BIODIVERSITY BUILDING BLOCKS FOR POLICY

D4.2: Report on deep learning development

28/02/2025

Author(s): Alexis Joly, Diego Marcos, Maxime Ryckewaert

Funded by the European Union

This project receives funding from the European Union's Horizon Europe Research and Innovation Programme (ID No 101059592). Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the EU nor the EC can be held responsible for them.

Prepared under contract from the European Commission
Grant agreement No. 101059592
EU Horizon Europe Research and Innovation Action

Project acronym:	**B3**
Project full title:	**Biodiversity Building Blocks for Policy**
Project duration:	01.03.2023 – 31.08.2026 (42 months)
Project coordinator:	Dr. Quentin Groom, Agentschap Plantentuin Meise (MeiseBG)
Call:	HORIZON-CL6-2021-GOVERNANCE-01
Deliverable title:	Landscape analysis
Deliverable n°:	D4.2
WP responsible:	WP 4
Nature of the deliverable:	Report
Dissemination level:	Sensitive
Lead partner:	INRIA
Recommended citation:	Ryckewaert, M. Marcos, D. Joly, A. (2025). ***Report on deep learning development***. B3 project deliverable D4.2.
Due date of deliverable:	Month n° 24
Actual submission date:	Month n° 24

Deliverable status:

Version	Status	Date	Author(s)
1.0	Final	28 Feb 2025	Maxime Ryckewaert, Diego Marcos and Alexis Joly

Table of contents

Key takeaway messages

- DeepMaxent extends the MaxEnt framework (based on the principle of maximum entropy) to neural networks, enabling multi-species modeling without the need for separate models per species.
- DeepMaxent and other deep learning approaches, including Cross-Entropy and Binary Cross-Entropy, were evaluated using the NCEAS and GeoPlant datasets.
- Biases related to heterogeneous sampling effort affect all species distribution modeling approaches, including deep learning-based methods.
- To address these biases, two strategies were implemented: direct modeling of the sampling effort and the Target Group Background correction.
- DeepMaxent demonstrates better predictive performance than traditional models (Maxent, Boosted Regression Trees [BRT], and Ensemble methods).
- A simulation platform, creating virtual species from real-data, was used to develop or to evaluate these corrections using ground truth data at the continent scale.

Executive summary

This report investigates the application of deep learning techniques within the B-CUBED project, aimed at enhancing species distribution modeling (SDM) using citizen science and opportunistic data. It highlights the inherent challenges of observation biases in opportunistic datasets, which impact SDM outcomes and are particularly relevant in the context of Deep-SDM methods. Despite these challenges, deep learning offers significant potential for handling large and complex biodiversity datasets effectively.

Key innovations discussed in this report include strategies to address spatial sampling bias, such as modeling sampling effort directly and applying the Target Group Background method. A novel model called DeepMaxent extends the Maxent framework, leveraging maximum entropy principles within a neural network architecture. DeepMaxent surpasses both classical approaches (Maxent, Boosted Regression Trees, Ensemble) and other deep learning methods (Cross-Entropy, Binary Cross-Entropy), as demonstrated using the NCEAS dataset.

An additional contribution is the development of a simulation platform that generates virtual species distributions tied to real-world data, providing a robust environment to evaluate deep learning algorithms against known ground truths. This combination of innovative methods and practical applications underscores the potential of Deep-SDM for advancing biodiversity monitoring and analysis.

Non-technical summary

This document presents how advanced artificial intelligence techniques, specifically deep learning, are being applied to improve predictions of where different species might be found.

The work was carried out as part of the B-CUBED project, which integrates data collected by citizen scientists and other public observations.
In this context, these observations represent *opportunistic* data, collected in a non-systematic manner, often relying on chance encounters or voluntary contributions rather than a structured, planned methodology. Such data, common in citizen science, is gathered whenever and wherever individuals happen to make observations, rather than through a rigorous scientific sampling protocol. As a result, some areas or species may be overrepresented while others are underrepresented, leading to uneven sampling and potential biases in the data.

To meet this challenge, the report explores methods for correcting the sampling effort and improving the reliability of the models. However, beyond reducing bias, a major difficulty lies in managing the growing volume and complexity of the data. With data sets such as GBIF continuing to grow and incorporating an increasingly large set of covariates, extracting meaningful information is becoming increasingly difficult. Large datasets introduce non-linearity, heterogeneity, missing values and biases, requiring more sophisticated modelling approaches. To solve these problems, the report presents DeepMaxent, a powerful new model that combines traditional statistical techniques with modern deep learning. Unlike conventional models, which are widely used in species distribution modelling but often require separate models for each species and struggle to handle large and complex datasets, DeepMaxent can handle several species simultaneously, offering a methodological perspective.

The report also describes the creation of a simulation platform for testing these methods using synthetic species data linked to real-world conditions. This framework allows researchers to measure how well new algorithms perform when the ground truths are known, helping to improve predictive accuracy for real species. Together, these advancements make significant progress in understanding and modeling biodiversity in a rapidly changing world.

List of aList of abbreviations

EU	European Union
SDM	Species Distribution Modelling
Deep-SDM	Deep Learning Species Distribution Modelling
NCEAS	National Centre for Ecological Analysis and Synthesis
GBIF	Global Biodiversity Information Facility
PO	Presence Only
PA	Presence Absence
AUC	Area Under the Curve

1.Introduction

1.1. Citizen science, opportunistic data and modelling

The Global Biodiversity Information Facility (GBIF) database is enriched by a combination of probabilistic and opportunistic samples, known as preferred samples. Probabilistic samples are selected at random using statistical methods, providing an impartial and generalisable representation of biodiversity in a given region. Opportunistic samples, on the other hand, often come from unsystematic collections by researchers or amateurs via citizen science applications (Bonnet et al., 2020; Callaghan et al., 2022). Opportunistic data may be influenced by the accessibility of sites, the season, or species of particular interest. The combination of these two types of sampling enables GBIF to maximise the quantity and diversity of the data collected, while mitigating the biases inherent in each method taken in isolation. In this way, this integrative approach offers a more complete and nuanced view of the world's biodiversity.

However, despite this integration, several biases persist in biodiversity datasets derived from citizen science initiatives. These include operator preference, where specific species or habitats are favored; accessibility bias, which skews data collection toward easily reachable areas; and seasonal bias, as observations are often concentrated in periods of favorable weather conditions. Additionally, modeling biodiversity using such heterogeneous, complex, and often incomplete data presents significant challenges. There is a pressing need to develop methodologies that not only enhance predictive performance and better approximate real-world biodiversity patterns but also incorporate debiasing techniques to minimize the impact of identified biases. Moreover, the validation of these methods is typically conducted on smaller, more restricted areas with significantly fewer validation data compared to the calibration dataset, adding another layer of complexity to model reliability and generalizability.

In biodiversity modeling, multiple covariates are used alongside species occurrence data. These covariates include terrestrial observations such as climate data, remote sensing images from satellites, or soil properties. Additionally to species dimension, the spatial (latitude and longitude) and temporal dimensions are considered within frameworks like the B-CUBED grid system, which organizes data into spatiotemporal cubes for improved analysis. At a given location and time, various types of data can be linked, including satellite images, tabular environmental variables, and time series data.
Other potential covariates include land use and land cover classifications, elevation, vegetation indices (e.g., NDVI), and anthropogenic impact indicators such as urbanization and infrastructure development. Integrating these diverse data sources enables a more holistic understanding of biodiversity patterns, but also necessitates advanced data harmonization techniques to account for inconsistencies and scale mismatches between datasets.

1.2. Methods for SDM (SDM)

Species Distribution Modeling (SDM) employs a variety of statistical and machine learning techniques to predict the occurrence, abundance, or habitat suitability of species across geographic landscapes based on environmental variables (Li & Wang, 2013; Valavi et al., 2022).

Traditional methods include Generalized Linear Models (GLMs), which extend linear regression by allowing for non-normal response distributions using link functions, and Generalized Additive Models (GAMs) (McCullagh, 2019), which add flexibility by incorporating smooth functions to capture non-linear relationships. Regularized regression approaches, like Lasso (L1) and Ridge (L2) regression, address overfitting by penalizing large coefficients, improving model stability with high-dimensional data. Multivariate Adaptive Regression Splines (MARS) (Friedman, 1991) offer a non-parametric regression technique that models relationships through piecewise linear splines, useful for capturing complex interactions between predictors.

More advanced machine learning models are also widely applied. MaxEnt (Maximum Entropy) developed by (Phillips et al., 2006) and its modern extension, MaxNet (Phillips et al., 2017), are popular for species presence-only (PO) data, estimating the probability distribution of a species with the least bias given known constraints. Boosted Regression Trees (BRT) and Gradient Boosting Machines (GBM) iteratively combine weak predictive models to improve accuracy, while Random Forests (RF) and Conditional Inference Forests (cforest) construct ensembles of decision trees to reduce variance and bias. Extreme Gradient Boosting (XGBoost) builds upon GBM with optimized performance, making it highly efficient for large datasets. Support Vector Machines (SVMs) maximize the margin between classes for classification and regression, particularly useful when data are sparse or have complex boundaries. Lastly, ensemble methods, including those implemented in the *biomod2* R package (Thuiller et al., 2009), combine multiple models to enhance robustness and prediction reliability. These methods encompass both traditional ensemble techniques, such as bagging and stacking, as well as boosted methods like Gradient Boosting Machines (GBM) and Boosted Regression Trees (BRT), which iteratively refine predictions by minimizing errors. By integrating diverse modeling approaches, ensemble methods provide consensus predictions that reduce individual model biases and improve overall performance. Together, these techniques provide a diverse toolbox for modeling species distributions, each with strengths depending on data structure, ecological processes, and study objectives.

2.Deep learning algorithms

2.1. General definition

Deep learning algorithms are a subset of machine learning methods that use artificial neural networks, generally composed at least of three layers: an input layer, a hidden layer and an output layer, to model and solve complex problems involving large amounts of data. A neural network is a computational model inspired by the structure consisting of layers of interconnected nodes or "neurons." Each neuron receives inputs, processes them through a non-linear activation function, and passes the output to the next layer. Neural networks are typically organized into three types of layers: an input layer that receives the data, hidden layers where the learning and feature extraction occur, and an output layer that generates the final prediction. The power of deep learning lies in its ability to iteratively learn hierarchical features from raw data by optimizing a statistical loss function, making it highly effective for tasks such as image

classification, natural language processing, and speech recognition. Unlike traditional machine learning models that often rely on hand-crafted features, deep learning models optimize both feature extraction and prediction simultaneously through backpropagation, where the model adjusts its parameters based on a loss function using gradient-based optimization. This flexibility and scalability have made deep learning foundational to modern artificial intelligence applications.

2.2. Deep learning for SDM

Deep learning models have become increasingly prominent in the field of species distribution modelling. These models are capable of processing vast amounts of biodiversity data, effectively capturing the intricate, non-linear relationships between various environmental factors and the presence or absence of species (Deneu et al., 2021; Estopinan et al., 2024). By leveraging environmental and remote sensing variables, deep learning techniques can uncover patterns that traditional methods might miss (Kellenberger et al., 2024).

However, this adaptability also means that deep learning models can inadvertently incorporate and magnify existing biases in the data. When working with datasets that are biased or unbalanced in terms of species representation, the models might produce skewed predictions. This issue underscores the importance of improving the robustness of deep learning methodologies in species distribution modelling.

To address these challenges, researchers are exploring advanced techniques and strategies to mitigate biases and enhance model reliability. Efforts are focused on developing more sophisticated approaches to handle imbalanced data, ensuring that the predictions are more accurate and generalizable. As the field evolves, the integration of robust deep learning models promises to significantly advance our understanding of species distribution and support more effective conservation efforts (Beery et al., 2021).

2.3.Matching loss functions to Machine Learning Objectives

Selecting an appropriate loss function is essential for aligning a machine learning model with its predictive objective (Ciampiconi et al., 2023), particularly in ecological modeling and biodiversity studies. Cross-entropy loss is a widely used function for multi-class classification problems (Demirkaya et al., 2020), such as predicting species list from data. It calculates the divergence between predicted class probabilities and the true class labels, ensuring the model correctly distinguishes between multiple species categories. In contrast, binary cross-entropy loss is suited for binary classification tasks, often used in PA models where the objective is to determine if a species is present in a given location or absent. This loss function measures the performance of a model predicting probabilities between two classes. For modeling event counts or spatial point processes (Renner et al., 2015), Poisson loss is appropriate, particularly when the data involves count-based responses such as the number of sightings of a species in a specific region. It assumes a Poisson distribution for the target variable, making it well-suited for ecological applications like species abundance modeling or predicting rare event distributions. Choosing the correct loss function not only improves model performance but also ensures that the predictions are meaningful and interpretable in the context of ecological phenomena.

2.4. Tailoring neural networks to fit model input covariates

In deep learning, various neural network architectures are designed to handle specific data types and objectives, much like how diverse ecosystems are adapted to distinct environmental conditions. Multi-Layer Perceptrons (MLPs), composed of fully connected layers, are suited for structured tabular data. Convolutional Neural Networks (CNNs) have shown particular promise in processing spatial patterns (Botella et al., 2018; Deneu et al., 2021) that are valuable for tasks like species modelling through image from remote sensing data. More recently, Transformer architectures, originally developed for natural language processing, have found applications in large-scale biodiversity research, where complex relationships between species and environmental variables require attention mechanisms. Transformers provide an alternative method for extracting complex patterns by dynamically weighting the importance of different inputs. This ability allows them to model ecological interactions, seasonal patterns, and spatial heterogeneity more effectively. Selecting the right architecture is akin to choosing the appropriate ecological model: the success of predictions or classifications hinges on how well the model structure aligns with the properties of the input data and the complexity of interactions being studied.

3. Algorithms development in B-CUBED

3.1. Applying maximum entropy principle to neural networks (DeepMaxent)

Useful links:

- Preprint paper: https://doi.org/10.48550/arXiv.2412.19217 submitted to Methods in Ecology and Evolution (under review)
- Source code: https://github.com/RYCKEWAERT/deepmaxent

3.1.1. Motivations

Maxent (Phillips et al., 2006) is a widely used method for modeling species distributions from PO data (Elith* et al., 2006; Valavi et al., 2022; Warren & Seifert, 2011). Maxent estimates a probability distribution over a given area by maximizing entropy while enforcing constraints on environmental features. It optimizes the entropy of this distribution across sites based on predefined transformations of variables, known as features. Maxent is classified as a single-species model, meaning it models the distribution of one species at a time. As a result, it requires the selection of background points, which act as pseudo-absences to contrast observed presences. The choice of these points is crucial, as it influences model accuracy and bias.

Moreover, the key challenge in PO-based species distribution models (SDMs) is spatial sampling bias, caused by the uneven distribution of recorded occurrences, often concentrated in more accessible regions. This bias can distort predictions (Fithian et al., 2015; Phillips et al., 2009a; Yackulic et al., 2013). To address this, Phillips et al, 2009 proposed the Target-Group

Background (TGB) correction, which selects background points from sampled areas based on co-occurring species, reducing false absences. Maxent is mathematically equivalent to Poisson regression and inhomogeneous Poisson Point Processes (PPP) (Renner & Warton, 2013), allowing theoretical validation of TGB's robustness against sampling bias (Botella et al., 2020; Fithian et al., 2015).

However, despite its effectiveness, Maxent faces limitations when dealing with heterogeneous or high-dimensional data. The model relies on handcrafted feature transformations, which may not fully capture complex, non-linear relationships between environmental variables and species distributions. This limitation becomes particularly problematic when incorporating diverse input data sources, such as:

- Remote sensing imagery, which contains high-dimensional spatial information.
- Temporal data, where species responses vary across different seasons or years.
- Multimodal data, integrating environmental, climatic, and ecological interactions.

In such cases, manually defining effective feature transformations is challenging, and traditional Maxent models struggle to scale efficiently. Deep learning provides a natural extension by allowing feature extraction to be learned directly from data, rather than manually engineered. Neural networks can capture complex spatial and temporal dependencies, making them well-suited for multi-species SDMs that incorporate large and heterogeneous datasets.

This motivates the development of DeepMaxent, which extends Maxent's entropy maximization principle within a deep learning framework. By learning latent representations, DeepMaxent aims to improve predictive performance while addressing sampling bias and computational challenges associated with high-dimensional ecological data.

3.1.2. Methodology

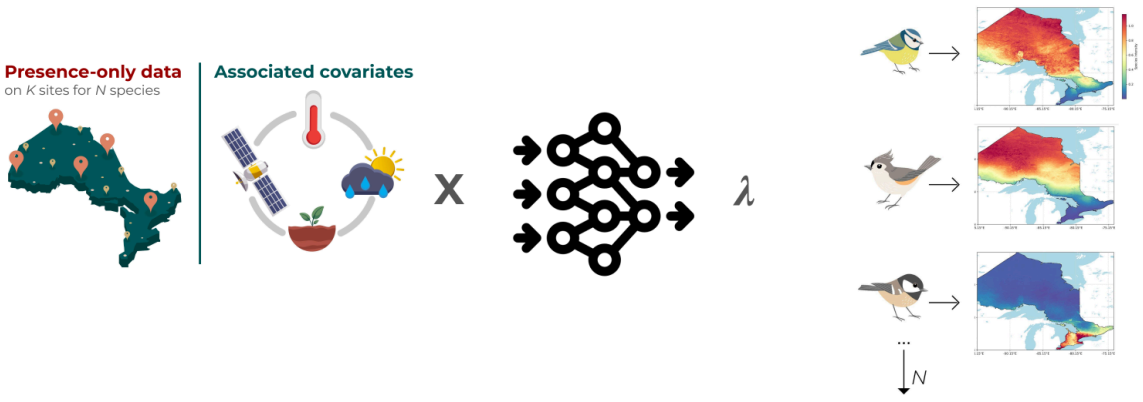

Figure 1: DeepMaxent: a deep learning based method for species distribution modelling using PO data.

We developed a method called DeepMaxent (see figure 1), a deep learning approach integrating Maxent's maximum entropy principle with bias correction mechanisms. DeepMaxent models species distributions as log-linear functions of shared latent features, using a loss function inspired by Maxent and PPP theory:

$$\mathcal{L}_{\mathcal{H},w}(\tilde{\lambda}, y) = -\frac{1}{KN} \sum_{i=1}^{K} w_j \sum_{j=1}^{N} \left(\frac{y_{ij}}{\sum_{k=1}^{K} y_{kj}} \right) \log \left(\frac{\lambda_{ij}}{\sum_{k=1}^{K} \lambda_{kj}} \right)$$

Where K and N represent the number of sites and species, respectively, w_j is a weighting function based on species. y_{kj} denotes the number of occurrences, and λ_{kj} refers to the species intensity at a given location k for a specific species j. We demonstrate that loss computation over site batches preserves global minimization, enabling efficient training via stochastic gradient descent (SGD). Additionally, we incorporate TGB correction within DeepMaxent to mitigate sampling bias. In Maxent, the probability of a species being present at a given site is determined by a set of environmental variables. These variables are transformed into a feature vector using predefined mathematical functions. In DeepMaxent, we extend this approach by replacing the predefined feature transformations with a neural network that learns features directly from the data.

Instead of manually defining how environmental variables are processed, DeepMaxent uses a neural network to extract a shared latent representation across species. This function captures complex, non-linear relationships between environmental conditions and species presence, potentially identifying patterns that traditional methods might overlook.

One of the main advantages of DeepMaxent is its scalability. The neural network calculates a single feature representation per site for all species. This means that, whatever the number of species, the computational cost remains constant. To achieve this, the model processes a multi-dimensional input tensor, where each site corresponds to a feature vector representing all species, allowing the model to efficiently handle additional species without increasing the overall computational load.

This approach is particularly useful when working with deep neural networks capable of handling high-dimensional inputs, such as spatio-temporal data, or when using architectures with multiple hidden layers to capture intricate environmental relationships.

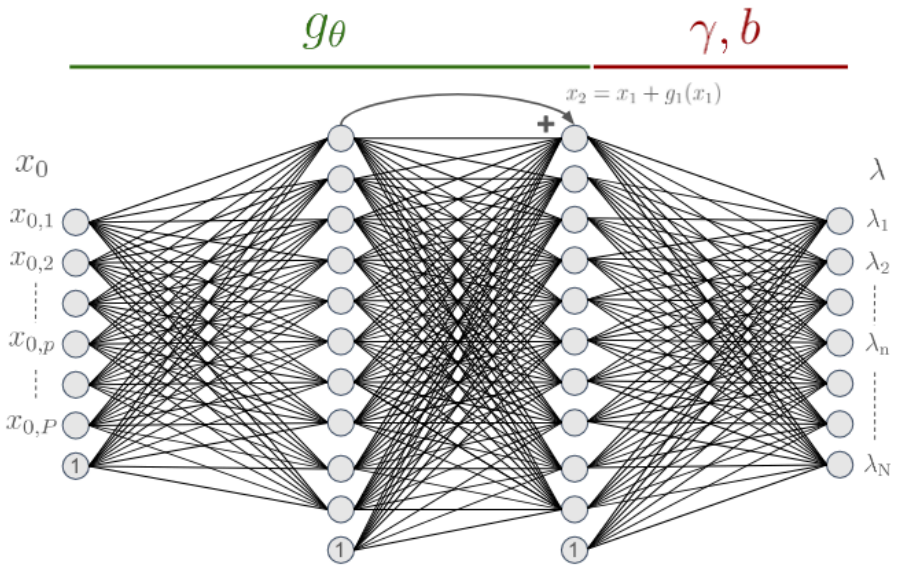

Figure 2: The residual neural network to estimate the intensity from variable input x_0, where P as is the variable number input, C is the number of hidden layer nodes, and N denotes the number of species (or target categories). The illustrated case involves two hidden layers. In the special case where there is only one hidden layer, no residual addition is applied.

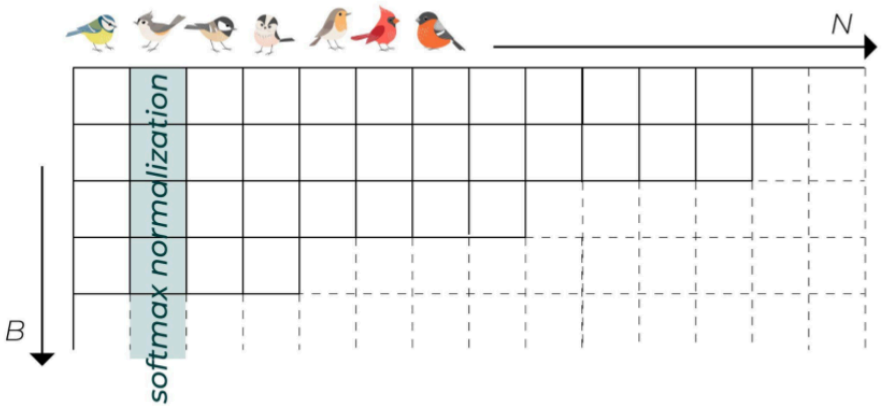

Figure 3: Normalisation is performed per batch using softmax along the 'site' direction for a given species.

Within this framework, we are implementing different DeepMaxent architectures, adapting them to the nature of the input data. Depending on the type of data, we use an MLP (see figure 2), ResNet18 or Transformer to extract significant features before applying the Maxent-inspired output layer. Data varies across domains, including structured tabular data (e.g. categorical and numerical features), temporal data (e.g. sequential sensor readings, stock market trends) and high-dimensional representations (e.g. spectral data, image patches or text embeddings). To extract meaningful features, we use an MLP, ResNet18 or Transformer, depending on the data

type. The MLP is suitable for tabular datasets, while the ResNet18 is used for image or spectral inputs, and the Transformer is suitable for processing sequential patterns, such as time series or textual features. This customised approach ensures efficient feature extraction before applying the Maxent-inspired output layer. The MLP processes structured tabular data, while the ResNet18 and Transformer architectures allow for more complex feature extraction, particularly when dealing with sequential patterns. This flexible approach ensures that Deep Maxent remains adaptable across different input modalities, leveraging deep learning to enhance species distribution modeling.

3.2. Disentangling spatial effort from species intensities

3.2.1.Motivations

Despite their potential, PO data have limitations because they only indicate where a species has been observed, without providing information about where the species is absent. These data are typically derived from opportunistic observations or occurrence records. However, using such observation data introduces several inherent challenges. One major issue is the bias arising from imperfect detection; not all individuals of a species present in an area are observed or recorded. Additionally, variations in sampling efforts across different regions and times can further skew the data. The subjective perspectives of individual observers also contribute to inconsistencies, as some species may be more likely to be reported than others. These factors collectively impact the reliability of species distribution models (SDM) that are trained using PO data (Fithian et al., 2015; Komori et al., 2020; Phillips et al., 2009).

To overcome the limitations of PO data, researchers have devised various methodologies centred around the concept of pseudo-absences. Pseudo-absences, often referred to as background or pseudo-negative points, involve designating certain geographic locations as negative samples to compensate for the absence data. One common approach involves sampling these pseudo-absences uniformly across the geographic space, creating random background points. Another strategy selects pseudo-absences from locations where other species, which are subject to similar sampling biases, have been observed, known as target-group background points. These techniques aim to provide a more balanced dataset, thereby enhancing the accuracy and reliability of species distribution models (SDMs) that are trained with these augmented datasets

3.2.2. Target-Group Background

When occurrence concentration is biased by spatial variations in sampling effort, a common SDM correction is the Target Group Background (TGB) method (Phillips et al., 2009a), originally designed for MaxEnt. TGB approximates spatial sampling effort using the distribution of occurrences from a Target Group (TG) of species, assigning background points to MaxEnt where TG species have been reported. This approach is effective when TG species are recorded alongside the target species, such as in citizen science programs.

A refinement, the TGOB approach (Botella et al., 2020), extends TGB by weighting background points based on the number of TG occurrences per site, better accounting for variations in

sampling effort. Under standard sampling assumptions, TGOB is theoretically robust to spatial bias and, given sufficient data, approximates the focal species' distribution if TG occurrences are well distributed across the study area.

Since MaxEnt is equivalent to a Poisson process on a spatial grid (Renner & Warton, 2013), applying TGB or TGOB in this framework should yield the same bias correction properties. Multiplying the predicted intensity by the number of background points per site in the loss is equivalent to providing the same number of background points in a Poisson loss.

3.2.3.Estimating the sampling effort

Rather than approximating sampling effort a priori using target-group occurrences, we adopt an alternative approach that jointly estimates sampling effort as a function of spatial covariates alongside species intensities. This correction strategy is widely used in Poisson-process SDMs (Botella et al., 2021; Renner et al., 2015; Saigusa et al., 2024; Warton et al., 2013).

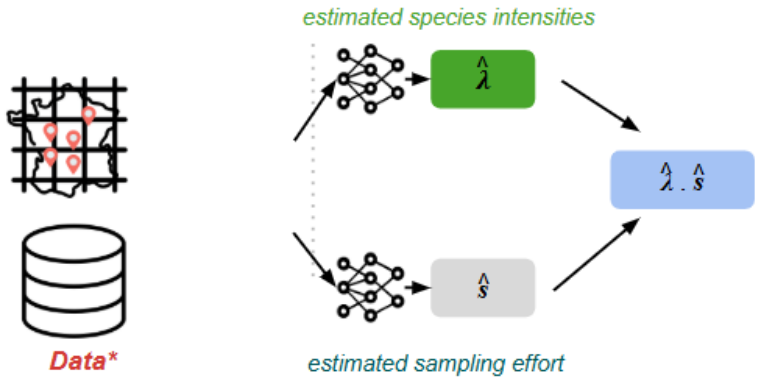

Figure 4: Proposed architecture for separately estimating sampling effort and species intensities.

In this framework, the predicted occurrence count of species is modeled as the product of species intensity and sampling effort, which is shared across species. Estimating the sampling effort directly can lead to identifiability issues (see figure 4), as variations in observed counts may stem from either sampling effort or species intensity. To resolve this, we introduce an additional loss term. This explicit modeling of **s** ensures a well-defined loss, separating sampling effort from species intensity, improving model stability, and preventing overfitting.

$$\mathcal{L}_{\mathcal{P}}(\lambda s, y) = \frac{1}{K.N} \sum_{i=1}^{K} \sum_{j=1}^{N} \left(\lambda_{ij} s_i - y_{ij} \log(\lambda_{ij} s_i) \right)$$

In the baseline scenario, the model only estimates species intensity, assuming that the sampling effort (denoted as **s**) is constant and equal to 1 across all locations. This simplification avoids the direct estimation of sampling effort and treats variations in observed counts solely as a result of species intensity.

$$\mathcal{L}_{\mathcal{P}}(\lambda, y) = \frac{1}{K.N} \sum_{i=1}^{K} \sum_{j=1}^{N} \left(\lambda_{ij} - y_{ij} \log \lambda_{ij}\right)$$

4. Associated open dataset

4.1. Observed Species Occurrence Data

4.1.1. B-CUBED data for species classification (Belgium, 2010)

The dataset was generated using the tools available in the B-CUBED project. This dataset is a typical biodiversity dataset in Belgium. It represents a subset from the year 2010, extracted from a more comprehensive dataset. The data is organised into spatial cubes to facilitate detailed biodiversity analysis for that year. For access to this dataset, please refer to the following resource: https://doi.org/10.15468/dl.e3j5kv.

This dataset is a structured collection of plant occurrence records in Belgium, extracted from the Global Biodiversity Information Facility (GBIF) database. Each row in the dataset represents a unique species occurrence, grouped by year-month (YYYY-MM format) and spatial grid cells (EEA reference grid cells), enabling a structured analysis of plant distributions over time. The dataset includes key attributes such as geographic coordinates (latitude, longitude), species taxonomic information (family, species, speciesKey, familyKey), and occurrence count per grid cell and time unit. To ensure data reliability, the query applies several quality filters:

- Geographic validity: Excludes records with missing or erroneous coordinates (e.g., zero coordinates, out-of-range values, or mismatches with country boundaries).
- Temporal constraints: Includes only records from the year 2000 onward, ensuring relevance for modern ecological analysis.
- Spatial precision: Restricts results to records with a coordinate uncertainty below 100 meters, ensuring high location accuracy.
- Taxonomic focus: Filters for the Plantae kingdom, ensuring only plant species are included.
- Presence-only records: Ensures that only occurrences marked as PRESENT are considered.

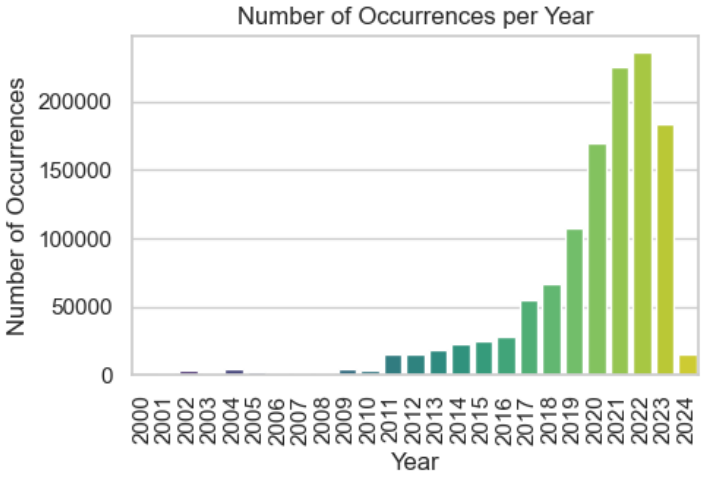

Figure 5: Distribution of 1,209,567 PO occurrences in Belgium, colored by year

The GBIF_EEARGCode function assigns each occurrence to an EEA (European Environment Agency) grid cell, facilitating spatial aggregation and large-scale analysis. The final dataset is grouped by yearMonth, spatial grid cell, and species information, with an occurrence count per species per grid cell per time unit. Results are ordered chronologically (most recent first) and then by grid cell and species.

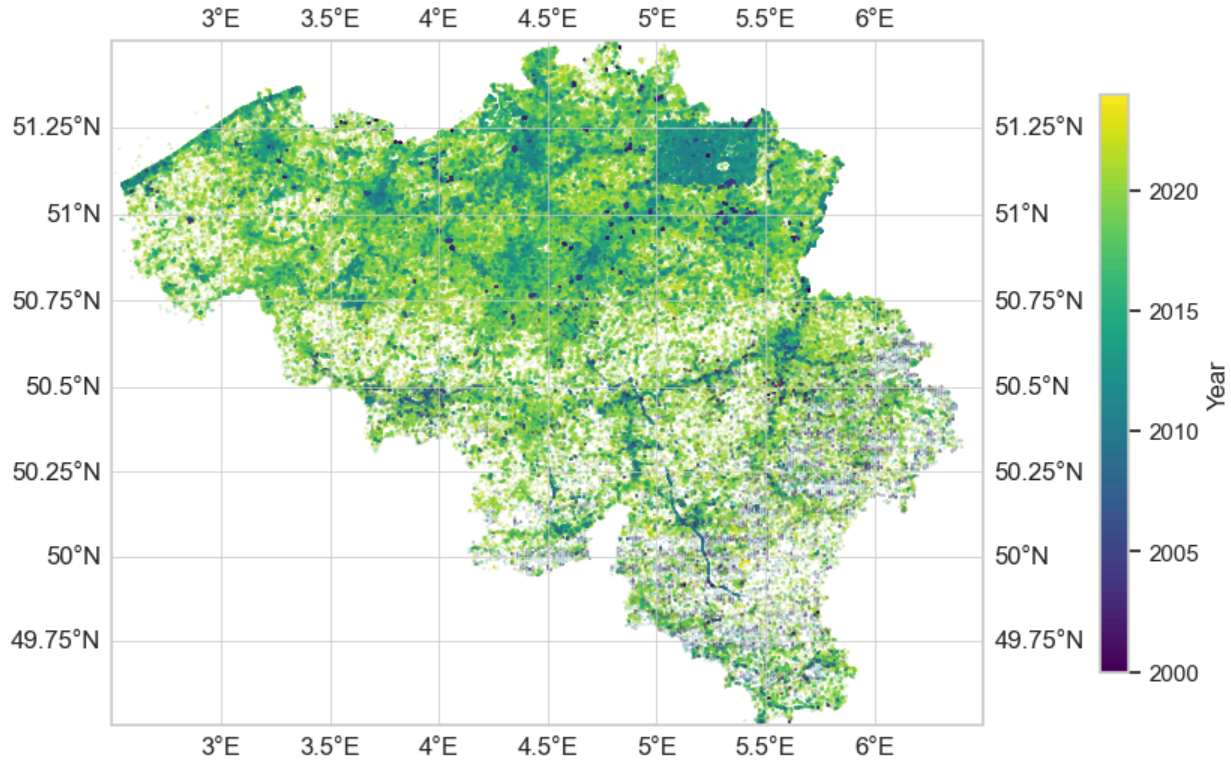

Figure 6: location of 1,209,567 PO occurrences in Belgium, colored by year.

The environmental covariates used in this study consist of 19 bioclimatic rasters derived from the WorldClim and CHELSA databases. These rasters capture key environmental variables such as temperature, precipitation, and altitude, providing essential context for species distribution modeling. The dataset is based on the high-resolution climatologies developed by (Karger et al., 2017), which offer detailed climate data for terrestrial areas worldwide. These bioclimatic variables are crucial for understanding species-environment relationships, as they influence plant distribution and ecological dynamics. The rasters represent various environmental factors such as temperature, precipitation, and altitude. The full dataset is available : https://chelsa-climate.org/bioclim/ and the related paper describing the dataset can be found here: https://doi.org/10.1038/sdata.2017.122

4.1.2.NCEAS dataset: A Benchmark Dataset for Species Distribution Modelling

Data from the National Centre for Ecological Analysis and Synthesis (NCEAS) have been openly released recently (Elith et al., 2020). This dataset includes PO and PA data from six global regions: Australian Wet Tropics (AWT), Canada (CAN), New South Wales (NSW), New Zealand (NZ), South America (SA), and Switzerland (SWI). It comprises data for 226 anonymized species from different biological groups (see table 1 and figure 7 and 8). The dataset contains different environmental predictive variables for each region, including climatic, soil variables or location information (more details in Elith et al., 2020).

This dataset has been used to evaluate and compare various methods (Elith* et al., 2006; Phillips et al., 2009a; Valavi et al., 2022), allowing for comparisons with existing SDM methods. All the species in each biological group in each region are used to form models based on presence data only. The models are then evaluated with PA data using the Area Under Curve (AUC) criterion. Finally, AUC values are averaged by region or for all regions.

Table 1: Details of the Elith dataset where each line corresponds to the data used to create a model.

Code	Location	Biological Group	Species number	Occurrences number(PO)	Occurrences number(PA)
AWT	Australian wet tropics	bird	40	3105	340
AWT	Australian wet tropics	plant	40	701	102
CAN	Ontario, Canada	bird	20	5063	14571
NSW	New South Wales	bate	54	187	570
NSW	New South Wales	bird	54	1781	1839
NSW	New South Wales	plant	54	680	5329
NSW	New South Wales	reptile	54	675	1008
NZ	New Zealand	plant	52	3088	19120
SA	South America	plant	30	2220	152
SWI	Switzerland	tree	30	35105	10013

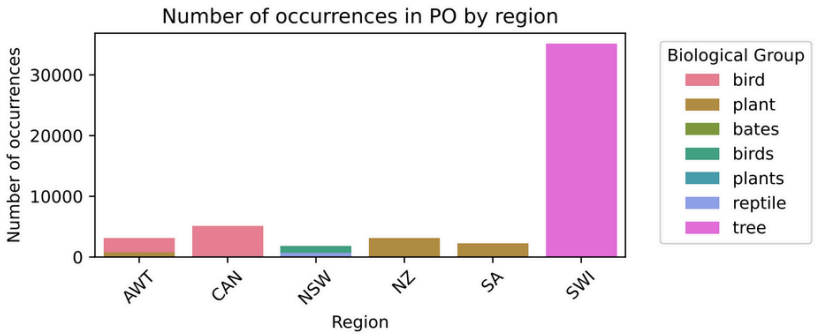

Figure 7: Total number of occurrences by region in PO data

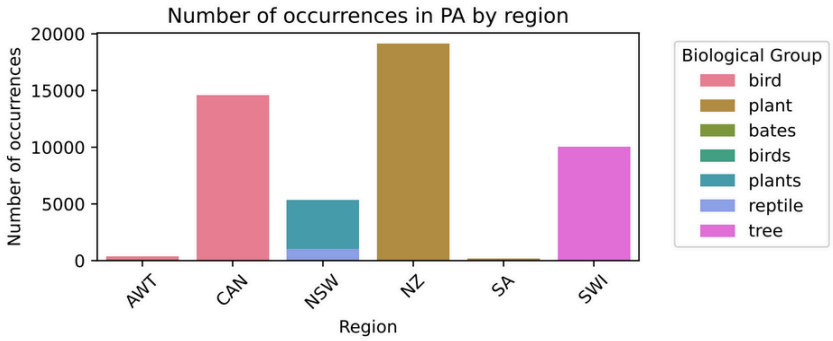

Figure 8: Total number of occurrences by region in PA data

4.1.3. GeoPlant: A Large-Scale Dataset

The GeoPlant dataset [3] is a large-scale dataset focused on plant distribution. It comprises 5,079,797 PO observations sourced from the Global Biodiversity Information Facility (GBIF, www.gbif.org), along with 88,987 PA survey records from the European Vegetation Archive (EVA).

GeoPlant integrates covariates from remote sensing and environmental variables:

- Sentinel-2 Imagery: High-resolution image patches providing spectral data for vegetation analysis.
- Landsat Time Series: Multi-temporal data capturing long-term vegetation dynamics, including spectral bands (R, G, B, NIR, SWIR1, SWIR2).
- Bioclimatic Variables: 19 variables from CHELSA WorldClim, describing temperature and precipitation patterns that shape vegetation distribution.
- Human Footprint Index: Measures anthropogenic impacts, including land use intensity and infrastructure development.

4.1.4. A simulation framework from real dataset

Evaluating deep learning methods for species distribution modelling (SDM) on a large scale, such as at a continental level, presents significant challenges due to the lack of reliable ground truth data. Real-world field observations are often biased and incomplete. Species occurrence data typically come from collaborative databases or survey campaigns, where sampling effort is rarely uniform across space. This leads to sampling biases influenced by factors such as human population density, site accessibility, and researchers' interest in particular regions or species. These biases complicate the validation of predictive models and the accurate assessment of their performance.

To address these limitations, a simulation platform has been developed that generates synthetic data simulating virtual species distributions and associated sampling processes. This approach provides several key benefits:

1. Known ground truth: Unlike real data, where the true distribution or intensity of species is unknown, simulation allows complete knowledge of the underlying parameters determining species presence or absence. This provides a clear benchmark for objectively evaluating the performance of deep learning algorithms.

2. Control over sampling biases: Simulation enables the intentional introduction of specific sampling biases to study their effects on predictive accuracy. In our case, we focus on the impact of spatially heterogeneous sampling effort. By manipulating variables such as the density of sampling points relative to human infrastructure or protected areas, we can assess the robustness of deep learning models to these variations in sampling coverage.

3. Standardised model comparison: The simulation framework provides a controlled environment where different algorithms can be evaluated under identical conditions, ensuring that performance differences arise from the algorithms themselves rather than from unknown biases in real-world data.

To do this, a simulation framework has been developed to build and evaluate deep learning algorithms. This solution has been proposed in order to obtain ground truth. To this end, a database was used to generate virtual species. The means and standard deviations of the real species are used to generate the virtual species that have relationships between the input variables. For a virtual species n, we construct a virtual *ground truth* intensity function. To do so, species intensity is built as a multivariate Gaussian of climatic variables:

$$f(\mathbf{x}; \mu_n, \Sigma_n) = \frac{1}{\sqrt{(2\pi)^k \det(\Sigma_n)}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_n)^T \Sigma^{-1} (\mathbf{x} - \mu_n)\right)$$

Where $\mathbf{x}$ are the bioclimatic variables used to simulate ground truth response and μ_n and Σ_n are estimated by sampling from one randomly selected real species.

5. Model evaluation

5.1. Comparison of methods and criteria

For the Belgium case study, we evaluate the use of deep learning for predicting species lists. To achieve this, we compare Cross Entropy (CE) and Binary Cross Entropy (BCE), as they are better suited for this specific task. Deep Maxent loss was not considered, as it is not appropriate for this type of prediction due to its scale-independent nature.

To assess model performance, we compute F1-score, Precision, and Recall. However, due to the lack of PA data, we rely on a spatial split strategy to divide the dataset into calibration, validation, and test sets. On the other hand, for NCEAS and GeoPlant datasets, all models were trained on PO data and evaluated using the Area Under the Curve (AUC) metric on PA data. Unlike the classification metrics used so far, AUC measures how well the values are ranked. The closer it is to 1, the better the ordering of the values aligns with the expected ranking.

For the NCEAS dataset, we compared Cross Entropy (CE), Binary Cross Entropy (BCE), and Poisson Loss and DeepMaxent losses to classical species distribution models (SDMs) commonly used in the literature, such as Maxent and Boosted Regression Trees (BRT). Additionally, we evaluated these implemented loss functions with and without TGB correction to analyze their impact on model performance.

The experiments were conducted using Landsat-derived time-series data, enabling us to assess how temporal patterns influence model performance across different loss formulations. To capture these temporal dependencies, we tested the ResNet18 and Transformer architectures for feature extraction.

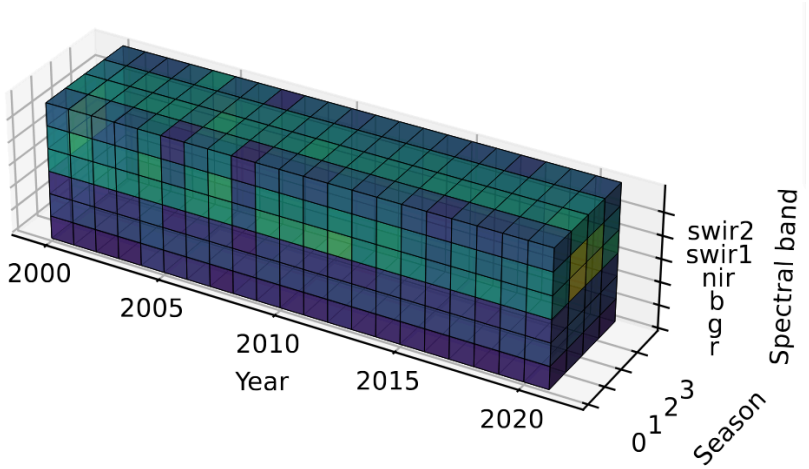

Figure 9: A cube representation of Landsat time series for a single observation at a specific location (latitude, longitude). The cube has dimensions of 6 spectral bands × 4 seasons × 21 years, where each value represents the seasonal median of a given spectral band for a specific year.

5.2. Results and discussion

5.2.1.Belgium case

Table 2 presents metrics from models trained using Cross Entropy (CE) and Binary Cross Entropy (BCE).

Table 2: Comparison of classification metrics for predicting species list.

Method	Precision	Recall	F1-score
CE	0.184	0.336	0.233
BCE	0.149	0.314	0.201

The results show that CE performs better than BCE across all metrics, precision, recall, and F1-score. The relatively low precision values for both methods indicate a high number of false positives, suggesting challenges in differentiating between classes. This could be due to the complexity of species co-occurrence patterns, biases in the training data, or limitations in the model's capacity to capture intricate relationships within the data. Additionally, evaluating models on a PO dataset concentrated in similar regions does not necessarily provide a reliable estimate of true species absences. In this case, the results incorporate the target group background (TGB) correction, as the model is trained only on the available data, meaning that absences correspond to the non-observation of other species within the same target group.
The lower recall of BCE suggests that it may fail to detect certain species as effectively as CE, possibly due to differences in how the loss function penalizes misclassifications.

5.2.2.NCEAS

Table 3 presents the performance of various standard comparative methods, including Maxent, Boosted Regression Tree (BRT) with or without TGB correction (Elith et al., 2020; Phillips et al., 2009b), as well as the multi-species neural network model proposed from literature (Zbinden et al., 2024). These methods are evaluated based on the average AUC per region and the overall average, as detailed in the studies. Additionally, the table includes the performance of our implemented losses including Cross-Entropy (CE), Binary-Cross-Entropy (BCE), Poisson and DeepMaxent, both with and without TGB correction.

Table 3: Comparison of method performance by region-averaged AUC and general averaged AUC over all regions.

Method	**Region**						
	AWT	CAN	NSW	NZ	SA	SWI	AVG

MaxEnt	0.686	0.587	0.700	0.738	0.804	0.809	0.721
BRT	0.681	0.577	0.701	0.735	0.795	0.816	0.718
RF	0.675	0.572	0.715	0.746	0.813	0.818	0.723
Ensemble	0.683	0.580	0.710	0.749	0.806	0.812	0.723
Zbinden	0.704	0.714	0.719	0.741	0.815	0.838	0.755
CE	0.701	0.661	0.732	0.724	0.772	0.793	0.731
BCE	0.656	0.600	0.718	0.736	0.804	0.799	0.719
Poisson	0.658	0.599	0.714	0.737	0.804	0.799	0.719
DeepMaxent	0.654	0.593	0.718	0.744	0.803	0.810	0.720
Using target-group background----------							
MaxEnt	0.732	0.716	0.741	0.738	0.798	0.837	0.760
BRT	0.700	0.728	0.738	0.740	0.792	0.842	0.757
CE	0.727	0.708	0.739	0.732	0.771	0.792	0.745
BCE	0.723	0.726	0.743	0.739	0.803	0.846	0.763
Poisson	0.712	0.727	0.732	0.731	0.800	0.846	0.758
DeepMaxent	0.714	0.732	0.752	0.754	0.803	0.850	0.767

Without TGB sampling bias correction, the performances of the different methods remain close, ranging from 0.718 to 0.723 in terms of general average AUC, with the notable exception of the CE loss, which performs significantly better (0.731). Aside from CE, we observe no overall performance gain from the tested deep learning losses (BCE, Poisson, DeepMaxent, ranging from 0.719 to 0.720) compared to traditional methods such as Maxent (0.721) or the best SDM ensemble (0.723). The TGB correction consistently enhances performance, increasing the average AUC across all regions for both existing methods from the literature and our implementations. However, the impact of the correction varies across approaches. For example, Maxent and BRT each gain 0.039 in general average AUC with TGB. Among our baseline losses, TGB leads to an AUC improvement of 0.011 for CE, 0.044 for BCE, and 0.039 for Poisson. Notably, DeepMaxent benefits the most, with a gain of 0.047, achieving the highest general AUC (0.767). These results highlight that DeepMaxent is also well-suited to this bias correction technique. The most significant region-specific AUC gains occurred in CAN and AWT, where spatial sampling bias is strongest. Additionally, the ‘Zbinden’ method, which reached a general average AUC of 0.755, incorporated both random and TGB points as absences in its BCE loss, demonstrating the crucial role of TGB points in achieving this performance. DeepMaxent-TGB achieved the highest AUC in four of the six regions (CAN, NSW, NZ, SWI), showcasing its robustness across different regions and biological groups (NSW includes four biological groups, see Table 1). BCE-TGB ranks as the second-best method overall, with a general AUC of 0.763. In contrast, CE-TGB and Poisson-TGB show lower general AUC values (0.745 and 0.758, respectively) compared to Maxent-TGB (0.760) and BRT-TGB (0.759).

5.2.3.GeoPlant

Table 4: Comparison of the average AUC across all species for two architectural types and all loss functions.

Method	Architecture	
	ResNet18	Transformer
CE	0.864	0.862
BCE	0.887	0.882
Poisson	0.867	0.863
DeepMaxent	0.885	0.882

BCE achieves the highest average AUC with both architectures (0.887 with ResNet18, 0.882 with Transformer). DeepMaxent follows closely (0.885 with ResNet18, 0.882 with Transformer). Poisson (0.867, 0.863) and CE (0.864, 0.862) show lower AUC values. ResNet18 slightly outperforms Transformer across all loss functions.
ResNet18 currently achieves the highest average AUC, suggesting it is the most effective architecture for this task. However, this advantage could also stem from the choice of hyperparameters rather than an inherent superiority of the model itself. The relatively small difference in AUC values between ResNet18 and Transformer indicates that both architectures are viable for capturing temporal patterns in PA data. This suggests that convolutional approaches effectively extract spatial features, particularly local patterns, while their ability to capture broader dependencies can be enhanced through techniques such as dilated convolutions or global pooling. Transformer-based models, which are inherently well-suited for capturing long-range dependencies, may require further optimization to fully leverage their potential in this context. The performance of BCE as the best loss function aligns with its suitability for binary classification tasks like PA modeling, but its effectiveness may also depend on class imbalance handling strategies such as weighted losses or sampling adjustments. The results highlight the importance of further investigating hyperparameter tuning, as variations in learning rate, batch size, or regularization could impact the observed differences. Additionally, testing these architectures on different types of data, such as remote sensing imagery or species abundance datasets, could provide valuable insights into their adaptability and potential for broader ecological applications.

5.2.4.Simulation case study

Figure 10 illustrates the different components of the simulation process. The left panel represents the definition of sampling effort, based on a total of 5 million occurrences distributed across the study area. The middle panel shows the simulated species intensities for a virtual species, serving as the ground truth. Finally, the right panel presents the resulting occurrence distribution, obtained by multiplying the species intensities by the sampling effort, following a

Poisson-distributed observation process. All values are displayed on a logarithmic scale, except for the virtual species intensities, due to the peak values in the sampling effort.

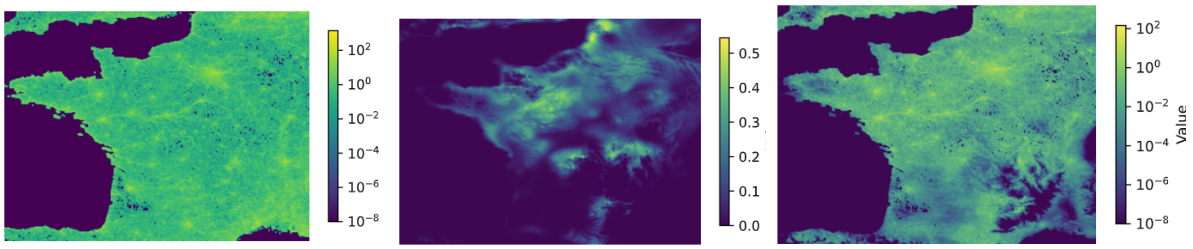

Figure 10: (Left) Definition of sampling effort based on a total of 5 million occurrences. (Middle) Simulated species intensities for a virtual species (ground truth). (Right) The resulting occurrence distribution, obtained by multiplying species intensities by sampling effort. All scales are logarithmic, except for virtual species intensities, due to the peak value in sampling effort.

Figure 11 highlights the impact of accounting for sampling effort when predicting species intensities. The left panel shows the predicted species intensities using a Poisson loss without explicitly considering sampling effort, where patterns related to sampling bias are still present. The middle panel displays the results obtained using the proposed disentangling method, which effectively reduces the influence of sampling bias, leading to more accurate species intensity estimates. The right panel represents the estimated sampling effort based on input variables, providing valuable information for future work. This confirms both the importance of addressing sampling biases in species distribution modeling and the effectiveness of the disentangling approach. Additionally, having an estimate of sampling effort opens the possibility of developing a dual-head DeepMaxent model, specifically designed to separate ecological patterns from sampling-related biases.

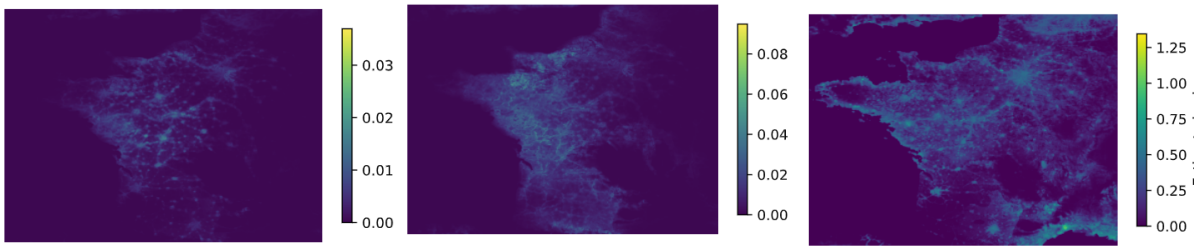

Figure 11: (Left) Predicted species intensities using Poisson loss without explicitly accounting for sampling effort. (Middle) Predicted species intensities using the proposed disentangling method. (Right) Estimated sampling effort based on input variables.

6. Conclusion and perspectives

Neural networks and deep learning enable models to learn complex patterns from large datasets. Deep learning, a subset of machine learning, refers to neural networks with multiple hidden layers that allow hierarchical feature extraction, making them particularly interesting for ecological modelling. In the context of species distribution modeling, deep learning architectures, such as convolutional neural networks (CNNs), transformers, and multi-layer perceptrons (MLPs), can effectively capture spatial, temporal, and spectral patterns, providing new opportunities for biodiversity research and conservation planning.

DeepMaxent, the proposed extension of the Maxent framework to neural networks, has demonstrated superior predictive performance compared to traditional models such as Maxent, Boosted Regression Trees (BRT), and Ensemble methods. Unlike these approaches, which require separate models for each species, DeepMaxent operates as a multi-species model, making it more scalable and efficient for biodiversity modeling. When evaluated on the NCEAS and GeoPlant datasets, deep learning-based methods, including DeepMaxent, Cross-Entropy (CE), and Binary Cross-Entropy (BCE), showed sensitivity to biases introduced by heterogeneous sampling effort, highlighting the need for bias correction techniques. Two key strategies: direct modeling of sampling effort and the Target Group Background (TGB) approach were implemented to mitigate these biases, improving model reliability.

The choice of the most suitable architecture and loss function depends on the modeling objective. BCE loss is better suited for predicting species PA lists, while CE is more appropriate for capturing species mixtures at a given site. Poisson loss aligns well with plant abundance modeling, whereas a Maxent-type loss is preferable when estimating relative probability densities across sites. ResNet and Transformer architectures excel at extracting spatial, temporal, and spectral patterns, while Multi-Layer Perceptrons (MLPs) remain well-suited for tabular data. A hybrid approach combining these architectures could be a promising direction, but its effectiveness is likely data-dependent, varying with dataset quality and structure.

An important methodological advancement is the use of data aggregation in cubes, allowing species occurrence data to be matched more effectively with remote sensing and environmental variables. By structuring data into grid-based representations, this approach enhances data availability and facilitates integration with satellite imagery and other large-scale datasets, making it particularly relevant for deep learning-based species distribution modeling. Recent developments such as DeepMaxent and deep SDM frameworks demonstrate the potential of leveraging these aggregated data structures for improved biodiversity modeling at continental scales.

7. Acknowledgements

We sincerely appreciate the contributions of the WP4 partners and the B3 partners for their valuable feedback. We would also like to thank all members who collaborated throughout the study, whether in the review process or in drafting the report. A special thank you to the B3 internal reviewers, Niels Billiet and Rocio Beatriz Cortes Lobos, for their dedicated efforts.

8. References

Beery, S., Cole, E., Parker, J., Perona, P., & Winner, K. (2021). Species Distribution Modeling for Machine Learning Practitioners: A Review. *ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS)*, 329–348. https://doi.org/10.1145/3460112.3471966

Bonnet, P., Joly, A., Faton, J., Brown, S., Kimiti, D., Deneu, B., Servajean, M., Affouard, A., Lombardo, J., Mary, L., Vignau, C., & Munoz, F. (2020). How citizen scientists contribute to monitor protected areas thanks to automatic plant identification tools. *Ecological Solutions and Evidence*, *1*(2), Article 2. https://doi.org/10.1002/2688-8319.12023

Botella, C., Joly, A., Bonnet, P., Monestiez, P., & Munoz, F. (2018). Species distribution modeling based on the automated identification of citizen observations. *Applications in Plant Sciences*, *6*(2), Article 2. https://doi.org/10.1002/aps3.1029

Botella, C., Joly, A., Bonnet, P., Munoz, F., & Monestiez, P. (2021). Jointly estimating spatial sampling effort and habitat suitability for multiple species from opportunistic presence-only data. *Methods in Ecology and Evolution*, *12*(5), Article 5. https://doi.org/10.1111/2041-210X.13565

Botella, C., Joly, A., Monestiez, P., Bonnet, P., & Munoz, F. (2020). Bias in presence-only niche models related to sampling effort and species niches: Lessons for background point selection. *PLOS ONE*, *15*(5), Article 5. https://doi.org/10.1371/journal.pone.0232078

Callaghan, C. T., Mesaglio, T., Ascher, J. S., Brooks, T. M., Cabras, A. A., Chandler, M., Cornwell, W. K., Cristóbal Ríos-Málaver, I., Dankowicz, E., & Urfi Dhiya'ulhaq, N. (2022). The benefits of contributing to the citizen science platform iNaturalist as an identifier. *PLoS Biology*, *20*(11), Article 11.

Ciampiconi, L., Elwood, A., Leonardi, M., Mohamed, A., & Rozza, A. (2023). *A survey and*

taxonomy of loss functions in machine learning (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2301.05579

Demirkaya, A., Chen, J., & Oymak, S. (2020). Exploring the Role of Loss Functions in Multiclass Classification. *2020 54th Annual Conference on Information Sciences and Systems (CISS)*, 1–5. https://doi.org/10.1109/CISS48834.2020.1570627167

Deneu, B., Servajean, M., Bonnet, P., Botella, C., Munoz, F., & Joly, A. (2021). Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment. *PLOS Computational Biology*, *17*(4), Article 4. https://doi.org/10.1371/journal.pcbi.1008856

Elith, J., Graham, C., Valavi, R., Abegg, M., Bruce, C., Ford, A., Guisan, A., Hijmans, R. J., Huettmann, F., Lohmann, L., Loiselle, B., Moritz, C., Overton, J., Peterson, A. T., Phillips, S., Richardson, K., Williams, S., Wiser, S. K., Wohlgemuth, T., & Zimmermann, N. E. (2020). Presence-only and Presence-absence Data for Comparing Species Distribution Modeling Methods. *Biodiversity Informatics*, *15*(2), Article 2. https://doi.org/10.17161/bi.v15i2.13384

Elith*, J., H. Graham*, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., J. Hijmans, R., Huettmann, F., R. Leathwick, J., Lehmann, A., Li, J., G. Lohmann, L., A. Loiselle, B., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., McC. M. Overton, J., Townsend Peterson, A., … E. Zimmermann, N. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*(2), Article 2. https://doi.org/10.1111/j.2006.0906-7590.04596.x

Estopinan, J., Bonnet, P., Servajean, M., Munoz, F., & Joly, A. (2024). *Modelling Species Distributions with Deep Learning to Predict Plant Extinction Risk and Assess Climate Change Impacts* (No. arXiv:2401.05470; Issue arXiv:2401.05470). arXiv. http://arxiv.org/abs/2401.05470

Fithian, W., Elith, J., Hastie, T., & Keith, D. A. (2015). Bias correction in species distribution models: Pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, *6*(4), Article 4. https://doi.org/10.1111/2041-210X.12242

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, *19*(1), 1–67.

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

Kellenberger, B., Winner, K., & Jetz, W. (2024). *The Performance and Potential of Deep Learning for Predicting Species Distributions*. https://doi.org/10.1101/2024.08.09.607358

Li, X., & Wang, Y. (2013). Applying various algorithms for species distribution modelling. *Integrative Zoology*, *8*(2), 124–135. https://doi.org/10.1111/1749-4877.12000

McCullagh, P. (2019). *Generalized linear models*. Routledge. https://www.taylorfrancis.com/books/mono/10.1201/9780203753736/generalized-linear-models-mccullagh

Phillips, S. J., Anderson, R. P., Dudík, M., Schapire, R. E., & Blair, M. E. (2017). Opening the black box: An open-source release of Maxent. *Ecography*, *40*(7), 887–893. https://doi.org/10.1111/ecog.03049

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*(3), 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., & Ferrier, S. (2009a). Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. *Ecological Applications*, *19*(1), Article 1. https://doi.org/10.1890/07-2153.1

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., & Ferrier, S. (2009b). Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. *Ecological Applications*, *19*(1), 181–197. https://doi.org/10.1890/07-2153.1

Renner, I. W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S. J., Popovic, G., & Warton, D. I. (2015). Point process models for presence-only analysis. *Methods in Ecology and Evolution*, *6*(4), Article 4. https://doi.org/10.1111/2041-210X.12352

Renner, I. W., & Warton, D. I. (2013). Equivalence of MAXENT and Poisson Point Process Models for Species Distribution Modeling in Ecology. *Biometrics*, *69*(1), 274–281. https://doi.org/10.1111/j.1541-0420.2012.01824.x

Saigusa, Y., Eguchi, S., & Komori, O. (2024). Robust minimum divergence estimation in a spatial Poisson point process. *Ecological Informatics*, *81*, 102569. https://doi.org/10.1016/j.ecoinf.2024.102569

Thuiller, W., Lafourcade, B., Engler, R., & Araújo, M. B. (2009). BIOMOD – a platform for ensemble forecasting of species distributions. *Ecography*, *32*(3), 369–373. https://doi.org/10.1111/j.1600-0587.2008.05742.x

Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J. J., & Elith, J. (2022). Predictive performance of presence-only species distribution models: A benchmark study with reproducible code. *Ecological Monographs*, *92*(1), Article 1. https://doi.org/10.1002/ecm.1486

Warren, D. L., & Seifert, S. N. (2011). Ecological niche modeling in Maxent: The importance of model complexity and the performance of model selection criteria. *Ecological Applications*, *21*(2), 335–342. https://doi.org/10.1890/10-1171.1

Warton, D. I., Renner, I. W., & Ramp, D. (2013). Model-based control of observer bias for the analysis of presence-only data in ecology. *PloS One*, *8*(11), e79168.

Yackulic, C. B., Chandler, R., Zipkin, E. F., Royle, J. A., Nichols, J. D., Campbell Grant, E. H., &

Veran, S. (2013). Presence-only modelling using MAXENT: When can we trust the inferences? *Methods in Ecology and Evolution*, *4*(3), 236–243. https://doi.org/10.1111/2041-210x.12004

Zbinden, R., van Tiel, N., Kellenberger, B., Hughes, L., & Tuia, D. (2024). On the selection and effectiveness of pseudo-absences for species distribution modeling with deep learning. *Ecological Informatics*, *81*, 102623. https://doi.org/10.1016/j.ecoinf.2024.102623

9. Annex

9.1.Links

9.1.1.Articles

DeepMaxent:	https://doi.org/10.48550/arXiv.2412.19217
NCEAS dataset description:	https://doi.org/10.17161/bi.v15i2.13384
GeoPlant:	https://doi.org/10.48550/arXiv.2408.13928

9.1.2.Data repositories

Belgium case ‘B-CUBED data’:	https://doi.org/10.15468/dl.e3j5kv
Belgium case ‘Covariates’	https://chelsa-climate.org/bioclim
Elith’ dataset:	https://osf.io/kwc4v
GeoPlant data:	https://lab.plantnet.org/seafile/d/59325675470447b38add/:

9.1.3.Source code

Belgium case and NCEAS:	https://github.com/RYCKEWAERT/b-cubed_deep-sdm :
Deepmaxent method:	https://github.com/RYCKEWAERT/deepmaxent