Combining specialized Sentinel-2 time series features with AlphaEarth Foundations for forest type mapping

Benedikt Hiebl[1] *, Nicola Alessi[2] , Giacomo Calvia[3] , Alessandro Bricca[3] , Gianmaria Bonari[4] , Giulio Zangari[3] , Stefan Zerbe[5,6], Martin Rutzinger[1]

[1] Department of Geography, University of Innsbruck, Innsbruck, Austria - Benedikt.Hiebl@uibk.ac.at
[2] Italian Institute for Environmental Protection and Research, Rome, Italy - nicola.alessi@isprambiente.it
[3] Faculty of Agricultural, Environmental and Food Sciences, Free University of Bozen-Bolzano, Italy - alessandro.bricca@unibz.it
[4] Department of Life Sciences, University of Siena, Italy - gianmaria.bonari@unisi.it
[5] University of Applied Sciences and Arts (HAWK), Göttingen, Germany - stefan.zerbe@hawk.de
[6] University of Hildesheim, Hildesheim, Germany - stefan.zerbe@uni-hildesheim.de

Keywords: satellite remote sensing, forest mapping, alpha earth foundations, sentinel-2.

Abstract

Accurate mapping of forest types and vegetation characteristics is essential for monitoring biodiversity and forest dynamics. Traditional Deep Learning (DL) models trained on Sentinel-2 time series achieve high performance, but require extensive preprocessing and sensor-related fine-tuning. In this study, we evaluate the recently introduced AlphaEarth Foundations (AEF) embeddings, which is a global, multi-modal feature representation of the earths surface, for forest mapping in Italy. We compare a) a Random Forest model trained on Sentinel-2 and climate time series features, b) a Multi-Layer Perceptron trained on AEF, c) a Time-Series Transformer trained on Sentinel-2 and climate annual time series, and d) a Cross-Attention fusion model combining both feature sets. Using 5-fold cross-validation in a regression and a classification task on two datasets (evergreen broad-leaved tree cover ETC, forest vegetation type FVT) we find that the combined model consistently outperforms the single-source approaches (RMSE = 0.161, Acc = 0.757). AEF-based models achieve comparable accuracy to the Sentinel-2-based models, while reducing extensive time series preprocessing and training time by an order of magnitude. Feature attribution using integrated gradients reveals that AEF provides stable baseline representations, while Sentinel-2 inputs add phenology-related detail. The results show, that integrating generalized embeddings with specialized spectral-temporal features improves predictive performance for forest mapping.

1. Introduction

Large scale mapping of vegetation characteristics on the landscape scale, such as biodiversity variables (Skidmore et al., 2021), is essential for biodiversity conservation, change detection, hazard monitoring and land use (Torres et al., 2021; Pettorelli et al., 2018). This requires on the one hand handling large amounts of Remote Sensing (RS) data from diverse sensors and platforms, and on the other hand collecting and processing vegetation plot observations as ground truth.

Especially in forest mapping, models integrating phenology dynamics proved to be more efficient than non-time-aware models, as species differentiation from spectral data requires a temporal component (Kollert et al., 2021; Chabalala et al., 2023; Grabska-Szwagrzyk et al., 2024). However, building specialized RS time series (TS) models for specific classification or regression tasks is complex and is limited by the sensor and mission characteristics. Sensor fusion approaches, e.g. for multi-spectral and Synthetic Aperture Radar (SAR) sensors, are used to combine the strengths of different sensors, but require heavy preprocessing of RS data (Chen et al., 2024; Vogeler et al., 2023). This limits not just the training side, but also the applicability for inference on large-scale mapping. Yet, this is still the state-of-the-art approach for operational applications as all components of the workflow are controlled by the user and performance is especially high for regional studies, when ground truth data is available locally.

With the rise of Deep Learning (DL) methods in differing research fields, pretraining and foundation models have become an effective approach for modeling tasks with limited availability of labeled data (Ge et al., 2023; Safonova et al., 2023). Especially in Earth Observation (EO) applications, large amounts of RS data stand against sparse ground truth plot data for training and validation, which often leads to overfitting issues during the training process. High quality ground truth data collection often requires cost and time intensive field work and cannot cover large spatial or temporal extents Kattenborn et al. (2021). Foundation models help by leveraging the power of self-supervised pretraining on large amounts of RS data for better generalization capabilities of models in large-scale mapping tasks (Yuan et al., 2022). These foundation models, often trained on single sensor data, and other contextually pretrained DL models are adapted to specific applications by fine-tuning them on specific datasets, e.g. for forest mapping or crop monitoring. This improves spatio-temporal generalizability and robustness of model predictions (Tseng et al., 2024; Ma et al., 2024; Hiebl et al., 2025). However, these models come with major limitations. a) They are trained on single RS sensor data such as Sentinel-2 without temporal context and therefor only accept similar feature input variables, and b) they still have to be fine-tuned, which requires DL expertise and GPU-accelerated hardware for training DL models. Recently Google DeepMind published their Alpha Earth Foundations (AEF) dataset, which is an annual, multi-modal EO feature embedding based on different sensors and temporal resolutions (Brown et al., 2025). These embeddings are a representation of the earths surface, that is specifically developed for environmental mapping. This promises new possibilities for global mapping of environmental variables as heavy preprocessing of multi-sensor time series and

* Corresponding author

fine-tuning large DL models becomes obsolete. Conceptually, this brings the power of DL to a wider range of users and applications. All three approaches a) specialized time series extrinsic regression and classification models, b) pretrained foundation models, and c) generalized feature embeddings such as AEF, have found their way into environmental monitoring and forest mapping (Sun et al., 2022). The first two already being well tested approaches for forest mapping, while the potential of AEF has still to be discovered.

In this study we evaluate the potential of generalized AEF embeddings for forest type and cover mapping against a specialized RS TS model. Furthermore, we will investigate the potential of a combined generalized plus specialized embeddings approach. The main research objectives in this study are a) evaluating the potential of AEF for plot-based forest mapping against a specialized model, and b) developing a model to combine the strengths of both approaches. We are using two different forest vegetation datasets as reference in classification (Forest Vegetation Type FVT) and regression tasks (Evergreen-broadleaved Tree Cover ETC). A Random Forest model (RF) and a Time-Series-Transformer (TST) are used with Sentinel-2 L2A (S2) annual time series and CHELSA climate data (Karger et al., 2020) to build specialized prediction models. We compare these to a standard Multi-Layer Perceptron model (MLP) with two layers trained solely on AEF. Additionally an approach for combining AEF with RS time series is investigated. To evaluate model performances a 5-fold cross-validation scheme is built on both datasets.

2. Methods

2.1 Data

2.1.1 Forest Vegetation Data The general study area are forested regions in Italy, where we utilize two different datasets for the classification/regression tasks (Fig. 1). For classification (FVT) an Italian-wide Forest Vegetation Database (VDB) is used (Alessi et al., 2019; Bonari et al., 2019). The dataset contains plot observations from 1995 to 2020 from different studies, which leads to a total of 9854 plot observations. We considered 6 different target classes: boreal, mediterranean broad-leaved, mediterranean needle-leaved, submediterranean, temperate broad-leaved, and temperate needle-leaved. For regression of ECT, Vegetation Plot Observations (VPO2025) conducted between 2023 and 2025 were used. The plots were sampled in 7 different areas scattered along a latitudinal gradient over Italy and contain 2016 single plot observations (see also Hiebl et al. (2025)). Both datasets comprise a diverse set of different forest vegetation from mediterranean to temperate regions, also covering mixed species and functional type woodlands.

2.1.2 Sentinel-2 L2A annual time series For each plot location in VDB and VPO2025 the Sentinel-2 L2A observations with a buffer of 10 m are extracted from Microsoft Planetary Computer (Microsoft-Open-Source et al., 2022) starting from 01.01.2022 to 31.12.2024. Invalid observations were masked using the Sen2Cor Scene Classification Layer (SCL). The three year time series were aggregated by day-of-year medians into a single synthetic year to increase observation density and stabilize the phenological signal. An additional outlier detection step was performed using the inter-quartile range to remove residual outliers after SCL masking and prevent overfitting. The resulting S2 time series has a sparse character with missing values along the 365 day annual time axis where observations are missing or had to be removed. With the Normalized Difference Vegetation Index (NDVI) a single vegetation index was calculated for the S2 time series and appended to the feature set.

2.1.3 Climate annual time series As additional climate data has proven to enhance prediction capabilities in environmental mapping (Grabska-Szwagrzyk et al., 2024), we integrated high-resolution CHELSA data into the set of predictors. CHELSA is a downscaled 1 km global climate dataset (Karger et al., 2020). We computed monthly climatologies for precipitation (pr), temperature-above-surface (tas), minimum temperature-above-surface (tasmin) and maximum temperature-above-surface (tasmax). We calculated the monthly average for 2022 to 2024 with a reference time range ranging from 2000 to 2015 using the chelsa-cmip6 python module (ScenarioMIP: ssp245) (Karger, 2024). Missing values were interpolated linearly for better convenience with the combined Attentionheads of the TST model and sparse S2 time series.

2.1.4 Alpha Earth Foundations AEF are precomputed 64-dimensional feature embeddings based on multi-modal remote sensing, climate and text data (Brown et al., 2025). They are available annually from 2017 to 2024, but are computed on spatio-temporal data, which gives them temporal context awareness, mandatory for many environmental mapping tasks. AEF embeddings for the plot observations in the two datasets were downloaded from GEE for 2022 to 2024. The three annual values were averaged to ensure relative stability against changes. As recommended in Brown et al. (2025) no further preprocessing steps were taken.

2.2 Model and training scheme

2.2.1 Model architectures We utilized 4 different model architectures. For AEF we opted for a simple 2-layer Perceptron Model (MLP_{AEF}). For training on S2/CHELSA data, a TST with a 2-layer Perceptron model head, was selected (TST_{S2}). A learnable time positional encoding based on the time step indices is added to the sequence before the attention mechanism to enable the model to capture time context. The spectral-temporal feature space of the Attention layer is mapped to a linear 64-dimensional embedding, which serves as input to the model heads. The third model ($TST_{AEF,S2}$) uses AEF, S2 and CHELSA as input (Fig. 2). We assume that AEF is the more stable feature representation compared to the raw time series inputs, due to its rich multi-modal, multi-temporal representation from self-supervised learning. Therefore we constructed a Cross Attention (CA) based TST model, where AEF attends to the S2/CHELSA time series, after projection into a common feature space. Conceptually this assumes that AEF is querying the S2/CHELSA time series features for additional or complementary information. This way the model can rely on the stable and generalized AEF features, while adding information from the detail-rich time series data. A skip connection, which is a linear layer that directly connects input features and output, from AEF to the model head and a head-initialization close to 0.0 ensures, that the baseline is a MLP based on AEF, if S2/CHELSA contains no enhancing additional information. Initializing the classification/regression model head weights and biases near 0.0 ensures that the model head starts unbiased, so input features and skip connection rather than random initialization determine the initial outputs. Basing the model on AEF keeps $TST_{AE,S2}$ relatively lightweight and fast in comparison to TST_{S2}. The TST models are loosely based on the

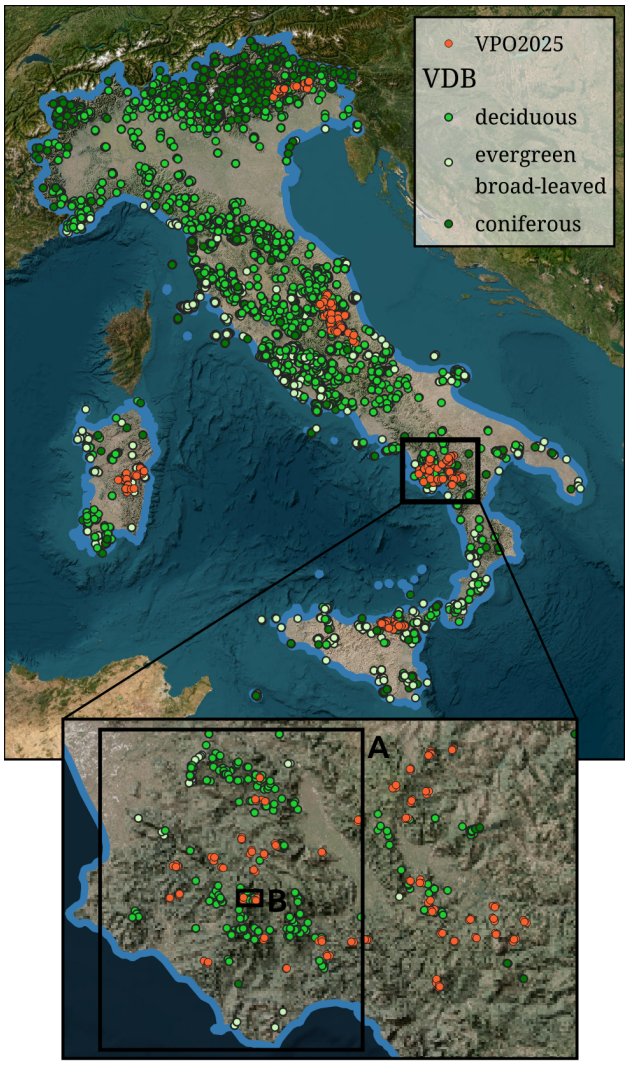

Figure 1. Distribution of VDB and VPO2025 data sources across Italy. A denotes the finally mapped area of Cilento National Park and B denotes the area that was selected for detailed comparison.

original vanilla transformer mentioned in Vaswani et al. (2023) and a pytorch time series implementation by Oguiza (2023). As a commonly used baseline in environmental mapping, we compared these three DL approaches to a S2/CHELSA based RF model (RF_{S2}). Monthly medians of S2/CHELSA annual time series were calculated, resulting in 168 features, that are used to train RF_{S2}.

2.2.2 Training procedure We used Cross Entropy Loss with class weighting (CELoss) for classification and Mean Squared Error Loss (MSELoss) for regression in combination with an Adam optimizer for training the models over 200 epochs with early stopping based on the validation loss metric (Mao et al., 2023). Class weights were calculated for CELoss to account for imbalanced class distributions. To improve generalizability in the regression/classification task we used data augmentation techniques with a augmentation probability based on the inverse kernel density (IKD) of the target variable. Therefor we calculated the square root of the IKD estimate of the training data label distribution, normalized to unit mean and clipped to avoid excessive influence of outliers. This way training data with labels, that appear less often in the data and are sampled more often during training, are more likely to be augmented. Inspired by Yuan et al. (2025), we used a combination of smoothing/interpolating with a Whittaker Algorithm, shifting the sparse observation mask randomly, and adding random noise to the smoothed data for augmenting the S2 time series. As CHELSA are interpolated TS, we only used a TS shifting augmentation. For AEF augmentation, random noise from a normal distribution ($SD = 0.1 * \overline{AEF}$) was added to a randomly selected features of AEF. The used augmentation techniques are available under `https://git.uibk.ac.at/rslab/sattstools/`. We did not apply augmentation in RF_{S2} training.

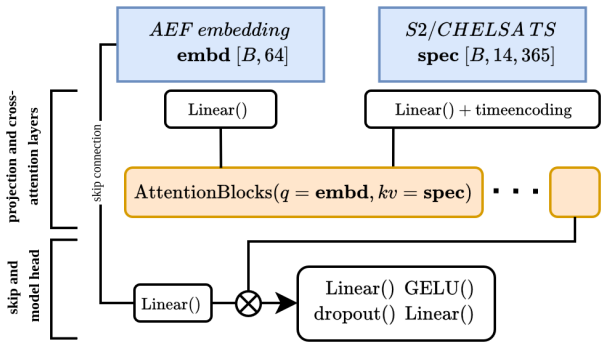

Figure 2. Base architecture of the AEF/S2/CHELSA Cross-Attention fusion model used for both tasks.

2.3 Evaluation

To evaluate the model performances we use a 5-fold cross-validation (CV) scheme, in which 20% of the dataset are used for testing (hold-out) and 80% flow into the training process. This way each sample appears once in the test dataset. Within the training dataset 25% of data is used as validation for hyperparameter tuning and model selection.
Error metrics for each model and CV-fold are calculated on the hold-out dataset. For FVT (classification task) overall accuracy (Acc), weighted F1-Score (F1w), and Macro F1-Score (F1m) are used for comparison, while Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and R²-Score (R²) are reported for ETC.
To evaluate the models from a user perspective a selected area in Italy is mapped using the three input data approaches to identify areas, with high error rates for each model. The chosen area in Cilento National Park in southern Italy has heterogeneous forest characteristics and landscapes. For a model perspective, we examine the cross-attention attributions using the Integrated Gradients (IG) of the model for several different samples to investigate which approach is more meaningful under certain circumstances. IG are calculated from the gradients of the DL models layers and can be used to explain feature contributions in a DL model.

3. Results and Discussion

3.1 Model performance

We tested all four models over 5 test datasets per target task in the 5-fold CV scheme (Fig. 3). For ETC, $TST_{AEF,S2}$ outperformed the other two models across all folds with a mean MAE of 0.11, mean RMSE of 0.161, and mean R² of 0.724. RF_{S2}, MLP_{AEF} and TST_{S2} had similar results for mean RMSE (0.179; 0.18; 0.177) and R² (0.691; 0.687; 0.698),

while RF_{S2} and TST_{S2} exhibited a slightly better mean MAE (0.127; 0.134; 0.125) (Table 1). The classification task with FVT showed a similar pattern with $TST_{AEF,S2}$ outperforming the other two models in all folds (Acc: 0.757, F1m: 0.712, F1w: 0.747). MLP_{AEF} and TST_{S2} had again quite similar performances (Acc: 0.734 vs. 0.736, F1w: 0.733 vs. 0.735), except for F1m (0.678 vs. 0.706). RF_{S2} underperformed on this task with $Acc = 0.712$, $F1m = 0.662$ and $F1w = 0.717$. The significant difference between F1m and F1w shows, that all four models perform poorly on minority classes. With RF_{S2} and MLP_{AEF} exhibiting the highest differences. As all DL models were trained with a class-weighted CELoss, this might stem from data augmentation, that we used quite rigorously on S2/CHELSA data, but less so on AEF (see sec. 2). The observed performance decline of RF_{S2} from ETC to FVT indicates, that conventional ML models, such as RF, remain competitive under limited dataset sizes ($n = 2016$), whereas DL models increasingly benefit from larger training datasets ($n = 9854$).

The performance increase of a combined model of +2.6% in Acc and -0.018 in RMSE compared to the stand-alone models, shows the potential of the approach. The CA mechanism seems to capture the additional information contained in the detailed S2 time series and leverages it to improve the AEF-based basemodel. Using the AEF features as baseline in $TST_{AEF,S2}$ proved to keep the model performance ”at least” as good as the MLP_{AEF}, with enhancement from the time series when applicable. We did not test other options of combinatory CA, such as combining AEF and S2/CHELSA self-attention with the CA approach or running parallel CA for AEF $\mapsto$ S2/CHELSA and S2/CHELSA $\mapsto$ AEF. But from our understanding, the results show that the idea of using the stable multi-modal embeddings that are more robust against error sources such as high cloud cover, and rely on a larger set of datasources, as a baseline and enhance this base with more detailed spectral-temporal information proved to be reliable.

With p-values of 0.0042 (MAE at ETC) and 0.0035 (Acc at FVT) in a Friedman-Test statistically significant differences can be detected between the models. We performed a pair-wise directional Wilcoxon signed-rank test under the assumption that $TST_{AEF,S2}$ always ranks better than the other three models. All pairs exhibit the lowest possible p-value of 0.031. Although results of statistical tests with a sample size of $n = 5$ should be treated with care, the consistency of performance metrics underline the statistical tests.

The high variability of performance metrics in both datasets between the CV-folds raises the question of training stability in the current setting. Splitting the relatively small datasets into training (60%), validation during training (20%), and testing (20%) can lead to a high amount of unseen data in the training datasets, which can impose low performance on the test dataset. While on the other hand, a random split might lead to spatial autocorrelation between the splits, which leads to a overestimation of model capabilities. Both problems are hard to come by in small datasets, which leads to a bias in the metrics values. Investigations into the label distributions between train/val/test data showed high fluctuations between the folds, which likely caused the high intra-model differences in performance metrics across the folds. Folds with close distribution match between train and test data in general led to better test-time performance of the models. Nevertheless, since the relative results across folds, datasets and between the four tested models are consistent, the overall assumption that $TST_{AEF,S2}$ outperforms the other models is well supported.

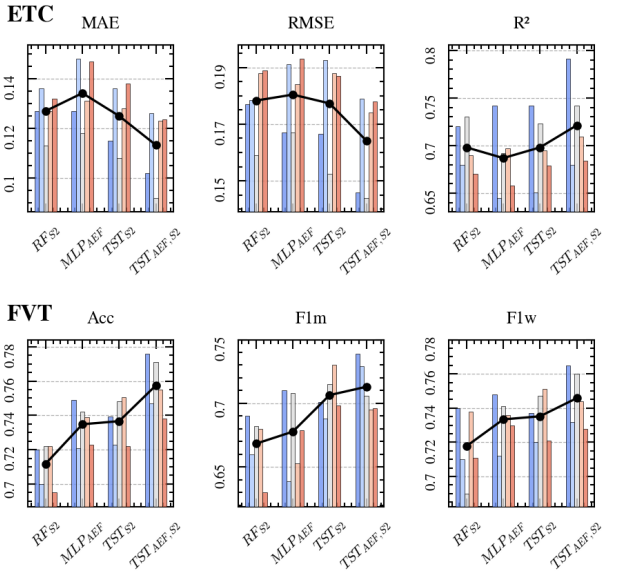

Figure 3. Test metrics of the three models per fold and per target of Evergreen broad-leaved Tree Cover (ETC) and Forest Vegetation Type (FVT).

3.2 Training procedure and feature attribution

One major advantage of AEF is the drastically reduced training time. The MLP used with AEF takes 30 s compared to $\sim$ 12 min for TST_{S2} and $\sim$ 5 min for $TST_{S2,AEF}$ with same batch size (64), 200 epochs and a dataset size of 5912 samples. This increase in time arises from, a) more complex and computing intensive data augmentation, b) larger amount of data to process, and c) the higher model complexity of TST. This does not take into account the amount of time necessary to preprocess/produce S2 and CHELSA time series data with cloud masking, outlier detection, and annual aggregation. And this also does not include the time needed for model development and adaptation of a meaningful time series DL model being it convolutional, recurrent, or transformer style (see e.g. Ismail Fawaz et al. (2020)).

Integrating a CA mechanism between AEF and S2/CHELSA gives the model the ability to learn a sample-wise weighting of the features, ultimately improving the prediction performance. This way the model can rely on AEF, when input S2 time series with low reliability, due to e.g. high winter cloud cover, are detected. We showed that this CA fusion yields promising results for fusing AEF as queries and other single-sensor RS data as key/values, specifically processed for a certain task. This is in line with other studies that use CA for data fusion purposes (Wang et al., 2025). However, if such a model is not trained carefully feature contributions of the RS data might be going towards 0. We analyzed IG of the final model to investigate, how much the different features contribute to the final output class/label. For FVT prediction at class ”mediterranean broad-leaved” in CV-fold 1 of the test dataset, we found that the sample-averaged attributions are 8.0 for S2/CHELSA and 22.1 for AEF (Fig. 4A). S2/CHELSA attributions were summed per feature and across all time steps. While there is some variability, the general picture is the same for all CV-folds and target classes. This means, that AEF contributes in this setting approximately $\sim$ 3x more to the output than S2/CHELSA time series. Nevertheless the contribution of S2/CHELSA time series is enough to improve overall model performance continu-

Model	ETC			FVT		
	MAE	RMSE	R^2	Acc	F1m	F1w
RF_{S2}	0.127 ± 0.008	0.179 ± 0.011	0.691 ± 0.023	0.712 ± 0.011	0.662 ± 0.022	0.717± 0.019
MLP_{AEF}	0.134 ± 0.012	0.180 ± 0.011	0.687 ± 0.034	0.734 ± 0.012	0.678 ± 0.029	0.733± 0.012
TST_{S2}	0.125 ± 0.012	0.177 ± 0.015	0.698 ± 0.032	0.736 ± 0.014	0.706 ± 0.016	0.735 ± 0.014
$TST_{AEF,S2}$	**0.110** ± 0.014	**0.161** ± 0.016	**0.724** ± 0.041	**0.757** ± 0.016	**0.712** ± 0.018	**0.747** ± 0.016

Table 1. Mean ± standard deviation of ETC performance metrics across 5 CV-folds.

ously over all tested datasets. While average attribution over all test samples is around 0.36, per-sample attributions range from a minimum of 0.14 to a maximum of 1.15. A standard deviation of 0.16 shows, that there are significant differences between the sampled inputs. This might highlight the fact, that - while being quite informative - there are quality/stability differences between the time series samples. In Fig. 5 the 25 input S2 time series in the test dataset are depicted, where the difference of averaged attributions for AEF embeddings and S2/CHELSA are highest and lowest, respectively. This shows that the AEF embeddings attend less to S2/CHELSA features, when the density of observations in the S2 time series is generally low. There are probably several other reasons that were not investigated here, such as e.g. label noise, why AEF attention to S2/CHELSA varies across samples.

Despite its high overall reliance on AEF, the CA learns meaningful patterns within the time series data. Fig. 4B shows that the per-feature attributions of annual spectral and climatic time series follow the phenological phases of the forest vegetation. Positive and negative contributions of features alternate across seasons. E.g. a high $tasmax$ during peak growth season in summer contributes to a high probability that the sample is assigned to the class of ”mediterranean broad-leaved” and high $B05$ values during senescence and leaf-off season lower the probability. This contributes to the assumption, that S2/CHELSA - if correctly attended to by the CA mechanism - can add meaningful information to the model and output. However, this raises the question, if the time and computation intensive preprocessing of RS time series with a mean contribution of 0.3 to the model output is worth the effort for a small improvement in model performance. Especially for large scale investigations on a continental or global scale this might be an issue and depends on the objectives of the practitioners.

However, the CA approach has its limitations as it does not use both data sources to its full potential. MLP_{AEF} and TST_{S2} both showed similar results in the CV scheme, which rises the question if a CA approach automatically underrepresents the S2/CHELSA dataset. Another possibility to combine both worlds would be to use a hierarchical pseudo-labeling approach, utilizing the highly generalized AEF to build a large pseudo-label database for specialized DL models (Zhang et al., 2024). In a study setting like ours with limited dataset sizes (9854 and 2026 plot observations) this could be a powerful approach to enhance the dataset sizes, with e.g. uncertainty-based resampling on the newly generated data.

3.3 Target mapping

To analyse the mapping results spatially ETC task was chosen and, a) the average ETC in % across CV-folds, b) the intra-model SD across the CV-folds as a measure of uncertainty for identification of error-prone regions, and c) the pair-wise inter-model differences of averaged predictions were calculated. As AEF is a relatively new dataset, that has not been used much in practice yet, this qualitative assessment by evaluating the produced maps of ETC, that goes beyond the metrics of ground truth data, is valuable to our future work.

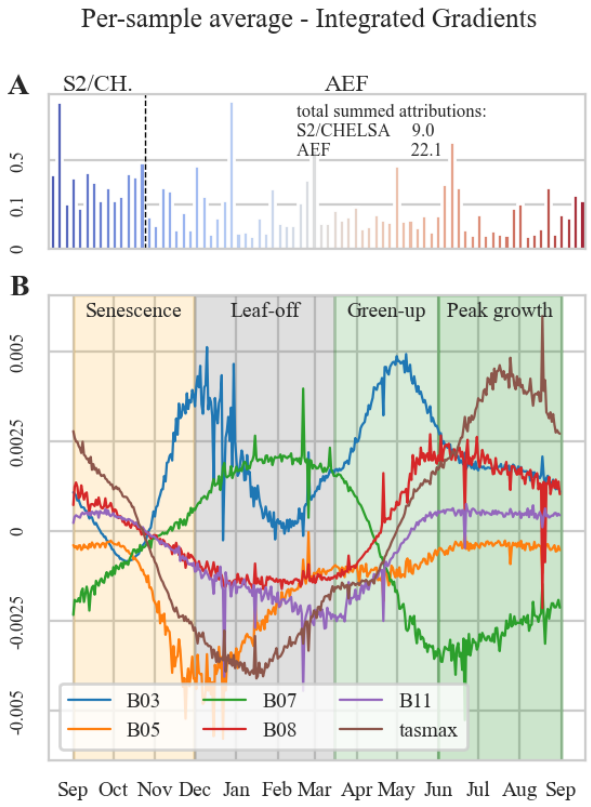

Figure 4. Sample-average Integrated gradients of $TST_{S2,AEF}$ for fold $n = 1$ at FVT class ”mediterranean broad-leaved”. A) Per feature attributions summed over time-dimension and split in S2/CHELSA and AEF. B) S2/CHELSA attributions for selected features across annual phenocycle.

Compared to MLP_{AEF}, TST_{S2} maps inherit higher pixel noise. In a way this is expected as the 10 m resolution per pixel mapping based on the S2 times series is the direct sensor signal information. In contrary, AEF data contains information from several sensors and datasets with lower resolution (Landsat - 30 m; ERA5 - $\sim$ 11 km), which spatially smooths the input data for MLP_{AEF}. Smoothing the time series, using e.g. a Whittaker Smoother, does not affect the spatial noise in the same way (Hiebl et al., 2025). The integration of topographic variables into AEF leads to a clear distinction between expositions in mountainous terrain (Fig. 6). $TST_{AEF,S2}$ maps are clearly a product of both datasources. In some areas with good illumination the noisy but more detailed patterns of S2 based modeling predominate, while in mountainous regions smoother, probably AEF dominated predictions are occuring. The sharp edges at borders between forest types or forest-nonforest, that are blurred in MLP_{AEF} and the reduced noise in combination with the topographic awareness are kept in $TST_{AEF,S2}$. The SD of MLP_{AEF} maps show, that the models are quite uncertain in north-facing areas across folds. This could in-

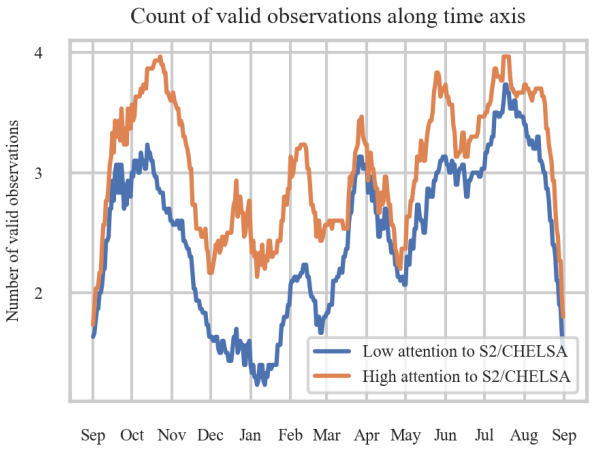

Figure 5. Density of valid observations in S2 time series for output generation that relies more on AEF embeddings and output that relies more on S2/CHELSA embeddings.

herit from a bias in the VPO2025 dataset that does not contain enough north-facing plots for proper modeling, or from misleading topographic feature information in the AEF embeddings feature vector. TST_{S2} exhibits similar high SD only for extremely steep north-facing terrain probably due to a lack of signal strength during winter and therefor large gaps with invalid data. For moderately steep forest slopes the predictions are relatively consistent across folds and where SD shows no continuous topographic pattern. $TST_{AEF,S2}$ SD values seem to be more randomly distributed, coming from input data noise (similar to TST_{S2}), but also in some parts topography. But all three models exhibit across-fold SD of values up to 25%. Which contributes to the issue, that good test data metrics do not mean good spatial mapping results.
The inter-model differences between TST_{S2} and MLP_{AEF} in the mapped region of Cilento exceed $\pm 20\%$ ETC in mountainous regions. On south-facing slopes MLP_{AEF} predicts relatively low ETC values compared to TST_{S2}. While north-facing slopes inherit relatively high values and negative ΔETC. $TST_{AEF,S2}$ seems to predict values that lie in-between the two other models, hence the difference patterns look similar but with lower amplitude. The large inter-model differences highlight the problems, that might often be overlooked in environmental mapping from sparse ground truth data. Despite the relatively comparable MAE, RMSE, and R^2 results (especially for TST_{S2} and MLP_{AEF}), the differences contribute to the problem, that test data can either inflate or deflate model performance depending highly on the chosen/available ground truth and training datasets, while spatial generalizability is difficult to achieve and especially test. Mapping efforts in mountainous regions remain challenging, due to the complex illumination and topographic features appearing in resulting maps, with ground truth data not sufficiently covering all expositions and characteristics of the terrain.

4. Conclusion

In this study, four different combinations of data sources and models to map forest variables in Italy were investigated. Alpha Earth Foundations (AEF), Sentinel-2/CHELSA (S2/CHELSA) and a combination of both were compared in a 5-fold cross-validation experiment on two different datasets with a regression (evergreen broad-leaved cover ETC) and a classification (forest vegetation type FVT) task. The results showed, that in general the lightweight AEF-based models yield similar performances than the S2/CHELSA based Random Forest and Time Series Transformer models. The main differences here are the drastically reduced training time and the obsolete preprocessing of large amounts of remote sensing time series data. This is a major advantage in large-scale mapping on a continental or global scale. The study further investigates a combination of both datasets by using a Cross Attention (CA) mechanism, where the AEF data as baseline attends to the S2/CHELSA time series for additional, complementary information. This approach outperformed the stand-alone models on both tasks across all metrics. It also showed that it can overcome common issues in RS time series models, such as prediction errors from cloud cover gaps or shaded areas in mountainous regions. All models inherit common problems with mapping environmental variables from plot observations, as large differences in mapped variables occur between model predictions, but also intra-model between CV-folds.
In conclusion, the exploratory study highlights the potential of combining generalized, pretrained (AEF) with specialized, detail-rich (S2/CHELSA) datasets for environmental mapping applications. Further investigation into combining the generalized AEF embeddings with specialized models for various applications is necessary and could be achieved by using e.g., bi-directional CA mechanisms or hierarchical pseudo-labeling.

Data and Code availability

Model architectures and training code are made available at `https://git.uibk.ac.at/c7161037/ae_training`. Training data and corresponding maps can be downloaded from `https://10.5281/zenodo.18375305`.

Acknowledgements

This research has been conducted within the project “TRACEVE - Tracing the evergreen broad-leaved species and their spread” (I 6452-B) funded by the Austrian Science Fund (FWF). GB was funded under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.4—Call for tender No. 3138 of 16 December 2021, rectified by Decree n. 3175 of 18 December 2021 of Italian Ministry of University and Research funded by the European Union – NextGenerationEU. Project code CN_00000033, Concession Decree No. 1034 of 17 June 2022 adopted by the Italian Ministry of University and Research, CUP B63C22000650007“, Project title “National Biodiversity Future Center—NBFC”.

References

Alessi, N., Těšitel, J., Zerbe, S., Spada, F., Agrillo, E., Wellstein, C., 2019. Ancient refugia and present-day habitat suitability of native laurophylls in Italy. *Journal of Vegetation Science*, 30(3), 564–574.

Bonari, G., Knollová, I., Vlčková, P., Xystrakis, F., Çoban, S., Sağlam, C., Didukh, Y. P., Hennekens, S. M., Acosta, A. T. R., Angiolini, C., Bergmeier, E., Bertacchi, A., Costa, J. C., Fanfarillo, E., Gigante, D., Guarino, R., Landi, M., Neto, C. S., Pesaresi, S., Rosati, L., Selvi, F., Sotiriou, A., Stinca, A., Turcato, C., Tzonev, R., Viciani, D., Chytrý, M., 2019. CircumMed Pine Forest Database: an electronic archive for Mediterranean

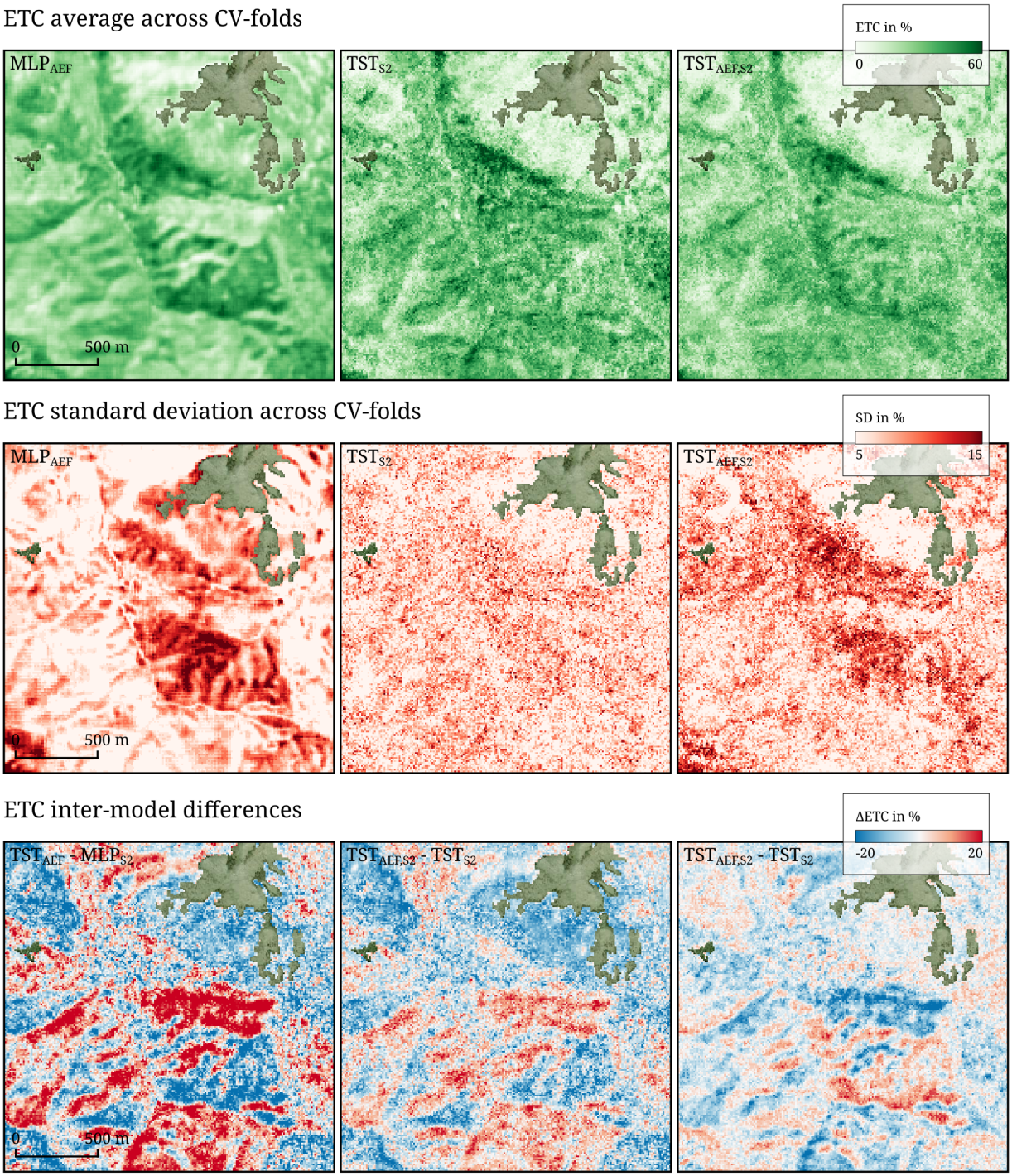

Figure 6. ETC mapping results for a selected mountainous area in Italys Cilento National Park. Average and standard deviation (SD) are calculated across CV-folds per-pixel. Inter-model differences are calculated based on the averaged ETC values of each model.

and Submediterranean pine forest vegetation data. *Phytocoenologia*, 311–318. Publisher: Schweizerbart’sche Verlagsbuchhandlung.

Brown, C. F., Kazmierski, M. R., Pasquarella, V. J., Rucklidge, W. J., Samsikova, M., Zhang, C., Shelhamer, E., Lahera, E., Wiles, O., Ilyushchenko, S., Gorelick, N., Zhang, L. L., Alj, S., Schechter, E., Askay, S., Guinan, O., Moore, R., Boukouvalas, A., Kohli, P., 2025. AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data. arXiv:2507.22291 [cs].

Chabalala, Y., Adam, E., Kganyago, M., 2023. Mapping fruit tree dynamics using phenological metrics from optimal Sentinel-2 data and Deep Neural Network. *CABI Agric Biosci*, 4(1), 51.

Chen, Y., , Ruyin, C., , Shuaijun, L., , Longkang, P., , Xuehong, C., , Chen, J., 2024. A new deep learning-based model for reconstructing high-quality NDVI time-series data in heavily cloudy areas: fusion of Sentinel 1 and 2 data. *International Journal of Digital Earth*,

17(1), e2407941. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/17538947.2024.2407941.

Ge, S., Antropov, O., Häme, T., McRoberts, R. E., Miettinen, J., 2023. Deep Learning Model Transfer in Forest Mapping Using Multi-Source Satellite SAR and Optical Images. *Remote Sensing*, 15(21), 5152. Number: 21 Publisher: Multidisciplinary Digital Publishing Institute.

Grabska-Szwagrzyk, E., Tiede, D., Sudmanns, M., Kozak, J., 2024. Map of forest tree species for Poland based on Sentinel-2 data. *Earth System Science Data*, 16(6), 2877–2891. Publisher: Copernicus GmbH.

Hiebl, B., Alessi, N., Calvia, G., Bricca, A., Bonari, G., Zangari, G., Dorigo, W., Zerbe, S., Rutzinger, M., 2025. Advancing forest mapping: Pretraining strategies and deep-ensemble based uncertainty for predicting evergreen broad-leaved cover from Sentinel-2 time series. *International Journal of Applied Earth Observation and Geoinformation*, 142, 104734.

Ismail Fawaz, H., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D. F., Weber, J., Webb, G. I., Idoumghar, L., Muller, P.-A., Petitjean, F., 2020. Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery*, 34(6), 1936–1962.

Karger, D. N., 2024. chelsa-cmip6: package containing functions to create monthly high-resolution climatologies.

Karger, D. N., Schmatz, D. R., Dettling, G., Zimmermann, N. E., 2020. High-resolution monthly precipitation and temperature time series from 2006 to 2100. *Sci Data*, 7(1), 248. Publisher: Nature Publishing Group.

Kattenborn, T., Leitloff, J., Schiefer, F., Hinz, S., 2021. Review on Convolutional Neural Networks (CNN) in vegetation remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 173, 24–49.

Kollert, A., Bremer, M., Löw, M., Rutzinger, M., 2021. Exploring the potential of land surface phenology and seasonal cloud free composites of one year of Sentinel-2 imagery for tree species mapping in a mountainous region. *International Journal of Applied Earth Observation and Geoinformation*, 94, 102208.

Ma, Y., Chen, S., Ermon, S., Lobell, D. B., 2024. Transfer learning in environmental remote sensing. *Remote Sensing of Environment*, 301, 113924.

Mao, A., Mohri, M., Zhong, Y., 2023. Cross-entropy loss functions: Theoretical analysis and applications.

Microsoft-Open-Source, McFarland, M., Emanuele, R., Morris, D., Augspurger, T., 2022. microsoft/planetarycomputer: October 2022.

Oguiza, I., 2023. tsai - A state-of-the-art deep learning library for time series and sequential data.

Pettorelli, N., Schulte to Bühne, H., Tulloch, A., Dubois, G., Macinnis-Ng, C., Queirós, A. M., Keith, D. A., Wegmann, M., Schrodt, F., Stellmes, M., Sonnenschein, R., Geller, G. N., Roy, S., Somers, B., Murray, N., Bland, L., Geijzendorffer, I., Kerr, J. T., Broszeit, S., Leitão, P. J., Duncan, C., El Serafy, G., He, K. S., Blanchard, J. L., Lucas, R., Mairota, P., Webb, T. J., Nicholson, E., 2018. Satellite remote sensing of ecosystem functions: opportunities, challenges and way forward. *Remote Sensing in Ecology and Conservation*, 4(2), 71–93.

Safonova, A., Ghazaryan, G., Stiller, S., Main-Knorn, M., Nendel, C., Ryo, M., 2023. Ten deep learning techniques to address small data problems with remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 125, 103569.

Skidmore, A. K., Coops, N. C., Neinavaz, E., Ali, A., Schaepman, M. E., Paganini, M., Kissling, W. D., Vihervaara, P., Darvishzadeh, R., Feilhauer, H., Fernandez, M., Fernández, N., Gorelick, N., Geijzendorffer, I., Heiden, U., Heurich, M., Hobern, D., Holzwarth, S., Muller-Karger, F. E., van de Kerchove, R., Lausch, A., Leitão, P. J., Lock, M. C., Mücher, C. A., O’Connor, B., Rocchini, D., Roeoesli, C., Turner, W., Vis, J. K., Wang, T., Wegmann, M., Wingate, V., 2021. Priority list of biodiversity metrics to observe from space. *Nat Ecol Evol*, 5(7), 896–906.

Sun, Z., Sandoval, L., Crystal-Ornelas, R., Mousavi, S. M., Wang, J., Lin, C., Cristea, N., Tong, D., Carande, W. H., Ma, X., Rao, Y., Bednar, J. A., Tan, A., Wang, J., Purushotham, S., Gill, T. E., Chastang, J., Howard, D., Holt, B., Gangodagamage, C., Zhao, P., Rivas, P., Chester, Z., Orduz, J., John, A., 2022. A review of Earth Artificial Intelligence. *Computers & Geosciences*, 159, 105034.

Torres, P., Rodes-Blanco, M., Viana-Soto, A., Nieto, H., García, M., 2021. The Role of Remote Sensing for the Assessment and Monitoring of Forest Health: A Systematic Evidence Synthesis. *Forests*, 12(8), 1134.

Tseng, G., Cartuyvels, R., Zvonkov, I., Purohit, M., Rolnick, D., Kerner, H., 2024. Lightweight, Pre-trained Transformers for Remote Sensing Timeseries.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I., 2023. Attention Is All You Need. arXiv:1706.03762 [cs].

Vogeler, J. C., Fekety, P. A., Elliott, L., Swayze, N. C., Filippelli, S. K., Barry, B., Holbrook, J. D., Vierling, K. T., 2023. Evaluating GEDI data fusions for continuous characterizations of forest wildlife habitat. *Front. Remote Sens.*, 4. Publisher: Frontiers.

Wang, H., Huang, Y., Huang, H., Wang, Y., Li, J., Gui, G., 2025. Uncertainty-Aware Dynamic Fusion Network with Criss-Cross Attention for multimodal remote sensing land cover classification. *Information Fusion*, 123, 103249.

Yuan, Y., Lin, L., Liu, Q., Hang, R., Zhou, Z.-G., 2022. SITS-Former: A pre-trained spatio-spectral-temporal representation model for Sentinel-2 time series classification. *International Journal of Applied Earth Observation and Geoinformation*, 106, 102651.

Yuan, Y., Lin, L., Xin, Q., Zhou, Z.-G., Liu, Q., 2025. An Empirical Study on Data Augmentation for Pixelwise Satellite Image Time-Series Classification and Cross-Year Adaptation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18, 5172–5188.

Zhang, J., You, S., Liu, A., Xie, L., Huang, C., Han, X., Li, P., Wu, Y., Deng, J., 2024. Winter Wheat Mapping Method Based on Pseudo-Labels and U-Net Model for Training Sample Shortage. *Remote Sensing*, 16(14).