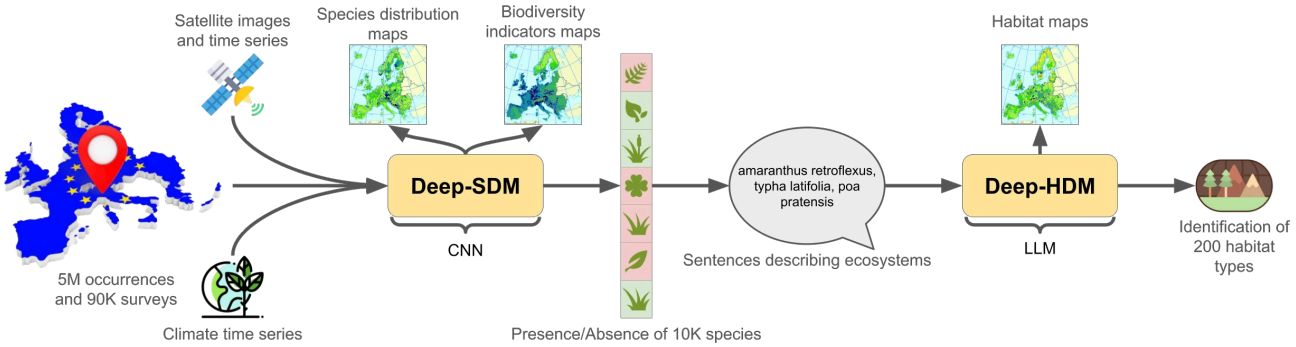
Mapping biodiversity at very-high resolution in Europe

César Leblanc[1], Lukas Picek[1], Benjamin Deneu[2], Pierre Bonnet[3],
Maximilien Servajean[4], Rémi Palard[3], and Alexis Joly[1]
[1] INRIA, [2] WSL, [3] CIRAD, and [4] LIRMM



Abstract

This paper describes a cascading multimodal pipeline for high-resolution biodiversity mapping across Europe, integrating species distribution modeling, biodiversity indicators, and habitat classification. The proposed pipeline first predicts species compositions using a deep-SDM, a multimodal model trained on remote sensing, climate time series, and species occurrence data at 50×50m resolution. These predictions are then used to generate biodiversity indicator maps and classify habitats with Pl@ntBERT, a transformer-based LLM designed for species-to-habitat mapping. With this approach, continental-scale species distribution maps, biodiversity indicator maps, and habitat maps are produced, providing fine-grained ecological insights. Unlike traditional methods, this framework enables joint modeling of interspecies dependencies, bias-aware training with heterogeneous presence-absence data, and large-scale inference from multi-source remote sensing inputs.

1. Introduction

Mapping biodiversity at high spatial resolution is essential for monitoring ecosystem health, assessing species distributions, and guiding conservation policies [24, 27, 44]. Effective biodiversity mapping enables the early detection of habitat loss, ecosystem degradation, and climate-induced changes in species ranges, providing crucial information for ecological research and decision-making [3, 40, 43].

However, generating such maps at a continental scale with fine spatial detail remains a significant challenge due to the limited availability of structured *in situ* data (i.e., data collected directly in the field), spatial biases in species observations, and the complex relationships between species and environmental factors [14, 18, 37].

A standard approach to tackle these challenges is integrating publicly available species occurrence data, ecological surveys, and remote sensing datasets. Citizen science platforms such as GBIF, Pl@ntNet, and iNaturalist provide large-scale species presence records [4, 12], while comprehensive biological surveys like EVA offer detailed vegetation data, including species composition and habitat characteristics. Additionally, remote sensing data from satellites such as Sentinel and Landsat enable large-scale biodiversity assessments by capturing environmental variables (e.g., precipitation, temperature, and soil) [32] at high spatial and temporal resolutions. To *convert* these sources to biodiversity maps, Species Distribution Models (SDMs) are widely used [23]. They predict species occurrence by analyzing the relationship between observed records and environmental conditions. Traditional approaches, such as MAXENT [45] and Random Forest [49] models, rely on statistical correlations but face challenges (e.g., spatial biases, low resolution, and an inability to model species interactions). Recent advances in deep learning-based SDMs (deep-SDMs) overcome these limitations by integrating multi-source data and capturing complex ecological dependencies, resulting in more accurate and scalable biodiversity predictions [15].

This work introduces a cascading multimodal pipeline that integrates SDM and Habitat Distribution Modeling (HDM) to generate high-resolution European biodiversity maps. Our approach leverages a deep-SDM, a multimodal model trained on remote sensing (Sentinel-2, Landsat), climate time series, and in situ species observations to predict species compositions at a 50×50m resolution. These predictions form the foundation for computing biodiversity indicator maps, capturing key ecological metrics. Finally, we apply Pl@ntBERT [35], a transformer-based species-to-habitat classifier, to infer habitat types based on species assemblages, improving habitat mapping beyond traditional remote sensing-based approaches. Unlike conventional SDMs, which treat species independently and rely on handcrafted environmental features, our deep-SDM models interspecies dependencies, mitigates spatial biases, and enables large-scale inference using heterogeneous presence-absence data. By incorporating HDM, our method extends beyond species distributions to produce detailed habitat maps, providing a more comprehensive view of ecosystem dynamics. This framework offers a scalable and fine-grained solution for biodiversity monitoring, delivering high-resolution species distribution, biodiversity indicators, and habitat maps at a continental scale.

2. Related Work

Accurate biodiversity and habitat mapping have traditionally relied on habitat suitability models [23] or direct classification from remote sensing data [1]. However, these methods are often constrained by limited spatial resolution [19], outdated reference datasets, and the inability to model interspecies relationships. Some rare studies combine deep learning, citizen science data, and remote sensing to track plant species changes [22]. Nevertheless, they are usually geographically restricted to a country.

Since mapping requires models that can predict species distributions and classify habitats. Traditional SDMs estimate species occurrence probabilities using environmental variables, while HDMs focus on habitat classification by analyzing species composition. This section summarizes key facts about SDMs, deep-SDMs, and HDMs, highlighting their strengths, limitations, and relevance to our approach.

Species distribution models (SDMs) predict where species are likely to occur by analyzing relationships between species observations and environmental conditions. Traditional SDMs approaches, i.e., MAXENT and Random Forests, rely on statistical correlations but face limitations, including spatial biases, low resolution, and the inability to model interactions between species [7, 46]. These weaknesses limit their effectiveness, especially when working with large-scale and complex ecosystems. To overcome these challenges, deep-SDMs integrate remote sensing, climate data, and species occurrences to improve prediction accuracy [5, 14, 17, 47, 51]. Unlike traditional SDMs, deep-SDMs can learn complex spatial patterns and ecological relationships using CNNs or transformers. This enables higher-resolution predictions at a large scale, making species distribution modeling more precise and scalable.

Habitat distribution models (HDMs) traditionally rely on expert systems [41] and machine learning [26]. Expert systems, though widely used [53], often overfit, making classification sensitive to minor plot variations, and sometimes require external criteria beyond species composition [13]. Machine learning models (i.e., NNs) [34], capture complex species composition patterns [8] but treat all species as equally different, failing to model ecological interdependencies [42]. While classical approaches are interpretable, they struggle with high-dimensional data. Deep learning, particularly transformers [55], has shown promise in biology, e.g., protein structure prediction [31], but remains underexplored in vegetation classification. Their ability to model global dependencies makes them a promising alternative for habitat classification.

3. Methodology

Dataset. To construct the maps, we use GeoPlant [47], a new European-scale dataset (see Fig. 1) designed for high-resolution species distribution modeling. GeoPlant covers over 11,000 plant species, i.e., most of the European flora, and is based on 5 million opportunistic Presence-Only (PO) records from GBIF and 90,000 exhaustive Presence-Absence (PA) surveys from the European Vegetation Archive (EVA). Besides, for each plant species observations, Sentinel-2 RGB and NIR satellite images with 10m resolution, a 20-year time series of climatic variables (i.e., precipitation and mean, min, and max month temperature), and satellite time series from the Landsat program (i.e., R, G, B, NIR, and SWIR1+2) are provided. Coordinates were not used as we want to reflect habitat suitability (i.e., learn a relationship between environment and occurrences) [11].

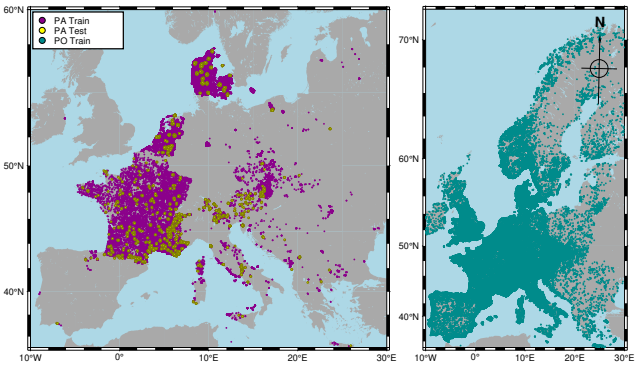


Figure 1. Geo spatial scale of the dataset (from [47]). The 5M PO occurrences (9,709 species) span all of Europe, but the 90K PA surveys (5,016 species) are primarily in France and Denmark.

The provided PO and PA data were aggregated into a 50×50m spatial grid, consistent with the resolution used during inference. This aggregation combines both data types into a single site occupancy dataset, where each grid cell contains a 1 or 0 per species, representing presence or absence. This setup allows using a Binary Cross-Entropy loss function, which is better suited for presence-absence probability estimation than Categorical Cross-Entropy. Additionally, a target group background approach [46] was applied to partially correct sampling bias [2]. To achieve this, training is restricted to grid cells containing at least one recorded species, ensuring that pseudo-absence points are sampled only from locations where other species have been observed. This method helps compensate for the lack of explicit absence data, improving the ecological relevance of background points.

Species Distribution Modeling. Our approach is based on deep multi-modal models, which have been shown to outperform classical SDMs [28–30]. The mapping process consists of two main phases: (i) training a deep-SDM using *in situ* observations combined with spatialized environmental and remote sensing data and (ii) inferring the trained model to predict species distributions across Europe.

We use a multi-modal ensemble approach (see Fig. 2), building on previous work [6, 33, 48], based on a modified ResNet-6 architecture with three separate branches for different input data types: (i) Sentinel-2 RGB+NIR imagery (128×128 patches at 10m resolution), (ii) Climate time series encoded as three-dimensional data cubes (year, month, and variables such as precipitation and temperature), and (iii) Landsat remote sensing time series, structured similarly with spectral bands (R, G, B, NIR, and SWIR1+2). Each input modality is encoded by a dedicated CNN encoder with six residual blocks, a design choice that improves performance over larger off-the-shelf architectures [47]. The extracted embeddings are concatenated and passed through a fully connected classifier which computes species presence probabilities using one fully connected layer with a sigmoid activation function. The model is trained using Stochastic Gradient Descent (SGD) with binary cross-entropy loss. The training code is available on GeoPlant GitHub.

Biodiversity Indicator Calculation. The biodiversity indicators are extracted from the species assemblages predicted by the SDM at a 50×50m resolution across Europe. These indicators summarize ecological properties such as species richness and the presence of specific taxonomic or functional groups, providing valuable insights into biodiversity patterns. To derive these assemblages, the species probabilities predicted by the SDM are thresholded using a conformal prediction approach [21]. This method ensures a low probability of omitting truly present species, even if

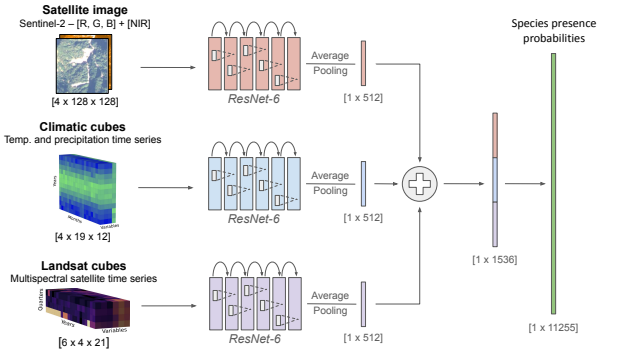


Figure 2. Selected SDM architecture (from [47]). This multi-modal ensemble model processes each modality (i.e., satellite images, climatic cubes, and Landsat cubes) through a lightweight 6-layer residual encoder (i.e., ResNet-6). The embeddings are then concatenated and passed to a final classification layer.

it results in some false positives. This conservation-focused strategy prioritizes minimizing omission errors and reducing the risk of underestimating species distributions, which is critical for biodiversity assessments.

We define seven biodiversity indicators from the predicted species assemblages to assess ecosystems' conservation status. These indicators capture key ecological and regulatory aspects, e.g.,

- **Species richness**: number of species.
- **EU directive**: number of species from the list provided by the EU Habitat directive.
- **Threatened species**: number of IUCN Red List species.
- **Most threatened**: IUCN threatened species status[1].
- **Tree species**: number of species from the lifeform "woody" of the Plan Of the World Online database.
- **Invasive species**: number of species from the CABI list.
- **Specialist species**: number of species estimated to be present with a very low probability of presence elsewhere.

Consequently, any indicator $n_S(x)$ relying on a number of present species among $|S|$ species of a particular type, can be modelled as a statistical variable following a Poisson binomial distribution (i.e., a sum of independent Bernoulli trials that are not necessarily identically distributed). Thus, the mean of $n_S(x)$ can be estimated as

$$\mu_S(x) = \sum_{i \in S} p(y_i = 1|x), \tag{1}$$

where S is the set of species of interest (e.g., endangered species) and x is a particular point of the map (i.e., a cell of $50 \times 50m$). The variance of $n_S(x)$ can be computed as

$$\sigma_S(x)^2 = \sum_{i \in S} p(y_i = 1|x) \cdot (1 - p(y_i = 1|x)), \tag{2}$$

[1] In the case of missing IUCN status, they were inferred using an automated method [56] also based on neural networks.

from which we can derive a confidence interval for each point of the map, e.g., through the 2-sigma rule:

$$\delta_S(x) = 2\sigma_S(x). \tag{3}$$

For an indicator based on $|S| = 10$ species with probabilities $p(y_1 = 1|x) = 0.9$, $p(y_2 = 1|x) = 0.8$, $p(y_3 = 1|x) = 0.1$, and $p(y_i = 1|x) = 0$ for $i \in [4, 10]$, we obtain $\mu_S(x) = 1.8$ and $\delta_S(x) = 1.1$, giving

$$n_S(x) = 1.8 \pm 1.1. \tag{4}$$

Thus, we can build a confidence interval map for almost all indicators (using $\delta_S(x)$ as the value for each point). Only the IUCN status of the most threatened species does not follow this pattern. For this one, we want to estimate the probability that at least one species of a particular IUCN status is present. If, for instance, we consider the set $S = EN$ of species with status ENDANGERED, the probability that at least one of them is present is equal to

$$p(n_{EN}(x) > 1|x) = 1 - \prod_{i \in EN} (1 - p(y_i = 1|x)). \tag{5}$$

If we have $|EN| = 10$ species and $p(y_1 = 1|x) = 0.9$, $p(y_2 = 1|x) = 0.8$, $p(y_3 = 1|x) = 0.1$ and $p(y_i = 1|x) = 0$ for $i \in [4, 10]$, then the probability that at least one ENDANGERED species is present is 98.2%.

Habitat Identification. Unlike traditional approaches that train models on satellite imagery labeled with EUNIS habitat types[2] [52], we infer habitats from the species assemblages predicted by the deep-SDM. Direct habitat classification from remote sensing is limited by the scarcity and outdated nature of labeled datasets, as most available EUNIS labels come from EVA surveys with a mean collection year of 1992. Many labeled sites have undergone significant ecological changes due to land-use transformation and climate change, making direct mapping unreliable. We follow the latest version of the EUNIS classification [10] and focus on levels 1, 2, and 3, the last being the most detailed.

Instead, since the primary value of EVA lies in its plant species assemblage data, we take a different approach: training a supervised model to predict EUNIS habitat types based on species composition. This method is less affected by temporal shifts in habitat labels because species assemblages remain a strong predictor of habitat type, even when direct habitat labels become outdated [34]. If the deep-SDM accurately predicts species assemblages at a given site, habitat types can be inferred with high confidence.

[2]The EUNIS habitat classification [39] is a hierarchical system for the categorization of natural and semi-natural habitats in Europe developed to support biodiversity management, conservation, and sustainable use.

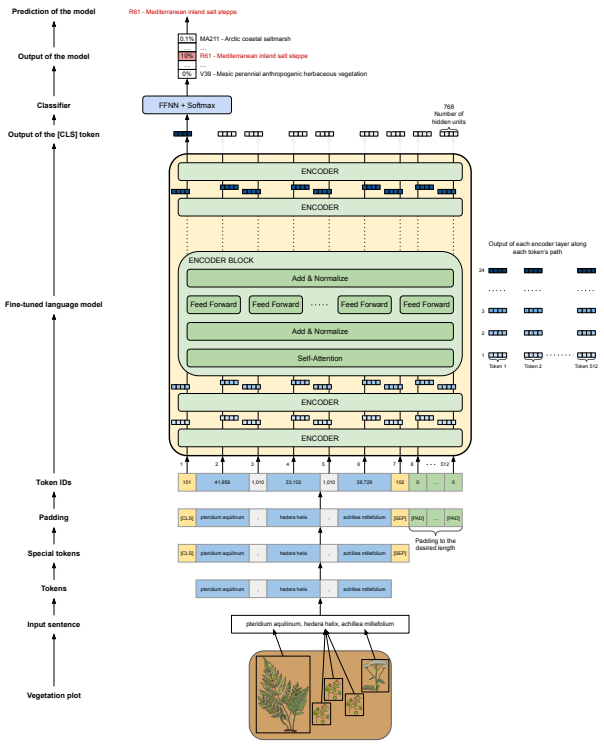


Figure 3. Pl@ntBERT HDM (from [35]) processes the input (list of species predicted by the deep-SDM) through multiple encoder layers, with the [CLS] token representation passed to a classifier to predict the most likely habitat type.

To implement this approach, we use Pl@ntBERT, a Python-based framework for training, sharing, and evaluating species-to-habitat classification models. Pl@ntBERT leverages large language models (LLMs), which have demonstrated strong performance in modeling plant species relationships [36]. It is built upon BERT, originally designed for natural language understanding [16], but adapted to capture latent dependencies between plant species in different ecosystems [38]. The model is trained in two stages:

Species-to-Species prediction: Given a predicted species assemblage, Pl@ntBERT learns to recover missing species by training on incomplete species lists. This step refines its understanding of species co-occurrence patterns [25].

Species-to-Habitat Classification: Fine-tuned on species assemblages from the deep-SDM, the model predicts the most probable EUNIS habitat type based on a sorted list of species by estimated spatial coverage (see Fig. 3).

Pl@ntBERT provides an efficient and scalable solution for habitat classification, leveraging the predictive power of species assemblages rather than relying on direct but potentially outdated habitat labels. The source code for training and inference is available on Pl@ntBERT GitHub.

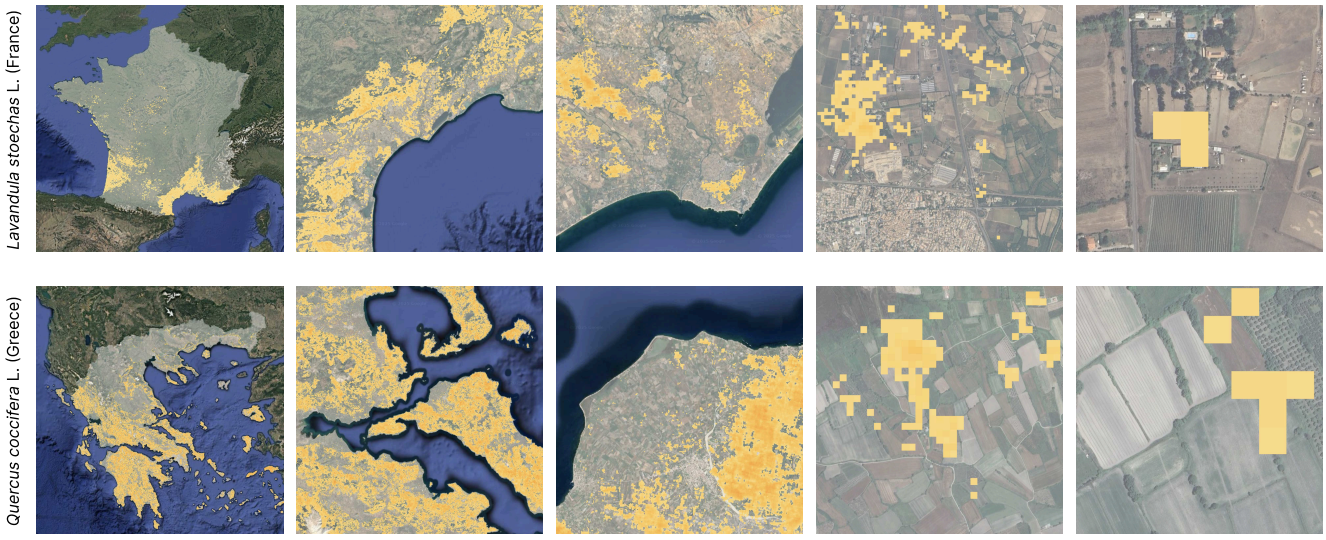


Figure 4. Example species distribution maps for two selected species occurring in France and Greece at different zoom levels. These maps are produced by the deep-SDM all over Europe for over 5,500 plant species at a 50×50m resolution.

4. Inference Details and Results

Species Distribution Maps. The trained SDM was used to generate high-resolution species distribution maps across Europe (see Fig. 4). To ensure scalability, the study area was divided into 25×25km meta-tiles, each processed independently. Within each tile, species predictions were made at 50×50m grid, totaling 5.5 billion cells. If a cell's center fell in water, it was moved to the nearest terrestrial point.

Inference was done for the year 2021, using environmental data averaged between March 21 and December 1, 2021. Unlike in training, where data cubes were extracted based on observation dates (hence capturing seasonal or interannual variability), inference used a fixed reference period. At each inference point, the model predicted presence probabilities for 11,255 species, which were then thresholded to retain only likely present species, significantly reducing storage requirements. The threshold was optimized on the validation set to maximize the F-score.

Maps were generated *just* for 5,558 out of 11,255 species. This does not necessarily indicate species absence but rather that their predicted probability remained below the confidence threshold. On average, a species was predicted in 132.8 million grid cells, covering 332,000 km² (2.4% of Europe). The most widespread species, *Agrostis capillaris L.*, appeared in 3.23 billion grid cells (58.6%).

To evaluate model performance, a spatial block hold-out split (10×10km grid) was used to mitigate spatial autocorrelation and assess generalization [50]. This approach was chosen as a realistic test of spatial interpolation based on species occurrence distribution. Each input modality was also evaluated separately in order to demonstrate their own predictive power, with Landsat data resulting in the highest value. See Tab. 1 for detailed evaluation.

The multi-modal model achieves a high AUC score [20] of 0.931, indicating strong performance in ranking true presence sites higher than true absence sites. This suggests that the predicted species distribution maps closely align with actual species occurrences. However, the F-score [54], which requires the model to predict the exact species assemblage for each test plot, is relatively low at 0.338.

A major limitation arises from the scale mismatch between the test vegetation plots and the predicted grid cells. The targeted resolution is 50×50m (2,500m^2), whereas test plots average 100m^2, meaning they contain significantly fewer species. As a result, many species predicted by the model may be considered false positives (i.e., are over-predicted) at the test plot scale, even if they are present at the full 2,500m^2 resolution.

Note: *The full workflow required approximately 30,000 GPU hours on Nvidia A100 GPUs, producing 15TB of data.*

Table 1. Evaluation of the SDM. The multimodal ensemble approach achieves a considerable performance improvement compared to the single modality models in terms of all metrics.

Branch	AUC	F-score	Recall@50	Recall@250
Sentinel	0.898	0.258	0.524	0.848
Bio	0.891	0.273	0.544	0.872
Landsat	0.920	0.312	0.595	0.873
All	**0.931**	**0.338**	**0.639**	**0.908**

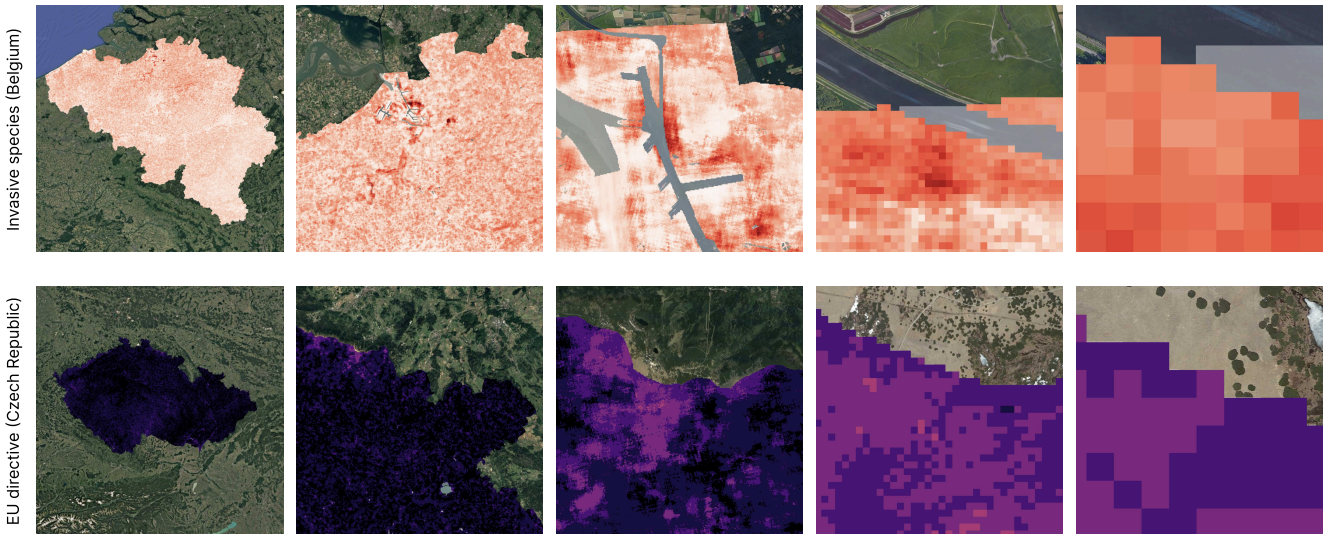


Figure 5. Example biodiversity indicator maps for two selected indicators occurring in Belgium and the Czech Republic at different zoom levels. These maps are produced with the output of the deep-SDM all over Europe for seven biodiversity indicators at a 50×50m resolution.

A more precise evaluation of recall and precision would require a complete ground-truth dataset at the 2,500m^2 scale, which is impossible due to the extreme effort required for manual surveys, or to predict species compositions at a 10×10m resolution, which approximately means multiplying the number of grid cells by 25. Instead, we use Recall@K to evaluate the model's ability to recover species despite the resolution mismatch. Since the exact number of species in a 50×50m cell is unknown, K=50 and K=250 serve as proxies for low- and high-diversity areas. The model retrieves nearly two-thirds of species for K=50 and over 90% for K=250, demonstrating strong recall despite the spatial scale limitations. Favoring recall at the expense of false positive ensures species are not missed (key in conservation), even if precision drops.

Biodiversity Indicators Maps. The workflow to create the high-resolution indicator maps at the European scale (see Fig. 5) is closely related to the one used for producing the species distribution maps based on the SDM. The meta-tiles of size 25×25km are processed one by one (in parallel), and within each tile, the indicators are computed for each point of the 50×50m grid based on the species assemblage predicted by the SDM. For most indicators, the two main operations are (i) filtering the species of interest for the targeted indicator and (ii) counting the number of filtered species. This can be implemented very efficiently on a GPU through the use of binary masks and the sum of tensor values. Only the indicator "IUCN status of the most threatened species in the assemblage" requires a slightly different process, but that was efficiently implemented by encoding status as integers and using look-up tables and a max operator. So far, all 7, i.e., (i) Species richness, (ii) EU directive, (iii) Threatened species, (iv) Most threatened, (v) Tree species, (vi) Invasive species, and (vii) Specialist species, biodiversity indicator maps have been produced.

Habitat Maps. The habitat maps (see Fig. 6) were inferred following a workflow similar to that used for biodiversity indicators. The study area was divided into 25×25km meta-tiles, which were processed in parallel. Within each tile, the model classified each 50×50m grid cell based on the species probabilities predicted by the SDM. The classifier directly assigns EUNIS Level 3 habitat types, while Levels 1 and 2 are inferred from the hierarchy.

In total, 200 habitat maps were generated at EUNIS Level 3, covering 60.4% of all habitat types at this level. On average, a habitat was mapped across 27.77 million grid cells, corresponding to 0.49% of Europe's total area. The most widespread habitat, R22: "Low and medium altitude hay meadow", was predicted in 681.5 million grid cells, covering 12.27% of Europe.

Experiments have shown that Pl@ntBERT, through its ability to model complex inter-species relationships, is able to outperform expert systems [9] (+5.54%) and tabular deep learning [34] (+1.14%) methods. Overall, the measured accuracy was 76% at level 1 of the EUNIS classification (8 broad habitat groups covered), 63% at level 2 (34 habitat groups covered), and 45% at level 3 (200 habitat types covered). In Table 2, we report the full performance evaluation with respect to the number of species that has been kept from the SDM predictions.

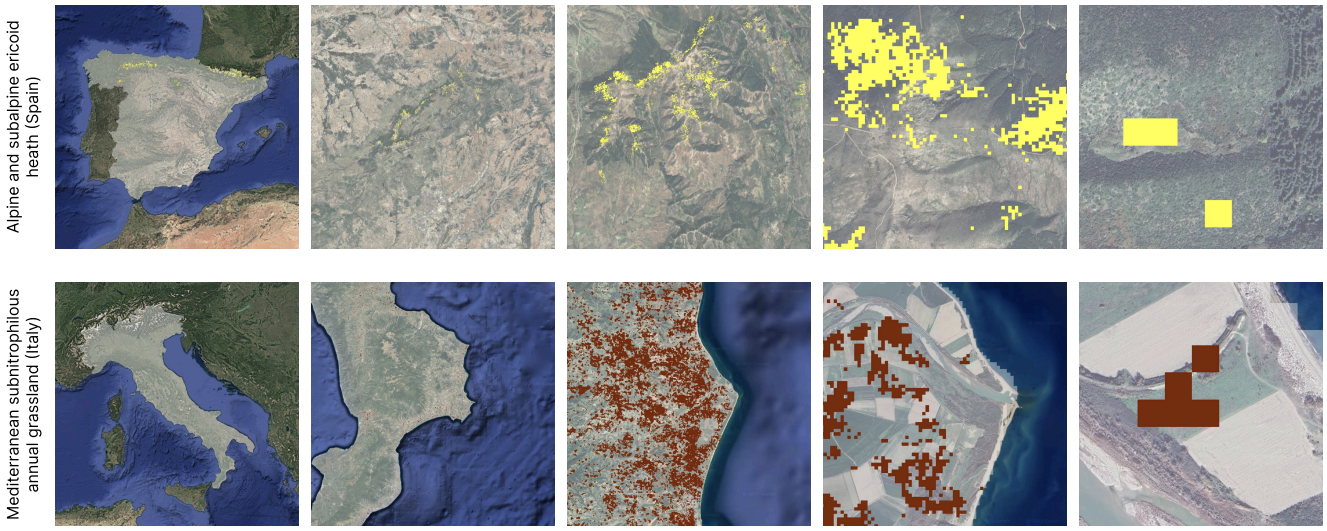


Figure 6. Example habitat maps for two selected habitats occurring in Spain and Italy at different zoom levels. These maps are produced by Pl@ntBERT with the output of the deep-SDM all over Europe for over 200 habitat types at a 50×50m resolution.

Table 2. Evaluation of the HDM. Accuracy is reported at all three hierarchical levels of EUNIS habitat classification (level 1 being the broader and level 3 the finer). Retaining more species from the deep-SDM predictions slightly improves classification performance across all levels.

Top-SDM predictions	Level 1	Level 2	Level 3
First 50 species	75.05%	61.29%	42.78%
First 100 species	**76.30%**	**62.68%**	**44.72%**

The model benefits from the fact that the SDM provides richer information, i.e., a calibrated softmax. Those probabilities are used directly as input in Pl@ntBERT, with predicted species being ordered in descending probability order in each sentence. This is a "reciprocal rank encoding method" but uses the probability score as the ranking function instead of the spatial coverage.

5. Conclusion

This work presents a multi-modal deep learning framework based on species distribution modeling (SDM), biodiversity indicators calculation, and habitat classification for high-resolution biodiversity mapping across Europe. Using remote sensing, climate variables, and species occurrence data, we provide a comprehensive, fine-scale view on species distributions, ecosystem diversity, and habitat types at an unprecedented 50×50m resolution at this scale. Our approach enables previously infeasible large-scale ecological assessments, offering new tools for biodiversity monitoring, conservation planning, and land-use management.

The Species Distribution Maps, generated using deep-SDM, effectively predict species occurrences by combining satellite imagery, climate time-series, and species records from GBIF and EVA. These maps provide baseline data for over 5,5k plant species, supporting efforts to track species distributions, monitor ecological shifts, and guide conservation policies.

The Biodiversity Indicator Maps provide insights into species richness, the presence of endangered or invasive species, as well as other key ecological metrics. These maps help identify biodiversity hotspots, vulnerable ecosystems, and priority areas for conservation.

The Habitat Maps, created by coupling SDM predictions with Pl@ntBERT, classify EUNIS habitat types across Europe. While these maps enhance the understanding of ecosystem distributions and habitat changes, challenges remain in classifying habitats at EUNIS Level 3, partly due to inconsistencies in expert-labeled training data.

Despite large contributions, several limitations remain. The reliance on species occurrence data from citizen science platforms and surveys introduces spatial biases, as certain regions and species are better documented than others (e.g, PO data are biased toward appealing species and PA data have limited geographic coverage). Additionally, prediction uncertainties persist, particularly in areas with low observation density or rapidly changing environmental conditions. The classification of habitats is further constrained by potential inconsistencies in EUNIS labeling, impacting the reliability of fine-scale habitat predictions. Finally, the multimodal nature and the size of the dataset require considerable computational resources for model training.

Acknowledgments

The research described in this paper was funded by the European Commission through the GUARDEN (safeGUARDing biodivErsity aNd critical ecosystem services across sectors and scales) and MAMBO (Modern Approaches to the Monitoring of BiOdiversity) projects. These projects received funding from the European Union's Horizon Europe research and innovation programme under grant agreements 101060693 (start date: 01/11/2022; end date: 31/10/2025) and 101060639 (start date: 01/09/2022; end date: 31/08/2026), respectively. Further models developed based on this methodology will directly meet the needs of the European biodiversity strategy for 2030 through those projects. They will be used in particular to enhance the biodiversity maps at the European scale. The content of this paper reflects the views only of the authors, and the European Commission cannot be held responsible for any use which may be made of the information contained therein.

The authors are grateful to the OPAL infrastructure from Université Côte d'Azur for providing resources and support. This work was granted access to the high-performance computing resources of IDRIS (Institut du Développement et des Ressources en Informatique Scientifique) under the allocation 2023-AD010113641R1 made by GENCI (Grand Equipement National de Calcul Intensif).

Our major thanks go to thousands of European vegetation scientists of several generations who collected the original vegetation-plot data in the field and made their data available to others and those who spent myriad hours digitizing data and managing the databases in the EVA. Vegetation plots data for this study were provided by Sylvain Abdulhak, Alicia Acosta, Emiliano Agrillo, Pierangela Angelini, Iva Apostolova, Olivier Argagnon, Fabio Attorre, Svetlana Aćić, Christian Berg, Ariel Bergamini, Erwin Bergmeier, Idoia Biurrun, Maxim Bobrovsky, Steffen Boch, Gianmaria Bonari, Anne Bonis, Zoltán Botta-Dukát, Jan-Bernard Bouzillé, Helge Bruelheide, Vanessa Bruzzaniti, Juan Antonio Campos, Andraž Čarni, Maria Laura Carranza, Laura Casella, Alessandro Chiarucci, Andrei Chuvashov, Milan Chytrý, János Csiky, Mirjana Krstivojević Ćuk, Renata Ćušterevska, Olga Demina, Jürgen Dengler, Panayotis Dimopoulos, Dmytro Dubyna, Tetiana Dziuba, Alexei Egorov, Rasmus Ejrnæs, Franz Essl, Jörg Ewald, Giuliano Fanelli, Federico Fernández-González, Úna FitzPatrick, Xavier Font, Gianpietro Giusso del Galdo, Emmanuel Garbolino, Itziar García-Mijangos, Rosario G Gavilán, Jean-Michel Genis, Michael Glaser, Valentin Golub, Friedemann Goral, Jean-Claude Gégout, Behlül Güler, Rense Haveman, Stephan Hennekens, Adrian Indreica, Maike Isermann, Ute Jandt, Jan Jansen, Florian Jansen, John Janssen, Anni Kanerva Jašková, Borja Jiménez-Alfaro, Martin Jiroušek, Veronika Kalníková, Ali Kavgacı, Larisa Khanina, Ilona Knollová, Vitaliy Kolomiychuk, Łukasz Kozub, Daniel Krstonošić, Helmut Kudrnovsky, Anna Kuzemko, Filip Küzmič, Zygmunt Kącki, Flavia Landucci, Igor Lavrinenko, Jonathan Lenoir, Armin Macanović, Corrado Marcenò, Aleksander Marinšek, Marco Massimi, Ruth Mitchell, Jesper Erenskjold Moeslund, Pavel Novák, Vladimir Onipchenko, Robin Pakeman, Hristo Pedashenko, Tomáš Peterka, Remigiusz Pielech, Vadim Prokhorov, Ricarda Pätsch, Aaron Pérez-Haase, Valerius Rašomavičius, Maria Pilar Rodríguez-Rojo, John Rodwell, Iris de Ronde, Eszter Ruprecht, Solvita Rūsiņa, Michele De Sanctis, Joop Schaminée, Joachim Schrautzer, Ingrid Seynave, Pavel Shirokikh, Jozef Šibík, Urban Šilc, Željko Škvorc, Desislava Sopotlieva, Angela Stanisci, Milica Stanišić-Vujačić, Zora Dajić Stevanović, Danijela Stešević, Jens-Christian Svenning, Grzegorz Swacha, Irina Tatarenko, Ioannis Tsiripidis, Ruslan Tsvirko, Pavel Dan Turtureanu, Domas Uogintas, Emin Uğurlu, Milan Valachovič, Kiril Vassilev, Roberto Venanzoni, Sophie Vermeersch, Risto Virtanen, Denys Vynokurov, Lynda Weekes, Wolfgang Willner, Thomas Wohlgemuth, Svitlana Yemelianova, and Dominik Zukal.

References

[1] Meisam Amani, Fatemeh Foroughnia, Armin Moghimi, Sahel Mahdavi, and Shuanggen Jin. Three-dimensional mapping of habitats using remote-sensing data and machine-learning algorithms. *Remote Sensing*, 15(17):4135, 2023. 2

[2] Robert A Barber, Stuart G Ball, Roger KA Morris, and Francis Gilbert. Target-group backgrounds prove effective at correcting sampling bias in maxent models. *Diversity and Distributions*, 28(1):128–141, 2022. 3

[3] Céline Bellard, Cleo Bertelsmeier, Paul Leadley, Wilfried Thuiller, and Franck Courchamp. Impacts of climate change on the future of biodiversity. *Ecology letters*, 15(4):365–377, 2012. 1

[4] Pierre Bonnet, Antoine Affouard, Jean-Christophe Lombardo, Mathias Chouet, Hugo Gresse, Vanessa Hequet, Remi Palard, Maxime Fromholtz, Vincent Espitalier, Hervé Goëau, et al. Synergizing digital, biological, and participatory sciences for global plant species identification: enabling access to a worldwide identification service. *Biodiversity Information Science and Standards*, 7, 2023. 1

[5] Christophe Botella, Benjamin Deneu, Diego Marcos, Maximilien Servajean, Joaquim Estopinan, Théo Larcher, César Leblanc, Pierre Bonnet, and Alexis Joly. The geolifeclef 2023 dataset to evaluate plant species distribution models at high spatial resolution across europe. *arXiv preprint arXiv:2308.05121*, 2023. 2

[6] Christophe Botella, Benjamin Deneu, Diego Marcos, Maximilien Servajean, Théo Larcher, César Leblanc, Joaquim Estopinan, Pierre Bonnet, and Alexis Joly. Overview of geolifeclef 2023: Species composition prediction with high spatial resolution at continental scale using remote sensing. In *CLEF 2023 Working Notes-24th Conference and Labs of the Evaluation Forum*, pages 1954–1971, 2023. 3

[7] Philipp Brun, Dirk N Karger, Damaris Zurell, Patrice Descombes, Lucienne C de Witte, Riccardo de Lutio, Jan Dirk Wegner, and Niklaus E Zimmermann. Multispecies deep

learning using citizen science data produces more informative plant community models. *Nature Communications*, 15 (1):4421, 2024. 2

[8] Lenka Cerna and Milan Chytrý. Supervised classification of plant communities with artificial neural networks. *Journal of Vegetation Science*, 16(4):407–414, 2005. 2

[9] Milan Chytrỳ, Lubomír Tichỳ, Stephan M Hennekens, Ilona Knollová, John AM Janssen, John S Rodwell, Tomáš Peterka, Corrado Marcenò, Flavia Landucci, et al. Eunis habitat classification: Expert system, characteristic species combinations and distribution maps of european habitats. *Applied Vegetation Science*, 23(4):648–675, 2020. 6

[10] Milan Chytrỳ, Marcela Řezníčková, Petr Novotnỳ, Dana Holubová, Zdenka Preislerová, Fabio Attorre, Idoia Biurrun, Petr Blažek, Gianmaria Bonari, Dariia Borovyk, et al. Floraveg. eu—an online database of european vegetation, habitats and flora. *Applied vegetation science*, 27(3):e12798, 2024. 4

[11] Elijah Cole, Grant Van Horn, Christian Lange, Alexander Shepard, Patrick Leary, Pietro Perona, Scott Loarie, and Oisin Mac Aodha. Spatial implicit neural representations for global-scale species mapping. In *International conference on machine learning*, pages 6320–6342. PMLR, 2023. 2

[12] Matteo Contini, Victor Illien, Mohan Julien, Mervyn Ravitchandirane, Victor Russias, Arthur Lazennec, Thomas Chevrier, Cam Ly Rintz, Léanne Carpentier, Pierre Gogendeau, et al. Seatizen atlas: a collaborative dataset of underwater and aerial marine imagery. *Scientific Data*, 12(1):67, 2025. 1

[13] Miquel De Caceres, Milan Chytry, Emiliano Agrillo, Fabio Attorre, Zoltan Botta-Dukat, Jorge Capelo, Balint Czucz, Jurgen Dengler, Jorg Ewald, Don Faber-Langendoen, et al. A comparative framework for broad-scale plot-based vegetation classification. *Applied Vegetation Science*, 18(4):543–560, 2015. 2

[14] Benjamin Deneu, Maximilien Servajean, Pierre Bonnet, Christophe Botella, François Munoz, and Alexis Joly. Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment. *PLoS computational biology*, 17(4), 2021. 1, 2

[15] Benjamin Deneu, Alexis Joly, Pierre Bonnet, Maximilien Servajean, and François 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*, 13:839279, 2022. 1

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 4

[17] Johannes Dollinger, Philipp Brun, Vivien Sainte Fare Garnot, and Jan Dirk Wegner. Sat-sinr: High-resolution species distribution models through satellite imagery. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 10:41–48, 2024. 2

[18] Jane Elith and John R Leathwick. Species distribution models: ecological explanation and prediction across space and time. *Annual review of ecology, evolution, and systematics*, 40(1):677–697, 2009. 1

[19] Joaquim Estopinan, Maximilien Servajean, Pierre Bonnet, Alexis Joly, and François Munoz. Mapping global orchid assemblages with deep learning provides novel conservation insights. *Ecological Informatics*, 81:102627, 2024. 2

[20] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006. 5

[21] Matteo Fontana, Gianluca Zeni, and Simone Vantini. Conformal prediction: a unified review of theory and new challenges. *Bernoulli*, 29(1):1–23, 2023. 3

[22] Lauren E Gillespie, Megan Ruffley, and Moises Exposito-Alonso. Deep learning models map rapid plant species changes from citizen science and remote sensing data. *Proceedings of the National Academy of Sciences*, 121(37): e2318296121, 2024. 2

[23] Antoine Guisan and Wilfried Thuiller. Predicting species distribution: offering more than simple habitat models. *Ecology letters*, 8(9):993–1009, 2005. 1, 2

[24] Antoine Guisan, Reid Tingley, John B Baumgartner, Ilona Naujokaitis-Lewis, Patricia R Sutcliffe, Ayesha IT Tulloch, Tracey J Regan, Lluis Brotons, Eve McDonald-Madden, Chrystal Mantyka-Pringle, et al. Predicting species distributions for conservation decisions. *Ecology letters*, 16(12): 1424–1435, 2013. 1

[25] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*, 2020. 4

[26] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009. 2

[27] Walter Jetz, Jana M McPherson, and Robert P Guralnick. Integrating biodiversity distribution knowledge: toward a global map of life. *Trends in ecology & evolution*, 27(3): 151–159, 2012. 1

[28] Alexis Joly, Christophe Botella, Lukáš Picek, Stefan Kahl, Hervé Goëau, Benjamin Deneu, Diego Marcos, Joaquim Estopinan, César Leblanc, Théo Larcher, et al. Overview of lifeclef 2023: evaluation of ai models for the identification and prediction of birds, plants, snakes and fungi. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 416–439. Springer, 2023. 3

[29] Alexis Joly, Lukáš Picek, Stefan Kahl, Hervé Goëau, Vincent Espitalier, Christophe Botella, Benjamin Deneu, Diego Marcos, Joaquim Estopinan, Cesar Leblanc, et al. Lifeclef 2024 teaser: Challenges on species distribution prediction and identification. In *European Conference on Information Retrieval*, pages 19–27. Springer, 2024.

[30] Alexis Joly, Lukáš Picek, Stefan Kahl, Hervé Goëau, Vincent Espitalier, Christophe Botella, Diego Marcos, Joaquim Estopinan, Cesar Leblanc, Théo Larcher, 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*, pages 183–207. Springer, 2024. 3

[31] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021. 2

[32] Dirk Nikolaus Karger, Dirk R Schmatz, Gabriel Dettling, and Niklaus E Zimmermann. High-resolution monthly precipitation and temperature time series from 2006 to 2100. *Scientific data*, 7(1):248, 2020. 1

[33] César Leblanc, Alexis Joly, Titouan Lorieul, Maximilien Servajean, and Pierre Bonnet. Species distribution modeling based on aerial images and environmental features with convolutional neural networks. In *CLEF (Working Notes)*, pages 2123–2150, 2022. 3

[34] César Leblanc, Pierre Bonnet, Maximilien Servajean, Milan Chytrỳ, Svetlana Aćić, Olivier Argagnon, Ariel Bergamini, Idoia Biurrun, Gianmaria Bonari, Juan A Campos, et al. A deep-learning framework for enhancing habitat identification based on species composition. *Applied Vegetation Science*, 27(3):e12802, 2024. 2, 4, 6

[35] César Leblanc, Pierre Bonnet, Maximilien Servajean, and Alexis Joly. Pl@ ntbert: leveraging large language models to enhance vegetation classification through species composition analysis. Università di Bologna, 2024. 2, 4

[36] Diego Marcos, Robert van de Vlasakker, Ioannis N Athanasiadis, Pierre Bonnet, Hervé Goeau, Alexis Joly, W Daniel Kissling, César Leblanc, André SJ van Proosdij, and Konstantinos P Panousis. Fully automatic extraction of morphological traits from the web: utopia or reality? *arXiv preprint arXiv:2409.17179*, 2024. 4

[37] Carsten Meyer, Patrick Weigelt, and Holger Kreft. Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology letters*, 19(8):992–1006, 2016. 1

[38] Peter J Morin. *Community ecology*. John Wiley & Sons, 2009. 4

[39] Dorian Moss. Eunis habitat classification–a guide for users. *European Topic Centre on Biological Diversity*, 2008. 4

[40] Tim Newbold, Lawrence N Hudson, Samantha LL Hill, Sara Contu, Igor Lysenko, Rebecca A Senior, Luca Börger, Dominic J Bennett, Argyrios Choimes, Ben Collen, et al. Global effects of land use on local terrestrial biodiversity. *Nature*, 520(7545):45–50, 2015. 1

[41] IR Noble. The role of expert systems in vegetation science. *Vegetatio*, 69:115–121, 1987. 2

[42] Julian D Olden, Joshua J Lawler, and N LeRoy Poff. Machine learning methods without tears: a primer for ecologists. *The Quarterly Review of Biology*, 83(2):171–193, 2008. 2

[43] Camille Parmesan and Gary Yohe. A globally coherent fingerprint of climate change impacts across natural systems. *nature*, 421(6918):37–42, 2003. 1

[44] Nathalie Pettorelli, Martin Wegmann, Andrew Skidmore, Sander Mücher, Terence P Dawson, Miguel Fernandez, Richard Lucas, Michael E Schaepman, Tiejun Wang, Brian O'Connor, et al. Framing the concept of satellite remote sensing essential biodiversity variables: challenges and future directions. *Remote sensing in ecology and conservation*, 2(3):122–131, 2016. 1

[45] Steven J Phillips, Robert P Anderson, and Robert E Schapire. Maximum entropy modeling of species geographic distributions. *Ecological modelling*, 190(3-4):231–259, 2006. 1

[46] Steven J Phillips, Miroslav Dudík, Jane Elith, Catherine H Graham, Anthony Lehmann, John Leathwick, and Simon Ferrier. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological applications*, 19(1):181–197, 2009. 2, 3

[47] Lukas Picek, Christophe Botella, Maximilien Servajean, César Leblanc, Rémi Palard, Theo Larcher, Benjamin Deneu, Diego Marcos, Pierre Bonnet, and alexis joly. Geoplant: Spatial plant species prediction dataset. In *Advances in Neural Information Processing Systems*, pages 126653–126676. Curran Associates, Inc., 2024. 2, 3

[48] Lukáš Picek, Christophe Botella, Maximilien Servajean, César Leblanc, Remi Palard, Théo Larcher, Benjamin Deneu, Diego Marcos, Joaquim Estopinan, Pierre Bonnet, et al. Overview of geolifeclef 2024: Species composition prediction with high spatial resolution at continental scale using remote sensing. In *CLEF 2024-Working Notes of the 25th Conference and Labs of the Evaluation Forum*, pages 1966–1977. CEUR, 2024. 3

[49] Anantha M Prasad, Louis R Iverson, and Andy Liaw. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, 9: 181–199, 2006. 1

[50] David R Roberts, Volker Bahn, Simone Ciuti, Mark S Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40 (8):913–929, 2017. 5

[51] Srikumar Sastry, Subash Khanal, Aayush Dhakal, Adeel Ahmad, and Nathan Jacobs. Taxabind: A unified embedding space for ecological applications. *arXiv preprint arXiv:2411.00683*, 2024. 2

[52] Sara Si-moussi, Joaquim Estopinan, and Wilfried Thuiller. Earth observation foundation models for large-scale biodiversity modelling. In *BioSpace25-Biodiversity insight from Space*, 2025. 4

[53] Lubomir Tichý, Milan Chytrý, and Flavia Landucci. Grimp: A machine-learning method for improving groups of discriminating species in expert systems for vegetation classification. *Journal of Vegetation Science*, 30(1):5–17, 2019. 2

[54] Cornelis J Van Rijsbergen. Information retrieval, 2nd edn. newton, ma, 1979. 5

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[56] Alexander Zizka, Tobias Andermann, and Daniele Silvestro. Iucnn–deep learning approaches to approximate species' extinction risk. *Diversity and Distributions*, 28(2):227–241, 2022. 3