Overview of GeoLifeCLEF 2024: Species Composition Prediction with High Spatial Resolution at Continental Scale using Remote Sensing

Lukas Picek[1], Christophe Botella[1], Maximilien Servajean[3], César Leblanc[1,2], Rémi Palard[1], Théo Larcher[1], Benjamin Deneu[1,2], Diego Marcos[1], Joaquim Estopinan[1,2], Pierre Bonnet[2] and Alexis Joly[1]

[1]*INRIA, LIRMM, Univ Montpellier, CNRS, Montpellier, France*
[2]*AMAP, Univ Montpellier, CIRAD, CNRS, INRAE, IRD, Montpellier, France*
[3]*LIRMM, AMIS, Univ Paul Valéry Montpellier, Univ Montpellier, CNRS, France*

Abstract

Understanding the spatiotemporal distribution of species is a cornerstone of ecology and conservation. Pairing species observations with geographic and environmental predictors allows us to model the relationship between an environment and the species present at a given location. In light of that, we organize an annual competition, GeoLifeCLEF, which focuses on benchmarking and advancing state-of-the-art species distribution modeling using available bioclimatic and remote sensing data. The GeoLifeCLEF 2024 dataset spans across Europe and encompasses most of its flora. The species observation data comprises over 5 million Presence-Only (PO) occurrences and approximately 90 thousand Presence-Absence (PA) surveys. Those data are paired with various high-resolution rasters, including remote sensing imagery, land cover, and elevation, and are combined with coarse-resolution data such as climate, soil, and human footprint variables. In this paper, we present (i) an overview of the GeoLifeCLEF 2024 competition, (ii) a description of the provided data, (iii) an overview of approaches used by the participating teams, and (iv) the main results analysis.

Keywords

LifeCLEF, biodiversity, environmental data, species distribution, prediction, evaluation, benchmark, methods comparison, presence-only data, presence-absence, model performance, remote sensing

1. Introduction

Global changes are transforming ecosystems at an alarming rate [1]. Land use changes and biological invasions are among the main drivers of the ongoing biodiversity decline, with their impacting processes often operating at a fine spatial grain. Monitoring species composition at high spatial resolution (10–50m) coherently at a continental scale would help characterize these impacts and facilitate the development of more effective conservation policies. Unfortunately, such comprehensive monitoring is largely impractical due to the vast areas involved, the resources required, and the complexity of ecosystems.

Species distribution models (SDMs) can be used to fill the spatial gaps of field monitoring by relying on the growing mass of spatial biodiversity data worldwide coupled with high-resolution remote sensing data. Indeed, using spatiotemporal satellite data along with coarser geographic predictors to improve SDM predictions has shown a potential to capture fine-scale patterns of species communities and improve their prediction, notably through the use of deep learning-based SDM (deepSDMs) [2, 3, 4, 5].

However, implementing SDMs at this resolution faces significant obstacles. The scarcity, imbalance, and heterogeneity of available species observations and environmental data are major challenges. Despite the mass of available Presence-Only (PO) records, notably contributed by global crowdsourcing platforms (e.g., Pl@ntNet and iNaturalist), important sampling biases arise in their collection and hamper the quality of SDM built from them [6, 7, 8].

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

lukas.picek@inria.fr (L. Picek); alexis.joly@inria.fr (A. Joly)
0000-0002-6041-9722 (L. Picek); 0000-0002-2828-4389 (P. Bonnet); 0000-0002-2161-9940 (A. Joly)

Last year's edition of GeoLifeCLEF [9, 10] illustrated it by introducing an SDM evaluation scheme based on standardized Presence-Absence (PA) surveys, and further highlighted the value of PA surveys in model training. However, the spatial coverage of PA data being spatially very limited, properly combining PA and PO data can leverage the extensive coverage of PO data while correcting its biases [9, 11, 12]. Even with comprehensive data, modeling a diverse biological group as diverse as plants is challenging. Europe has over 11,000 plant species, most of which are rare, leading to a strong class imbalance in machine learning. This imbalance makes it difficult to develop accurate models, especially for predicting rare species distributions. Satellite data is crucial for characterizing the spatio-temporal context of plant communities at a high spatial resolution, and could potentially capture specific contexts of rare species. However, the integration of satellite data into SDMs is relatively recent [4]. Satellite data brings specific challenges within SDM regarding their integration with coarser but complementary geographic descriptors (e.g., climate, soil) and the integration of its spatial and temporal dimensions.

For the 2024 edition of the GeoLifeCLEF[1], we have assembled an extensive dataset at a European scale to investigate these issues. This dataset is designed to facilitate the evaluation of multi-label prediction of species composition at high spatial resolution, based on standardized Presence-Absence (PA) data balanced across bio-regions of Europe. The campaign aims to address several key challenges:

1. Multi-label learning with massive single positive labels (i.e., PO data) and multi-label data.
2. Managing strong class imbalance; a lot of rare species with few samples and a large number of samples for few categories.
3. Handling large-scale data and learning from multiple types of predictors, including multi-band satellite images and time-series data.

By focusing on these challenges, the GeoLifeCLEF 2024 aims to advance the field of species distribution modeling. The integration of diverse data types and the development of robust models capable of accurate, high-resolution predictions will provide valuable insights into species distribution patterns. These advancements will support more effective biodiversity conservation and environmental management practices.

2. Dataset and Evaluation Protocol

The GeoLifeCLEF 2024 dataset contains species observation data, including Presence-Only occurrences and Presence-Absence surveys, paired with various environmental predictors. This dataset offers a rich array of environmental rasters, Sentinel-2 satellite images, a 20-year climatic time series, and satellite time-series point values. Building on the dataset from the previous edition [15], we retained the majority of the provided Presence-Only (PO) occurrences (5 million) and expanded the number of the Presence-Absence (PA) survey records up to 90,000.

As in the previous year, the PA data was divided into training and test sets (95/5) using a spatial block hold-out procedure [16], employing a spatial grid with 10×10 km cells. This method ensures a thorough evaluation of the models by randomly selecting test cells to maintain balance across biogeographical regions. To allow easy use of the data, we provided all environmental predictors as pre-extracted scalar values in separate CSV files. Additionally, the time-series data were formatted as 3D cubes (torch tensors) for ease of use in machine-learning workflows.

2.1. Species Observation Data

The species observation data includes approximately 5 million **Presence-Only** (PO) occurrences and around 90 thousand **Presence-Absence** (PA) survey records. PO data, commonly collected without protocol, is widespread across Europe but prone to various sampling biases (see Figure 1). Reporters (citizen scientists) may miss species due to seasonal visibility, misidentification, or lack of interest. Below is a brief description of both PO and PA data.

[1] The GeoLifeCLEF 2024 competition took place in the CVPR–FGVC11 and LifeCLEF workshops [13, 14].

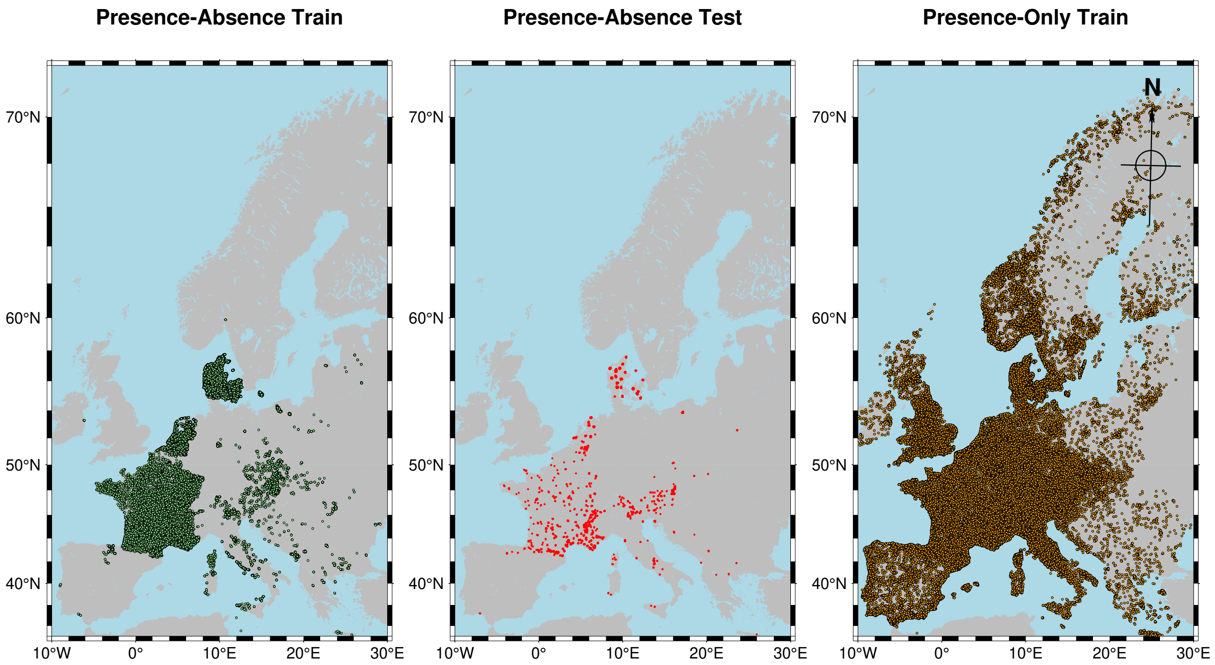

Figure 1: Geo spatial scale of the GeoLifeCLEF 2024 dataset. The Presence-Only (PO) data spans all of habitable Europe. The Presence-Absence (PA) sites are primarily from France, Denmark, Switzerland, Czechia, and Italy.

Presence-Absence Surveys. Conducted by experienced botanists, PA surveys exhaustively report plant species in small spatial plots (10–400 sq meters). Species not observed are likely absent. The data originates from 29 datasets hosted in the European Vegetation Archive (EVA) and includes sources like Denmark Naturdata, IGN National Forest Inventory, and Belgium INBOVEG. Despite the dataset's size (93,703 surveys), it covers only 5,016 species (about half of Europe's flora) with highly imbalanced species distribution. The training and test splits (95/5) were created using a spatial block hold-out procedure [16] with 10×10 km cells, resulting in 88,987 training and 4,716 test surveys, ensuring biogeographical balance.

Presence-Only Occurrences. PO records are geolocated species observations with unknown sampling protocols, offering no information on the absence of other species. Sampling effort varies widely across space, time, and species, often concentrating in populated areas and focusing on charismatic species. Despite biases, PO data helps fill gaps in PA surveys when sampling biases are controlled in model calibration [11, 12]. The PO data includes 5 million records of 9,709 species reported between 2017 and 2021 from 13 datasets extracted from the GBIF [17].

2.2. Enviromental Predictors

The spatialized geographic and environmental predictor data are crucial for precise predictive modeling. Therefore, for each species observation (PO and PA), we provide: (i) a four-band 128×128 satellite image at 10 m resolution around the occurrence location, (ii) time series of past values for six satellite bands at the point location, (iii) various environmental rasters at the European scale (e.g., climate, soil, elevation, land use, and human footprint variables), and (iv) monthly time series of four climatic variables from 2000 to 2019. Given the dataset's diverse sources and significant preprocessing requirements, we briefly describe the acquisition process and data details below. Besides, we summarize all available predictors in Table 1.

Table 1
Environmental predictors summary. Description and link to all provided predictors (available for all species observation data, i.e., PO and PA), as well as their spatial resolution.

Name	Description	Source	Resolution
Climate	19 rasters of historical bioclim data (1981–2010)	CHELSA	$\sim$1km
Monthly Climate	4 variables from January 2000 to December 2019	CHELSA	$\sim$1km
Soil	9 pedological rasters	Soilgrids	$\sim$1km
Elevation	Elevation above sea level	ASTER	$\sim$30m
Land cover	According to IGBP classification (17 classes)	MODIS	$\sim$500m
Human footprint	7 pressures on the environment for 1993 and 2009	Venter et al.	$\sim$1km
Satellite imagery	RGB and NIR patches centered on each observation and taken the same year	Sentinel-2	10m
Satellite time series	Time series of six quarterly satellite bands values since winter 1999	Landsat	30m

2.2.1. Environmental Rasters

Species observations are paired with various environmental rasters, including bioclimatic, soil, elevation, land cover, and human footprint data. These rasters, provided as .TIF files in WGS84 (EPSG:4326) coordinates, cover Europe from $(-32.26, 26.63)$ to $(35.58, 72.18)$.

Land cover. We provide a medium-resolution (500m) multi-band land cover raster for Europe, extracted from the MODIS Terra+Aqua dataset [18]. This GeoTIFF includes 13 layers of land cover classifications and confidence levels. The data was processed and reprojected to WGS84, with each band describing land cover class predictions or confidence levels. The IGBP (17 classes) and LCCS (43 classes) layers are recommended for species distribution modeling.

Human footprint data includes 16 low-resolution (1km) rasters summarizing human pressures, and 14 detailed rasters for seven pressures across two periods (1993–2009). These were derived from Venter et al.'s global human footprint rasters [19], reprojected to WGS84. Variables include built environment, population density, electrical infrastructure, cropland, pastureland, roads, railways, and navigable waterways.

Elevation data, crucial for modeling plant distribution, is provided as a GeoTIFF and in CSV format. It was extracted from the ASTER Global Digital Elevation Model V3 via NASA portal.

Soilgrids. Nine low-resolution (1 km) soil rasters covering a depth of 5 to 15 cm were integrated from SoilGrids 2.0. These rasters, derived from resampling the original 250m resolution data, include key soil properties like pH and granulometry.

2.3. Satellite Images

We provide Sentinel-2 RGB and Near-Infrared (NIR) satellite images (128×128) at 10m resolution (see Figure 2), centered on observation locations and captured in the same year. Images are from pre-processed rasters with cloud and shadow removal, available on the Ecodatacube platform. Values are thresholded at 10,000, re-scaled to [0,1], gamma corrected with $\gamma = 2.5$, re-scaled to [0,255], and encoded as *uint8*.

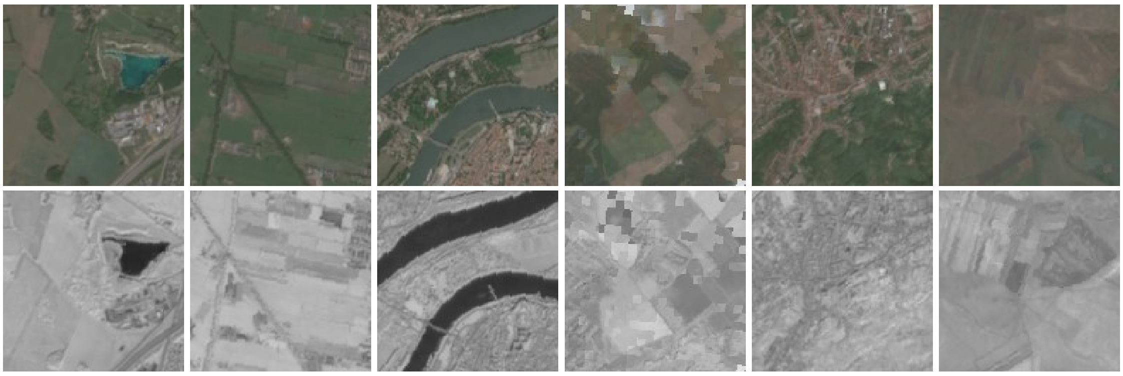

Figure 2: Sentinel-2 satellite images. First row RGB. Second row Near-Infrared (NIR).

2.4. Satellite Time-Series

We provide satellite time-series data spanning over 20 years, obtained from the Landsat ARD program and pre-processed by EcoDataCube. Each location is linked to quarterly median values of six bands (R, G, B, NIR, SWIR1, SWIR2) ranging from 1999 to 2020 and capturing environmental changes. Data points are aggregated into CSV files and converted into 3D tensors [BAND, QUARTER, YEAR].

2.5. Climatic Variables

We provide Monthly and Average Climatic rasters from CHELSA [20]. Monthly rasters include four variables (mean, min, max temperature, total precipitation) from Jan 2000 to Dec 2019 (960 rasters, 30 arcsec resolution). The rasters consist of 19 variables averaged from 1981 to 2010. Values for PO and PA records are pre-extracted into CSV files and aggregated into 3D tensors [RASTER, YEAR, MONTH].

2.6. Evaluation Metric

As in the previous edition [10], the evaluation metric used was the sample-averaged F1 score (F_1). The F_1-score measures the agreement between predicted and actual species composition in a given area and time. In ecological surveys, such as those in Protected Areas (PAs), each survey instance i has a ground-truth set of labels Y_i representing the plant species identified by experts within a grid. Given this setup and a list of predicted labels $\widehat{Y}_{i,1}, \widehat{Y}_{i,2}, \ldots, \widehat{Y}_{i,R_i}$, the micro F_1-score is computed as follows:

$$\mathrm{F}_1 = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + (FP_i + FN_i)/2}, \tag{1}$$

$$\text{where} \begin{cases} TP_i = \text{\# of correctly predicted species, i.e., } |\widehat{Y}_i \cap Y_i|. \\ FP_i = \text{\# of species predicted but not observed, i.e., } |\widehat{Y}_i \setminus Y_i|. \\ FN_i = \text{\# of species not predicted but present, i.e., } |Y_i \setminus \widehat{Y}_i|. \end{cases} \tag{2}$$

This formulation encapsulates the precision and recall elements crucial for assessing the accuracy of predictive models in ecological studies.

3. Participants and methods

A total of 83 participants / 51 teams participated[2] in this year's edition of the GeoLifeCLEF challenge and submitted 1,184 entries, with an average of 23 entries per participant and a maximum of 175 entries by the top-ranked participant. In Figure 3, we report the private leaderboard performance for all participants' methods alongside the organizers' baseline methods. Hereafter, we provide a short overview of the teams' methods, which are further elaborated in working notes [21, 22, 23, 24, 25, 26].

[2] All teams with less than 2 submissions and/or submitted only duplicated baselines were filtered out; approximately 50 teams.

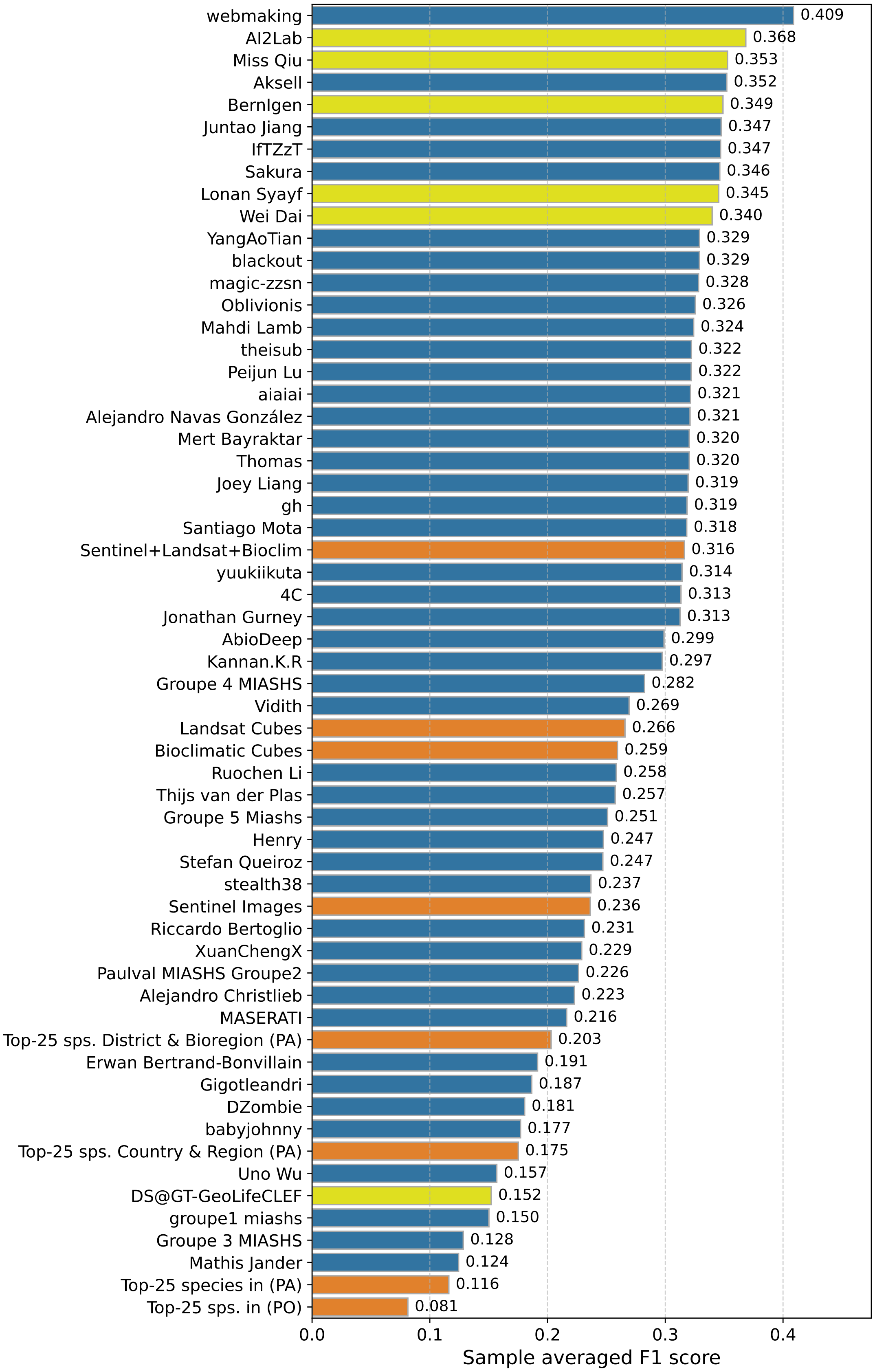

Figure 3: Official GeoLifeCLEF 2024 results. All 51 teams and their private scores in terms of sample-averaged F_1. Orange depicts baselines and yellow teams who submitted working notes.

3.1. Baselines

We provide a variety of weak and strong baselines for all participants to allow a good starting point, continual performance increase, and working with different modalities. All the baselines are using PA data only and were provided in the form of Jupyter Notebooks that could be run (both training and inference) directly on Kaggle. Considering the significant extent to which this baseline's performance can be enhanced, we encouraged participants to experiment with various techniques, architectures, losses, etc. Below, we briefly describe all baselines:

Naive baselines. With the extensive and numerous observational data available, one can naively predict species presence by identifying the most common species within specific administrative or biogeographical regions. For example, by predicting the top 25 most common species based on Presence-Absence (PA) data, the resulting sample-averaged F_1 score is 11.6%. In contrast, applying this same method to Presence-Only (PO) data yields an F_1 score of 8.1%. This discrepancy highlights a distribution shift between the PA and PO data, reflecting differences in species occurrence patterns captured by each data collection method.

Small Residual Convolutional Neural Networks for data cubes. Using a ResNet-18 architecture [27] as a feature extractor, we have designed a model specifically tailored to fit the small cube Landsat and Bioclimatic data (i.e., the input size of $19\times12\times4$ for the bioclimatic time series and $21\times4\times6$ for the Landsat time series). The original ResNet-18 model was chosen for its balance between depth and computational efficiency, making it a suitable candidate for our purposes. Our modifications aimed to maintain the robustness of the original ResNet-18 model while optimizing it for the constraints of GLC's data dimensions. These adjustments involved reconfiguring the input layers to accommodate the specific dimensions of the bioclimatic and Landsat datasets, ensuring that the model can effectively capture the spatiotemporal patterns inherent in the data. The adapted models, optimized with Binary Cross Entropy (BCE) loss, displayed significant performance by achieving sample-averaged F_1 scores of 0.259 if trained on the bioclimatic time series data and 0.266 for the Landsat time series data[3]. These results underscore the effectiveness of our tailored model in dealing with GLC's unique data structure, suggesting that even with reduced complexity, the model retained strong predictive capabilities. Furthermore, these findings emphasize the importance of customizing neural network architectures to align with the specific characteristics of the input data. This ensures optimal performance without unnecessary computational overhead and enhances model efficiency, setting a precedent for future efforts to adapt established and state-of-the-art architectures for specialized datasets, such as GeoLifeCLEF.

Swin tranformer for the Sentinel-2 images. We made slight modifications to the architecture of a Swin-v2-t [28] model to enable it to accept input from all four modalities of Sentinel-2 data, containing RGB (Red, Green, Blue) and NIR (Near Infra Red) channels rather than the conventional three-channel input. This adaptation was crucial to fully leverage the rich information provided by the Sentinel-2 dataset, which includes essential spectral data captured in the infrared range, often critical for tasks involving vegetation and land cover analysis. Such a model achieved a sample-averaged F_1 score of 0.235. Although this score is lower compared to other models trained on different modalities, it reflects the challenges and complexities involved in integrating and effectively utilizing the multimodal data within the same model architecture. The reduction in F_1 score clearly indicates potential areas for further optimization. This includes fine-tuning the model's hyperparameters, experimenting with different training strategies, or incorporating additional preprocessing steps. Despite the lower performance, this baseline provides valuable insights into the feasibility and limitations of adapting transformer-based models like Swin-v2-t for multiband remote sensing data. It suggests that while straightforward architectural modifications can enable the use of richer data inputs, achieving optimal performance may require more sophisticated approaches to fully harness the potential of all data.

[3] Additionally, using the bioclimatic cubes and ResNet-18, ResNet-34, and ResNet-50, we attained sample-averaged F_1 scores of 0.251, 0.245, and 0.252, respectively.

multimodal model. Building upon the single-modality models described earlier, we created a multimodal model that combines them. The model integrates the ResNet-18-based models for the bioclimatic and Landsat time series data and the modified Swin-v2-t model for the Sentinel-2 data. A Multi-Layer Perceptron (MLP) is used as a head, enabling the straightforward combination of features extracted from all three models. This multimodal model achieved a notable sample-averaged F_1 score of 0.316. This significant improvement over the individual models' performances highlights the inherent multimodality of the task, demonstrating that combining different data sources can lead to more accurate and robust predictions.

3.2. Participants and Results

webmaking (Top1): The best-performing participant developed and combined four types of algorithms: (i) Predicting the N most frequent species per country, small rectangular regions, biogeographical regions and their combination (reaching a maximum $F_1 = 0.21$ with this approach only), (ii) Random Forests using the previous geographic features along with bioclimatic variables (maximum $F_1 = 0.25$ with the ensemble (i+ii)), (iii) XGBoost (maximum $F_1 = 0.37$ with (i+ii+iii)) and (iv) our multimodal baseline (maximum $F_1 = 0.41$ with (i+ii+iii+iv)). Notably, the performance gains obtained by ensembling were not only due to the combination of model types but also to the optimization of the predicted number of species per survey and how he combined the different predicted species sets. For instance, the ensemble of (i+ii) improved from 0.25 to 0.3 just by optimizing the weighting of models in the ensemble prediction. Besides, by exploring species weighting schemes to reinforce species predicted by several models or down-weight the predicted species pairs unlikely to co-occur among PAs, he gained 0.02.

AI2Lab team (Top2) [21]: This team started from the multimodal model provided as the baseline by the organizers, to which they made several significant improvements: (i) addition of a fourth modality (i.e., tabular environmental data encoded with an MLP), (ii) use of PO data samples through a pseudo-labeling procedure, (iii) use of an improved encoder for the Sentinel-2 images (pre-trained with self-supervised learning on an external dataset), (iv) use of an ensemble of models optimized on different folds, and (v) optimization of the detection threshold. They finally got an F_1 score of 0.368 on the private leaderboard. The most significant gains have been obtained by the use of the ensemble of models (+0.021), the optimization of the detection threshold (+0.012), and the use of PO data samples through pseudo-labeling (+0.008).

Miss Qiu (Top3) [24]: This team initially reused the multimodal model baseline, but opted for a different fusion method utilizing cross-attention instead of MLP. This adjustment resulted in a slight improvement in performance. They incorporated several enhancements, some of which were similar to the AI2Lab team's approach (e.g., utilizing an ensemble of k-fold models), while also introducing unique methods, such as (i) enriching predictions with species commonly found in neighboring PA and PO samples, (ii) optimizing the number of returned species, and (iii) employing various data augmentation techniques, including mixup. Their final F_1 score on the private leaderboard was 0.353.

BernIgen (Top5) [23]: This team started working primarily on a model using only tabular data based on the XGBoost method (known to work very well on classical species distribution models). They have previously reduced the dimensionality of the input data with a PCA (Principal Component Analysis) and the number of output species by keeping only the most likely species (about 10%). This model alone already delivers pretty good performance (F_1 score of 0.31). They improved prediction performance by adaptively predicting the number of species to return for each test plot using a regression model (also based on XGBoost). This strategy led to a significant improvement in the F_1 score, gaining an additional point. Finally, we combined this model with the multimodal model provided by the organizers, achieving an F_1 score of 0.349 on the private leaderboard.

Lonan Syayf (Top9) [26]: This participant started with the provided multimodal baseline and modified the architecture on the Landsat time-series, substituting the ResNet-based model in the baseline by a 3-layer MLP that takes the Landast time-series and bioclimatic variables as a single input vector. This improves the private F_1 from 0.316 to 0.323, with a small additional improvement to 0.329, by removing all species with less than 10 observations. In addition, the number of species reported is made variable by predicting as present those that have a score higher than a tunable threshold, allowing to further improve the score to 0.342.

Wei Dai (Top10) [22]: This team used the provided multimodal baseline and focused primarily on the least performing modality – Sentinel-2 image patches. They tried different architectures (e.g., ConvNeXt [29], MaxVit [30], and Swin-v2 [28]) and their ensembling. By ensembling four models, they reduced the relative error of the provided baseline by around 15%; the best single model – MaxVit-t – reduced the relative error by 12.5%.

DS@GT-GeoLifeCLEF (Top48) [25]: This team explored various methods, including (i) efficient nearest neighbor search using Locality Sensitive Hashing, (ii) training convolutional neural networks on DCT coefficients instead of raw pixels, and (iii) utilizing Tile2Vec [31], a self-supervised learning technique that generates embeddings of satellite imagery tiles. They encountered and reported various difficulties with model convergence and could not achieve results surpassing an F_1 score of 0.16.

4. Discussion and Conclusion

The main outcomes we can derive from the GeoLifeCLEF 2024 are the following:

Provided baselines had a positive impact on overall performance. Compared to last year, with a single team outperforming the best baseline, this year, 25 participants achieved similar or better performance than the provided baselines. We assume that the continuous process of publishing better baselines increased the participation engagement since allowed continuous improvements.

Proactive engagement with the community and continual release of better baselines increased the impact. Similarly, as in the previous case, the proactive engagement with the community through the Kaggle forum allowed crowd-sourcing of methods and continuous incremental performance increase. Compared to other LifeCLEF and FGVC competitions, the GeoLifeCLEF 2024 competition got 10–100 times more participants and submissions.

Multimodal is more than single-modal. The multimodal models were the key to success. All participants' working notes reported large performance gains by combining models based on different modalities, regardless of the type of algorithm used to exploit these modalities (e.g., random forests, XGBoost, CNNs). Yet, the potential of the high dimensional remote sensing data was only exploited by deep learning-based models.

Accounting for species community capacity is key. Our baseline development showed that the sum of species presence probabilities predicted independently of each other largely over-predicted the actual number of species. Most top teams developed techniques to constrain or predict the number of present per species community before applying a probability ranking rule [32]. Given the high spatial resolution considered here, it is coherent with the important local constraints of species assemblages, such as competition.

High amount of Presence-Absence (PA) training data positively influence the results. Methods based on the PA data consistently outperformed the ones based solely on the PO data. Some minor gains were reported in combining PA and PO data through specific methods accounting for sampling biases. The provision of more PA data in the training dataset may have contributed to a much higher performance compared to last year's edition (for which the best F_1 score was 0.27). Many challenges remain in developing data integration methods compatible with machine learning modeling pipelines.

For the future, it seems important to understand why improving performance with Presence-Only data is difficult, even though it is much larger. The presence of observation bias is clearly a plausible reason (some species are observed more than others), but it seems the spatial scale of the test set's plots may also be an issue. They are indeed quite small (10×10m on average) and do not necessarily reflect the presence of all the species in larger areas such as the one considered by the models. Moreover, the locations of these plots themselves follow specific protocols, which may introduce observation biases different from those of Presence-Only data.

Interestingly, the development of multimodal models highlights the importance of leveraging diverse data inputs in environmental monitoring and analysis. This approach not only sets a promising precedent but also opens up exciting possibilities for future work in the field, suggesting that multimodal data fusion can substantially enhance the performance of predictive models in complex, real-world tasks. It also highlights the inherent difficulty in balancing the contributions of each modality to the overall prediction accuracy, especially when dealing with high-dimensional and diverse data inputs. We recognize these challenges and are committed to addressing them in our future work.

Acknowledgement

The research described in this paper was funded by the European Commission via the MAMBO (http:doi.org/10.3030/101060639) and GUARDEN (http:doi.org/10.3030/101060693) projects, which have received funding from the European Union's Horizon Europe research and innovation program under grant agreements 101060693 and 101060639.

References

[1] E. S. Brondizio, J. Settele, S. Diaz, H. T. Ngo, Global assessment report on biodiversity and ecosystem services of the intergovernmental science-policy platform on biodiversity and ecosystem services, IPBES secretariat, Bonn, Germany. (2019).

[2] C. Botella, A. Joly, P. Bonnet, P. Monestiez, F. Munoz, A deep learning approach to species distribution modelling, Multimedia Tools and Applications for Environmental & Biodiversity Informatics (2018) 169–199.

[3] B. Deneu, M. Servajean, P. Bonnet, C. Botella, F. Munoz, A. Joly, Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment, PLoS computational biology 17 (2021) e1008856.

[4] B. Deneu, A. Joly, P. Bonnet, M. Servajean, F. Munoz, Very high resolution species distribution modeling based on remote sensing imagery: how to capture fine-grained and large-scale vegetation ecology with convolutional neural networks?, Frontiers in plant science 13 (2022) 839279.

[5] J. Estopinan, M. Servajean, P. Bonnet, F. Munoz, A. Joly, Deep species distribution modeling from sentinel-2 image time-series: a global scale analysis on the orchid family, Frontiers in Plant Science 13 (2022) 839327.

[6] E. H. Boakes, P. J. McGowan, R. A. Fuller, D. Chang-qing, N. E. Clark, K. O'Connor, G. M. Mace, Distorted views of biodiversity: spatial and temporal bias in species occurrence data, PLoS biology 8 (2010) e1000385.

[7] N. J. Isaac, M. J. Pocock, Bias and information in biological records, Biological Journal of the Linnean Society 115 (2015) 522–531.

[8] T. Mesaglio, C. T. Callaghan, An overview of the history, current contributions and future outlook of inaturalist in australia, Wildlife Research 48 (2021) 289–303.

[9] C. Botella, B. Deneu, D. M. Gonzalez, M. Servajean, T. Larcher, C. Leblanc, J. Estopinan, P. Bonnet, A. Joly, Overview of geolifeclef 2023: Species composition prediction with high spatial resolution at continental scale using remote sensing, in: CLEF 2023: Conference and Labs of the Evaluation Forum, 2023.

[10] C. Botella, B. Deneu, D. Marcos Gonzalez, M. Servajean, T. Larcher, C. Leblanc, J. Estopinan, P. Bonnet, A. Joly, Overview of GeoLifeCLEF 2023: Species composition prediction with high spatial resolution at continental scale using remote sensing, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, 2023.

[11] W. Fithian, J. Elith, T. Hastie, D. A. Keith, Bias correction in species distribution models: pooling survey and collection data for multiple species, Methods in Ecology and Evolution 6 (2015) 424–438.

[12] D. A. Miller, K. Pacifici, J. S. Sanderlin, B. J. Reich, The recent past and promising future for data integration methods to estimate species' distributions, Methods in Ecology and Evolution 10 (2019) 22–37.

[13] A. Joly, L. Picek, S. Kahl, H. Goëau, V. Espitalier, C. Botella, B. Deneu, D. Marcos, J. Estopinan, C. Leblanc, T. Larcher, M. Šulc, M. Hrúz, M. Servajean, et al., Overview of LifeCLEF 2024: Challenges on species distribution prediction and identification, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2024.

[14] A. Joly, L. Picek, S. Kahl, H. Goëau, V. Espitalier, C. Botella, B. Deneu, D. Marcos, J. Estopinan, C. Leblanc, et al., Lifeclef 2024 teaser: Challenges on species distribution prediction and identification, in: European Conference on Information Retrieval, Springer, 2024, pp. 19–27.

[15] C. Botella, D. Benjamin, M. Diego Gonzalez, S. Maximilien, L. Théo, E. Joaquim, L. César, P. Bonnet, A. Joly, The GeoLifeCLEF 2023 Dataset to evaluate plant species distribution models at high spatial resolution across Europe, 2023. URL: https://hal.science/hal-04152362, the full dataset is freely available at the link below (perennial repository) for academic use or other non-commercial use: https://lab.plantnet.org/seafile/d/936fe4298a5a4f4c8dbd/.

[16] D. R. Roberts, V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, J. J. Lahoz-Monfort, B. Schröder, W. Thuiller, et al., Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure, Ecography 40 (2017) 913–929.

[17] GBIF.Org User, Occurrence download, 2022. URL: https://www.gbif.org/occurrence/download/0144742-220831081235567. doi:10.15468/DL.8WVZQF.

[18] M. A. Friedl, D. Sulla-Menashe, B. Tan, A. Schneider, N. Ramankutty, A. Sibley, X. Huang, Modis collection 5 global land cover: Algorithm refinements and characterization of new datasets, Remote sensing of Environment 114 (2010) 168–182.

[19] O. Venter, E. W. Sanderson, A. Magrach, J. R. Allan, J. Beher, K. R. Jones, H. P. Possingham, W. F. Laurance, P. Wood, B. M. Fekete, et al., Global terrestrial human footprint maps for 1993 and 2009, Scientific data 3 (2016) 1–10.

[20] D. N. Karger, O. Conrad, J. Böhner, T. Kawohl, H. Kreft, R. W. Soria-Auza, N. E. Zimmermann, H. P. Linder, M. Kessler, Climatologies at high resolution for the earth's land surface areas, Scientific data 4 (2017) 1–20.

[21] Y. Chen, T. Peng, W. Li, C.-S. Chen, Combining present-only and present-absent data with pseudo-label generation for species distribution modeling, in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, 2024.

[22] Z. Cheng, W. Dai, J. Sun, Multi-modal feature fusion networks for geolifeclef 2024, in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, 2024.

[23] T. Chopard, D. Rawlings, Exploring biodiversity: A multi-model approach to multi-label plant species prediction, in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, 2024.

[24] H. Liu, Z. Tao, P. Jiang, Q. Sun, M. Wan, Tighnari: Multi-modal plant species prediction based on hierarchical cross-attention using graph-based and vision backbone-extracted features, in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, 2024.

[25] A. Miyaguchi, P. Aphiwetsa, M. McDuffie, Tiled compression and embeddings for multilabel classification in geolifeclef 2024, in: Working Notes of CLEF 2024 - Conference and Labs of the

Evaluation Forum, 2024.

[26] A. Syayfetdinov, Multimodal networks for species distribution modeling, in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, 2024.

[27] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[28] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, et al., Swin transformer v2: Scaling up capacity and resolution, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 12009–12019.

[29] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 11976–11986.

[30] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, Y. Li, Maxvit: Multi-axis vision transformer, in: European conference on computer vision, Springer, 2022, pp. 459–479.

[31] N. Jean, S. Wang, A. Samar, G. Azzari, D. Lobell, S. Ermon, Tile2vec: Unsupervised representation learning for spatially distributed data, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 3967–3974.

[32] M. D'Amen, A. Dubuis, R. F. Fernandes, J. Pottier, L. Pellissier, A. Guisan, Using species richness and functional traits predictions to constrain assemblage predictions from stacked species distribution models, Journal of biogeography 42 (2015) 1255–1266.