https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

# Benchmarking Earth Observation Embeddings for Pan-EU Aboveground Biomass Mapping

Yu-Feng Ho[1,2], Madlene Nussbaum[2], Léa Enguehard[3], Qian Song[3], Mustafa Serkan Isik[1], Jordy de Lange[4], Carmelo Bonannella[1], Leandro Parente[1], Derek Karssenberg[2], Martin Herold[3], Ichsani Wheeler[1], and Tomislav Hengl[1]

[1]OpenGeoHub Foundation, Doorwerth, The Netherlands
[2]Department of Physical Geography, Faculty of Geosciences, Utrecht University, Utrecht, The Netherlands
[3]Remote Sensing and Geoinformatics, GFZ Helmholtz Centre for Geosciences, Potsdam, Germany
[4]Spheer, Groningen, The Netherlands

**Correspondence:** Yu-Feng Ho (yu-feng.ho@opengeohub.org)

**Abstract.** The paper presents a comprehensive evaluation of pre-trained Earth Observation (EO) embeddings compared to traditional remote sensing approaches trained with standard multi-sensor satellite data (Sentinel-1, Sentinel-2, and ALOS PALSAR-2, both with and without the addition of multi-scale terrain derivatives) for mapping aboveground biomass density (AGBD). The rank result was evaluated on 4 criteria: point-estimate accuracy, spatial pattern reconstruction, saturation effects, and uncertainty quantification. EO features were specifically used with tree-based (Quantile Random Forest Regression; QRFR) and deep learning (Convolutional Neural Network; CNN) models to map AGBD across varied biogeographical region in Finland (≈ 73,000 $\mathrm{km}^2$), Spain (≈ 273,000 $\mathrm{km}^2$), and the Netherlands (≈ 2,200 $\mathrm{km}^2$). A proportional stratified sampling design using airborne laser scanning (ALS) estimates of AGBD was implemented to generate sample tiles for each country, and country-specific models were trained. The benchmarking results show that EO embeddings achieve predictive accuracy that is highly competitive when compared with models that use Sentinel-1, Sentinel-2, and PALSAR: the lowest RMSE values were recorded by AlphaEarth – QRFR with terrain variables in the Netherlands (42.1 Mg $\mathrm{ha}^{-1}$), SpheerFM – CNN in Spain (17.5 Mg $\mathrm{ha}^{-1}$), and TESSERA – CNN in Finland (13.4 Mg $\mathrm{ha}^{-1}$). Pairing these embeddings with a CNN model yielded clear performance gains: TESSERA improved across all three test countries, while SpheerFM showed distinct enhancements in Spain and Finland. This strong performance was further corroborated by spatial pattern reconstruction metrics. In addition, a rigorous uncertainty assessment of both the QRFR and CNN models using empirical metrics, Prediction Interval Coverage Probability (PICP) and Mean Prediction Interval (MPI), showed that both models derive unbiased, well-calibrated prediction intervals. Ultimately, TESSERA utilized CNN as downstream model comes as the highest-ranked configuration. The result from different predictors set is also marginal, implying no single set dominates. The choice of applying EO embeddings for large-scale AGBD mapping requires consideration also data licensing and transparency, data volume, and their existing spatial temporal coverage.

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

## 1 Introduction

Above-Ground Biomass Density (AGBD), measured in Mg ha$^{-1}$, is a critical metric of terrestrial carbon stocks and fluxes. The mapping of AGBD provides key spatial information to guide forest logging and long-term resource management. Beyond that, it serves as an essential carbon baseline for carbon removal efforts and carbon farming practices. Aboveground biomass and soil organic matter represent the primary mechanisms of carbon sequestration within the terrestrial carbon cycle (Carlson et al., 2001). Although the major carbon stock is located in sediments under the sea, plant growth and decay of terrestrial vegetation contribute a relatively large portion of the carbon flux. Forest ecosystems store the majority of terrestrial biomass above ground worldwide (Post et al., 1990), where live biomass (above and below ground) has approximately 40%, second only to mineral soil (to 1 m depth) that contains 44% in forest ecosystems (Sedjo, 1993). Thus, estimating forest aboveground biomass, especially woody biomass, is crucial to monitoring carbon fluxes and for forest resource management. Detailed maps of AGBD can help highlight key hotspots for forest ecosystem productivity, biodiversity, and carbon storage (Ma et al., 2024), directly forming decisions about forest conservation, logging practices, and climate policy (Kumari et al., 2026). Monitoring forest AGBD is especially important in the European Union as the baseline required to implement European Carbon Removal and Carbon Farming (CRCF) initiatives (Chiti et al., 2024). By tracking biomass dynamics over time, continuous AGB monitoring supports the quantification of business-as-usual (BAU) carbon fluxes and distinguishes CRCF interventions that satisfy the QU.A.L.ITY framework criteria: quantification, additionality, long-term storage, and sustainability (Jensen, 2026).

Combining forest measurements on-site with air and space remote sensing is essential to establish carbon baselines and support subsequent certification frameworks (European Parliament and Council of the European Union, 2024). Forest monitoring On-site has been conducted for a long time through National Forest Inventories (NFIs) (Tomppo et al., 2010). NFIs rely on terrestrial field measurements, which are costly, labor-intensive and time-intensive, require specialized expertise, and are typically constrained to small sample plots. Increasingly, forest monitoring at the state and national levels incorporates airborne laser scanning (ALS). For example, Tanase et al. (2026) integrated Spanish NFI data with ALS, using dense point clouds to reconstruct canopy height and structural attributes. This research demonstrated that forest monitoring can be effectively upscaled in larger regions by modeling. However, ALS remains financially demanding and budget-reliant, suffering from limited revisit capabilities, partial spatial coverage, and uneven availability across regions and countries (Hancock et al., 2021). Consequently, constructing a consistent pan-EU carbon baseline remains challenging. Satellite remote sensing such as based on the open Sentinel-1 and Sentinel-2 images bridges these gaps, providing the scalable, continuous data needed to monitor forest AGB at regional, continental, and global scales.

Integrating multi-sensor remote sensing data with ground-truth measurements via machine learning (ML) and deep learning (DL) algorithms enables the generation of spatially explicit, high-resolution AGB maps across diverse ecosystems—substantially advancing carbon cycle research and sustainable forest management (Altarez et al., 2024). However, despite the capacity of EO-ML frameworks to produce detailed, high-frequency AGBD estimates over broad regions, deploying EO at large scale comes with significant technical challenges. The production of large-scale high-resolution AGBD products requires substantial infrastructure for data storage and pre-processing. For instance, processing annual Sentinel-2 (optical) Level-2A data requires

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

raw cloud-masking pipelines handling multi-petabyte volumes, whereas Sentinel-1 A/B requires rigorous radiometric and terrain corrections to accurately map vegetation structure (Gazzea et al., 2023). These infrastructural bottlenecks effectively limit high-detail, consistent and large-scale biomass mapping to be delivered, ultimately delaying environmental policy implementation and climate action.

Recently, earth observation foundation model embeddings (EO embeddings) have emerged as a highly efficient alternative, compressing massive high-resolution EO time-series while claiming to maintain competitive predictive accuracy (Stewart et al., 2026). The state-of-the-art EO embeddings in the context of AGB estimation, such as TESSERA (Feng et al., 2026) and AlphaEarth (Brown et al., 2025), have demonstrated competitive performance while providing accessible, annual, pan-European coverage. This enables large-scale downstream AGB mapping with accuracies comparable to conventional machine learning pipelines trained directly on satellite ARD archives. Beyond their computational efficiency, EO embeddings offer greater architectural flexibility than Geospatial Foundation Models (GFMs) (Yang et al., 2025). They allow for incorporating auxiliary environmental covariates excluded during pre-training. As an example, Fang et al. (2026) demonstrated that augmenting EO embeddings with additional contextual covariates substantially improves downstream predictive performance compared to using EO embeddings alone.

Local environmental drivers, such as topography, soil properties, and anthropogenic disturbances, significantly shape landscape-level biomass patterns (Liu et al., 2017). Terrain variables derived from Digital Terrain Models (DTMs) can address the environmental variation that current EO embeddings overlook. DTMs quantify spatial variability across hydrological, geomorphological, and ecological processes, making them essential covariates in hydrological monitoring, soil surveys, and predictive vegetation mapping. In the context of AGB mapping, topographical metrics such as slope and elevation serve as critical predictors of spatial heterogeneity, largely because anthropogenic land-use and vegetation dynamics are topographically constrained (Salinas-Melgoza et al., 2018). Recent work by Lemke et al. (2025) further underscores the strong relationship between AGB distribution and complex terrain variables, including landform position, topographic wetness, and potential solar radiation. However, these studies focus primarily on localized or relatively small study areas, leaving the impact of terrain variables on large-scale continuous AGB mapping largely unexplored. Furthermore, the integration of these terrain derivatives with novel EO embeddings has not yet been investigated.

While supplementing embeddings with external features is a promising step, effective downstream predictive models remain essential to translate these predictors (EO embeddings and external features) into meaningful environmental variables. However, the comparative performance of different downstream model architectures operating on EO embeddings remains underexplored. Although recent studies have begun to evaluate specific model families, such as Ishikawa et al. (2026) comparing multilayer perceptions (MLPs) against Random Forests (RF) for tree-species mapping, or Silva and Gonçalves (2026) benchmarking RF, Gradient Boosting Decision Tree (GBDT), and Support Vector Machines (SVM) for burned-area mapping, direct comparisons between convolutional deep learning models and traditional machine learning methods remain untested in EO embeddings. Comprehensive evaluations spanning a wider spectrum of downstream architectures are notably lacking. As an example of experimentation with ML models for AGBD estimation, Sialelli et al. (2025) conducted a landmark large-scale benchmark comparing GBDT, fully convolutional networks (FCN), UNet and a customized CNN based on the Xception ar-

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

chitecture, addressing both the point estimation and the spectral saturation effect of AGBD estimation in EO-ML workflows. Nevertheless, for operationalization of AGBD predictions as a certified carbon baseline in CRCF frameworks, point-estimate metrics and evaluation of spectral saturation effect alone are insufficient. What equally vital are rigorous uncertainty quantification and spatial structural fidelity(European Parliament and Council of the European Union, 2024).

To address these gaps, the primary objective of this study is to evaluate the utility of state-of-the-art EO embeddings for continental-scale AGBD mapping, assessing potential trade-offs between data compression and predictive accuracy compared to standard multi-sensor ARD baselines. Furthermore, we investigate the extent to which integrating auxiliary terrain covariates enhances downstream prediction fidelity. Finally, we systematically compare the performance of tree-based machine learning ensembles against convolutional deep learning architectures. In summary, this research focuses on three connected questions:

1. How does predictive performance of EO embeddings compares to training models from mutli-sensor satellite ARD for predicting AGBD?
2. Which model (RF or CNN) performs better with regard to point estimate, spatial pattern reconstruction, saturation effects, and uncertainty quantification?
3. Do additional terrain variables affect the accuracy of AGBD prediction according to the evaluation criteria mentioned above?

## 2 Method and Material

### 2.1 General Framework

The general framework describing the workflow from EO and AGBD reference data acquisition, train-validation-test splitting, and models training, to test tile inference and comprehensive model assessment is shown in Figure. 1.

In this specific study, we used data from three (3) purposively selected countries representing southern (Spain), central (the Netherlands), and northern (Finland) Europe. The reference data was produced through regression model that scale up forest inventories or sample plots with national ALS. Aftermath, a proportional stratified sampling design was implemented to generate sample tiles for each country and then country-specific models were trained. Each predictor (feature) set, AlphaEarth, TESSERA, SpheerFM, and multi-sensor ARD from Sentinel-1, Sentinel-2, and PALSAR, was trained using two competing models: Random Forest (RF) and Convolutional Neural Network (CNN). Additionally, each predictor setup was paired with multiscale terrain variables to evaluate the impact of topographic features on AGB estimation. The 10 m resolution layers (SpheerFM, TESSERA, AlphaEarth, and Sentinel-2) are native 10 m spatial support, while multiscale terrain variables and PALSAR were downscaled using cubic spline interpolation to ensure spatial consistency and smooth predictions in the final product.

To ensure a fair comparison between the RF and CNN models in each country, the same overall sampling design was applied with minor model-specific adaptations. Firstly, both RF and CNN models are assigned the same train, validation and test tile

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

sets, derived from the proportional sampling strategies. The validation set is used in the modeling processes for validation the intermediate training result. The test set is not used in any step of the model building and thus for independent evaluation. Secondly, to ensure robust quantification of prediction uncertainty, we adopted the Quantile Random Forest Regressor (QRFR) (Meinshausen and Ridgeway, 2006) as our RF model, and a tailored CNN based on the Xception architecture adapted from Lang et al. (2023) and Sialelli et al. (2025) as our CNN model. These models implement scalable approaches to generate predictive distributions. For model fitting, we first utilized the calibration subset to tune the Random Forest hyperparameters. The final QRFR model was subsequently trained using the resulting optimal hyperparameters. In contrast, the CNN model was trained using fixed hyperparameters and the deep learning architecture developed by Sialelli et al. (2025).

Once trained, the models were used to predict the AGBD values on the test tiles, which were left alone prior to the training process. The model prediction outputs provided both point estimates (a single value per pixel) and uncertainty estimates (a distribution array per pixel). We then used these tile-based predictions to compute spatial pattern metrics (SSIM, SIM, SIP, and SIV) (Eqs. 6–9) as well as point-estimate accuracy metrics (RMSE, MAE, IDR, and CCC) (Eqs. 2–5). Finally, uncertainty quantification was evaluated using PICP (Eq. 10) for reliability testing alongside MPI (Eq. 11).

### 2.2 Reference Data

The reference AGBD maps were derived from national forest inventories (or sample plots) and ALS data (Song et al., 2026). These reference datasets span vastly different biogeographical regions across Europe (European Environment Agency, 2002): Mediterranean in Spain, Atlantic in the Netherlands, Boreal in Finland, and Alpine in both Spain and Finland. We assume that a model that performs the best in these three areas can be successfully scaled-up and used to pan-EU mapping.

In each country, ALS returns were used to generate a canopy height model (CHM), and AGB was modeled from plot-level CHM statistics and co-located NFI measurements using either linear or log-linear regression; the best-performing form was retained per country. In Spain and the Netherlands, the models were fitted separately for each dominant leaf type (DLT) (Langanke et al., 2017), since both countries contain a substantial share of both coniferous and broadleaf forests. In Finland, stratification was not applied, as the finish forest is largely coniferous and the broadleaf fraction is too small to support a separate model. or all three-country models we assess performance by using 5-fold cross-validation, and relative RMSE (rRMSE) is reported as RMSE divided by the mean observed plot AGBD. The results of Song et al. (2026) are briefly summarized here:

- **Finland**: AGBD maps are generated from 1 m CHM with a log-linear regression model $log(AGBD) = a + b * log(\bar{H})$, where $\bar{H}$ is the mean canopy height of the plot. The coefficients $a$ and $b$ are estimated from 9,631 pairs of mean CHM matched at the spatiotemporal level and AGBD value at the plot-level. The AGBD values are derived from national sample plots from 2019 to 2021. The year difference between the airborne lidar and the field inventory data is less than one year. The cross-validated performance is $R^2$ = 0.71, with a RMSE of 29.19 Mg/ha, and rRMSE = 31.19%. The AGBD reference covers forest plots for an entire country account for approximately 73,000 $\mathrm{km}^2$ of woody biomass.

- **Spain**: AGBD maps are generated with a linear regression model $AGBD = a + b * \bar{H}$, fitted using the NFI and the 1 m CHM data from 2016 to 2021. The model is fitted separately for each DLT, with at most one year between the ALS and

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

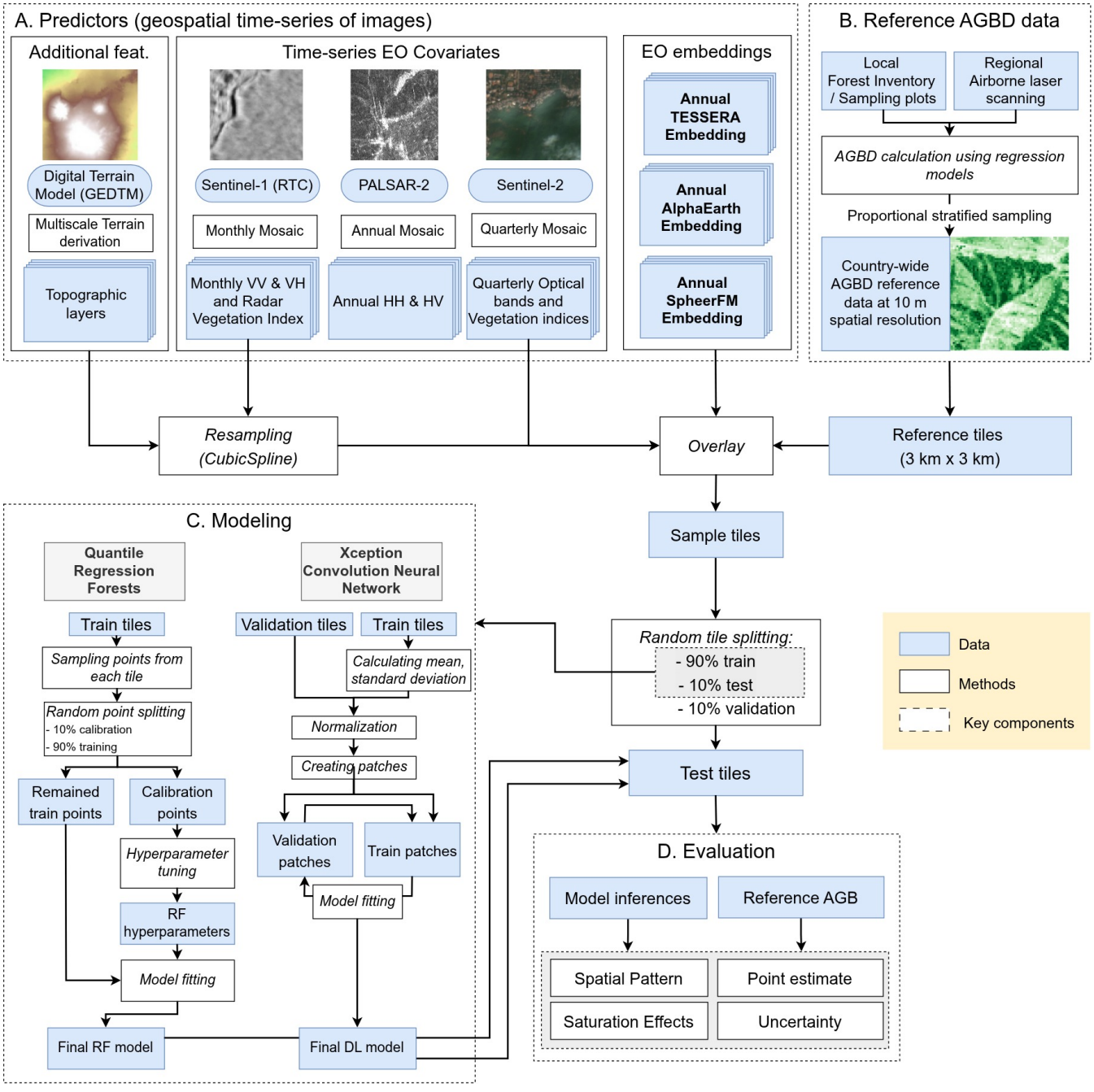



**Figure 1.** The general workflows, key inputs (A: predictors; B: reference AGBD data), modeling framework (C) and perforamnce evaluation (D) for benchmarking large-scale AGBD modeling of EO embeddings.

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

NFI data. In total, 4863 plot–CHM pairs were retained, of which 2,308 are coniferous-dominated and 2,555 broadleaf-dominated. In 5-fold cross-validation the coniferous model reaches $R^2 = 0.45$ (RMSE 46.7 Mg ha$^{-1}$, rRMSE 57.4%) and the broadleaf model $R^2 = 0.60$ (RMSE 46.1 Mg ha$^{-1}$, rRMSE 55.0%). The maps cover several provinces, including Castile and León, Catalonia, Castile-La Mancha, Aragon, and Extremadura, in total approximately 273,000 km$^2$ of woody biomass. The range of reference biomass spans Spain, covering flat and high-altitude plains, elevated plateau, hills, and mountainous terrain; continental and Mediterranean climate.

- **Netherlands**: AGBD maps are generated with a linear regression model $AGBD = a + b * \bar{H}$ fitted using the NFI and the 0.5 m CHM data from 2016 to 2019. The model is fitted separately for each DLT. Because ALS and NFI acquisitions rarely coincide in time, the temporal offset between a plot and its CHM is up to two years in this case. In total, 307 plot–CHM pairs were retained, of which 150 are coniferous-dominated and 157 broadleaf-dominated. In 5-fold cross-validation the broadleaf model reaches $R^2 = 0.49$ (RMSE 86.7 Mg ha$^{-1}$, rRMSE 41.3%) and the conifer model $R^2 = 0.58$ (RMSE 51.7 Mg ha$^{-1}$, rRMSE 36.9%). The total AGBD reference covers in total approximately 2,200 km$^2$ of woody biomass in the Netherlands.

### 2.3 Multi-sensor Satellite ARD

The following EO data sources have been used in this study:

- **Sentinel-1**: We obtained Sentinel-1 data from the Level-3 monthly mosaics provided by the Copernicus Data Space Ecosystem (CDSE) (Copernicus Data Space Ecosystem, 2026). CDSE generates these monthly mosaics by applying radiometric calibration, thermal noise removal, and radiometric terrain correction using the Copernicus DEM. Valid observations acquired within each calendar month are subsequently combined using local-resolution-weighted multi-temporal compositing, in which observations with higher effective local spatial resolution receive greater weight and observations affected by radar shadow are excluded (Small, 2012). This procedure reduces the speckle and residual effects of topography and viewing geometry, resulting in spatially more homogeneous backscatter composites. We used the 20 m monthly VV and VH mosaics as input covariates. Additionally, we calculated the Radar Vegetation Index (RVI) from the monthly Sentinel-1 VV and VH backscatter mosaics to provide an additional covariate describing vegetation scattering characteristics. We used the dual-polarization formulation adapted for Sentinel-1 (Mandal et al., 2020).

- **ALOS PALSAR-2**: We obtained annual L-band Synthetic-Aperture Radar (SAR) data from the JAXA Global PALSAR/PALSAR-2 Yearly Mosaic (Version 2.6.0), accessible through the JAXA Forest/Non-Forest dataset portal (Japan Aerospace Exploration Agency, 2024). The product consists of seamless annual images generated from PALSAR-2 SAR image strips at 25 m spatial resolution. For each year and location, JAXA preferentially selects observations from acquisitions showing relatively low sensitivity to surface-moisture conditions, and only observations from the corresponding year are included in the mosaic. The source imagery is orthorectified and radiometrically slope corrected using the ALOS World 3D 30 m digital surface model (AW3D30). Radiometric balancing and destriping are also applied between neighboring image

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

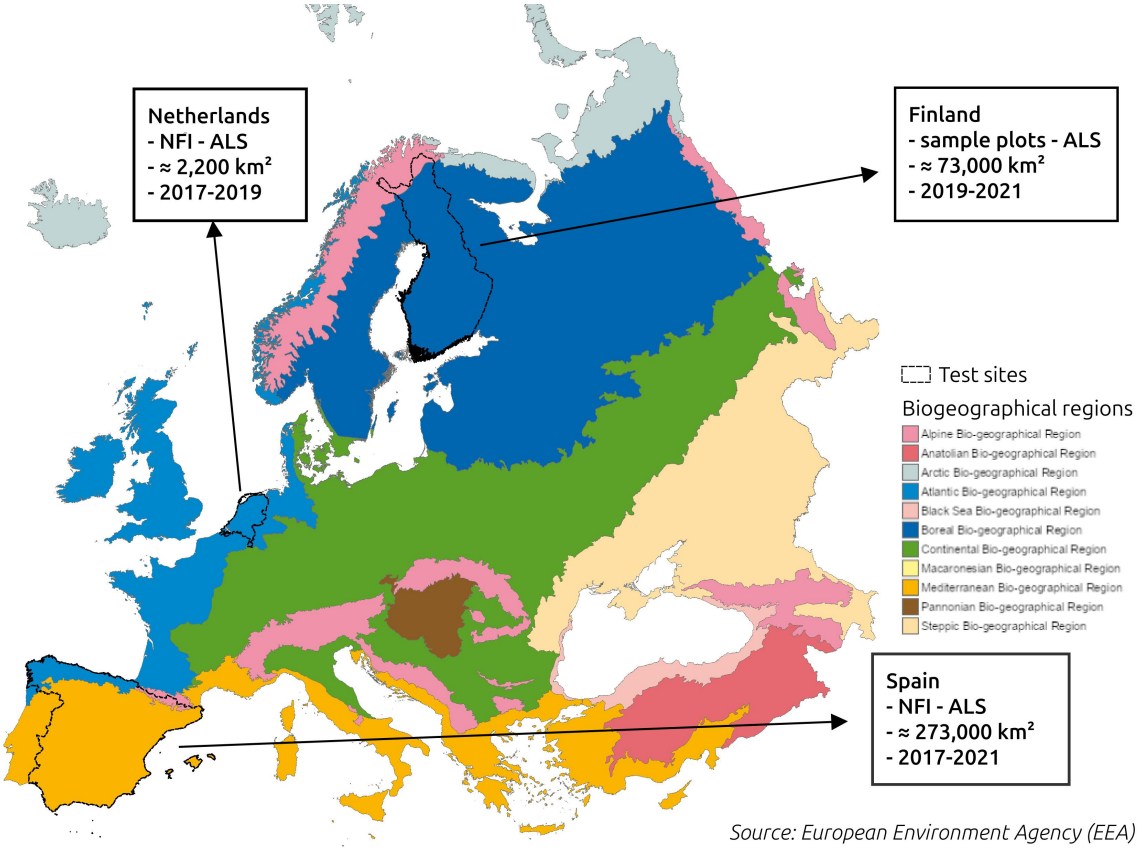



**Figure 2.** Overview of three countries used for model benchmarking and key characteristics overlaid with biogeographical region of Europe (European Environment Agency, 2002). The base map was designed and developed by Esri | Powered by Esri.

strips to reduce discontinuities in backscatter intensity (Shimada et al., 2014). We used the HH and HV polarization bands, supplied as 16-bit digital numbers (DN), and converted them to terrain-normalized gamma-nought backscatter in decibels according to $\gamma_0 = 10\log_{10}(\mathrm{DN}^2) - 83.0$ dB given by Shimada et al. (2014).

– **Sentinel-2**: Sentinel-2 imagery was obtained as Sentinel-2 Level-3 quarterly mosaics via the Copernicus Data Space Ecosystem (CDSE) at 10 resolution. Rather than the standard 13 spectral bands, this quarterly mosaic product comprises four spectral bands—B02 (blue), B03 (green), B04 (red), and B08 (near-infrared)—along with an observation count layer for valid pixels. During product generation, cloud masking was first applied to the Sentinel-2 Level-2A input data. Subsequently, the first quartile (25th percentile) of the surface reflectance distribution was extracted to suppress residual bright pixels misclassified as cloud-free. Pixels lacking valid observations within the quarterly timeframe were assigned a NaN value (European Union, 2024). Additionally, we derived a suite of optical vegetation indices, including the Ratio Vegetation Index (RVI), Normalized Difference Vegetation Index (NDVI), Difference Vegetation Index (DVI),

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

Enhanced Vegetation Index (EVI), and Optimized Soil Adjusted Vegetation Index (OSAVI). We selected these indices in accordance with state-of-the-art AGBD mapping research to maximize predictive accuracy under the premise of using limited spectral bands from the CDSE (Moradi et al., 2022; Mutanga et al., 2023).

### 2.4 EO Embeddings

Satellite embeddings are used in a wide range of foundational model products in different formats and specifications (Stewart et al., 2026). In this research, we explicitly choose embeddings that are in numerical vector representation at 10 m resolution to embody the benefits of EO Embeddings for downstream tasks:

- **AlphaEarth Foundation Satellite Embeddings**: AlphaEarth is a 64-band numerical vector embedding developed by Google and Google DeepMind (Brown et al., 2025). The model was trained using self-supervised deep learning on multi-modal remote sensing inputs, including PALSAR-2, Sentinel-1 A/B, and Sentinel-2 Level-1C data. The embeddings were accessed and processed using the pipeline provided by Taylor Geospatial (2025).

- **TESSERA v1.1 Embeddings**: TESSERA v1.1 is an open-source 128-band numerical vector embedding developed at the University of Cambridge (Feng et al., 2026). The self-supervised model was trained on Sentinel-1 Radiometrically Terrain Corrected (RTC) and Sentinel-2 Level-2A imagery. The dataset download pipeline and instructions are documented in the official GitHub repository (Cambridge Centre for Earth Observation (ucam-eo), 2026).

- **SpheerFM embeddings**: SpheerFM is a commercial foundation model developed by Spheer (Spheer, 2026). Comprising 0.7 million parameters, the self-supervised model was trained exclusively on a dataset of 200 million Sentinel-2 pixels. The training framework geographically prioritized the Netherlands and its adjacent regions, deliberately targeting natural ecosystems and areas of extensive agriculture to capture relevant vegetation dynamics.

### 2.5 Terrain model and variables

Large-scale terrain features were extracted from GEDTM30, a 30 m global digital terrain model, and a data set of multiscale variables (Ho et al., 2025). The dataset comprises digital terrain height models alongside multiscale topographic variables designed to capture geomorphometric and hydrological complexity, including landform morphology, solar illumination (light and shadow), surface curvature, topographic position, and hydrology. To derive these multiscale variables, the GEDTM terrain height data were reprojected from WGS84 to an equal-area projected coordinate system before being reprojected back to WGS84 to generate a seamless and consistent global mosaic. This processing workflow mitigates pixel oversampling and spatial distortion in high-latitude regions (Ho et al., 2025).

#### Sampling design

We implemented a tile-based standardized spatial sampling scheme. The reference AGBD layer was partitioned into 300×300 pixel tiles or about 3 km×3 km in size. Specifically, a 3 km×3 km grid was generated over the AGBD reference datasets

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

**Table 1.** Specifications of input satellite imagery, EO embeddings, and DTM-derived topographic variables used for biomass mapping.

| Dataset | Variables | Spatial Res. (m) | Temporal Res. | Reference |
|---|---|---|---|---|
| AlphaEarth Foundations | 64 numerical vectors | 10 | Yearly | Brown et al. (2025); Taylor Geospatial (2025) |
| TESSERA v1.1 | 128 numerical vectors | 10 | Yearly | Feng et al. (2026) |
| SpheerFM | 100 numerical vectors | 10 | Yearly | Spheer (2026) |
| Sentinel-2 Level 3 | Red (B04), Green (B03), Blue (B02), Near-Infrared (B08), NDVI[a], DVI[b], RVI[c], EVI[d], OSAVI[e] | 10 | Quarterly | European Union (2024) |
| Sentinel-1 Mosaics | VH, VV, RVI[f] | 20 | Monthly | Copernicus Data Space Ecosystem (2026); Small (2012) |
| JAXA Global PALSAR/PALSAR-2 Mosaic v2.5.0 | HH, HV | 25 | Yearly | Japan Aerospace Exploration Agency (2024) |
| GEDTM30 terrain variables | Terrain height, Diff. mean elev.[g], Normal SD[h], Multidirectional hillshade, LS factor, Curvatures[i], Slope in degree, SCA[j], TWI[k] | 30, 60, 120, 240, 480, 960 | Static | Ho and Hengl (2026); Ho et al. (2025) |

[a] Normalized Difference Vegetation Index
[b] Difference Vegetation Index
[c] Ratio Vegetation Index
[d] Enhanced Vegetation Index
[e] Optimized Soil Adjusted Vegetation Index
[f] Radar Vegetation Index
[g] Difference from mean elevation
[h] Spherical standard deviation of the Normals
[i] Maximal, minimal, profile, and tangential curvatures
[j] Specific catchment area
[k] Topographic wetness index

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

in Spain, Finland, and the Netherlands, yielding a total of 9616, 54769, and 1484 tiles, respectively. The tiles were then partitioned into training, validation, and test sets. Then, stratified proportional sampling was implemented with a spatial buffer (*"moat"*). Firstly, the mean AGBD (Mg ha$^{-1}$) was calculated, enabling classification into five distinct biomass strata for each tile: low ($[0–50]$), medium-low ($(50–100]$), medium ($(100–150]$), medium-high ($(150–200]$), and high ($> 200$). Based on these biomass strata, we randomly sampled 20%, 15%, and 4% of the available tiles for Spain, the Netherlands, and Finland, respectively. This initial sampling minimizes the dense clustering of reference ALS data while accurately reflecting regional biomass distributions. Next, we randomly partitioned these subsets into training (80%), validation (10%), and test (10%) splits. To prevent spatial leakage from coarse-resolution predictors (e.g., multiscale terrain variables), we applied a one-tile spatial buffer (*"moat"*) around each validation and test tile, strictly isolating them from the training data. Because spatial tile density varied by country, this moating procedure resulted in the removal of a slightly different proportion of tiles across the three regions. Consequently, the final dataset comprised 1424 training, 178 validation, and 178 testing tiles for Spain; 1598 training, 166 validation, and 166 testing tiles for Finland; and 145 training, 19 validation, and 19 testing tiles for the Netherlands (Figure 3).

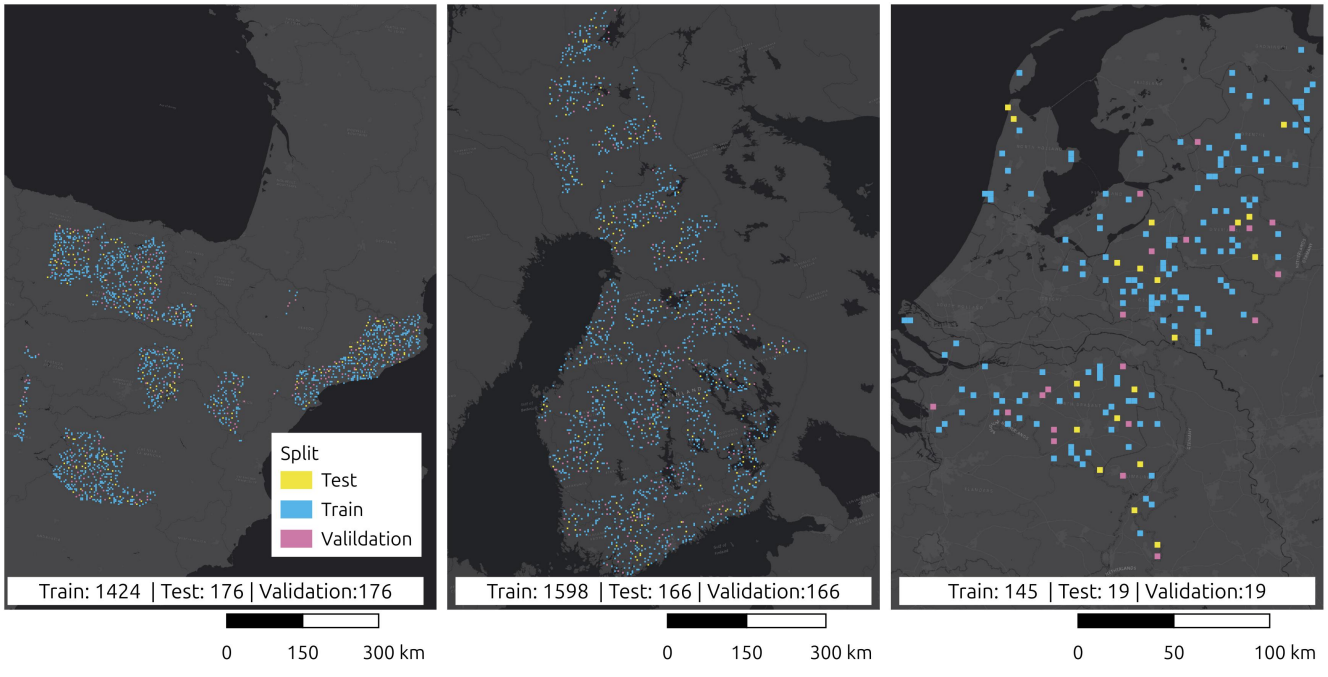



**Figure 3.** Spatial distribution of training (blue), validation (pink), and testing (yellow) tiles selected across Spain (left), Finland (middle), and the Netherlands (right) using a proportional stratified sampling scheme with a one-tile spatial buffer. Each tile is 3 km×3 km in size. The base map was designed and developed by Esri | Powered by Esri.

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

## Models

We focus primarily on scalable models to enable annual 10 m high-resolution pan-European AGBD mapping alongside uncertainty quantification. Two distinct modeling approaches were chosen: Quantile Random Forest Regression (QRFR) (Meinshausen and Ridgeway, 2006) and the Xception Convolutional Neural Network (Xception CNN) (Chollet, 2017):

- **Quantile Random Forest Regressor (QRFR)**: QRFR serves as our baseline tree-based machine learning model for stable point estimation and non-parametric uncertainty quantification. It has proven highly scalable for continental-scale and global-scale spatiotemporal mapping (Tian et al., 2025; Hengl et al., 2026). While sharing the standard RF structure, QRFR retains the full conditional distribution of target variables in leaf nodes rather than merely averaging predictions, thereby deriving predictive quantiles to represent uncertainty (Meinshausen and Ridgeway, 2006). We implement the model using the quantile-forest Python package (Zillow Group, 2022), an extension compatible with scikit-learn. From each sampled training tile, we randomly selected 20% of the pixels to serve as training points. These points were then aggregated per country to construct a tabular dataset suitable for tree-based modeling. During the model fitting phase, this dataset was partitioned into a 90% training subset and a 10% calibration subset. The calibration subset was utilized for hyperparameter tuning to optimize the maximum depth, maximum features, minimum samples per split, and minimum samples per leaf. Finally, the resulting optimal hyperparameters were used to train the final QRFR model.

- **Xception Convolutional Neural Network**: The Xception CNN is a deep learning architecture optimized for spatial feature extraction via Depthwise Separable Convolutions (SepConv). The model architecture reduces parameter overhead by 1×1 convolutions rather than performing standard 3D convolutions, effectively mapping cross-channel correlations (Chollet, 2017; Bendersky, 2018). This architecture was adapted for environmental mapping by Lang et al. (2019), and was used in particular to produce the first global 10 m canopy height map with pixel-level uncertainty (Lang et al., 2023). In recent benchmarks, Sialelli et al. (2025) demonstrated that Xception CNN outperforms Gradient Boosted Decision Trees (GBDT), Fully Convolutional Networks (FCN), and UNet architectures in AGBD prediction. We build upon the open-source Xception CNN architecture provided by Sialelli et al. (2025), incorporating four key modifications into our benchmarking pipeline: (1) training directly on 10 m reference AGBD imagery, (2) spatial data augmentations (random horizontal flips, vertical flips, and cutout regularization), (3) adopting Gaussian Negative Log-Likelihood (NLL) as the loss function for direct parametric variance estimation, and (4) integrating two-dimensional spatial dropout layers. Following Lang et al. (2023), the use of Gaussian NLL enables unbiased estimation of predictive uncertainty. Unlike the QRFR approach, which requires tabular data, our model takes the sampled tiles directly as input. We then normalized by subtracting the mean and dividing by the standard deviation both the training and validation sets into $z$-scores. We derived the mean and standard deviation exclusively from the training set. Subsequently, the 300×300 pixel training and validation tiles were cropped into 25×25 pixel patches with a stride of 25 pixels.

Aboveground biomass observations typically exhibit a strongly positively skewed and zero-inflated distribution. To effectively train the models, we adopted model-specific strategies to account for these distributional characteristics. For the CNN

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

model, where the Gaussian NLL loss function assumes normally distributed residuals to generate unbiased mean and variance estimates, the target AGBD values were transformed via a square-root function. However, to prevent back-transformation bias when converting predictions back to the original unit scale (Mg ha$^{-1}$), we then applied the unbiased back-transformation (Nussbaum et al., 2018):

$$\widetilde{Y}(\mathbf{s}) = \hat{f}(\mathbf{s})^2 + \hat{\sigma}^2 - \mathrm{Var}[\hat{f}(\mathbf{s})], \tag{1}$$

where $\widetilde{Y}(\mathbf{s})$ denotes the unbiased AGBD prediction, $\hat{f}(\mathbf{s})$ represents the prediction on the square-root scale, $\hat{\sigma}^2$ is the estimated residual variance of the fitted model, and $\mathrm{Var}[\hat{f}(\mathbf{s})]$ is the variance of $\hat{f}(\mathbf{s})$ between the realizations of the model.

For the QRFR model, we addressed target skewness natively by fitting decision trees using a Poisson deviance criterion. By assuming that the variance scales proportionally with the mean ($\mathrm{Var}(Y) = \mu$), the Poisson objective naturally accounts for higher variance at elevated biomass levels and accommodates right-skewed, zero-inflated distributions without requiring explicit mathematical target transformations (Parente et al., 2026).

**Evaluation metrics of model performance**

We used four (4) point-estimate accuracy metrics to evaluate model performance on test tiles by countries. For the following equations, let $n$ represent the total number of observations, $y_i$ the observed reference value, $\hat{y}_i$ the predicted value, and $\bar{y}$ the mean of the observed values. The four metrics include:

- **Root Mean Square Error (RMSE):** RMSE quantifies the square root of the average squared differences between predicted and observed values. It is particularly sensitive to outliers due to the squaring of error terms, making it an indicator of large model deviations.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{2}$$

- **Mean Absolute Error (MAE):** This metric calculates the average of the absolute differences between predictions and actual observations. Unlike RMSE, MAE acts as an L1 norm where all individual errors are weighted equally. This makes it more robust to outliers and provides a more stable representation of the typical model error.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{3}$$

- **Interdecile Range (IDR):** IDR is the difference between the 90th percentile ($P_{90}$) and the 10th percentile ($P_{10}$) of the errors. This metric evaluates the statistical spread and consistency of the model predictions. Taking the $P_{10}$ and $P_{90}$ prevents outliers to dominate the dispersion estimate.

$$\mathrm{IDR} = P_{90} - P_{10} \tag{4}$$

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

- **Concordance Correlation Coefficient (CCC):** This metric is designed to measure the agreement between predicted ($y$) and observed ($\hat{y}$) values, capturing both the scatter of the data around the $y = \hat{y}$ line and the distance between the mean prediction and the mean observation (Steichen and Cox, 2002).

$$\text{CCC} = \frac{2 \cdot \text{Cov}(y, \hat{y})}{\text{Var}(y) + \text{Var}(\hat{y}) + (\bar{y} - \bar{\hat{y}})^2} \tag{5}$$

Beyond point estimates, the spatial pattern of aboveground biomass has often been neglected, yet it plays a critical role in CRCF for field inspections. To evaluate pattern reconstruction across different models and predictor configurations, we adopt the spatial metrics proposed by Jones et al. (2016):

- **Similarity in Means (SIM):** This metric evaluates the similarity of average values or quantities between two maps — it ranges from 0 to 1, where 0 indicates complete dissimilarity and 1 indicates that the local means are perfectly identical.

$$\text{SIM} = \frac{2\mu_y\mu_{\hat{y}} + C_1}{\mu_y^2 + \mu_{\hat{y}}^2 + C_1} \tag{6}$$

- **Similarity in Variance (SIV):** This metric quantifies the degree of local variability — it ranges from 0 to 1, where a value of 1 indicates a perfect match in local variability between the two maps.

$$\text{SIV} = \frac{2\sigma_y\sigma_{\hat{y}} + C_2}{\sigma_y^2 + \sigma_{\hat{y}}^2 + C_2} \tag{7}$$

- **Similarity in Pattern (SIP):** This metric measures the covariance correlation — it tracks the alignment of underlying geometric patterns, gradients, or spatial shapes align between the maps. SIP spans from -1 to 1, where 1 indicates identical spatial features, 0 indicates no structural relationship, and -1 indicates a perfect negative correlation.

$$\text{SIP} = \frac{\sigma_{y\hat{y}} + C_3}{\sigma_y\sigma_{\hat{y}} + C_3} \tag{8}$$

- **Structural Similarity Index Measure (SSIM):** This composite metric is calculated as the product of the three aforementioned metrics. Spanning from -1 to 1, it evaluates the overall agreement between two spatial datasets by quantifying their simultaneous alignment in local means, variability, and structural patterns.

$$\text{SSIM} = \text{SIM} \times \text{SIV} \times \text{SIP} \tag{9}$$

These metrics are computed locally using an sliding Gaussian 11×11 pixel window and then averaged across all valid pixels in the spatial domain (in each tile). Let $\mu_y$ and $\mu_{\hat{y}}$ be the local means of the observed and predicted values, $\sigma_y$ and $\sigma_{\hat{y}}$ their respective local standard deviations, and $\sigma_{y\hat{y}}$ the local cross-covariance between the observed and predicted spatial windows. The stabilizing constants are defined as $C_1 = (K_1 L)^2$, $C_2 = (K_2 L)^2$, and $C_3 = C_2/2$, where $L$ represents the range of the AGBD observations (maximum minus minimum value), and the default scalar constants are set to $K_1 = 0.01$ and $K_2 = 0.03$.

Furthermore, AGB mapping inherently suffers from signal saturation, where spectral data fails to capture the full complexity of dense canopy structures, typically leading to the underestimation of high-biomass areas (Mutanga et al., 2023). To assess this

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

limitation, we evaluated prediction errors prediction error (the difference between observed, $y$, and predicted, $\hat{y}$, values) across all test tiles. The observed AGBD was categorized into five discrete ranges: low (0–50 ton/ha), low-medium (50–100 ton/ha), medium (100–150 ton/ha), high-medium (150–200 ton/ha), and high (>200 ton/ha). Aggregating the mean prediction error within each bin allowed us to identify which model configurations most effectively mitigate saturation effects across increasing forest densities.

Finally, we also incorporate metrics to quantify the accuracy of uncertainty estimations (accuracy of prediction errors), applying established empirical metrics following Tian et al. (2025). The uncertainty assessment evaluates whether the prediction intervals (PI) generated by the models accurately reflect empirical probabilities. To achieve this, we aggregated pixel-level predictions from test tiles across all countries to calculate the tile-based PICP (Eq. 10) and MPI (Eq. 11) for each model. This empirical approach examines whether the actual percentage of observed values falling within a given prediction interval matches the nominal target probability.

Let $l_i$ and $u_i$ represent the lower and upper bounds of the estimated prediction interval for observation $i$, and $n$ represent the total number of observations; the following two metrics are then derived using:

- **Prediction Interval Coverage Probability (PICP):** PICP represents the proportion of reference observations that fall within the estimated prediction bounds. An accurately calibrated uncertainty model will yield a PICP that perfectly matches the nominal confidence level (Malone et al., 2011).

$$\text{PICP} = \frac{1}{n}\sum_{i=1}^{n} c_i, \quad \text{where } c_i = \begin{cases} 1, & \text{if } y_i \in [l_i, u_i] \\ 0, & \text{if } y_i \notin [l_i, u_i] \end{cases} \tag{10}$$

- **Mean Prediction Interval (MPI):** Alternatively referred to as Prediction Interval Width (PIW), MPI calculates the average span between the upper ($u_i$) and lower ($l_i$) bounds across all prediction intervals. It serves as a straightforward indicator of model confidence: a wider interval denotes higher predictive uncertainty, whereas a narrower interval signifies greater precision (Kasraei et al., 2021).

$$\text{MPI} = \frac{1}{n}\sum_{i=1}^{n}(u_i - l_i) \tag{11}$$

From QRFR we derived layers of quantiles from 0.05 to 0.95 (in 0.05 increments) to construct pixel-wise prediction intervals. For instance, the 90% two-sided prediction interval is derived simply by subtracting the 5th percentile layer (0.05 quantile) from the 95th percentile layer (0.95 quantile). For the Xception CNN, the model optimizes a Gaussian NLL. The network is designed to predict the log-variance, denoted as $s = \log(\hat{\sigma}^2)$, rather than the raw variance directly (Nix and Weigend, 1994). Consequently, the predicted value at any given probability quantile $p$, denoted as $\hat{y}_p$, is derived analytically from the predicted mean and the exponentiated log-variance using Gaussian distributional assumptions:

$$\hat{y}_p = \hat{y} + Z_p \cdot \sqrt{\exp(s)} \tag{12}$$

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

where $\hat{y}$ is the predicted point estimate (mean), and the standard deviation is recovered via $\sqrt{\exp(s)}$. The scalar $Z_p = \Phi^{-1}(p)$ represents the $z$-score corresponding to the target probability $p$. Here, $\Phi$ denotes the Cumulative distribution function of the standard normal distribution, which computes the probability that a standard normal random variable is less than or equal to a given value. Its inverse, $\Phi^{-1}$ (the quantile function), maps the target cumulative probability $p \in (0, 1)$ back to the required standard deviations ($z$-score) from the mean. For instance, to extract the $5^{\text{th}}$ and $95^{\text{th}}$ percentiles for constructing a $90\%$ prediction interval, the model utilizes $Z_{0.05} = \Phi^{-1}(0.05) \approx -1.645$ and $Z_{0.95} = \Phi^{-1}(0.95) \approx 1.645$, respectively.

In order to identify the optimal modeling framework for regional AGBD mapping, simultaneous evaluation of point-estimate accuracy, spatial pattern reconstruction, uncertainty quantification, and saturation effects is essential. To synthesize these multidimensional metrics, we developed a *"multi-criteria scoreboard"* that assigns a consolidated ranking to every combination of predictor and model. This framework employs a rank-based scoring system where model configurations are assigned points proportional to their performance across individual metrics (from the worst to the best). For error-based metrics (MAE, RMSE, and IDR), lower values received higher points, whereas for agreement and structural metrics (CCC, SSIM, SIP, SIV, and SIM), higher values received maximum points. Saturation effects were assessed across the five discrete AGBD categories, rewarding configurations with mean prediction errors closest to zero. For uncertainty, both PICP and MPI were evaluated: PICP performance was scored by computing the mean absolute divergence between PICP and expected coverage across nominal confidence level spanning 10% to 90% (in 10% increments). To ensure fair MPI comparison despite varying coverage levels, raw MPI values were interpolated to standard expected coverage targets prior to averaging, with lower calibrated MPI values earning higher scores.

Finally, metrics were grouped into four equal-weight evaluation dimensions: point-estimate accuracy, spatial pattern reconstruction, saturation effects, and uncertainty quantification. A weighted final score was derived by averaging the points within each dimension to ensure equal voting power across all four domain criteria. Notably, while SSIM is a composite of SIP, SIV, and SIM, all four metrics were retained within the spatial dimension to equally reward models that achieve both individual component fidelity and overall structural agreement.

## 3 Result

The evaluation of performance shows produced differences in predictive point-estimate accuracy, spatial pattern reconstruction, uncertainty quantification and saturation of AGBD for various sets of predictors, the efficacy of the models (QRFR versus CNN) and the impact of including GEDTM. The metrics were calculated in the test tiles for the Netherlands (19 tiles), Spain (176 tiles), and Finland (166 tiles), with median values reported in the results at the tile-level. For uncertainty and saturation effects, the calculation was done pixel-based. Table 2, 3 and 4 summarize the tile-based results.

### 3.1 Point-estimate comparison

Firstly, we merely look into QRFR models. In Spain, AlphaEarth achieved the best overall performance with the lowest MAE, IDR and the best agreement (CCC) at 14.3 and 41.5, 0.629 respectively (Table 2). With addition of GEDTM, AlpheEarth

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

yields the lowest RMSE 20.0. In Finland (Table 3), AlphaEarth with the addition of GEDTM ranks the first of all predictor configurations, achieving the best performance across every metric (RMSE: 16.5, MAE: 13.0, IDR: 40.7, CCC: 0.812). In the Netherlands (Table 4), the AlphaEarth baseline produced the best performance of MAE and CCC at 33.1 and 0.736 respectively. With addition of GEDTM, it yield the lowest RMSE (42.1) and IDR (99.2).

Performance dynamics, however, changed when switching to CNN model. In Spain (Table 2), the rank of the models by metrics was mixed: SpheerFM achieved the best RMSE (17.5). SpheerFM+GEDTM produced the highest IDR (36.3), and TESSERA reached the lowest MAE (12.6) and the best agreement (CCC:0.737). In Finland (Table 3), TESSERA decisively outperformed all other configurations, recording the best results across all four metrics (RMSE: 13.4, MAE: 10.3, CCC: 0.893). With GEDTM, TESSERA reaches the lowest IDR at 32.2. However, in the Netherlands (Table 4), AlphaEarth remained the best results, reaching the lowest RMSE (42.4), MAE (33.1), IDR (100.4) and highest CCC (0.738). Overall, CNN model manages to yield the better performance (dominate the best performance) in three countries.

To evaluate the impact of adding terrain features, Figure 4, compiled results of three countries, shows adding GEDTM improves the prediction from QRFR in Finland and the Netherlands among TESSERA, SpheerFM and multi-sensor satellite ARD (S1+S2+PALSAR). Nevertheless, GEDTM does not improve the same model output in Spain, and also for AlphaEarth as predictor set. In addition, using CNN as model, GEDTM often worsen the prediction accuracy in terms of RMSE. On the other hand, the CNN model consistently outperforms the QRFR model across all three countries (Figure 4). While the choice of model has little effect on AlphaEarth's point-estimate accuracy, the CNN architecture notably enhances the performance of both SpheerFM and TESSERA. In Spain and Finland, this architectural benefit is prominent enough that TESSERA and SpheerFM surpass AlphaEarth when utilized with the CNN. Finally, EO embeddings, compared with multi-sensor satellite ARD, offer in general lower RMSE across three countries and models, except for TESSERA with QRFR in the Netherlands.

### 3.2 Result of spatial pattern reconstruction

Again, we first look at the QRFR model. In Spain (Table 2), the best rank for each metric was more evenly distributed in configurations: SpheerFM achieved the best overall SSIM (0.520), but AlphaEarth reached the highest SIP (0.968), TESSERA obtained the highest SIV (0.917), and AlphaEarth+GEDTM secured the best SIM (0.627). In Finland (Table 3), AlphaEarth+GEDTM leads in overall SSIM (0.663) SIP (0.991) and SIM (0.717), whereas TESSERA+GEDTM provided the highest SIV (0.944). In the Netherlands (Table 4), standalone AlphaEarth leads in SSIM (0.631) and SIM (0.689), while AlphaEarth+GEDTM attains the highest SIP (0.986). Meanwhile, S1+S2+PALSAR records the top SIV at 0.941. Transitioning to the CNN model, AlphaEarth consistently performed best in the Netherlands (Table 4), with the highest SSIM (0.662), SIP (0.987) and SIM (0.726). Meanwhile, S1+S2+PALSAR narrowly leads in SIV (0.940). In Spain (Table 2), SpheerFM recorded the best SSIM (0.648), and with the addition of GEDTM, it also achieved the highest SIV (0.926) and SIM (0.739); however, TESSERA provided the best SIP (0.978). Finally, in Finland (Table 3), TESSERA scores the best in SIP (0.995) and SIM (0.797). With the addition of GEDTM, It also leads in SSIM (0.752) and SIV (0.950).

Figure 5 illustrates the SSIM results across the three countries to assess how GEDTM and models influence spatial pattern reconstruction. Adding GEDTM marginally improves spatial patterns for the TESSERA, SpheerFM, and S1+S2+PALSAR

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

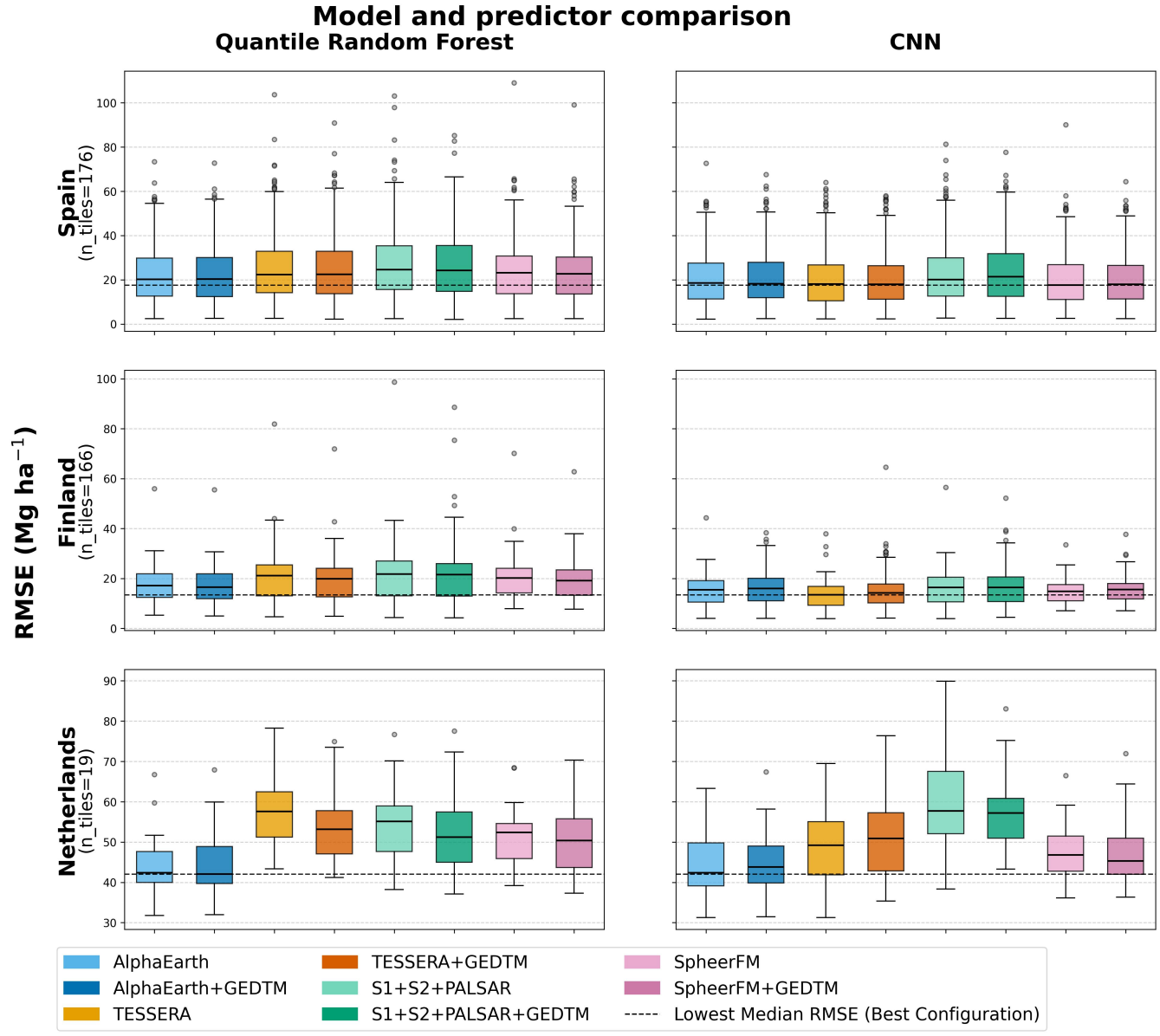



**Figure 4.** Cross comparison in point-estimate (RMSE) of predictor configurations, downstream models in Spain, Finland and the Netherlands. CNN: Xception convolutional neural network.

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

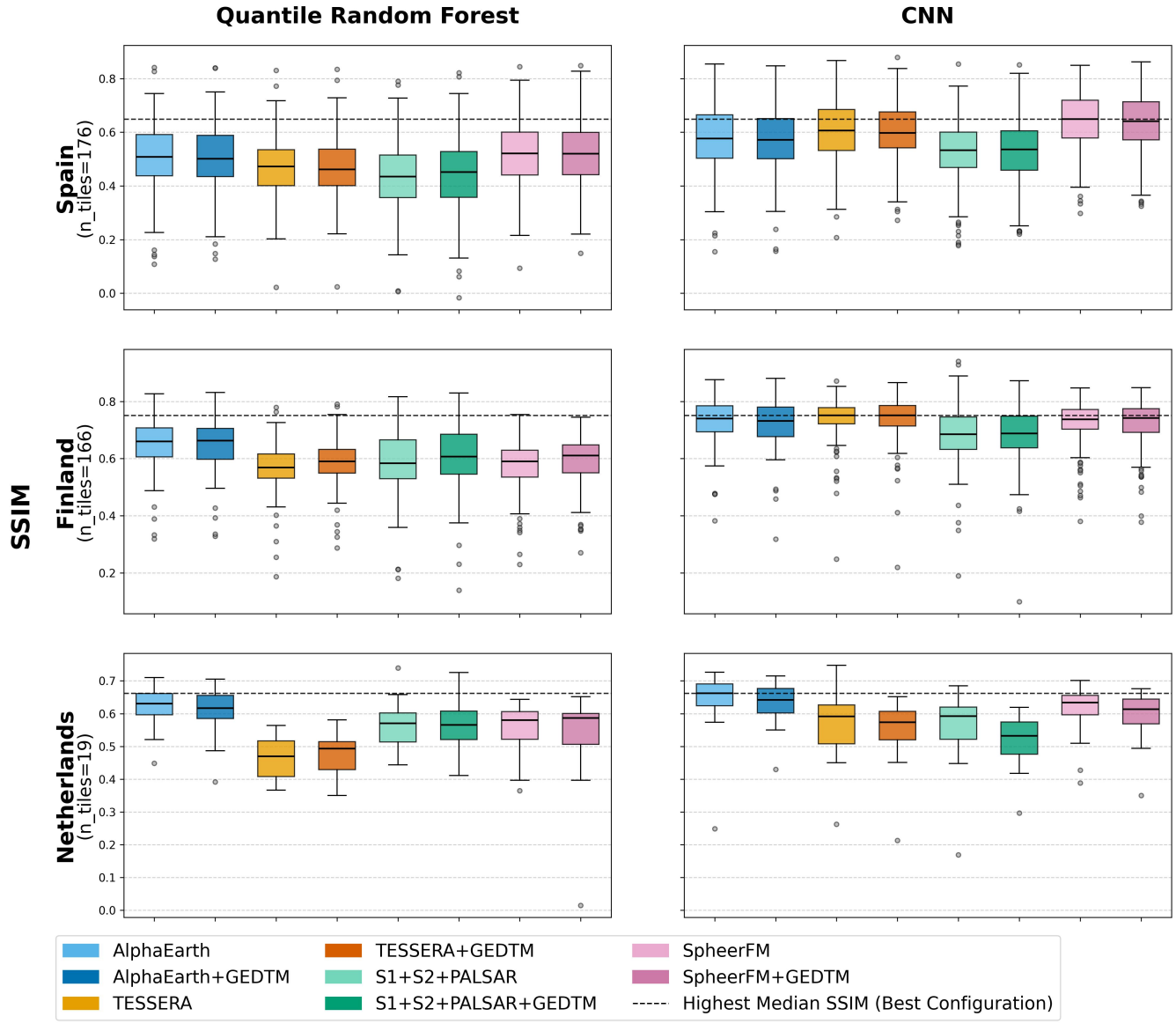



**Figure 5.** Cross comparison in spatial pattern reconstruction (SSIM) of predictor-model configurations in Spain, Finland and the Netherlands. CNN: Xception convolutional neural network.

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

under the QRFR model in Finland and the Netherlands. The AlphaEarth baseline sees negligible or even adverse effects from topographic data, regardless of the model. Towards the choice of models, transitioning from the QRFR to the CNN models gains a substantial improvement in structural similarity. For instance, in Finland, the interquartile range (IQR) of SSIM shifts from around 0.5–0.7 under the QRFR model to around 0.7–0.8 under the CNN model, a positive trend can consistently be observed in also both Spain and the Netherlands (Figure 5). Aligning with the earlier point-estimate evaluations, the TESSERA and SpheerFM predictor sets demonstrate marked performance gains when paired with the CNN model.

To evaluate the practical significance of spatial pattern reconstruction, Figure 6 illustrates the spatial AGBD predictions across predictor sets and models for a representative test area (Tile *049235_Soria_2019*). A key divergence appears between models: while AlphaEarth outputs coherent spatial features under both QRFR and CNN models, TESSERA, SpheerFM, and S1+S2+PALSAR fail to generate continuous spatial structures using QRFR. When integrated into the CNN, however, TESSERA and SpheerFM sharply resolve fine-scale landscape features, cleanly capturing the high-biomass strip in the lower-left quadrant and the central forest patch. However, while the S1+S2+PALSAR visually improves under the CNN, its spatial enhancement remains less refined than that achieved by TESSERA and SpheerFM.

### 3.3 Performance of the uncertainty quantification

Figure 7 illustrates the PICP (top row) and MPI (bottom row) for the QRFR (left column) and CNN (right column) models, explicitly visualizing the coverage error (calculated as the empirical PICP minus the expected coverage). The CNN models demonstrate highly accurate probability calibration, with coverage errors tightly constrained to approximately $\pm 2\%$. In contrast, the QRFR model slightly overestimates uncertainty, particularly when utilizing EO embedding-based predictors. Nevertheless, the maximum uncertainty quantification error of the QRFR model never exceeds 6%. Given by the PICP of each configuration, the MPI curves were drawed at the bottom part of Figure 7. Among the predictor setups, AlphaEarth (with and without GEDTM) achieves narrower uncertainty bounds. Within the CNN models specifically, EO embedding-based predictors achieve the MPIs (57–63 ton/ha) at around 90% PI, whereas S1+S2+PALSAR obtains above 65 Mg ha$^{-1}$. Nevertheless, CNN models achieve narrower uncertainty bounds than all QRFR models (above 70 Mg ha$^{-1}$ at 90% expected coverage). In general, these findings demonstrate that CNN consistently outperforms QRFR by generating better-calibrated and narrower uncertainty.

### 3.4 Saturation Effects

Figure 8 shows per-pixel prediction errors categorized by discrete above-ground biomass ranges (0–50, 50–100, 100–150, 150–200, and $>$200 Mg ha$^{-1}$) across the evaluated predictor configurations. The analysis compares the QRFR and CNN models using available pixels from all test tiles in Spain, Finland, and the Netherlands. Error bounds denote the 90th and 10th percentiles. As the result, in Spain and Finland, where reference AGBD falls predominantly within the low and low-medium ranges, the median of prediction errors across all QRFR predictors reveal an underestimation of high-biomass areas by approximately 50–100 ton/ha. Conversely, because the Dutch reference data is heavily concentrated in the medium to high ranges, the models mitigate this high-end saturation effect, restricting the underestimation to below 50 ton/ha. However, this skewed data distribution causes the Dutch models to overestimate low-biomass areas by approximately 50 ton/ha, where

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

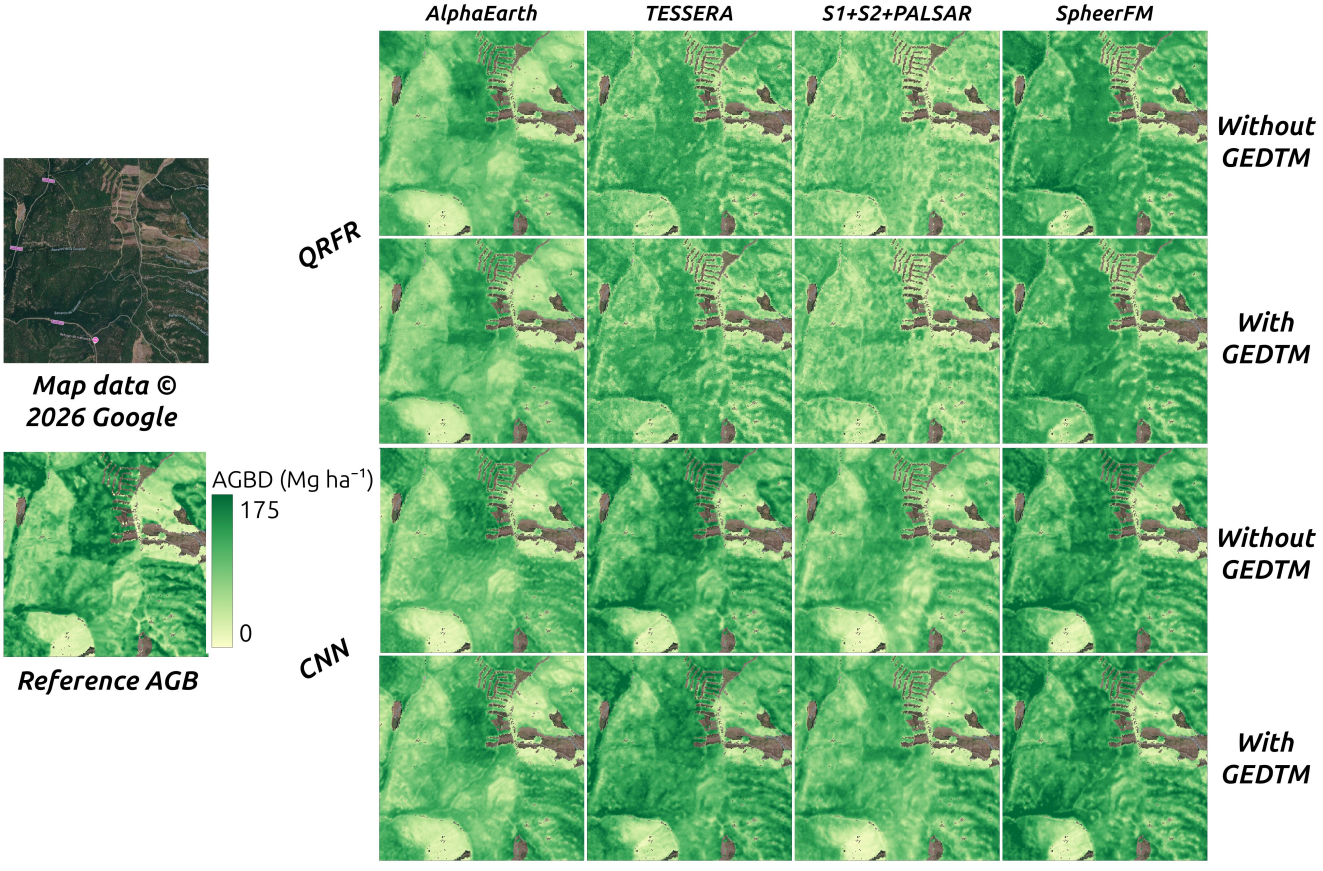



**Figure 6.** Spatial AGBD predictions generated across evaluated predictor configurations and models (QRFR and CNN) for a representative test area (Tile 049235, Soria, Spain, 2019; 41.8°N, 2.938°W). The tile designation indicates the tile ID (049235), target region (Soria, Spain), and observation year (2019). QRFR: Quantile random forest regressor. CNN: Xception convolutional neural network. Map data: © 2026 Google.

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

**Table 2.** Comparison of point-estimate accuracy and spatial pattern metrics across predictor-model configurations for Spain. Bold text indicates the best performance within a specific predictor set for a given model, while an asterisk (*) denotes the highest overall performance. QRFR: Quantile random forest regressor; CNN: Xception convolutional Neural Network; Base/+GEDTM: without/with adding features of terrain variables (GEDTM). Details of metrics refer to section 2.5.

| *Spain* | | | Point Estimate | | | | Spatial Pattern | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | RMSE | MAE | IDR | CCC | SSIM | SIP | SIV | SIM |
| *(n=176)* | | | | | | | | | | |
| QRFR | AlphaEarth | Base | 20.4 | **14.3** | **41.5** | **0.629** | 0.506 | **0.968** | 0.873 | 0.624 |
| | | +GEDTM | **20.0** | 14.6 | 41.6 | 0.618 | 0.499 | 0.968 | 0.856 | **0.627** |
| | TESSERA | Base | 22.1 | 15.9 | 44.5 | 0.542 | 0.472 | 0.960 | **0.917** | 0.545 |
| | | +GEDTM | 22.4 | 16.0 | 45.6 | 0.548 | 0.461 | 0.959 | 0.904 | 0.549 |
| | S1+S2+PALSAR | Base | 24.4 | 18.9 | 51.0 | 0.474 | 0.435 | 0.949 | 0.917 | 0.511 |
| | | +GEDTM | 24.2 | 17.5 | 49.2 | 0.508 | 0.451 | 0.951 | 0.908 | 0.520 |
| | SpheerFM | Base | 23.2 | 16.2 | 44.9 | 0.581 | **0.520** | 0.957 | 0.911 | 0.607 |
| | | +GEDTM | 22.6 | 15.8 | 44.1 | 0.582 | 0.520 | 0.958 | 0.899 | 0.609 |
| CNN | AlphaEarth | Base | 18.3 | 13.3 | 37.4 | 0.697 | 0.576 | 0.973 | 0.882 | 0.681 |
| | | +GEDTM | 18.2 | 13.1 | 37.6 | 0.683 | 0.570 | 0.972 | 0.879 | 0.689 |
| | TESSERA | Base | 18.0 | **12.6*** | 36.8 | **0.737*** | 0.605 | **0.978*** | 0.907 | 0.692 |
| | | +GEDTM | 17.9 | 12.9 | 36.5 | 0.722 | 0.597 | 0.975 | 0.906 | 0.689 |
| | S1+S2+PALSAR | Base | 20.0 | 14.4 | 41.3 | 0.638 | 0.532 | 0.965 | 0.874 | 0.649 |
| | | +GEDTM | 21.3 | 15.3 | 42.8 | 0.618 | 0.535 | 0.959 | 0.885 | 0.639 |
| | SpheerFM | Base | **17.5*** | 12.9 | 37.0 | 0.732 | **0.648*** | 0.972 | 0.920 | 0.735 |
| | | +GEDTM | 17.9 | 13.6 | **36.3*** | 0.727 | 0.640 | 0.971 | **0.926*** | **0.739*** |

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

**Table 3.** Comparison of point-estimate accuracy and spatial pattern metrics across predictor-model configurations for Finland. Bold text indicates the best performance within a specific predictor set for a given model, while an asterisk (*) denotes the highest overall performance. QRFR: Quantile random forest regressor; CNN: Xception convolutional Neural Network; Base/+GEDTM: without/with adding features of terrain variables (GEDTM). Details of metrics refer to section 2.5.

| *Finland* | | | Point Estimate | | | | Spatial Pattern | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | RMSE | MAE | IDR | CCC | SSIM | SIP | SIV | SIM |
| *(n=166)* | | | | | | | | | | |
| QRFR | AlphaEarth | Base | 17.1 | 13.5 | 41.7 | 0.806 | 0.660 | 0.990 | 0.932 | 0.714 |
| | | +GEDTM | **16.5** | **13.0** | **40.7** | **0.812** | **0.663** | **0.991** | 0.927 | **0.717** |
| | TESSERA | Base | 21.1 | 16.5 | 51.0 | 0.733 | 0.569 | 0.986 | 0.942 | 0.614 |
| | | +GEDTM | 19.9 | 15.3 | 46.0 | 0.761 | 0.590 | 0.988 | **0.944** | 0.639 |
| | S1+S2+PALSAR | Base | 21.7 | 17.2 | 52.0 | 0.718 | 0.584 | 0.984 | 0.944 | 0.628 |
| | | +GEDTM | 21.5 | 16.9 | 47.8 | 0.735 | 0.607 | 0.985 | 0.943 | 0.658 |
| | SpheerFM | Base | 20.2 | 15.7 | 48.4 | 0.740 | 0.590 | 0.985 | 0.924 | 0.647 |
| | | +GEDTM | 19.1 | 14.7 | 44.8 | 0.761 | 0.610 | 0.987 | 0.925 | 0.666 |
| CNN | AlphaEarth | Base | 15.5 | 12.0 | 35.9 | 0.852 | 0.740 | 0.992 | 0.943 | 0.792 |
| | | +GEDTM | 15.9 | 12.5 | 36.5 | 0.842 | 0.732 | 0.992 | 0.941 | 0.788 |
| | TESSERA | Base | **13.4*** | **10.3*** | 32.7 | **0.893*** | 0.751 | **0.995*** | 0.947 | **0.797*** |
| | | +GEDTM | 14.2 | 10.8 | **32.2*** | 0.884 | **0.752*** | 0.994 | **0.950*** | 0.795 |
| | S1+S2+PALSAR | Base | 16.4 | 12.7 | 38.8 | 0.844 | 0.685 | 0.992 | 0.928 | 0.741 |
| | | +GEDTM | 16.4 | 12.7 | 38.7 | 0.843 | 0.688 | 0.992 | 0.935 | 0.743 |
| | SpheerFM | Base | 14.9 | 11.5 | 35.3 | 0.870 | 0.738 | 0.992 | 0.945 | 0.787 |
| | | +GEDTM | 15.5 | 11.9 | 35.5 | 0.869 | 0.742 | 0.992 | 0.945 | 0.790 |

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

**Table 4.** Comparison of point-estimate accuracy and spatial pattern metrics across predictor-model configurations for the Netherlands. Bold text indicates the best performance within a specific predictor set for a given model, while an asterisk (*) denotes the highest overall performance. QRFR: Quantile random forest regressor; CNN: Convolutional Neural Network; Base/+GEDTM: without/with adding features of terrain variables (GEDTM). Details of metrics refer to section 2.5.

| *the Netherlands* | | | Point Estimate | | | | Spatial Pattern | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *(n=19)* | | | RMSE | MAE | IDR | CCC | SSIM | SIP | SIV | SIM |
| QRFR | AlphaEarth | Base | 42.4 | **33.1** | 101.7 | **0.736** | **0.631** | 0.986 | 0.925 | **0.689** |
| | | +GEDTM | **42.1*** | 33.3 | **99.2*** | 0.716 | 0.616 | **0.986** | 0.910 | 0.678 |
| | TESSERA | Base | 57.6 | 46.1 | 131.0 | 0.529 | 0.470 | 0.961 | 0.904 | 0.549 |
| | | +GEDTM | 53.2 | 41.6 | 121.5 | 0.558 | 0.493 | 0.969 | 0.891 | 0.570 |
| | S1+S2+PALSAR | Base | 55.1 | 42.7 | 122.5 | 0.629 | 0.570 | 0.973 | **0.941*** | 0.619 |
| | | +GEDTM | 51.2 | 40.9 | 116.8 | 0.621 | 0.566 | 0.974 | 0.932 | 0.627 |
| | SpheerFM | Base | 52.4 | 39.2 | 120.7 | 0.650 | 0.580 | 0.973 | 0.917 | 0.655 |
| | | +GEDTM | 50.4 | 38.3 | 112.6 | 0.633 | 0.586 | 0.977 | 0.901 | 0.651 |
| CNN | AlphaEarth | Base | **42.4** | **33.1*** | **100.4** | **0.747*** | **0.662*** | **0.987*** | 0.927 | **0.726*** |
| | | +GEDTM | 43.8 | 33.4 | 102.4 | 0.740 | 0.642 | 0.986 | 0.927 | 0.714 |
| | TESSERA | Base | 49.2 | 39.1 | 118.5 | 0.702 | 0.591 | 0.979 | 0.930 | 0.650 |
| | | +GEDTM | 50.9 | 38.2 | 117.0 | 0.653 | 0.574 | 0.979 | 0.903 | 0.650 |
| | S1+S2+PALSAR | Base | 57.7 | 46.3 | 131.8 | 0.636 | 0.592 | 0.968 | **0.940** | 0.648 |
| | | +GEDTM | 57.2 | 44.6 | 132.6 | 0.529 | 0.532 | 0.965 | 0.883 | 0.632 |
| | SpheerFM | Base | 46.8 | 35.1 | 111.4 | 0.704 | 0.634 | 0.981 | 0.939 | 0.701 |
| | | +GEDTM | 45.3 | 35.7 | 112.0 | 0.693 | 0.614 | 0.980 | 0.922 | 0.693 |

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

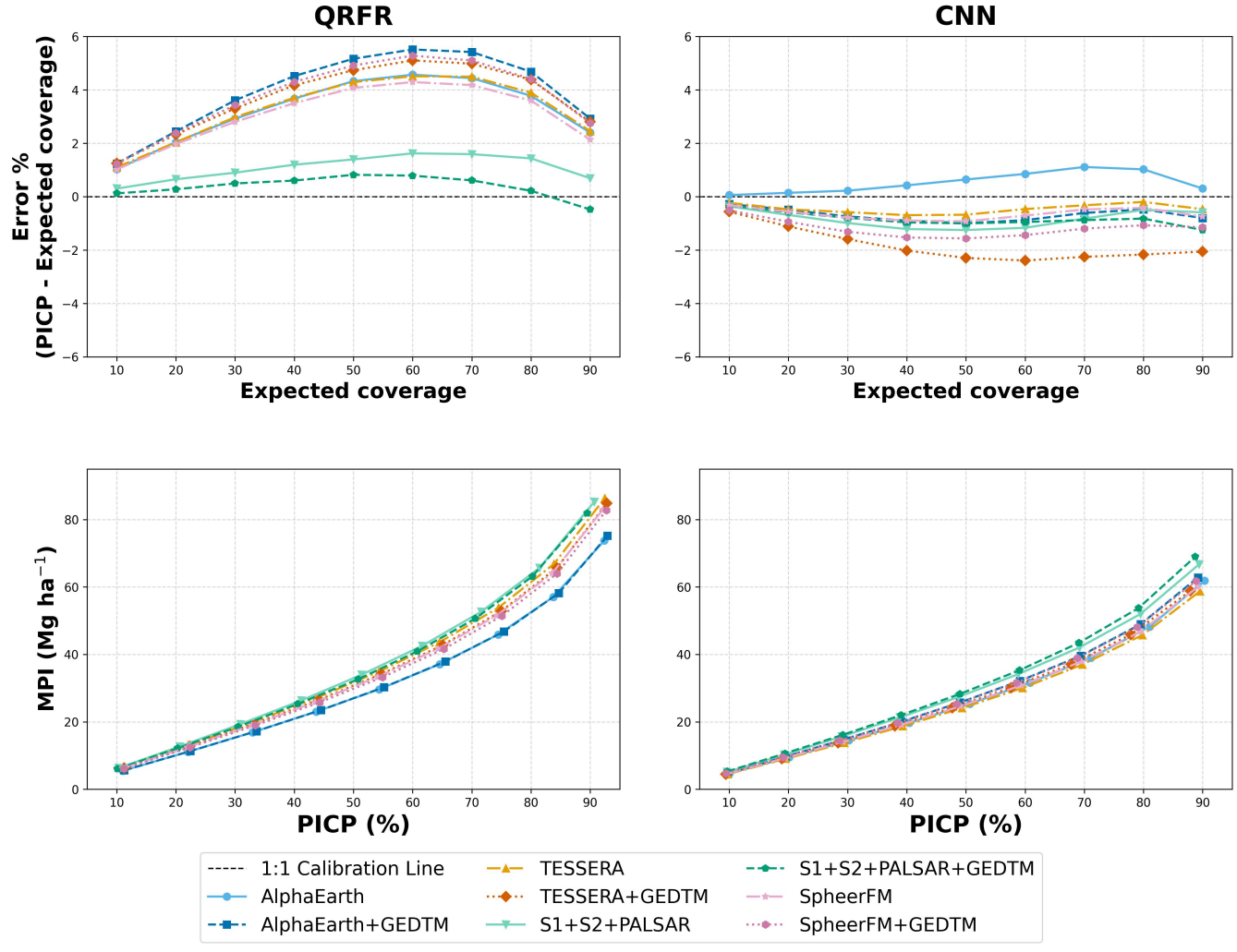



**Figure 7.** Uncertainty evaluation derived from per-pixel aggregations across the three test countries. The top row presents the PICP, and the bottom row displays the MPI for the CNN (left column) and QRFR (right column) models. QRFR: Quantile random forest regressor; CNN: Convolutional Neural Network.

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

the systemic bias that is entirely absent in the Spain and Finland results. Evaluating specific configurations, the AlphaEarth baseline yields the most balanced predictions, preserving the least bias in both over- and underestimation. While the integration of GEDTM has a largely neutral impact on the median bias across most configurations, it provides distinct structural benefits: it effectively reduces the underestimation of high-range biomass in Finland and narrows the overall residual variance across all three countries.

### 3.5 Ranking predictors and models

The summary results from all three countries, four-dimension metrics were compiled in a multi-criteria scoreboard (Figure 9). For each metric, the best performance earns 16 points whereas the worst one earns 1 point. The weighted and unweighted grand total is presented to indicate the overall winner (see Section 2.5 for weighting details). Based on this comprehensive scoring framework, the TESSERA configuration paired with the CNN model ranks first overall, whereas the TESSERA baseline utilizing the QRFR model ranks last. The aggregated scoreboard demonstrates that CNN models consistently outperform their QRFR counterparts. However, the inclusion of GEDTM marginally improves the overall fidelity of the AGBD mapping with QRFR but decreases in CNN model. Configurations utilizing pre-trained EO embeddings generally achieve higher rankings than those trained from multi-sensor satellite ARD, an advantage that is particularly pronounced when employing the CNN model. Remarkably, the only instance where a tree-based model surpassed a CNN configuration was AlphaEarth paired with QRFR, which outperformed the S1+S2+PALSAR – CNN setup.

## 4 Discussion

### 4.1 Summary findings

In this study, we evaluated the efficacy of three EO embeddings for large-scale AGBD mapping: AlphaEarth, TESSERA, and SpheerFM. Our methodology featured a comprehensive cross-evaluation spanning multiple predictor configurations (with and without addition of terrain features), target countries, and models. Overall, the results indicate that pairing TESSERA embeddings with a CNN model emerges as the highest-ranked configuration by a narrow margin. However, several nuanced observations worth further discussion.

Our comparative analysis yielded three primary observations regarding model-predictor configurations for large-scale AGBD mapping:

1. **Contribution of terrain variables**: While multiscale terrain variables capture complex hydrological and geomorphological relationships (Ho et al., 2025), they provided only marginal gains for the QRFR model and no benefit to the CNN model or AlphaEarth. We suspect that AlphaEarth inherently captures these features during pre-training (Brown et al., 2025), while the CNN model extracts sufficient spatial context directly from the SAR and optical layers.

2. **Response of pixel-wise and convolutional embeddings towards different models**: Switching the model from QRFR to CNN drastically improved the performance of pixel-wise self-superivised learning (SSL) models (TESSERA and

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

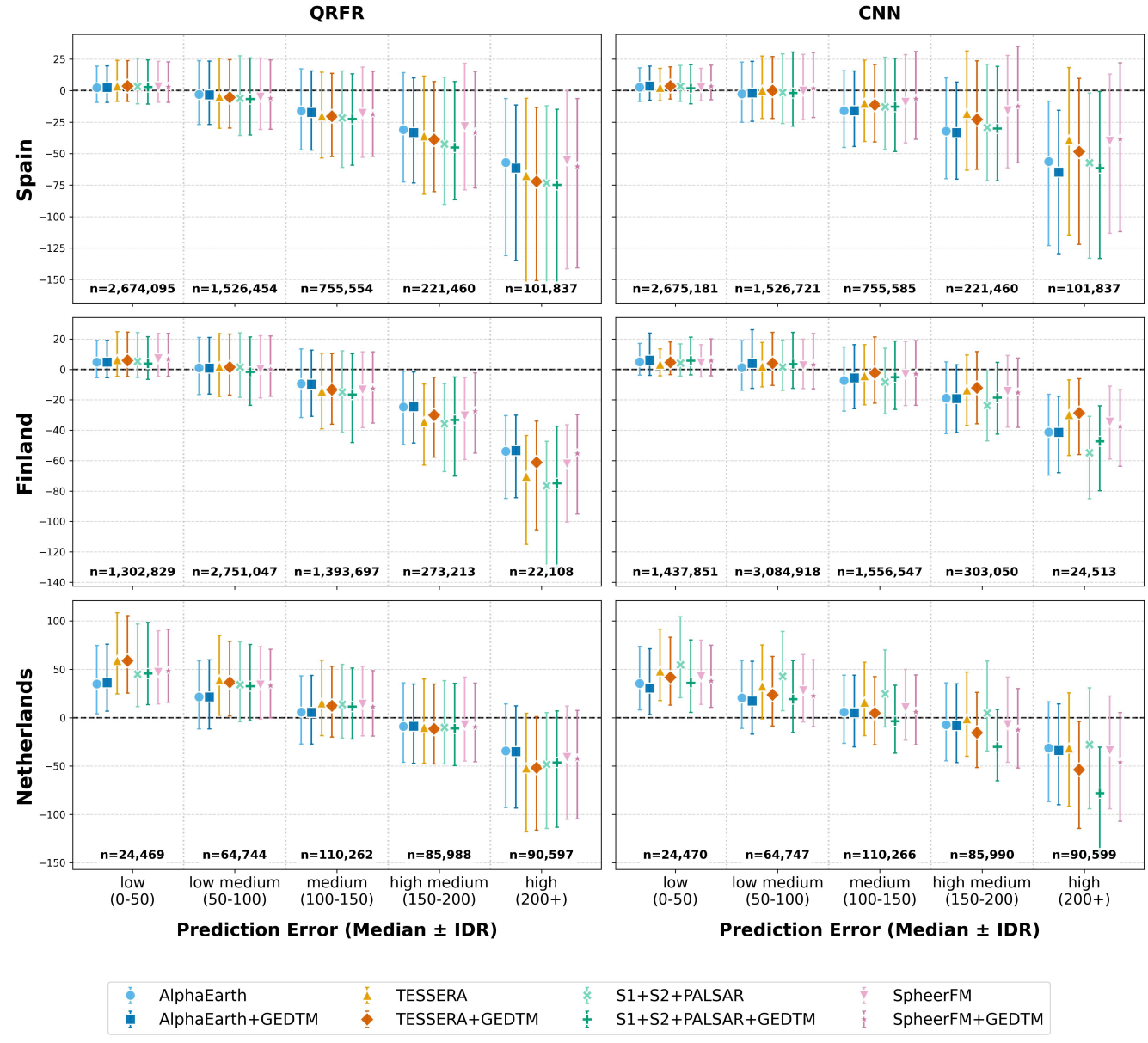



**Figure 8.** Per-pixel prediction error across discrete above-ground biomass ranges (0–50, 50–100, 100–150, 150–200, and >200 Mg ha$^{-1}$) for the evaluated predictor configurations using the QRFR and CNN models. Pixel data was aggregated from test tiles across Spain, Finland, and the Netherlands. The upper and lower bounds represent the 90th and 10th percentiles, respectively. QRFR: Quantile random forest regressor; CNN: Convolutional Neural Network.

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

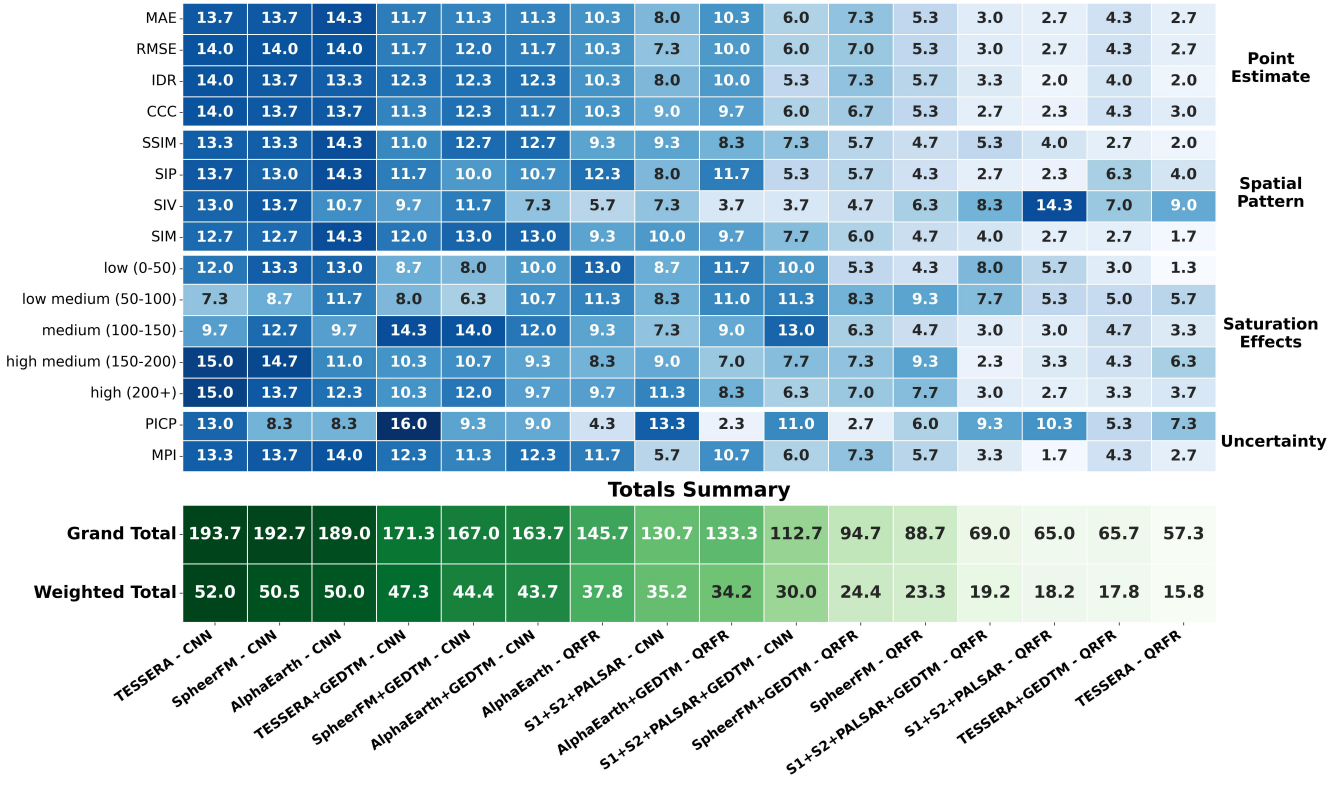



**Figure 9.** Multi-criteria scoreboard summarizing model performance across four evaluation dimensions (point-estimate accuracy, spatial pattern reconstruction, saturation effects, and uncertainty quantification). Columns display the 16 predictor–model configurations. For each metric, configurations are ranked from 1 (the worst) to 16 (the best). The final ranking is sorted by total weighted points. The darker color implies the better performance; the lighter color implies the worse performance.

SpheerFM) across all four evaluation dimensions. This demonstrates that pixel-wise EO embeddings can match or exceed the performance of convolution-based embeddings (AlphaEarth) when paired with a spatially aware downstream model.

3. **Impact of CNN model on spatial pattern reconstruction**: Comparing Figure 4 and Figure 5 reveals the true value of the CNN model. While it does not perform significantly better than the QRFR in terms of point-based accuracy, it deos dramatically improves spatial fidelity (median SSIM increased from 0.43–0.52 to 0.53–0.65 in Spain, 0.57–0.66 to 0.68–0.75 in Finland, and 0.47–0.63 to 0.59–0.66 in the Netherlands). We therefore recommend CNN models primarily for their significant enhancements to spatial pattern reconstruction rather than point-estimate accuracy.

Finally, we think that the variations in performance across Finland, Spain, and the Netherlands are due to differences in the characteristics of the reference-data and forest ecology. A major constraint of the Dutch evaluation is its limited reference dataset, comprising only approximately 150 training tiles after executing the sampling design. In contrast, Spain and Finland

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

offer substantially larger training pools with 1424 and 1598 tiles, respectively. This data sparsity poses a distinct challenge for model optimization in the Netherlands. Furthermore, the absolute magnitude of prediction errors reflects distinct biomass distributions driven by local ecology (Mo et al., 2024). Although Finland and Spain exhibit median RMSE values of approximately 20 Mg ha$^{-1}$, the median RMSE in the Netherlands exceeds 40 Mg ha$^{-1}$. The Dutch NFI plots represent dense, high-biomass temperate forests favored by a maritime climate, whereas biomass accumulation in Spain and Finland is constrained by seasonal water deficits (Mediterranean climate) and short growing seasons (boreal climate), respectively. Because the error scales with the total volume, RMSE disproportionately highlights performance gaps in high-density regions.

### 4.2 Design choice of EO embeddings

SpheerFM results in similar performance as TESSERA, where both are derived from per-pixel foundation models. However, TESSERA is trained on global data from Sentinel-1 and Sentinel-2 (approximately 800-million pixels), while SpheerFM is trained on only Dutch natural land from merely Sentinel-2 (approximately 200-million pixels). In addition, TESSERA is around 65 times bigger than SpheerFM during inference (46-million versus 0.7-million parameters). The bigger size of TESSERA's is only an advantage if the task at hand indeed requires a model of such complexity. Its input consists of around 500 data points, 40 time steps, 10 Sentinel-2 and 2 Sentinel-1 bands. All of these are highly correlated with each other, so the intrinsic dimensionality of the input is much lower than the input size would suggest. The strong performance of SpheerFM indicates that its smaller model size is sufficient for AGBD mapping. Version 2 of TESSERA compared model size with performance, supporting the increase in parameters creating a marginal effect as the model size grows (Feng et al., 2026).

Another explanation factor could be the focus of SpheerFM's training data and training objective. The objective of SpheerFM stimulates the extraction of features that are invariant to various noise which satellite capture incurs (clouds, irregular acquisitions, atmospheric effects, etc.). TESSERA's model is tasked (through Barlow Twins) with finding invariance between different views of the same year. These views were created by drawing two independent samples of 40 temporal satellite observations per year, using replacement when necessary (Feng et al., 2026). This alignment encourages similar embeddings for two distinct views of the same year. This representation learning scheme succeeds when the pretext task is sufficiently challenging, but not impossible. However, the two views likely share measurements. Reflecting the aforementioned choice of input, these measurements are highly correlated. This strong correlation could lead to ineffective training through an oversimplified learning objective, which prevents the model from extracting useful invariance. The underlying factors driving the performance similarity between large and small EO embeddings (TESSERA versus SpheerFM) require further investigation, as these frameworks differ substantially across multiple design choices, including model capacity, input modalities, and self-supervised training objectives.

**EO Embeddings: advantages and disadvantages**

Corley et al. (2026) established key guidelines for this emerging era of GFMs. A primary critique of current models is the lack of open-source weights and pretraining data, which are often withheld due to data privacy concerns. Under these constraints,

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

releasing the numerical embedding vectors serves as a crucial compromise, establishing a fixed, staged result and enabling the reproducibility and usability of the GFMs for downstream tasks.

A prominent advantage of EO embeddings is their SSL-enabled data compression. For instance, the TESSERA embedding compresses an annual time series of Sentinel-1 (2 bands) and Sentinel-2 (10 bands) data. Given a typical 5- to 6-day revisit time, a raw annual stack stores approximately 60 observations per year, requiring a single pixel to store over 600 distinct spectral and backscatter entries, let alone incorporating additional derived indices. In our evaluation, the EO embeddings successfully compressed this temporal density and ultimately outperformed the combined raw inputs of Sentinel-1, Sentinel-2, and PALSAR for aboveground biomass mapping. However, this performance comparison must be contextualized by the operational constraints of our large-scale mapping objective. The raw Sentinel-2 baseline data, sourced from CDSE) was limited to four bands (red, green, blue, and near-infrared) at a quarterly temporal resolution. The absence of the red-edge bands in this raw setup likely hindered its predictive accuracy, as previous literature discovered their critical role in AGB mapping (Moradi et al., 2022; Zhang et al., 2024). Consequently, expanding the availability of full-spectrum Sentinel-2 bands within CDSE temporal composites would significantly benefit future large-scale AGB mapping efforts.

GFMs are widely noted for their cost-efficiency, as they require significantly less labeled data to achieve high downstream task performance (Du et al., 2025; Zhu et al., 2026; Ball et al., 2026). This data efficiency presents a substantial advantage for large-scale mapping initiatives, where data scarcity is frequently a primary bottleneck. However, a key distinction and a major operational advantage of utilizing pre-computed EO embeddings rather than interacting directly with a GFM—is the ease of integrating auxiliary features. GFMs utilize SSL architectures, meaning downstream models are typically trained by replacing the pre-training output layer with task-specific heads. Consequently, incorporating a new feature (such as terrain variables) requires running full GFM inference to extract the latent embeddings before they can be concatenated with the additional covariates. In contrast, EO embeddings supply these latent representations directly as an machine-learning-ready product, entirely bypassing the GFM inference step and seamlessly streamlining the integration of additional features into the downstream model.

A notable disadvantage of EO embeddings is the black-box nature of their SSL architectures. Consequently, the resulting latent bands may lack the interpretable temporal granularity required for fine-scale, intra-annual change detection. For instances, although AlphaEarth is designed to encapsulate temporal dynamics, recent research by Brinkhoff and McGavin (2026) demonstrates that it underperforms compared to dedicated time-series models for crop event prediction. In constrast, these embeddings excel in overarching annual mapping tasks that rely on seasonal phenological signatures, frequently outperforming traditional methods in applications such as crop type classification (Guerrero-Viu et al., 2026) and tree species mapping (Ishikawa et al., 2026). Given these contrasting capabilities, further independent evaluations are required to determine whether EO embeddings can reliably replace established workflows in which specific operational contexts.

To our knowledge, this study represents the first attempt to systematically quantify predictive uncertainty using pre-trained EO embeddings. Our findings demonstrate that when evaluated under identical parameterizations, EO embeddings are fully compatible with established uncertainty quantification methods, supporting these calculations as robustly as models trained from scratch across both the QRFR and CNN models. This foundational capability opens new possibilities for operationalising

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

uncertainty quantification. Crucially, generating reliable uncertainty metrics at this scale supports the rigorous requirements of the CRCF regulation, suggesting that EO embedding frameworks could effectively serve as robust, supportive baseline maps for carbon accounting (European Parliament and Council of the European Union, 2024). However, further research is required to develop unbiased uncertainty aggregation for these fully machine learning based carbon baseline products (Wadoux and Heuvelink, 2023).

### 4.3 Visual comparison with existing AGBD maps

There are several existing EO-based biomass via various carbon-monitoring initiatives at pan-European scale. To contextualize our results, Figure 10 visually compares our best-performing TESSERA–CNN predictions against established continental- and global-scale AGBD products: ESACCI (Santoro and Cartus, 2026), CTrees (Yang et al., 2026), PlanetScope (Liu et al., 2023b), and ESA FCM (Santoro et al., 2026)). We evaluate these maps across three representative test tiles corresponding to distinct reference years: Soria (upper tile, 2019), Lapland (middle tile, 2021), and Helmond (bottom tile, 2017). Although the ESACCI, CTrees, and TESSERA–CNN predictions temporally align with these reference years, the ESA FCM data represent a 2021 baseline, and the PlanetScope maps utilize a public 30 m release (Liu et al., 2023a) with an unspecified observation year.

This visual assessment suggests that higher-resolution AGBD models more effectively resolve fine-scale spatial patterns and local biomass distributions. However, strict comparative conclusions are limited by the temporal mismatches across the products. Furthermore, existing maps are based on entirely different training, calibration, and reference datasets. A visual inspection alone cannot definitively establish relative predictive superiority. Instread, a rigorous, unbiased benchmarking of these products requires future validation against independent, third-party data collections, such as the Integrated Carbon Observation System (ICOS) sites (Franz et al., 2018) is needed. Furthermore, beyond single-year AGBD accuracy and prediction-interval calibration, operational carbon monitoring also requires temporally consistent estimates of biomass change, validation of trends and disturbances, and uncertainty propagation across years and spatial reporting units.

### 4.4 Large scale AGBD mapping with EO Embedding

Pre-processing raw satellite imagery across the entire European continent demands high-performance computational infrastructures. Although we acknowledge the availability of alternative GFMs such as Presto, DOFA, Prithvi and many more (Young et al., 2026), their standard SSL architectures present significant operational bottlenecks for deployment at this scale. Generating large-scale (e.g. pan-EU annual at 10 m) AGBD layers directly through these GFMs would require streaming billions of pixels through their inference pipelines to merely extract the latent embeddings. By utilizing pre-computed EO embeddings, our approach effectively bypasses this computationally massive inference step, offering a highly efficient and scalable framework for continental environmental mapping. Furthermore, selecting an operational framework for large-scale mapping must also weigh practical deployment factors: data licensing and transparency, data volume, and spatial temporal coverage.

Regarding licensing and transparency, both AlphaEarth and TESSERA provide their embeddings under permissive open licenses (CC-BY 4.0 and CC-0, respectively). However, AlphaEarth and SpheerFM retain closed-source model parameters and generation workflows. In contrast, TESSERA is a fully open-source project. This transparency gives TESSERA a clear

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

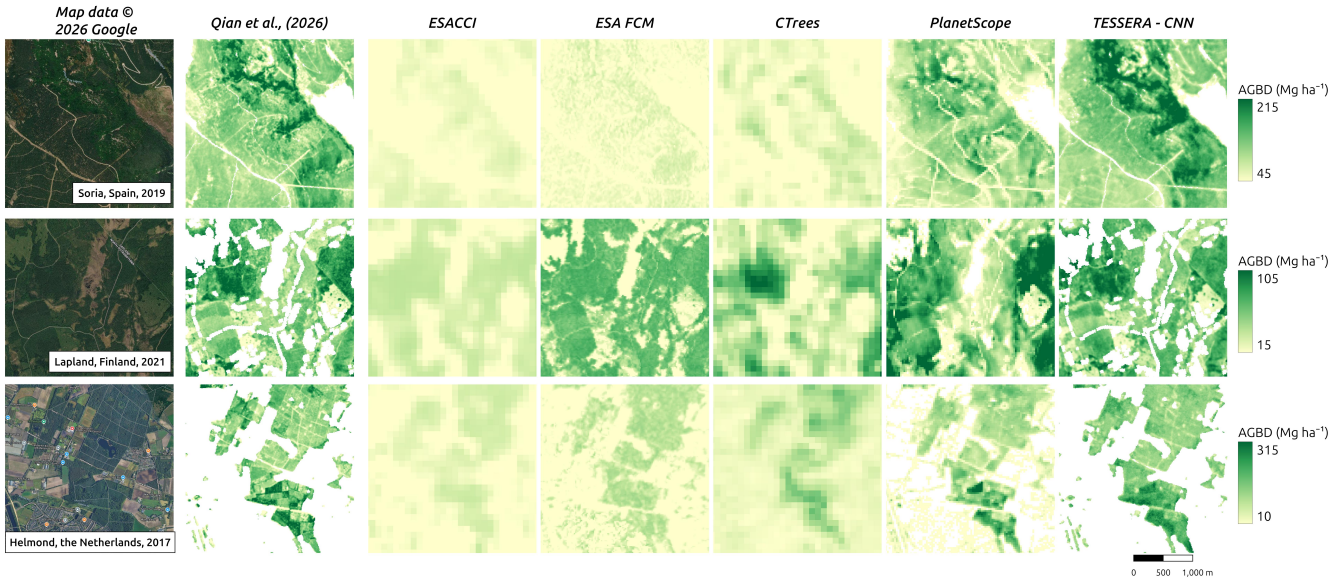



**Figure 10.** Visual comparison of AGBD predictions from ESACCI (100 m), ESA FCM (20 m), CTrees (100 m), PlantScope (30 m), and the best performing TESSERA–CNN model (10 m) alongside 10 m reference data. Example sites are drawn from Soria, Spain (2019), Lapland, Finland (2021), and Helmond, the Netherlands (2017). Please note that the ESA FCM map reflects 2017 data, and the PlanetScope map utilizes the public 30 m Zenodo version with an unknown observation year (Liu et al., 2023b). In addition, the products utilize different source of forest mask, high-resolution satellite imagery (left side) indicates the possible land cover. Map data: © 2026 Google.

strategic advantage for institutions that seek to independently deploy, compute, and maintain downstream AGBD mapping workflows on local infrastructure. From a data volume perspective, the winning TESSERA setup presents a notable computational burden. Because it utilizes 128 bands in a 32-bit floating-point format, it requires an 8-fold increase in the expected input data size compared to AlphaEarth's 64-band byte format. Fortunately, the recent release of TESSERA v2 mitigates this limitation by implementing Matryoshka representations, which provide highly efficient data compression while reportedly achieving competitive or superior performance relative to both TESSERA v1.1 and AlphaEarth (Feng et al., 2026). Finally, spatial and temporal availability remains a decisive operational factor. AlphaEarth, developed by Google DeepMind, currently possesses an unmatched operational scale, providing global time-series coverage from 2017 onward. By comparison, TESSERA v1.1 is currently restricted to European coverage and sparsely distributed global tiles, and SpheerFM embeddings are strictly limited to the Netherlands and sparse parcels around other European countries (Spheer, 2026).

## 5 Conclusions

In conclusion, we validate the capacity of Earth Observation (EO) foundation model embeddings to accurately map annual above-ground biomass density (AGBD) at scale through a rigorous, multi-dimensional assessment encompassing point-

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

estimate accuracy, spatial pattern reconstruction, saturation effect, and uncertainty quantification. By evaluating performance across three distinct European biogeographical regions: Mediterranean (Spain), Atlantic (the Netherlands), and Boreal (Finland)—our results demonstrate that EO embeddings perform comparably and even somewhat outperform the multi-sensor satellite ARD baseline (S1+S2+PALSAR) across all four dimensions in annual AGBD mapping. Furthermore, to our knowledge, this represents the first study to formally subject EO embeddings to uncertainty quantification, confirming their ability to generate robust and reliable prediction intervals within EO-based mapping workflows. The choice of model also proved critical; CNN substantially improved spatial pattern reconstruction compared to traditional tree-based RF model. Synthesizing these metrics across all three countries into a multi-criteria scoreboard, the TESSERA–CNN configuration emerged as the best ranked performing framework. However, fulfilling the ultimate objective of operational carbon monitoring requires more than high-resolution, large-scale mapping. Future research should prioritize temporally consistent estimations, the validation of long-term trends and forest disturbances, and rigorous propagation of uncertainty across both space and time.

*Code and data availability.* Source code implementation of the benchmark workflow are publicly available at https://codeberg.org/OGCR-project/agb-embeddings-benchmark, and is versioned via Zenodo: Ho, Y.-F. (2026). Workflow for Benchmarking EO Embeddings for Continental-scale Aboveground Biomass Mapping (Version 1.0.0) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.23537159. The detail of reference AGB data in Finland, Netherlands and Spain is avaiable via Zenodo: Song, Q., Enguehard, L., Kuschnerus, M., & Herold, M. (2026). Harmonized Airborne Laser Scanning-derived forest above ground biomass maps in selected regions of Finland, Spain, and the Netherlands (Version V 1.0) [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.23591088

*Author contributions.* YH: conceptualization, data curation, formal analysis, methodology, validation, visualization, software, writing (original draft preparation). MN: conceptualization, methodology, supervision. LE: data curation, conceptualization, writing (original draft preparation). QS: data curation, conceptualization, writing (original draft preparation). JL: data curation, conceptualization, writing (original draft preparation). MSI: data curation, writing (original draft preparation). CB: conceptualization, methodology, writing (original draft preparation). LP: conceptualization, methodology, supervision. DK: conceptualization, methodology, supervision. MH: conceptualization, methodology. IW: funding acquisition. TH: conceptualization, methodology, supervision, writing (original draft preparation).

*Competing interests.* The contact author has declared that none of the authors has any competing interests.

*Acknowledgements.* This research is supported by multiple grants. We would like to acknowledge The Intergenerational Open Geospatial Carbon Registry (OGCR), received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101218854; the project *FORWARDS* for sharing data: FORWARDS. H2020 project funded by the European Commission,

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.



number 101084481 call HORIZON-CL6-2022-CLIMATE-01-05. The support from ForestNavigator (grant agreement No. 101056875), and HGF HFMI 3D-ABC projects

We also knowledge the AlphaEarth Foundations Satellite Embedding dataset, produced by Google and Google DeepMind and TESSERA, produced by the University of Cambridge. For the support of this manuscript, we are grateful to Nico Lang for his insights into the ecological implications of above-ground biomass mapping at 10 m resolution; Mikhail Urbazaev for valuable discussions regarding structural similarity indices; Krzysztof Dyba for clarifying the conceptual distinctions between prediction and confidence intervals in machine learning uncertainty quantification; and Seyyed Ali Ahmadi and Sajed Sarabandi for fruitful brainstorming and assistance with debugging the deep learning models.

During the preparation of this manuscript/study, the author(s) used artificial intelligence chatbots to help improve the clarity and readability of this manuscript. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

## References

Altarez, R. D. D., Apan, A., and Maraseni, T.: Uncovering the Hidden Carbon Treasures of the Philippines' Towering Mountains: A Synergistic Exploration Using Satellite Imagery and Machine Learning, PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science, 92, 55–73, https://doi.org/10.1007/s41064-023-00270-3, 2024.

Ball, J. G. C., Wicklein, J. A., Feng, Z., Knezevic, J., Jaffer, S., Madhavapeddy, A., Atzberger, C., Dalponte, M., and Coomes, D. A.: Geospatial foundation models enable data-efficient tree species mapping in temperate mountain forests, Science of Remote Sensing, 13, 100 466, https://doi.org/10.1016/j.srs.2026.100466, 2026.

Bendersky, E.: Depthwise separable convolutions for machine learning, https://eli.thegreenplace.net/2018/depthwise-separable-convolutions-for-machine-learning/, 2018.

Brinkhoff, J. and McGavin, S. L.: Comparing AlphaEarth Foundations embeddings and satellite time-series for rice crop event prediction, SSRN Electronic Journal, https://doi.org/10.2139/ssrn.7228924, 2026.

Brown, C. F., Kazmierski, M. R., Pasquarella, V. J., Rucklidge, W. J., Samsikova, M., Zhang, C., Shelhamer, E., Lahera, E., Wiles, O., Ilyushchenko, S., et al.: AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data, arXiv preprint arXiv:2507.22291, https://doi.org/10.48550/arXiv.2507.22291, 2025.

Cambridge Centre for Earth Observation (ucam-eo): TESSERA: Temporal Embeddings of Surface Spectra for Earth Representation and Analysis, https://github.com/ucam-eo/tessera, gitHub repository, accessed October 16, 2026, 2026.

Carlson, C. A., Bates, N. R., Hansell, D. A., and Steinberg, D. K.: Carbon cycle, in: Encyclopedia of Ocean Sciences, pp. 477–486, Elsevier, 2nd edn., https://doi.org/10.1006/rwos.2001.0202, 2001.

Chiti, T., Rey, A., Abildtrup, J., Böttcher, H., Diaci, J., Frings, O., Lehtonen, A., Schindlbacher, A., and Zavala, M. A.: Carbon farming in the European forestry sector, Tech. Rep. 15, European Forest Institute, https://doi.org/10.36333/fs15, 2024.

Chollet, F.: Xception: Deep Learning with Depthwise Separable Convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1251–1258, https://doi.org/10.1109/CVPR.2017.195, 2017.

Copernicus Data Space Ecosystem: Sentinel-1 Level 3 Monthly Mosaics, https://documentation.dataspace.copernicus.eu/Data/Sentinel1.html, accessed: 2026-09-30, 2026.

Corley, I., Lehmann, N., Robinson, C., Tseng, G., Fuller, A., Alemohammad, H., Shelhamer, E., Marcus, J., and Kerner, H.: No One Knows the State of the Art in Geospatial Foundation Models, arXiv preprint arXiv:2605.12678, https://doi.org/10.48550/arXiv.2605.12678, 2026.

Du, A., Del Prete, R., Mousist, A., Manser, N., Marre, F., Barton, A., Seubert, C., Meoni, G., and Chin, T.-J.: First On-Orbit Demonstration of a Geospatial Foundation Model, arXiv preprint arXiv:2512.01181, https://doi.org/10.48550/arXiv.2512.01181, 2025.

European Environment Agency: Europe's biodiversity—biogeographical regions and seas, Tech. Rep. EEA Report 1/2002, European Environment Agency (EEA), https://www.eea.europa.eu/en/analysis/publications/report_2002_0524_154909, 2002.

European Parliament and Council of the European Union: Regulation (EU) 2024/3012 of the European Parliament and of the Council of 27 November 2024 establishing a Union certification framework for permanent carbon removals, carbon farming and carbon storage in products, Official Journal of the European Union, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202403012, 2024.

European Union: Sentinel-2 Level 3 Quarterly Cloudless Mosaic (Q2 2024), Copernicus Data Space Ecosystem, https://dataspace.copernicus.eu, product type: `S2MSI_L3__MCQ`. Accessed: 2026-08-27, 2024.

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

Fang, C., Buhendwa, B. M., Riaz, F., and Wang, J.: Performance Assessment and Application of AlphaEarth Embedding In Air Quality—Case of PM2.5 in Changchun, Remote Sensing Applications: Society and Environment, 34, 101 989, https://doi.org/10.1016/j.rsase.2026.101989, 2026.

Feng, Z., Atzberger, C., Jaffer, S., Knezevic, J., Sormunen, S., Young, R., Lisaius, M. C., Immitzer, M., Jackson, T., Ball, J., et al.: Tessera: Temporal embeddings of surface spectra for earth representation and analysis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 34 818–34 831, https://openaccess.thecvf.com/content/CVPR2026/papers/Feng_TESSERA_Temporal_Embeddings_of_Surface_Spectra_for_Earth_Representation_and_CVPR_2026_paper.pdf, 2026.

Franz, D., Acosta, M., Altimir, N., Arriga, N., Arrouays, D., Aubinet, M., Aurela, M., Ayres, E., López-Ballesteros, A., Barbaste, M., et al.: Towards long-term standardised carbon and greenhouse gas observations for monitoring Europe's terrestrial ecosystems: a review, International Agrophysics, 32, 439–455, https://doi.org/10.1515/intag-2017-0039, 2018.

Gazzea, M., Solheim, A., and Arghandeh, R.: High-resolution mapping of forest structure from integrated SAR and optical images using an enhanced U-net method, Science of Remote Sensing, 8, 100 093, https://doi.org/10.1016/j.srs.2023.100093, 2023.

Guerrero-Viu, J., López-Cifuentes, A., Pérez-Villar, I., and Pacifici, F.: Temporal Sensitivity Analysis of Tessera Embeddings, arXiv preprint arXiv:2608.27175, https://doi.org/10.48550/arXiv.2608.27175, 2026.

Hancock, S., McGrath, C., Lowe, C., Davenport, I., and Woodhouse, I.: Requirements for a global lidar system: spaceborne lidar with wall-to-wall coverage, Royal Society Open Science, 8, 211 166, https://doi.org/10.1098/rsos.211166, 2021.

Hengl, T., Consoli, D., Tian, X., Nauman, T. W., Nussbaum, M., Isik, M. S., Parente, L., Ho, Y.-F., Simoes, R., Gupta, S., et al.: OpenLandMap-soildb: global soil information at 30 m Spatial resolution for 2000–2022+ based on Spatiotemporal machine learning and harmonized legacy soil samples and observations, Earth System Science Data, 18, 989–1036, https://doi.org/10.5194/essd-18-989-2026, 2026.

Ho, Y.-F. and Hengl, T.: Global Ensemble Digital Terrain Model 30m (GEDTM30), Zenodo, https://doi.org/10.5281/zenodo.18887460, version v1.2.0, 2026.

Ho, Y.-F., Grohmann, C. H., Lindsay, J., Reuter, H. I., Parente, L., Witjes, M., and Hengl, T.: GEDTM30: Global ensemble digital terrain model at 30 m and derived multiscale terrain variables, PeerJ, 13, e19 673, https://doi.org/10.7717/peerj.19673, 2025.

Ishikawa, T., Bonannella, C., Lerink, B. J. W., and Rußwurm, M.: Assessing the effectiveness of deep embeddings for tree species classification in the Dutch Forest Inventory, Science of Remote Sensing, 13, 100 498, https://doi.org/10.1016/j.srs.2026.100498, 2026.

Japan Aerospace Exploration Agency: Global 25 m Resolution PALSAR/PALSAR-2 Forest/Non-Forest Map (FNF), Version 2.6.0, https://www.eorc.jaxa.jp/ALOS/en/dataset/fnf_e.htm, accessed: 2026-10-03, 2024.

Jensen, L.: Certifying EU permanent carbon removals: State of play in implementing the EU's Carbon Removal and Carbon Farming Regulation, Briefing PE 785.709, European Parliamentary Research Service (EPRS), European Parliament, https://data.europa.eu/doi/10.2861/230182, accessed: 2026-10-26, 2026.

Jones, E. L., Rendell, L., Pirotta, E., and Long, J. A.: Novel application of a quantitative spatial comparison tool to species distribution data, Ecological Indicators, 70, 67–76, https://doi.org/10.1016/j.ecolind.2016.05.051, 2016.

Kasraei, B., Heung, B., Saurette, D. D., Schmidt, M. G., Bulmer, C. E., and Bethel, W.: Quantile regression as a generic approach for estimating uncertainty of digital soil maps produced from machine-learning, Environmental Modelling & Software, 144, 105 139, https://doi.org/10.1016/j.envsoft.2021.105139, 2021.

Kumari, P., Lavania, P., Dobriyal, M., Srivastav, M., and Kumar, P.: Artificial intelligence and forest management, in: Advanced Geospatial and Ground Based Techniques in Forest Monitoring, pp. 255–272, Elsevier, https://doi.org/10.1016/B978-0-443-18949-4.00002-6, 2026.

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.

Lang, N., Schindler, K., and Wegner, J. D.: Country-wide high-resolution vegetation height mapping with Sentinel-2, Remote Sensing of Environment, 233, 111 347, https://doi.org/10.1016/j.rse.2019.111347, 2019.

Lang, N., Jetz, W., Schindler, K., and Wegner, J. D.: A high-resolution canopy height model of the Earth, Nature Ecology & Evolution, 7, 1778–1789, https://doi.org/10.1038/s41559-023-02206-6, 2023.

Langanke, T., Ramminger, G., Buzzo, G., and Berndt, F.: Copernicus Land Monitoring Service—High Resolution Layer Forest: Product Specifications Document - V11, https://land.copernicus.eu/user-corner/technical-library, accessed: 2023-12-31, 2017.

Lemke, D., Dimov, L., Czech, H., Knight, P., Finch, W., and Condit, R.: Relationship between topographic variables and live aboveground tree biomass on a large temperate forest plot, Forest Ecosystems, 14, 100 338, https://doi.org/10.1016/j.fecs.2025.100338, 2025.

Liu, K., Wang, J., Zeng, W., and Song, J.: Comparison and evaluation of three methods for estimating forest above ground biomass using TM and GLAS data, Remote Sensing, 9, 341, https://doi.org/10.3390/rs9040341, 2017.

Liu, S., Brandt, M., Nord-Larsen, T., Chave, J., Reiner, F., Lang, N., Tong, X., Ciais, P., Igel, C., Li, S., Mugabowindekwe, M., Pascual, A., Guerra-Hernandez, J., Saatchi, S., Yue, Y., Chen, Z., and Fensholt, R.: Canopy height and biomass map for Europe, https://doi.org/10.5281/zenodo.8154445, 2023a.

Liu, S., Brandt, M., Nord-Larsen, T., Chave, J., Reiner, F., Lang, N., Tong, X., Ciais, P., Igel, C., Pascual, A., et al.: The overlooked contribution of trees outside forests to tree cover and woody biomass across Europe, Science Advances, 9, eadh4097, https://doi.org/10.1126/sciadv.adh4097, 2023b.

Ma, T., Zhang, C., Ji, L., Zuo, Z., Beckline, M., Hu, Y., Li, X., and Xiao, X.: Development of forest aboveground biomass estimation, its problems and future solutions: A review, Ecological Indicators, 159, 111 653, https://doi.org/10.1016/j.ecolind.2024.111653, 2024.

Malone, B. P., McBratney, A. B., and Minasny, B.: Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes, Geoderma, 160, 614–626, https://doi.org/10.1016/j.geoderma.2010.11.013, 2011.

Mandal, D., Kumar, V., Ratha, D., Dey, S., Bhattacharya, A., Lopez-Sanchez, J. M., McNairn, H., and Rao, Y. S.: Dual polarimetric radar vegetation index for crop growth monitoring using sentinel-1 SAR data, Remote Sensing of Environment, 247, 111 954, https://doi.org/https://doi.org/10.1016/j.rse.2020.111954, 2020.

Meinshausen, N. and Ridgeway, G.: Quantile regression forests, Journal of Machine Learning Research, 7, 983–1006, https://www.jmlr.org/papers/volume7/meinshausen06a/meinshausen06a.pdf, 2006.

Mo, L., Crowther, T. W., Maynard, D. S., Van den Hoogen, J., Ma, H., Bialic-Murphy, L., Liang, J., De-Miguel, S., Nabuurs, G.-J., Reich, P. B., et al.: The global distribution and drivers of wood density and their impact on forest carbon stocks, Nature Ecology & Evolution, 8, 2195–2212, https://doi.org/10.1038/s41559-024-02564-9, 2024.

Moradi, F., Darvishsefat, A. A., Pourrahmati, M. R., Deljouei, A., and Borz, S. A.: Estimating aboveground biomass in dense Hyrcanian forests by the use of Sentinel-2 data, Forests, 13, 104, https://doi.org/10.3390/f13010104, 2022.

Mutanga, O., Masenyama, A., and Sibanda, M.: Spectral saturation in the remote sensing of high-density vegetation traits: A systematic review of progress, challenges, and prospects, ISPRS Journal of Photogrammetry and Remote Sensing, 198, 297–309, https://doi.org/10.1016/j.isprsjprs.2023.03.010, 2023.

Nix, D. A. and Weigend, A. S.: Estimating the mean and variance of the target probability distribution, in: Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94), vol. 1, pp. 55–60, IEEE, https://doi.org/10.1109/ICNN.1994.374138, 1994.

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaepman, M. E., and Papritz, A.: Evaluation of digital soil mapping approaches with large sets of environmental covariates, Soil, 4, 1–22, https://doi.org/10.5194/soil-4-1-2018, 2018.

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.



Parente, L., Ehrmann, S., Hengl, T., Fritz, S., Bonannella, C., Malek, Ž., Fischer, C. G., Perez, K., Stanimirova, R., Meyer, C., et al.: Global distribution of cattle, horses, goats, sheep and buffaloes at 1 km resolution for 2000–2022 based on subnational census data and spatiotemporal machine learning, PeerJ, 14, e21 494, https://doi.org/10.7717/peerj.21494, 2026.

Post, W. M., Peng, T.-H., Emanuel, W. R., King, A. W., Dale, V. H., and DeAngelis, D. L.: The global carbon cycle, American Scientist, 78, 310–326, https://www.jstor.org/stable/29774118, 1990.

Salinas-Melgoza, M. A., Skutsch, M., and Lovett, J. C.: Predicting aboveground forest biomass with topographic variables in human-impacted tropical dry forest landscapes, Ecosphere, 9, e02 063, https://doi.org/10.1002/ecs2.2063, 2018.

Santoro, M. and Cartus, O.: ESA Biomass Climate Change Initiative (Biomass_cci): Global datasets of forest above-ground biomass for the years 2005–2012 and 2015–2024, v7.0, https://doi.org/10.5285/6429D1AAFE1E43B9B414E4A5A7F8B903, 2026.

Santoro, M., Cartus, O., Araza, A., Herold, M., Miettinen, J., Rosenqvist, A., Kobayashi, K., Tadono, T., and Seifert, F. M.: Europe-wide maps of biomass density based on satellite remote sensing data for 2017, 2020, 2021 and 2023, Data in Brief, 65, 112 536, https://doi.org/10.1016/j.dib.2026.112536, 2026.

Sedjo, R. A.: The carbon cycle and global forest ecosystem, Water, Air, and Soil Pollution, 70, 295–307, https://doi.org/10.1007/BF01105003, 1993.

Shimada, M., Itoh, T., Motooka, T., Watanabe, M., Shiraishi, T., Thapa, R., and Lucas, R.: New global forest/non-forest maps from ALOS PALSAR data (2007–2010), Remote Sensing of Environment, 155, 13–31, https://doi.org/10.1016/j.rse.2014.04.014, 2014.

Sialelli, G., Peters, T., Wegner, J. D., and Schindler, K.: AGBD: A Global-scale Biomass Dataset, ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, X-G-2025, 829–838, https://doi.org/10.5194/isprs-annals-X-G-2025-829-2025, 2025.

Silva, L. and Gonçalves, J.: Evaluating AlphaEarth and Tessera geospatial embeddings for machine learning-based burned area mapping in Portugal, Earth Resources and Environment, in review, 2026.

Small, D.: SAR backscatter multitemporal compositing via local resolution weighting, in: 2012 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 4521–4524, https://doi.org/10.1109/IGARSS.2012.6350465, 2012.

Song, Q., Enguehard, L., Kuschnerus, M., and Herold, M.: Harmonized Airborne Laser Scanning-derived forest above ground biomass maps in selected regions of Finland, Spain, and the Netherlands, https://doi.org/10.5281/zenodo.23591088, version V 1.0, 2026.

Spheer: spheer-fm-embeddings dataset, https://huggingface.co/datasets/spheer/spheer-fm-embeddings, accessed: 2026-10-19, 2026.

Steichen, T. J. and Cox, N. J.: A note on the concordance correlation coefficient, The Stata Journal, 2, 183–189, https://doi.org/10.1177/1536867X0200200206, 2002.

Stewart, A. J., Fang, H., Corley, I. A., and Zhu, X. X.: Earth Embeddings, arXiv preprint arXiv:2608.03410, https://doi.org/10.48550/arXiv.2608.03410, 2026.

Tanase, M. A., Martini, J. P., García García, D., Zavala, M. Á., and Ruiz-Benito, P.: Forest structure and its changes from multi-temporal lidar data: a homogeneously derived database for peninsular Spain, Annals of Forest Research, 69, 15–34, https://doi.org/10.15287/afr.2026.4341, 2026.

Taylor Geospatial: AlphaEarth Foundations Satellite Embedding Dataset, https://source.coop/tge-labs/aef/v1/annual, accessed via Source Cooperative (Updated Mar 22, 2026), 2025.

Tian, X., de Bruin, S., Simoes, R., Isik, M. S., Minarik, R., Ho, Y.-F., Şahin, M., Herold, M., Consoli, D., and Hengl, T.: Spatiotemporal prediction of soil organic carbon density in Europe (2000–2022) using earth observation and machine learning, PeerJ, 13, e19 605, https://doi.org/10.7717/peerj.19605, 2025.

https://doi.org/10.5194/egusphere-2026-6643
Preprint. Discussion started: 17 November 2026
© Author(s) 2026. CC BY 4.0 License.



Tomppo, E., Gschwantner, T., Lawrence, M., and McRoberts, R. E.: National Forest Inventories: Pathways for Common Reporting, Springer, Dordrecht, https://doi.org/10.1007/978-90-481-3233-1, 2010.

Wadoux, A. M.-C. and Heuvelink, G. B.: Uncertainty of spatial averages and totals of natural resource maps, Methods in Ecology and Evolution, 14, 1320–1332, https://doi.org/https://doi.org/10.1111/2041-210X.14106, 2023.

Yang, L., Chen, N., Yue, J., Liu, Y., Ma, J., Ghamisi, P., Plaza, A., and Fang, L.: Survey of multimodal geospatial foundation models: Techniques, applications, and challenges, arXiv preprint arXiv:2510.22964, https://doi.org/10.48550/arXiv.2510.22964, 2025.

Yang, Y., Dalagnol, R., Lin, W., Kwon, N., Liu, Z., Saatchi, S., and Sagang, L. B.: From Snapshot Maps to Continuous Monitoring of Global Forest Carbon at 100 m Resolution (2000–2025), EarthArXiv, https://doi.org/10.31223/X5KJ4Q, 2026.

Young, R., Gabtraupov, A., Soror, K., and Keshav, S.: How Usable Are Geospatial Foundation Models? A Systematic Evaluation of 89 Models, arXiv preprint arXiv:2608.03804, https://doi.org/10.48550/arXiv.2608.03804, 2026.

Zhang, L., Yin, X., Wang, Y., and Chen, J.: Aboveground biomass mapping in semiarid forests by integrating airborne LiDAR with Sentinel-1 and Sentinel-2 time-series data, Remote Sensing, 16, 3241, https://doi.org/10.3390/rs16173241, 2024.

Zhu, X. X., Xiong, Z., Wang, Y., Stewart, A. J., Heidler, K., Wang, Y., Yuan, Z., Dujardin, T., Xu, Q., and Shi, Y.: On the foundations of Earth foundation models, Communications Earth & Environment, 7, 103, https://doi.org/10.1038/s43247-025-03127-x, 2026.

Zillow Group: quantile-forest: Quantile Regression Forests in Python, https://zillow.github.io/quantile-forest/, scikit-learn compatible implementation of Quantile Regression Forests, 2022.