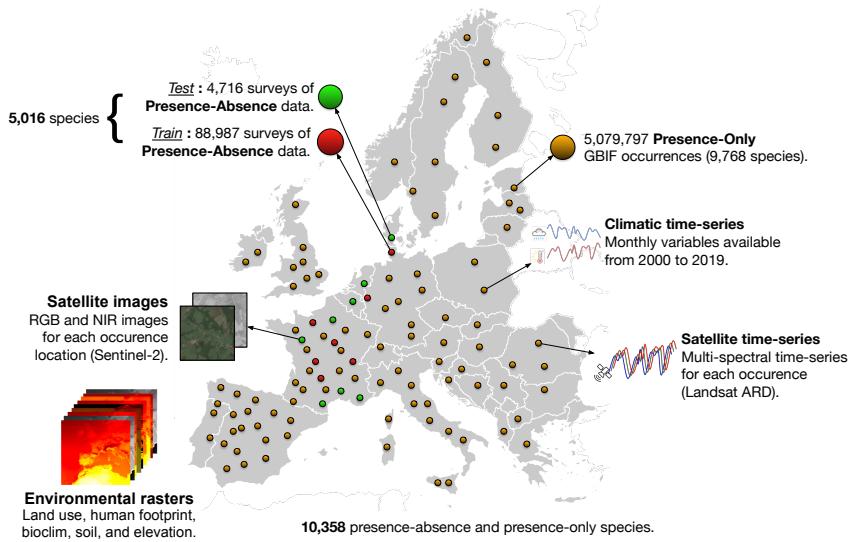


GeoPlant: Spatial Plant Species Prediction Dataset

Lukas Picek¹, Christophe Botella¹, Maximilien Servajean², César Leblanc¹, Rémi Palard¹

Théo Larcher¹, Benjamin Deneu¹, Diego Marcos^{1,3}, Pierre Bonnet⁴ and Alexis Joly¹

¹ INRIA, Montpellier, France ² Université Paul Valéry, Montpellier, France ³ Université de Montpellier, France and ⁴ CIRAD, UMR AMAP, Montpellier, France



Abstract

The difficulty of monitoring biodiversity at fine scales and over large areas limits ecological knowledge and conservation efforts. To fill this gap, Species Distribution Models (SDMs) predict species across space from spatially explicit features. Yet, they face the challenge of integrating the rich but heterogeneous data made available over the past decade, notably millions of opportunistic species observations and standardized surveys, as well as multi-modal remote sensing data. In light of that, we have designed and developed a new European-scale dataset for SDMs at high spatial resolution (10-50 m), including more than 10k species (i.e., most of the European flora). The dataset comprises 5M heterogeneous Presence-Only records and 90k exhaustive Presence-Absence survey records, all accompanied by diverse environmental rasters (e.g., elevation, human footprint, and soil) that are traditionally used in SDMs. In addition, it provides Sentinel-2 RGB and NIR satellite images with 10 m resolution, a 20-year time-series of climatic variables, and satellite time-series from the Landsat program. In addition to the data, we provide an openly accessible SDM benchmark (hosted on Kaggle), which has already attracted an active community and a set of strong baselines for single predictor/modality and multimodal approaches. All resources, e.g., the dataset, pre-trained models, and baseline methods (in the form of notebooks), are available on Kaggle, allowing one to start with our dataset literally with two mouse clicks.

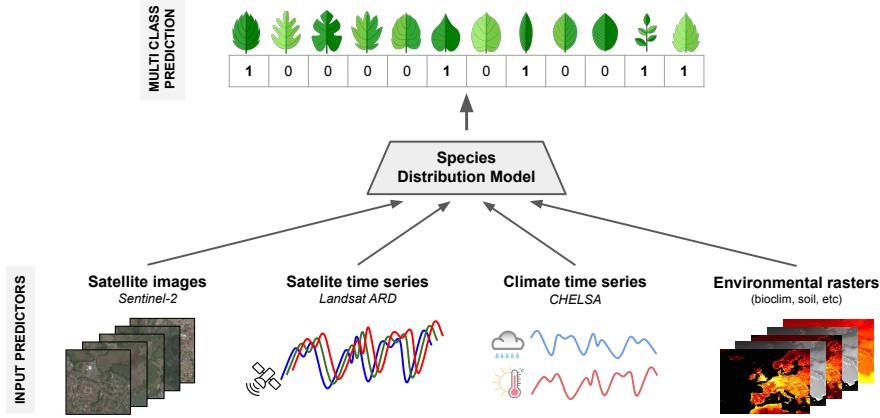


Figure 1: **Our view on Species Distribution Models (SDM).** The SDM utilizes multi-modal predictors (e.g., satellite, climate, and environmental data) for given GPS coordinates to predict multi-species compositions at that location.

1 Introduction

Global changes rapidly transform ecosystems, and their local impacts are context-dependent and hard to predict [17]. Monitoring species composition at high spatial resolution is crucial for understanding biodiversity responses and aiding decision-making, but has proven to be extremely challenging. Deep learning-based species distribution models (deep SDMs) [1, 7, 14] offer a promising venue by allowing to use high-resolution geographic predictors and remote sensing data to address sampling gaps [13, 22, 33] (see Figure 1). However, the heterogeneity, imbalance, bias and complexity of species observations and environmental data make model implementation challenging. Apart from that, standardized biodiversity data, i.e., exhaustive Presence-Absence (PA) surveys, are limited in coverage as they are time-consuming and costly to update and maintain. Even with a relatively large amount of PA data, it is difficult to model and map biological groups with large taxonomic diversity, such as plants, which have around 400k species to date, a vast majority of which are rare [21].

On the other hand, Presence-Only (PO) data from citizen-science platforms/initiatives (e.g., iNaturalist and Pl@ntNet), have emerged as valuable sources of large amounts of biodiversity data [3]. Even though those data have the potential to fill the distribution gap of the PA surveys, as they provide millions of geolocated records of tens of thousands of species annually, they are severely limited in that they do not indicate the absence of non-observed species and are heavily biased towards areas with a high density of observers [29]. Besides, the PO data represent a fraction of the species communities in regions with limited sampling and are biased toward common and/or appealing species [23]. As a result, incorporating PO data into SDMs risk introducing these biases [2, 39].

To allow a standardized use of available data and enable further research in ecological modeling, machine learning, and species distribution modeling, we have assembled a new European-scale dataset for Plant Species Prediction – **GeoPlant**. The dataset includes more than 5M heterogeneous PO records and 90k standardized PA surveys covering 10k+ species. All records are accompanied by (i) diverse environmental rasters (e.g., elevation, human footprint, and soil), (ii) Sentinel-2 RGB and NIR satellite images with 10 m resolution, (iii) a 20-year time series of climatic variables and (iv) satellite time series from the Landsat program. The GeoPlant dataset is the biggest dataset for species distribution modeling and the only dataset that includes satellite images and time series, climate time series, and environmental rasters. Besides, through the standardized PA data available in GeoPlant, model evaluation is made robust to the many biases of the PO data.

Following our successful long-term efforts in benchmarking SDM models [5, 35, 36, 6], we also provide an openly accessible benchmark (hosted on Kaggle), which has already attracted an active community and established strong baselines for various single and multi-modal approaches. With all needed resources (i.e., dataset, pre-trained models, and baseline methods) already publicly available, we create an ideal environment for benchmarking any new species distribution modeling approach.

2 Related Work

Species distribution models have relied for decades on geographic predictors at a spatial resolution of the order of a kilometer, such as bioclimatic [4], land cover [37] or human footprint variables [15]. At the same time, remote sensing data represent an unprecedented opportunity to provide high spatial resolution species distribution models with rich and globally consistent predictor variables describing the environment [31] and its temporal changes. However, its integration into species distribution models is recent and challenging [13, 22]. This data can complement the picture of the environmental landscape provided by the variables above at a coarser spatial scale. However, integrating variables at different spatial or temporal resolutions within deep learning architectures brings challenges.

Open datasets and benchmarks for species distribution modeling are still rare, and objective comparison of methods on existing datasets is limited [52]. The most cited benchmark [19], built in 2006, includes point location records for 226 anonymized species from six regions with accompanying predictor variables. Its key innovation was providing both PO and PA data to address spatial distribution biases in PA sites [45]. Our new dataset scales this approach up significantly, with about 100 times more occurrences and species, and includes more diverse predictors such as medium-resolution remote sensing data. **GeoPlant** allow evaluation of new SDMs, particularly those based on multi-modal deep neural networks, and help identify fundamental factors determining species distribution. Another benchmark [10], published in 2021, uses forest inventory data across the western US to explore SDM extrapolation limits. This dataset covers 286,551 plots and focuses on 108 tree species with 19 bioclimatic predictors at a coarse spatial resolution (1 km). In addition to dedicated benchmarks, more ecological studies are publishing their data openly, which allows their use for SDM evaluation [56, 54, 48]. However, these often suffer from small scale and basic predictors based on tabular data. Recent work on deep SDMs [14, 22, 9] incorporates more complex predictors like images or time series but usually relies on PO data alone, introducing significant evaluation biases.

The **GeoLifeCLEF** is a long-term SDM evaluation campaign [5, 8, 35, 36, 6], organized by the authors, attracting considerable participation. GeoLifeCLEF aims to evaluate SDMs with unprecedented scope in species coverage, predictor multi-modality, and spatial resolution (approximately 10 meters). Before 2023, the datasets were based solely on PO data with observational bias. In 2023, PA data were included as the main test set based on exhaustive survey data from the EVA [11]. The dataset presented here (GeoPlant) extends the 2023 dataset with additional PA data and new predictors, such as bioclimatic time series. It is unique in its scale (continental), spatial resolution (10m for the finer modality), and predictor diversity. From a machine learning perspective, it poses two major challenges: (i) multi-modality—how to select and combine numerous heterogeneous information sources, and (ii) distribution shift—how to train effective models on PO data to predict PA.

Available methods suitable for species distribution modeling are usually divided into four groups:

- *Traditional statistical methods*, like generalized linear models [25, 27, 55], logistic regression [43], and Generalized Additive Models (GAMs) [57], form the backbone of SDM, with frameworks like Hierarchical Modeling of Species Communities (HMSC) [41] integrating additional data sources.
- *Traditional machine learning* include boosted regression trees, random forests, support vector machines, and neural networks, address complex species-environment relationships, with CNNs [12, 18, 20, 42] and provide advanced feature extraction from spatial environmental arrays.
- *Presence-only methods*, such as MaxEnt [44], estimate species observation probabilities based on environmental covariates and often use pseudo negatives for enhanced accuracy. A recent comparative study evaluated 13 different models, including both statistical and ML models [52]. However, since the evaluation was based on the dataset of Elith et al. [19], it did not allow the evaluation of more complex models, i.e, deep SDMs.
- *Deep SDMs* is a generic term for the new generation of SDMs using deep learning methods as a means of improving predictive performance and better understanding the contribution of complex factors such as spatial and temporal structures [7, 14, 22, 9]. This type of model has proved to be the most successful in all editions of GeoLifeCLEF [5, 8, 35, 36, 6].

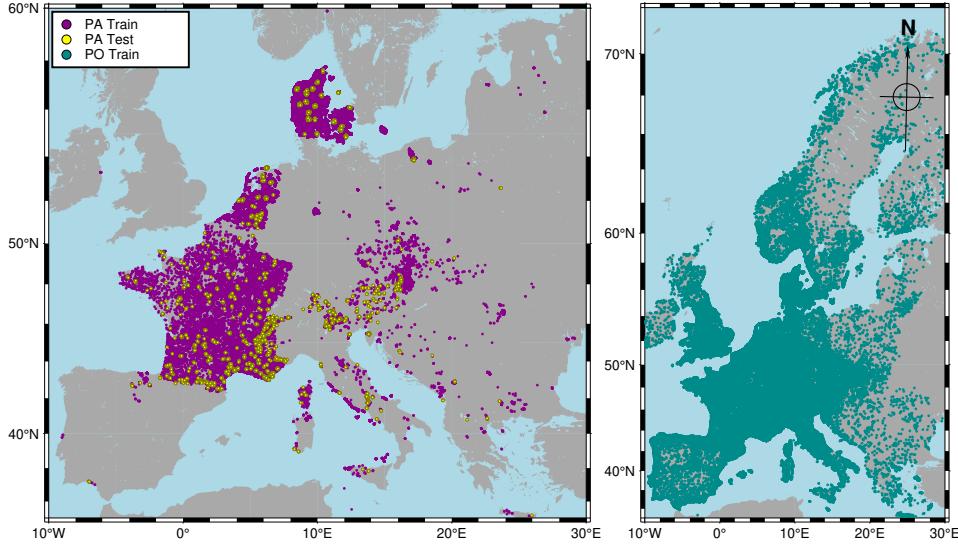


Figure 2: **Geo spatial scale of the dataset.** While the provided Presence-Only (PO) data spans all of habitable Europe, the Presence-Absence (PA) training and test sites are primarily from France, Denmark, Switzerland, and Czechia.

3 GeoPlant Dataset

The GeoPlant dataset comprises **Species Observation** (i.e., *Presence-Only* occurrences and *Presence-Absence* surveys) and various **Environmental Predictors** data and spans 38 European countries and eight bio-geographic regions, e.g., Alpine, Atlantic, and Boreal (see Figure 2) For each species observation, we provide: (i) diverse environmental rasters (e.g., elevation, human footprint, land use, and soil), (ii) Sentinel2-based RGB and Near-Infra-Red satellite images (128×128) with 10 m resolution, (iii) a 20-year time series of climatic variables, and (iv) satellite time-series point values for six satellite bands (R, G, B, NIR, SWIR1, and SWIR2) from the Landsat program. The data is highly diverse. Therefore, we provide a detailed description of each predictor below.

3.1 Species Observation Data

The species observation data comprises approximately 5 million **Presence-Only** (PO) occurrences and around 90 thousand **Presence-Absence** (PA) survey records. The PO data is the most commonly and widely available type of data and covers most European countries, but it has been sampled without any protocol, leading to various sampling biases, and the local observation of a species provides no information on the absence of others. A reporter (i.e., citizen scientist) might not have reported some species due to seasonal visibility, misidentification, or lack of interest. For both PO and PA data, we provide a short description below.

Presence-Absence (PA) surveys. A presence-absence survey is obtained by experienced botanists who report, as exhaustively as possible, the plant species in a given small spatial plot (usually 10–400 square meters). Hence, all species not observed during a PA survey are likely truly absent from the plot. The provided data originates from 29 source datasets hosted in the European Vegetation Archive (EVA, e.g., Denmark Naturdata, IGN National Forest Inventory, Belgium INBOVEG) with different spatial extents and targeted habitats. Despite the relatively large size of the PA dataset (93,703 surveys), it only covers 5,016 species—approximately half of the European flora. Besides, the distribution of these species is highly imbalanced, with most species only being observed once or twice among all the PA surveys. While constructing the training and test splits (95/5), we used a spatial block hold-out procedure [47] using a spatial grid with 10×10 km cells (see the spatial grid in Figure 2) and ended up with 88,987 surveys for training and 4,716 for testing. The 10×10 km test blocks were randomly selected to ensure balance in biogeographical regions.

Table 1: **Presence-Only dataset sources.** Selected GBIF datasets cover 38 European countries. "Uniq. species" indicates the number of unique species in each dataset compared to the rest.

GBIF Dataset Name	Records	Species	Uniq. species
(our) Pl@ntNet Observations + Pl@ntNet Occurrences	2,298,884	4,631	295
Danmarks Miljøportals Naturdatabase	691,313	1,457	14
iNaturalist Research-grade Observations	625,681	7,496	1,754
Norwegian Species Observation Service	601,101	2,243	167
Observation.org, Nature data from around the World	241,205	5,108	429
Non-native plant occurrences in Flanders and the Brussels	178,544	1,464	134
Artportalen (Swedish Species Observation System)	163,513	2,771	464
National Plant Monitoring Scheme U.K.	120,413	1,109	11
Vascular plant records verified via iRecord	103,213	2,179	99
Swiss National Databank of Vascular Plants	49,173	58	2
Invazivke - Invasive Alien Species in Slovenia	4,171	60	1
Masaryk University - Herbarium BRNU	2,586	1,321	122
GeoPlant Presence-Only data (<i>Combined</i>)	5,079,797	9,709	—

Presence-Only (PO) occurrences. A Presence-Only (PO) record is a geolocated species observation whose sampling protocol is unknown and which doesn't inform about the absence of other species. The sampling effort is highly heterogeneous in space, time, and across species. As most PO records originate from citizen-science platforms, they are generally concentrated in populated and easily accessible areas and focus on charismatic and easy-to-identify plant species. Despite this, PO data can help compensate for gaps in PA surveys when controlling for sampling biases in model calibration [24, 40]. The PO data comprises 5 million records of 9,709 plant species reported between 2017 and 2021. This data originates from 13 pre-selected datasets extracted from the Global Biodiversity Information Facility (GBIF) [28], listed in and referenced in Table 1.

3.2 Environmental predictor data

The spatialized geographic and environmental predictor data are crucial for precise predictive modeling. In light of that, we have developed the biggest publicly available dataset regarding available resources and their diversity. Each survey or species observation (PO and PA) is accompanied with: (i) A four-band 128×128 satellite image at 10 m resolution around the occurrence location. (ii) Time series of the past values for six satellite bands at the point location. (iii) Various environmental rasters at the European scale (e.g., climatic, soil, elevation, land use, and human footprint variables), and (iv) Monthly time series of four climatic variables for any observation, as we provide monthly climatic rasters from 2000 to 2019. As the dataset originates from various sources and requires significant preprocessing, we thoroughly describe the acquisition process and data description below.

3.2.1 Environmental rasters

As mentioned earlier, we associated species observations with diverse environmental rasters, e.g., bioclimatic, soil, elevation, land cover, and human footprint. Environmental rasters used include 19 low-resolution bioclimatic rasters for Europe, nine low-resolution soil rasters describing soil properties, a high-resolution elevation raster, a medium-resolution multi-band land cover raster, and 16 low-resolution human footprint rasters (14 detailed pressures for two time periods and two summary rasters). All the environmental rasters are provided as .TIF files in standardized longitude/latitude coordinates, reprojected to the WGS84 coordinate system, EPSG:4326, and with the same spatial extent¹, including all the species observation data.

Land cover. Land cover variables helped explain species distributions at all scales and significantly improved bioclimatic model performance at thinner spatial resolutions starting from 20 km [37]. Furthermore, the interactions between climate change and land cover change remain poorly

¹The bounding box extends from $(min_x, min_y)=(-32.26, 26.63)$ to $(max_x, max_y)=(35.58, 72.18)$, with the corners adjusted by ± 1 degree from the extreme species observations.

understood and could strongly modify both land cover change and the distribution of threats [38]. Following on that, we provide a medium-resolution multi-band land cover raster covering Europe. It is provided as a compressed GeoTIFF file with a resolution of 500m. We used the NASA earthdata portal to extract the 24 HDF raster tiles from the MODIS Terra+Aqua [26]). These HDF files stack 13 layers corresponding to various land cover classifications or encoding the class confidence detailed are provided in User Guide. We merged all the HDF files into a single multi-band GeoTIFF, reprojected it to WGS84, and cropped it to the extent of GeoPlant. Each band in the provided GeoTIFF describes either the land cover class prediction or its confidence under various classifications. We recommend using IGBP (17 classes) or LCCS (43 classes) layers, often used in species distribution modeling.

Human footprint. Human impact on biodiversity loss has been widely studied, resulting in 1 million species at risk of extinction (IPBES, 2019). Human pressure significantly influences species extinction risk and is a better predictor of species' geographic range than biological traits [15, 16]. To capture this impact, we provide summary rasters combining all human pressures and detailed rasters per pressure, preserving the original data integrity. These include (i) GeoTIFF with 16 low-resolution rasters for human footprint and (ii) GeoTIFF with 14 detailed rasters across seven environmental pressures (e.g., nightlight level, population density) for two time periods (1993–2009). Both rasters go with 30 arcsec resolution (1 km). To develop the data, we used global terrestrial human footprint rasters from Venter et al. [53], which provide reference data for human settlement and activities. Derived from remote sensing and surveys, these rasters measure direct and indirect human pressures at the kilometer scale across eight variables: 1) built environment, 2) population density, 3) electrical infrastructure, 4) cropland, 5) pastureland, 6) roads, 7) railways, and 8) navigable waterways. Except for roads and railways, each variable is available and consistent for two years: 1993 and 2009 [53]. The cumulative scores are normalized by biome, as described in [49], ensuring equal representation of different pressure levels across biomes. The rasters were reprojected from Mollweide (ESRI:54009) to the WGS84 geographic coordinate system and cropped consistently.

Elevation. Topography significantly affects plant species distribution by influencing light, moisture, and nutrient conditions. Including topography as a covariate in species distribution models (SDMs) has improved their performance significantly [50]. Since large-scale data on edaphic factors are still scarce, topography serves as an excellent proxy. Therefore, we provide Elevation for all available records in the form of a GeoTIFF file and in CSV as scalar values. The raster was extracted from the ASTER Global Digital Elevation Model V3 using NASA earthdata portal.

Soilgrids. Physico-chemical soil properties (e.g., Ph, granulometry) are crucial determinants of a plant species' survival ability. SoilGrids [46] is a system for global digital soil mapping that uses machine learning methods to map the spatial distribution of soil properties across the globe. SoilGrids prediction models are fitted at 250m resolution using over 230k soil profile observations from the WoSIS database and a series of environmental covariates. We integrated nine soil rasters corresponding to a depth of 5 to 15cm at a resolution of 30 arcsec (~ 1 km), i.e., the aggregated version of SoilGrids 2.0 rasters derived by resampling at 1km the mean initial predictions at 250m for each property. Nine pedologic low-resolution rasters were downloaded from isrig.org (already provided in WGS84) and cropped to the same extent.

3.3 Satellite Images

Remote sensing data is a rich and globally consistent predictor variable describing the surrounding environment [31] in high resolution. Therefore we provide Sentinel2-based RGB and Near-Infra-Red (NIR) satellite images (128×128) with 10m resolution centered on the geolocated spot and taken in the same year (see Figure 3). All images were extracted from the pre-processed rasters (composites of monthly rasters), with eliminated cloud coverage and shadows, available on the Ecodatacube platform. The values in extracted image patches are first thresholded at 10,000, rescaled to [0,1], and a gamma correction of 2.5 is applied (i.e., values are powered by $1/2.5$). Finally, the values are rescaled to [0,255] and rounded for uint8 encoding. This process avoids using a range for high reflectance values ($>10,000$ in uint16) and gives more range to values close to zero, which are the most common.



Figure 3: **Satellite image data.** 128×128 images from Sentinel2. First row RGB, Second row NIR.

3.3.1 Satelite time-series

In addition to Satellite Images, we provide comprehensive satellite time series data that spans over 20 years of satellite imagery. The data, obtained through the Landsat ARD program and pre-processed by EcoDataCube, covers a wide temporal range of quarterly values from 1999 to 2020, providing a detailed understanding of environmental changes over the past two decades. Each PO and PA location is linked to the time series of median point values over each season since winter 1999 for six satellite bands (R, G, B, NIR, SWIR1, and SWIR2), capturing high-resolution local signatures of seasonal vegetation changes, extreme natural events like fires, and land use changes. Due to the large size of the original rasters, data points from each spectral band were extracted for all PA and PO locations and aggregated into CSV files. A CSV file with 84 columns (representing 84 seasons from winter 1999 to autumn 2020) was created for each band. These CSV files were then aggregated into 3d tensors (cubes) with axes as [BAND, QUARTER, YEAR]. See the visualization in Figure 4.

3.3.2 Climatic variables

Previous GeoLifeCLEF editions have demonstrated that climatic conditions are vital for predicting plant and animal species [34, 33]. Hence, we provide Monthly and Average Climatic rasters available at CHELSA [32]. The Monthly Climatic rasters contain four climatic variables (mean, min, and max temperature, and total precipitation) from Jan. 2000 to Dec. 2019 (960 rasters with a pixel resolution of 30 arcsecs – 1 km). The Average Climatic rasters combine 19 rasters with various averaged variables calculated from 1981 to 2010, e.g., mean annual temperature, seasonality, and extreme or limiting environmental conditions. As for the satellite time series, we pre-extracted the scalar values for all the PO and PA records and provided them as CSV files, and we aggregated them into 3d tensors with axis [RASTER-TYPE, YEAR, MONTH]. See the cube visualization in Figure 4.

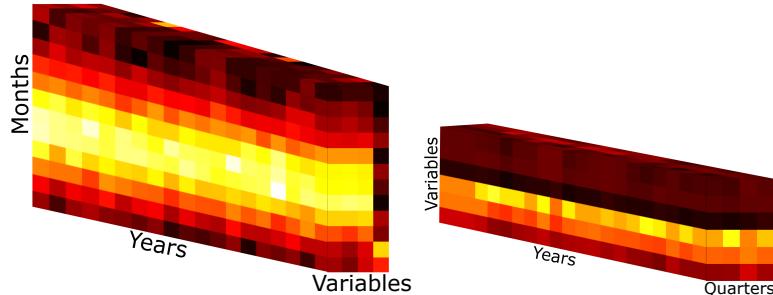


Figure 4: **Time-series data cubes.** (left), Monthly climatic cubes with variables “temperature mean”, “max” and “mean” and “precipitation”. One record per month for 19 years (2000-2019) are provided. The value corresponds to the pixel at the observation coordinate. (right), satellite time series with bands red, green, blue, near-infrared, short wave infrared 1 and 2. One value per quarter for 21 years are provided. The value corresponds to the central pixel of the satellite patch.

4 GeoPlant Benchmark

Following our positive previous experiences, we use Kaggle to host the GeoPlant benchmark. The main benefits of the platform include (i) easy dataset use and referencing, (ii) model sharing, (iii) code development via Jupyter with free GPU resources, and (iv) straightforward setup and community management.

Evaluation criteria. As the provided test set is based exclusively on multi-label data from the exhaustive Presence-Absence surveys, we define the primary evaluation metric as the sample-averaged F_1 -score. The F_1 -score is an average measure of overlap between the predicted and actual set of species present at a given location and time. Thus, for each test PA survey i associated with a set of ground-truth labels Y_i (i.e., the set of plant species reported by experts on a small grid), and provided list of predicted labels $\hat{Y}_{i,1}, \hat{Y}_{i,2}, \dots, \hat{Y}_{i,R_i}$ the micro F_1 -score is computed as

$$F_1 = \frac{1}{N} \sum_{i=1}^N \frac{TP_i}{TP_i + (FP_i + FN_i)/2},$$

where $\begin{cases} TP_i = \text{Number of predicted species truly present, i.e. } |\hat{Y}_i \cap Y_i|. \\ FP_i = \text{Number of species predicted but absent, i.e. } |\hat{Y}_i \setminus Y_i|. \\ FN_i = \text{Number of species not predicted but present, i.e. } |Y_i \setminus \hat{Y}_i|. \end{cases}$

Assets. Apart from the dataset, we also provide a wide variety of valuable assets for deep SDMs. Notably, we provide:

- *Malpolon*: A Pytorch-based framework designed for deep SDM training using various input covariates, such as bioclimatic rasters, remote sensing images, and land-use rasters. Malpolon provides examples and scenarios to guide users through different use cases, such as training models on their own datasets or participating in challenges like GeoLifeCLEF.
- *Dataloaders*: To allow easy use of the raw or preprocessed data, we provide data loaders for environmental predictors and species observations. The data loaders, available on GitHub, support online (while training) loading of pre-extracted visual patches, visual/bioclimatic rasters, and time series data, each paired with geolocation and predictor data identifiers.
- *Baseline notebooks*: Jupyter Notebooks with the implementation of single and multi-modal baselines. All are available in a form for direct use on Kaggle.
- *Pre-trained models*: All models trained using the baseline notebooks provided through Kaggle.

5 Some Weak and Strong Baselines

In this section, we briefly describe the PA data-based baselines. Additional experiments with PO data are available in the supplementary materials. Details about used hyperparameters, optimization strategy, etc., supporting reproducibility of all achieved and reported results, are also provided in supplementary materials and available Jupyter Notebooks.

Naive baselines. With the dense and numerous observation data, one can naively predict the species' presence just by selecting a set of the most common species within administrative or bio-geographical regions. In our initial experiments (see Supplementary), we show that selecting top-25 most common species from PA metadata based on district & bio-geographical zone resulted in a sample-averaged F_1 of 20.6%. Using the same approach but with the PO data resulted in an F_1 below 9%.

PA experiment. We evaluate three architectures and modalities over the PA observations to demonstrate the potential of different modalities and the importance of multimodal approaches. We take ResNet18 [30] (previously used in SDM) as a baseline and compare it to two custom and lighter residual architectures (a smaller, ResNet6-based custom CNN, and a simple Multi-Layer Perceptron (MLP)) on all modalities separately. Furthermore, we take the best-performing one (i.e., the custom CNN) and test its performance as encoders for different modalities. All models were trained to maximize the top-25 average samples F_1 using Binary Cross Entropy (BCE) and AdamW.

Table 2: **Baseline performance for selected custom architectures.** The small custom architectures, CNN with skip connection and 6 blocks, provides better performance in terms of all metrics, i.e., sample averaged F₁, Precision (Pre.), and Recall (Rec.).

	Clim. cubes (C)			Sat. cubes (S)			Sat. images (I)			C + S			C + S + I		
	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1
MLP	15.5	23.9	17.3	25.0	42.2	28.5	13.2	22.6	15.2	–	–	–	–	–	–
ResNet-18	23.2	37.5	25.9	22.0	37.4	25.1	16.4	27.6	18.7	–	–	–	–	–	–
CNN	24.4	38.6	27.0	26.0	44.3	29.7	18.9	31.6	21.5	28.4	46.8	32.0	28.6	47.1	32.2
CNN + Top-K	25.9	34.6	27.8	28.5	38.8	30.8	21.1	28.3	22.5	30.1	41.5	33.2	30.4	42.7	33.6

Estimating the number of species to predict per survey. Following upon the naive baselines, we have developed a straightforward approach for multi-label classification. Instead of finding a threshold to select present species, we add a separate regression step that estimates the number of species to predict per survey. Following [51] we do not train the model on a regression task. The output space (i.e., the number of species) is divided into 15 bins containing the same number of surveys within the training set (i.e., quantiles). The average number of species per survey within each bin is then computed. The output of our model thus becomes the expected number of species for the most likely species. In addition, the F₁ measure is not symmetrical, leading to a preference for the overestimation of K . For this reason, we predicted for each survey the K most likely species where K is the number of species estimated plus an offset, which we empirically set to five:

$$\hat{\Gamma}(\mathcal{A}) = \{\sigma_{\mathcal{A}}(k) : k \in \{1, \dots, \hat{\eta}(\mathcal{A}) + \text{offset}\}\}$$

$$\hat{\eta}(\mathcal{A}) \approx |S_n \cap \mathcal{A}|,$$

where \mathcal{A} is the survey field, S_n the test set presences, $|S_n \cap \mathcal{A}|$ the set of present species in the survey. Hence, $\hat{\eta}(x)$ approximates the number of species present in the survey. $\sigma_x(k)$ is k^{th} species in decreasing order of probability. This resulted in the best performances (bottom row of Table 2).

6 Conclusion

This paper presents the GeoPlant dataset, a unique compilation of species observation and environmental predictor data spanning 38 European countries. With its diverse range of predictors, including satellite imagery, climatic variables, and detailed environmental rasters, the dataset provides a basis for advancing large-scale species distribution modeling (SDM). The GeoPlant dataset aims to address previous SDM datasets’ limitations by offering: (i) a significantly larger and more diverse set of species occurrences, 5 million PO opportunistic observations, and 90 thousand PA survey records, which also allow for a meaningful evaluation of the SDM results, and (ii) a rich set of predictors, including not only traditional environmental factors, such as climatic variables, land cover and human footprint indices, but also medium-resolution satellite imagery and time series.

In addition to the dataset, a benchmark (hosted via a dedicated Kaggle competition) and a set of strong baselines are provided. All are easily accessible and publicly available through the Kaggle and GitHub repositories. Furthermore, all baselines are available on Kaggle in the form of Jupyter Notebooks that allow direct reproducibility for all provided baselines.

By providing an open and extensive benchmark for species distribution modeling, we hope to encourage the development of innovative modeling techniques and contribute to the broader understanding of species distribution patterns at a continental scale.

Acknowledgements

Further models developed from this dataset will directly meet the needs of the Europe biodiversity strategy for 2030 through the HORIZON Europe projects MAMBO (10.3030/101060639) and GUARDEN (10.3030/101060693). They will be used in particular as a basis to map biodiversity indicators at the European scale (e.g., presence of endangered species, invasive species, and habitat condition metrics). 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.

References

- [1] D. J. Benkendorf and C. P. Hawkins. Effects of sample size and network depth on a deep learning approach to species distribution modeling. *Ecological Informatics*, 60:101137, 2020.
- [2] E. H. Boakes, P. J. McGowan, R. A. Fuller, D. Chang-qing, N. E. Clark, K. O'Connor, and G. M. Mace. Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *PLoS biology*, 8(6):e1000385, 2010.
- [3] P. Bonnet, A. Affouard, J.-C. Lombardo, M. Chouet, H. Gresse, V. Hequet, R. Palard, M. Fromholtz, V. Espitalier, H. 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.
- [4] T. H. Booth, H. A. Nix, J. R. Busby, and M. F. Hutchinson. Bioclim: the first species distribution modelling package, its early applications and relevance to most current maxent studies. *Diversity and Distributions*, 20(1):1–9, 2014.
- [5] C. Botella, P. Bonnet, and A. Joly. Overview of GeoLifeCLEF 2018: location-based species recommendation. In *CLEF task overview 2018, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2018, Avignon, France.*, 2018.
- [6] C. Botella, B. Deneu, D. Marcos, M. Servajean, T. Larcher, C. Leblanc, J. Estopinan, P. Bonnet, and 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.
- [7] C. Botella, A. Joly, P. Bonnet, P. Monestiez, and F. Munoz. A deep learning approach to species distribution modelling. *Multimedia Tools and Applications for Environmental & Biodiversity Informatics*, pages 169–199, 2018.
- [8] C. Botella, M. Servajean, P. Bonnet, and A. Joly. Overview of geolifeCLEF 2019: plant species prediction using environment and animal occurrences. In *CLEF 2019-Conference and Labs of the Evaluation Forum*, volume 2380, 2019.
- [9] P. Brun, D. N. Karger, D. Zurell, P. Descombes, L. C. de Witte, R. de Lutio, J. D. Wegner, and N. E. Zimmermann. Rank-based deep learning from citizen-science data to model plant communities. *bioRxiv*, pages 2023–05, 2023.
- [10] N. D. Charney, S. Record, B. E. Gerstner, C. Merow, P. L. Zarnetske, and B. J. Enquist. A test of species distribution model transferability across environmental and geographic space for 108 western north american tree species. *Frontiers in Ecology and Evolution*, 9:689295, 2021.
- [11] M. Chytrý, S. M. Hennekens, B. Jiménez-Alfaro, I. Knollová, J. Dengler, F. Jansen, F. Landucci, J. H. Schaminée, S. Aćić, E. Agrillo, et al. European vegetation archive (eva): an integrated database of european vegetation plots. *Applied vegetation science*, 19(1):173–180, 2016.
- [12] D. R. Cutler, T. C. Edwards Jr, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler. Random forests for classification in ecology. *Ecology*, 88(11):2783–2792, 2007.
- [13] B. Deneu, A. Joly, P. Bonnet, M. Servajean, and 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*, 13:839279, 2022.
- [14] B. Deneu, M. Servajean, P. Bonnet, C. Botella, F. Munoz, and A. Joly. Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment. *PLoS computational biology*, 17(4):e1008856, 2021.
- [15] M. Di Marco and L. Santini. Human pressures predict species' geographic range size better than biological traits. *Global Change Biology*, 21(6):2169–2178, 2015.
- [16] M. Di Marco, O. Venter, H. P. Possingham, and J. E. Watson. Changes in human footprint drive changes in species extinction risk. *Nature communications*, 9(1):4621, 2018.
- [17] S. M. Díaz, J. Settele, E. Brondízio, H. Ngo, M. Guèze, J. Agard, A. Arneth, P. Balvanera, K. Brauman, S. Butchart, et al. The global assessment report on biodiversity and ecosystem services: Summary for policy makers. 2019.
- [18] J. M. Drake, C. Randin, and A. Guisan. Modelling ecological niches with support vector machines. *Journal of applied ecology*, 43(3):424–432, 2006.

- [19] J. Elith, C. Graham, R. Valavi, M. Abegg, C. Bruce, A. Ford, A. Guisan, R. J. Hijmans, F. Huettmann, L. Lohmann, B. Loiselle, C. Moritz, J. Overton, A. T. Peterson, S. Phillips, K. Richardson, S. Williams, S. K. Wiser, T. Wohlgemuth, and N. E. Zimmermann. Presence-only and Presence-absence Data for Comparing Species Distribution Modeling Methods. *Biodiversity Informatics*, 15(2):69–80, July 2020.
- [20] J. Elith, J. R. Leathwick, and T. Hastie. A working guide to boosted regression trees. *Journal of animal ecology*, 77(4):802–813, 2008.
- [21] B. J. Enquist, X. Feng, B. Boyle, B. Maitner, E. A. Newman, P. M. Jørgensen, P. R. Roehrdanz, B. M. Thiers, J. R. Burger, R. T. Corlett, et al. The commonness of rarity: Global and future distribution of rarity across land plants. *Science advances*, 5(11):eaaz0414, 2019.
- [22] J. Estopinan, M. Servajean, P. Bonnet, F. Munoz, and A. Joly. Deep species distribution modeling from sentinel-2 image time-series: a global scale analysis on the orchid family. *Frontiers in Plant Science*, 13:839327, 2022.
- [23] M. J. Feldman, L. Imbeau, P. Marchand, M. J. Mazerolle, M. Darveau, and N. J. Fenton. Trends and gaps in the use of citizen science derived data as input for species distribution models: A quantitative review. *PloS one*, 16(3):e0234587, 2021.
- [24] W. Fithian, J. Elith, T. Hastie, and D. A. Keith. Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, 6(4):424–438, 2015.
- [25] S. D. Foster and P. K. Dunstan. The analysis of biodiversity using rank abundance distributions. *Biometrics*, 66(1):186–195, 2010.
- [26] M. A. Friedl, D. Sulla-Menashe, B. Tan, A. Schneider, N. Ramankutty, A. Sibley, and X. Huang. Modis collection 5 global land cover: Algorithm refinements and characterization of new datasets. *Remote sensing of Environment*, 114(1):168–182, 2010.
- [27] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.
- [28] GBIF.Org User. Occurrence download, 2022.
- [29] T. Hastie and W. Fithian. Inference from presence-only data; the ongoing controversy. *Ecography*, 36(8):864–867, 2013.
- [30] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [31] K. S. He, B. A. Bradley, A. F. Cord, D. Rocchini, M.-N. Tuanmu, S. Schmidlein, W. Turner, M. Wegmann, and N. Pettorelli. Will remote sensing shape the next generation of species distribution models? *Remote Sensing in Ecology and Conservation*, 1(1):4–18, 2015.
- [32] D. N. Karger, O. Conrad, J. Böhner, T. Kawohl, H. Kreft, R. W. Soria-Auza, N. E. Zimmermann, H. P. Linder, and M. Kessler. Climatologies at high resolution for the earth’s land surface areas. *Scientific data*, 4(1):1–20, 2017.
- [33] C. Leblanc, A. Joly, T. Lorieul, M. Servajean, and P. Bonnet. Species distribution modeling based on aerial images and environmental features with convolutional neural networks. In *CLEF (Working Notes)*, pages 2123–2150, 2022.
- [34] T. Lorieul, E. Cole, B. Deneu, M. Servajean, P. Bonnet, and A. Joly. Overview of geolifeCLEF 2022: Predicting species presence from multi-modal remote sensing, bioclimatic and pedologic data. In *CLEF (Working Notes)*, pages 1940–1956, 2022.
- [35] T. Lorieul, E. Cole, B. Deneu, M. Servajean, and A. Joly. Overview of GeoLifeCLEF 2021: Predicting species distribution from 2 million remote sensing images. In *Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum*, 2021.
- [36] T. Lorieul, E. Cole, B. Deneu, M. Servajean, and A. Joly. Overview of GeoLifeCLEF 2022: Predicting species presence from multi-modal remote sensing, bioclimatic and pedologic data. In *Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum*, 2022.
- [37] M. Luoto, R. Virkkala, and R. K. Heikkilä. The role of land cover in bioclimatic models depends on spatial resolution. *Global ecology and biogeography*, 16(1):34–42, 2007.

- [38] C. S. Mantyka-Pringle, P. Visconti, M. Di Marco, T. G. Martin, C. Rondinini, and J. R. Rhodes. Climate change modifies risk of global biodiversity loss due to land-cover change. *Biological Conservation*, 187:103–111, 2015.
- [39] T. Mesaglio and C. T. Callaghan. An overview of the history, current contributions and future outlook of inaturalist in australia. *Wildlife Research*, 48(4):289–303, 2021.
- [40] D. A. Miller, K. Pacifici, J. S. Sanderlin, and B. J. Reich. The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution*, 10(1):22–37, 2019.
- [41] O. Ovaskainen, G. Tikhonov, A. Norberg, F. Guillaume Blanchet, L. Duan, D. Dunson, T. Roslin, and N. Abrego. How to make more out of community data? a conceptual framework and its implementation as models and software. *Ecology letters*, 20(5):561–576, 2017.
- [42] S. L. Özesmi and U. Özesmi. An artificial neural network approach to spatial habitat modelling with interspecific interaction. *Ecological modelling*, 116(1):15–31, 1999.
- [43] J. Pearce and S. Ferrier. An evaluation of alternative algorithms for fitting species distribution models using logistic regression. *Ecological modelling*, 128(2-3):127–147, 2000.
- [44] S. J. Phillips, R. P. Anderson, and R. E. Schapire. Maximum entropy modeling of species geographic distributions. *Ecological modelling*, 190(3-4):231–259, 2006.
- [45] S. J. Phillips, M. Dudík, J. Elith, C. H. Graham, A. Lehmann, J. Leathwick, and S. Ferrier. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological applications*, 19(1):181–197, 2009.
- [46] L. Poggio, L. De Sousa, N. Batjes, G. Heuvelink, B. Kempen, E. Ribeiro, and D. Rossiter. Soilgrids 2.0: producing soil information for the globe with quantified spatial uncertainty, soil, 7, 217–240, 2021.
- [47] 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(8):913–929, 2017.
- [48] J. S. Rousseau and M. G. Betts. Factors influencing transferability in species distribution models. *Ecography*, 2022(7):e06060, 2022.
- [49] E. W. Sanderson, M. Jaiteh, M. A. Levy, K. H. Redford, A. V. Wannebo, and G. Woolmer. The human footprint and the last of the wild: the human footprint is a global map of human influence on the land surface, which suggests that human beings are stewards of nature, whether we like it or not. *BioScience*, 52(10):891–904, 2002.
- [50] H. Sormunen, R. Virtanen, and M. Luoto. Inclusion of local environmental conditions alters high-latitude vegetation change predictions based on bioclimatic models. *Polar Biology*, 34:883–897, 2011.
- [51] L. Stewart, F. Bach, Q. Berthet, and J.-P. Vert. Regression as classification: Influence of task formulation on neural network features. In *International Conference on Artificial Intelligence and Statistics*, pages 11563–11582. PMLR, 2023.
- [52] R. Valavi, G. Guillera-Arroita, J. J. Lahoz-Monfort, and J. Elith. Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. *Ecological Monographs*, 92(1):e01486, 2022.
- [53] 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(1):1–10, 2016.
- [54] S. Vignali, A. G. Barras, R. Arlettaz, and V. Braunisch. Sdm tune: An r package to tune and evaluate species distribution models. *Ecology and Evolution*, 10(20):11488–11506, 2020.
- [55] Y. Wang, U. Naumann, S. T. Wright, and D. I. Warton. mvabund—an r package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution*, 3(3):471–474, 2012.
- [56] D. P. Wilkinson, N. Golding, G. Guillera-Arroita, R. Tingley, and M. A. McCarthy. A comparison of joint species distribution models for presence–absence data. *Methods in Ecology and Evolution*, 10(2):198–211, 2019.

- [57] S. N. Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(1):3–36, 2011.