategy, with the usual caveat that label availability and quality will likely be the limiting factor.

5 CONCLUSION

We presented a controlled, national-scale comparison between pre-computed foundation embeddings and handcrafted phenology features for deciduous–evergreen forest mapping in metropolitan France. Using 14.1 million labelled pixels and tile-based, GRECO-stratified spatial cross-validation, AlphaEarth embeddings consistently outperformed harmonic descriptors when paired with the same lightweight classifiers, with a +3.7 percentage point gain in mean macro-F1 across Logistic Regression, Linear SVM, and Random Forest. Beyond classification performance, embeddings produced better-calibrated probabilities and substantially more spatially coherent maps without any post-processing, which is critical for operational use.

Two additional diagnostics clarify practical deployment conditions. First, a label-source experiment (BD Forêt polygons versus in-situ/expert labels) reveals measurable source shift, supporting pooled training as

the most robust national strategy. Second, a simple fusion test shows limited but non-zero complementarity: adding harmonic descriptors to embeddings yields a small, consistent gain.

Because the main benchmark matches feature dimensionality (EMB-14 vs HARM-14) and reuses hyperparameters tuned on harmonics, the reported embedding advantage should be interpreted as a conservative lower bound; higher-dimensional embedding variants further improve performance (Supplementary Material). Overall, pre-computed embeddings can act as a drop-in replacement for handcrafted harmonic features in national deciduous–evergreen mapping, reducing dependence on local GPU training and shifting the main bottleneck toward label curation, harmonisation, and uncertainty-aware validation.

ACKNOWLEDGEMENTS

The author thanks Sarah Brood, Alexandre d'Aspremont, and Philippe Ciais for reviewing the manuscript and for methodological discussions that helped refine the study.

DATA AVAILABILITY

Research code, experiment configurations, and the working manuscript are available at `https://github.com/ArthurCalvi/S2-Tree-Phenology`. Source datasets remain subject to the access terms of their respective providers.

CODE AVAILABILITY

The public repository contains the data-preparation, feature-extraction, model-training, evaluation, and mapping workflows used in this study.

REFERENCES

[Dataset] (2025). Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. arXiv:2507.22291. Accessed 2025-08-29

Asner, G. P. (1998). Biophysical and biochemical sources of variability in canopy reflectance. *Remote Sensing of Environment* 64, 234–253. doi:10.1016/S0034-4257(98)00014-5

Belgiu, M. and Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing* 114, 24–31. doi:10.1016/j.isprsjprs.2016.01.011

Bolton, D. K., Gray, J. M., Melaas, E. K., Moon, M., Eklundh, L., and Friedl, M. A. (2020). Continental-scale land surface phenology from harmonized landsat 8 and sentinel-2 imagery. *Remote Sensing of Environment* 240, 111685. doi:10.1016/j.rse.2020.111685

[Dataset] Cha, K., Seo, J., and Lee, T. (2023). A billion-scale foundation model for remote sensing images

Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., et al. (2022). Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. In *Advances in Neural Information Processing Systems*. vol. 35, 197–211

[Dataset] Dionelis, N., Fibaek, C., Camilleri, L., Luyts, A., Bosmans, J., and Le Saux, B. (2024). Evaluating and benchmarking foundation models for earth observation and geospatial ai

[Dataset] European Union (2024). Copernicus land monitoring service: Dominant leaf type 2018. https://land.copernicus.eu/pan-european/high-resolution-layers/forests/dominant-leaf-type. Accessed: 2024-01-15

Francini, S., Chirici, G., Giannetti, F., D'Amico, G., Vangi, E., Travaglini, D., et al. (2024). Forest species mapping via sentinel-2 and inventory data: A machine learning approach at scale. *Remote Sensing of Environment* 301, 113920. doi:10.1016/j.rse.2023.113920

Gaydon, M. et al. (2024). Pureforest: A large-scale aerial lidar and aerial imagery dataset for tree species classification. *Remote Sensing* To be completed with full citation details

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Moore, R. (2017). Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment* 202, 18–27. doi:10.1016/j.rse.2017.06.031

Helman, D., Lensky, I. M., Tessler, N., and Osem, Y. (2015). A phenology-based method for monitoring woody and herbaceous vegetation in mediterranean forests from ndvi time series. *Remote Sensing* 7, 12314–12335. doi:10.3390/rs70912314

Inglada, J., Vincent, A., Arias, M., Tardy, B., Morin, D., and Rodes, I. (2017). Operational high resolution land cover map production at the country scale using satellite image time series. *Remote Sensing* 9, 95. doi:10.3390/rs9010095

Institut national de l'information géographique et forestière (2013). *Fiches descriptives des grandes régions écologiques (GRECO)*. Tech. rep., IGN France. Ecological region reference used for stratification

[Dataset] Institut national de l'information géographique et forestière (2024a). Bd forêt v2: Base de données forestière de l'ign. https://geoservices.ign.fr/bdforet. Accessed: 2024-01-15

[Dataset] Institut national de l'information géographique et forestière (2024b). Lidar hd: Lidar haute densité sur l'ensemble du territoire. https://geoservices.ign.fr/lidarhd. Accessed: 2024-01-15

Institut National de l'Information Géographique et Forestière (2024). *Masque Forêt version Beta v1.0*. Tech. rep., IGN France

[Dataset] Ishikawa, T., Bonannella, C., Lerink, B. J. W., and Rußwurm, M. (2025). Deep pre-trained time series features for tree species classification in the dutch forest inventory

Jönsson, P. and Eklundh, L. (2002). Seasonality extraction by function fitting to time-series of satellite sensor data. *IEEE Transactions on Geoscience and Remote Sensing* 40, 1824–1832. doi:10.1109/TGRS.2002.802519

Kennedy, R. E., Yang, Z., and Cohen, W. B. (2010). Landtrendr: A time-series segmentation approach for mapping forest disturbance and recovery. *Remote Sensing of Environment* To be completed with volume, pages, doi

[Dataset] Khanna, S., Liu, P., Zhou, L., Meng, C., Rombach, R., Burke, M., et al. (2023). Diffusionsat: A generative foundation model for satellite imagery

Li, X., Zhou, Y., Meng, L., Asrar, G. R., Lu, C., and Wu, Q. (2023). Mapping evergreen forests using a new phenology index, time series sentinel-1/2 and google earth engine. *Ecological Indicators* 149, 110157. doi:10.1016/j.ecolind.2023.110157

[Dataset] national de l'information géographique et forestière (IGN), I. (2018). BD Forêt version 2.0 – descriptif de contenu. Technical documentation, IGN. Available from https://geoservices.ign.fr/bdforet

[Dataset] Office National des Forêts (2024). Office national des forêts. https://www.onf.fr. Accessed: 2024-01-15

[Dataset] Singh, G., Ghosh, S., Kerner, H., Raskar, R., Brown, D., Cresswell, J. C., et al. (2023). Prithvi: A foundation model for earth observation. ArXiv:2310.08597

534 Ulrich, E. et al. (2021). *RENECOFOR: National Forest Monitoring Network*. Tech. rep., Office National
535 des Forêts. To be completed with full citation details
536 Verbesselt, J., Hyndman, R. J., Zeileis, A., et al. (2010). Detecting trend and seasonal changes in satellite
537 image time series. *Remote Sensing of Environment* To be completed with volume, pages, doi
538 Wilson, B. T., Knight, J. F., and McRoberts, R. E. (2018). Harmonic regression of landsat time series for
539 modeling attributes from national forest inventory data. *ISPRS Journal of Photogrammetry and Remote*
540 *Sensing* 137, 29–46. doi:10.1016/j.isprsjprs.2018.01.006
541 [Dataset] Xie, Y., Wang, Z., Chen, W., Li, Z., Jia, X., Li, Y., et al. (2024). When are foundation models
542 effective? understanding the suitability for pixel-level classification using multispectral imagery
543 Zhu, Z. and Woodcock, C. E. (2014). Continuous change detection and classification of land cover using
544 all available landsat data. *Remote Sensing of Environment* 144, 152–171. doi:10.1016/j.rse.2014.01.011