Received: 19 December 2024 | Accepted: 22 January 2026

DOI: 10.1111/2041-210x.70262

Methods in Ecology and Evolution

RESEARCH ARTICLE

Applying the maximum entropy principle to neural networks enhances multi-species distribution models

Maxime Ryckewaert[1,2] | Diego Marcos[2] | Christophe Botella[2] | Maximilien Servajean[3] | Pierre Bonnet[4,5] | Alexis Joly[2]

[1]AGAP Institut, CIRAD, INRAE, Institut Agro, Université de Montpellier, Montpellier, France

[2]Inria, Univ Montpellier, Montpellier, France

[3]LIRMM, AMIS, Univ Paul Valéry Montpellier, Univ Montpellier, CNRS, Montpellier, France

[4]AMAP, Univ Montpellier, CIRAD, CNRS, INRAE, IRD, Montpellier, France

[5]CIRAD, UMR AMAP, Montpellier, France

Correspondence
Maxime Ryckewaert
Email: maxime.ryckewaert@cirad.fr

Funding information
European Union under the Horizon Europe Research and Innovation Programme, Grant/Award Number: 101059592; MAMBO (Modern Approaches to the Monitoring of Biodiversity, Grant/Award Number: 101060639

Handling Editor: Huijie Qiao

Abstract

1. The increasing volume of presence-only (PO) data generated by citizen science initiatives has greatly expanded biodiversity databases, but the statistical use of these data in species distribution models (SDMs) remains limited by strong sampling biases and the absence of reliable absence information. Existing approaches based on Poisson point processes, such as Maxent, provide powerful tools, yet rely on predefined features that restrict their flexibility and scalability.
2. We introduce DeepMaxent, a new SDM framework that leverages neural networks to learn a shared, data-driven feature extractor across multiple species while remaining grounded in the maximum entropy principle of Maxent, enabling efficient learning even on very large datasets with thousands of species. DeepMaxent uses a normalized Poisson likelihood, which models the probability of choosing each site given a species, to estimate species-specific suitability surfaces directly from PO observations. In other words, the model predicts suitable locations for each species rather than predicting which species occurs at a given site.
3. We evaluate DeepMaxent on two contrasting datasets: the National Centre for Ecological Analysis and Synthesis (NCEAS) benchmark, containing six small case studies designed to evaluate the impact of spatial sampling biases, and the much larger GeoPlant, dataset covering the whole of Europe. Using PO data for calibration and independent presence–absence data for validation, DeepMaxent consistently outperforms Maxent and leading deep learning-based SDMs. Compared with Maxent, it achieves an area under the ROC curve of 0.768 versus 0.760 on NCEAS, 0.860 versus 0.823 on GeoPlant and enables the use of high-dimensional data modalities, such as satellite images, for which Maxent is unsuitable.
4. DeepMaxent combines the normalized Poisson formulation of Maxent with the learnable features, shared among species, of deep learning approaches. This results in better performance than either Maxent or previous deep learning methods, and lower compute requirements than single-species SDMs, while the

formulation makes the method compatible with the integration of survey data to further improve sampling bias correction.

KEYWORDS
deep learning, DeepMaxent, maximum entropy principle, neural networks, presence-only data, sampling bias, species distribution model, Target-Group Background

1 | INTRODUCTION

In recent years, the rapid growth of citizen science projects has contributed significantly to the expansion of biodiversity databases. Among the different types of data collected, a large amount consists of presence-only (PO) observations (Bonnet et al., 2020; Callaghan et al., 2022). PO records have been instrumental in improving our understanding of species distributions and helping inform conservation strategies (Carvalho et al., 2011; Guisan et al., 2013).

Maxent (Phillips et al., 2006) is one of the most widely used and effective methods for species distribution model (SDM) based on PO data (Elith et al., 2006, 2020; Valavi et al., 2022; Warren & Seifert, 2011). Maxent generates a relative probability of species occurrence across space as a function of environmental variables. This function is applied to various predefined transformations of input environmental variables. Maxent's output can be interpreted as the probability of observing the species in each site relative to the other sites and knowing that it has been observed once. This probability estimate of Maxent is actually equivalent to the one that is derived from a related Poisson regression or a spatially discretized Poisson process (Renner & Warton, 2013). Maxent's name arises from the fact that its formulation leads to finding, among all solutions that fit the PO training data, the one that maximizes the entropy of this spatial probability distribution. This behaviour, in which spatially smooth solutions are favoured for regions with sparse PO data, contributed to Maxent's robust performances against many other SDM methods under such data regimes and in diverse contexts (Elith et al., 2006; Valavi et al., 2022).

Yet, a major issue when calibrating SDMs using PO data is spatial sampling bias, which leads to clustering of PO records in areas with high sampling effort, typically of higher accessibility or greater human activity. Such bias can distort SDM outputs, leading to inaccurate species distribution estimates (Fithian et al., 2015; Phillips et al., 2009; Yackulic et al., 2013).

A wide range of strategies have been proposed to correct for spatial sampling bias in SDMs, including methods based on background points manipulation, spatial filtering of records or explicit bias modelling (Boria et al., 2014; Fithian et al., 2015; Phillips et al., 2009). Specifically for Maxent, Phillips et al. (2009) proposed the Target-Group Background correction (hereafter TGB), which restricts the background sites used by Maxent to those where at least one species was observed among a Target Group (TG) of species. This TG should contain species being sampled along with the focal one. Phillips et al. (2009) evaluated the TGB correction with Maxent on a large standardized dataset. This correction was also robust in later studies (Fourcade et al., 2014). Besides its simplicity and empirical robustness, the TGB correction also comes with theoretical guarantees. Indeed, with the assumptions that all species occurrences are drawn from independent Poisson point processes thinned by a same sampling bias, as it is often assumed (Botella et al., 2021; Fithian et al., 2015), and that the TG species cumulated intensities are constant, TGB yields an unbiased estimate of the species relative intensity across sites (Botella et al., 2020), which then applies within Maxent (Renner & Warton, 2013). More recent studies have proposed other bias corrections in the context of deep learning-based SDMs (deepSDMs), by adapting the loss function to weight presences and background points (Gillespie et al., 2024; Zbinden et al., 2024).

Feature design, that is, the definition of pre-defined transformations (features) of the input variables, is an important step in traditional SDMs, including Maxent (Komori et al., 2024; Phillips & Dudík, 2008). Deep learning is a family of data-driven methods that removes the need for feature design by allowing the model to learn arbitrary non-linear features from the data using neural networks, backpropagation and stochastic gradient descent (Goodfellow et al., 2016; Hornik et al., 1989; LeCun et al., 1989). While other approaches are also capable of learning non-linear relationships, such as Generalized Additive Models, Multivariate Adaptive Regression Splines or Boosted Regression Trees (BRT), and have been used for SDMs (Phillips et al., 2009), deep learning methods offer a broader and more flexible class of models that can automatically learn rich, hierarchical features while integrating efficient regularization strategies to mitigate overfitting (LeCun et al., 2015; Schmidhuber, 2015). Additionally, like other multi-species SDMs do, although typically restricted to the linear case (Ovaskainen, Tikhonov, Norberg, et al., 2017; van der Veen et al., 2023). deepSDMs can also learn shared features to simultaneously predict multiple species distributions. Furthermore, these learnt features tend to be more predictive and robust the more species are included (Botella et al., 2018; Chen et al., 2017), leading to a recent interest in deep learning for multi-SDMs (Kellenberger et al., 2024). In addition, deepSDM architectures can capture predictive features from structured and high-dimensional input data, such as remote sensing imagery (; Deneu et al., 2021; Estopinan et al., 2022, 2024). In spite of this, deepSDM performance have remained limited so far when using low-dimensional input variables (Zbinden et al., 2024). Besides, they remain susceptible to sampling biases (Zbinden et al., 2024), and potentially amplifying them due to their capacity of fitting arbitrary functions.

In this study, we propose DeepMaxent, a method that combines the Maxent principle of maximum entropy with the data-driven feature extraction capabilities of deep learning methods. The DeepMaxent model uses PO data to jointly learn shared latent features and the functions that map from these features to the probability distribution across sites for each species. We propose a loss function, hereafter the DeepMaxent loss, that generalizes Maxent for modelling the probability function with a wider class of functions, including neural networks, preserving the equivalence with the Poisson regression loss (Renner & Warton, 2013). In contrast to loss functions often used for deepSDMs, which attempt at modelling the relative probability of each species given a site, DeepMaxent aims at modelling the probability of selecting each site for an observation given a species, as done in Maxent (see Figure 1). Note that the latter is an easier objective, since modelling the relative probability of each species given a site requires capturing the relative abundances between species. Unlike the original Maxent loss, which is optimized accounting for the whole dataset at every optimization step, we adopt a mini batch-based approach to ensure scalability in terms of data and model size. We show that the global minimizer of such loss is the same as the one using the full dataset to inform each optimization step. Similarly to Maxent, we show how the TGB correction can be implicitly incorporated into DeepMaxent to efficiently mitigate spatial sampling bias. We evaluate DeepMaxent and compare it to alternative methods on two reference benchmarks (Elith et al., 2020; Picek et al., 2025), both encompassing PO data for SDM training and presence–absence (PA) data for evaluation. We carried an extensive

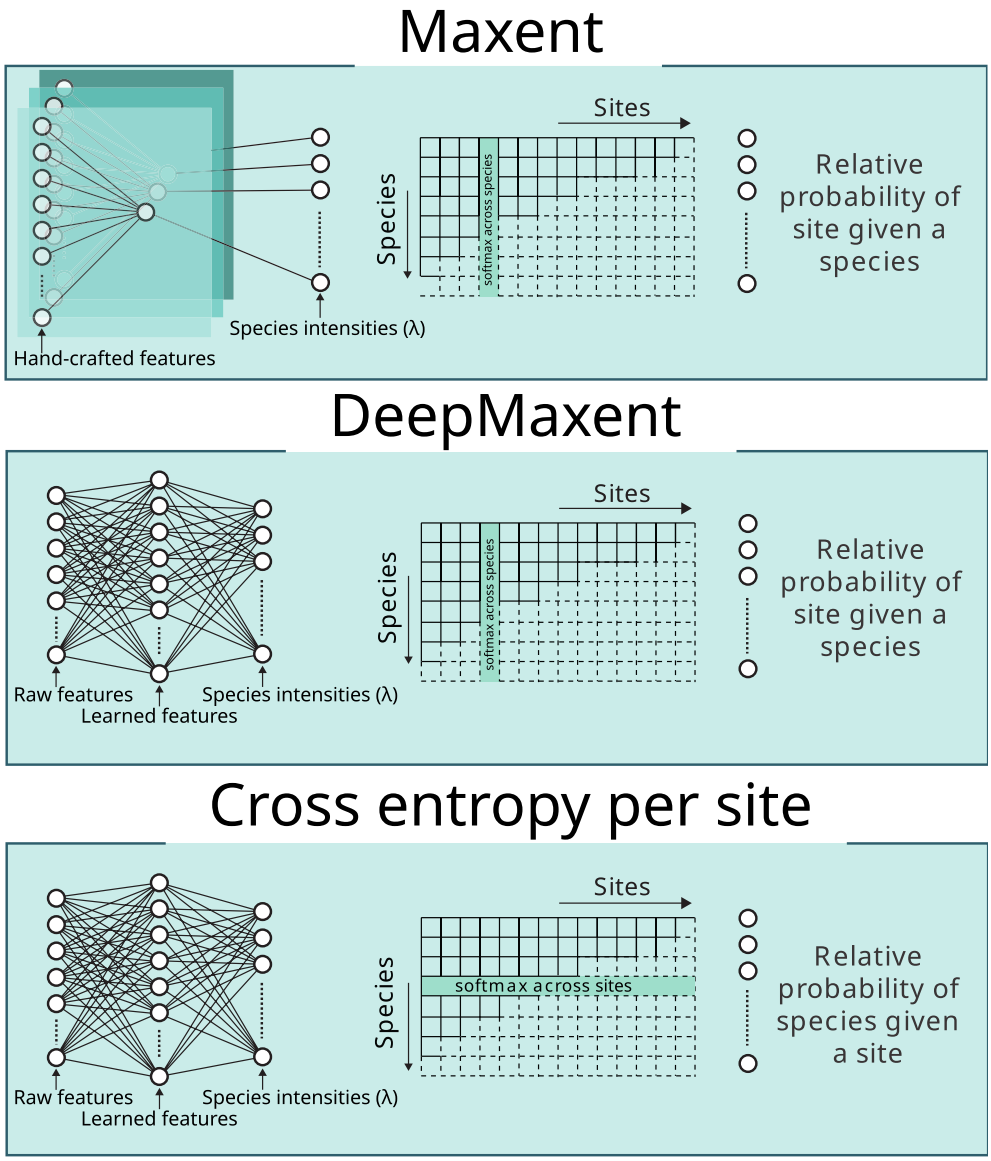

FIGURE 1 Illustration of three species distribution modelling approaches (Maxent, DeepMaxent, and a common cross-entropy [CE]). Maxent relies on handcrafted environmental features and trains an independent model for each species. DeepMaxent uses a single deep learning model to predict multiple species simultaneously, incorporating batch normalization across samples to standardize feature representations. In contrast, the commonly used CE loss approach applies normalization along the species dimension, focusing on predicting the presence of species at each site rather than modelling species jointly.

comparative evaluation on the National Centre for Ecological Analysis and Synthesis (NCEAS) dataset (Elith et al., 2020), comprising six distinct regions and different biological groups. We compare DeepMaxent to alternative loss functions (Poisson, Cross-Entropy [CE] across species, Binary Cross Entropy [BCE]) with or without the TGB correction, and to various state-of-the-art SDMs, notably Maxent and other multi-species deepSDM. We also conduct sensitivity and ablation studies on this dataset to assess the importance of DeepMaxent's hyper-parameters and components. We conduct additional experiments on the GeoPlant dataset (Picek et al., 2025), including an independent comparison of DeepMaxent and Maxent, and an illustration of how DeepMaxent can leverage remote sensing input data with an assessment of the related performance gains compared to tabular climatic data.

2 | MATERIALS AND METHODS

2.1 | DeepMaxent: Maximum entropy principle for SDMs based on neural networks

2.1.1 | A generalization of Maxent's loss function

We introduce here the statistical model underlying the investigated methods. We consider a geographic domain $\mathcal{D} \subset \mathbb{R}^2$ composed of K non-overlapping spatial areas $d_1, \dots, d_K \subset \mathcal{D}$ (e.g. a regular mesh), hereafter called sites. We consider $N \in \mathbb{N}^*$ species indexed by j, and note $y_{ij} \in \mathbb{N}$ the count of PO observations for species j in site $i \in [\![1, K]\!]$. For each site $i \in [\![1, K]\!]$, we have covariates $\mathbf{x}_i \in \mathbb{R}^P$ typically encoding environmental factors (e.g. climate or soil properties), as well as positive intensity values $\lambda_{ij} > 0$ for and species j. We further assume λ_{ij} to be a parametric function of the covariates of the form $\lambda_{ij} = \lambda_j(\mathbf{x}_i)$, where we explicitly define $\lambda_j : \mathbb{R}^P \to \mathbb{R}_+$ as a positive intensity function mapping site-level covariates to expected counts. This function is then parametrized as $\lambda_j(\mathbf{x}_i) = \exp\left(b_j + f_{\theta_j}(\mathbf{x}_i)\right)$ with real parameters b_j, θ_j. The function f_{θ_j} can be an arbitrary mapping $f_{\theta_j} : \mathbb{R}^P \to \mathbb{R}$. In a Poisson regression setting, the counts y_{ij} are modelled as Poisson random variables $y_{ij} \sim \mathcal{P}(\lambda_{ij})$ where these counts are independent between sites and species. This Poisson regression model notably corresponds to the discrete approximation of a spatial Inhomogeneous Poisson Process (IPP), and is very often used for SDMs (Renner et al., 2015) with a known equivalence to Maxent when f_{θ_j} is linear in x_i (Renner & Warton, 2013). Such discretization is common and handful to avoid the computational burden of fitting an IPP on a continuous spatial domain where covariates vary at high resolution, and it makes particular sense when covariates come from a stack of geographic rasters.

We note for convenience the intensity values (i.e. model predicted counts) across all sites and species by the matrix $\mathbf{\Lambda} = \{\lambda_{ij}\}_{i=1,j=1}^{K,N} \in \mathbb{R}_+^{K \times N}$, and similarly $\mathbf{Y} = \{y_{ij}\}_{i=1,j=1}^{K,N} \in \mathbb{N}_+^{K \times N}$ for the PO count data. The negative log-likelihood of the above described Poisson regression, hereafter the Poisson loss, is written in Equation (1).

$$\mathcal{L}_{\mathcal{P}}(\mathbf{\Lambda}, \mathbf{Y}) = \frac{1}{KN} \sum_{i=1}^{K} \sum_{j=1}^{N} \left(\lambda_{ij} - y_{ij} \log \lambda_{ij}\right), \tag{1}$$

The parameters' estimates for the Poisson loss (i.e. the maximum likelihood estimates of the Poisson regression) can then be noted $\left(\hat{b}_1^{\mathcal{P}}, \dots, \hat{b}_N^{\mathcal{P}}, \hat{\theta}_1^{\mathcal{P}}, \dots, \hat{\theta}_N^{\mathcal{P}}\right) = \underset{b_1,\dots,b_N,\theta_1,\dots,\theta_N}{\operatorname{argmin}} \mathcal{L}_{\mathcal{P}}(\mathbf{\Lambda}, \mathbf{Y})$. It is noteworthy that, due to the equivalence from Renner and Warton (2013), the sites included in the terms of Equation (1) can be interpreted as the background sites (or points) of Maxent (Phillips et al., 2009), as developed in the dedicated section below.

We introduce in Equation (2) a new loss $\mathcal{L}_{\mathcal{H}}$ derived from the Poisson loss of Equation (1), named DeepMaxent loss. We can see from this equation that the DeepMaxent loss is a modification of the Poisson loss where the counts y_{ij} and intensity values λ_{ij} are normalized by their sum over sites, and each species term is weighted by the species total PO count. The loss measures the discrepancy between the observed and predicted probability distributions of PO across sites for each species. Indeed, Equation (2) corresponds to a weighted sum of species-wise CE losses (or Kullback–Leibler divergences) between the empirical and predicted distribution of PO records over sites.

$$\begin{aligned}\mathcal{L}_{\mathcal{H}}(\mathbf{\Lambda}, \mathbf{Y}) &= -\frac{1}{KN} \sum_{j=1}^{N} \sum_{i=1}^{K} y_{ij} \log\left(\frac{\lambda_{ij}}{\sum_{k=1}^{K} \lambda_{kj}}\right) \\ &= -\frac{1}{KN} \sum_{j=1}^{N} \left(\sum_{k=1}^{K} y_{kj}\right) \sum_{i=1}^{K} \left(\frac{y_{ij}}{\sum_{k=1}^{K} y_{kj}}\right) \log\left(\frac{\lambda_{ij}}{\sum_{k=1}^{K} \lambda_{kj}}\right)\end{aligned} \tag{2}$$

We show in Appendix A.1 that the DeepMaxent loss generalizes the loss of Maxent to intensity models $\lambda_{ij} = \exp\left(b_j + f_{\theta_j}(x_i)\right)$ which are not necessarily log-linear, for example, when f is built on neural networks, and obviously to multiple species. That is, if λ_{ij} is log-linear, the DeepMaxent loss is equivalent to the one of Maxent. Furthermore, Appendix A.1 shows that we preserve the equivalence with the Poisson estimate from Renner and Warton (2013) for non log-linear intensities: The global minimizer of the DeepMaxent loss for the parameters of the probability distribution across sites $\left(\hat{\theta}_1^{\mathcal{H}}, \dots, \hat{\theta}_N^{\mathcal{H}}\right)$ is equal to the one of the Poisson loss $\left(\hat{\theta}_1^{\mathcal{P}}, \dots, \hat{\theta}_N^{\mathcal{P}}\right)$ of Equation (1), and the difference with the Poisson loss is that the latter provides an estimate for b_j. Although b_j is not identifiable under the DeepMaxent loss, as it cancels out in Equation (2), we consider it for consistency across all losses introduced and tested below, including the Poisson loss.

2.1.2 | Feature extraction using neural networks

In Maxent, the intensity is defined as a log-linear function of a feature vector $f(\mathbf{x})$, composed of pre-determined transformations of $\mathbf{x}$. We extend the principle of Maxent by replacing $f(\mathbf{x})$ with a feature extractor instantiated as a neural network $g_\theta : \mathbb{R}^P \mapsto \mathbb{R}^C$ parametrized by θ, where C is the dimensionality of the last hidden representation. The output $g_\theta(\mathbf{x}) \in \mathbb{R}^C$ is a shared latent representation across all species, as illustrated in Figure B1. The intensity function $\lambda_j(\mathbf{x})$ of species j in DeepMaxent is then given by:

$$\lambda_j(\mathbf{x}) = \exp\left(\sum_{c=1}^{C} \gamma_{jc} g_\theta(\mathbf{x})_c + b_j\right), \tag{3}$$

where $\gamma_j \in \mathbb{R}^C$ is the species-specific weight vector and $b_j \in \mathbb{R}$ is the species-specific intercept. Collectively, the weight vectors for all species form the matrix $\Gamma \in \mathbb{R}^{N \times C}$, and the intercepts form the vector $\mathbf{b} \in \mathbb{R}^N$. The function g_θ can automatically learn complex, non-linear relationships between environmental variables and the presence of multiple species from the data, potentially enabling the model to identify environmental patterns that cannot be captured by a linear mapping. With this multi-species architecture, similarly to many earlier deepSDM implementations, we have one shared g_θ across all species, which offers a key computational advantage compared to training one DeepMaxent model per species, given the complexity of g_θ. Indeed, at each gradient descent step, g_θ and its gradient are computed only once per site, independent of the number of species. Besides, sharing this representation of the environment across species yielded more robust performances in deepSDMs (Botella et al., 2018), especially for data-poor species.

2.1.3 | Batched algorithm and partition function approximation in DeepMaxent

One of the challenges of adapting Maxent to a deep learning framework is the computation of the partition functions $\sum_k \lambda_{kj}$ for species j, corresponding to the denominator in Equation (2), which normalizes the predicted intensity $\lambda_j(\mathbf{x})$ over the K sites. This may become problematic when the number of observed sites increases, which often happens when considering a large geographic domain or a finer spatial resolution. In this case, the number of terms K in the partition function becomes large, making it challenging to compute it exactly. To address this challenge efficiently, we leverage standard stochastic optimization techniques widely used in deep learning (e.g. mini-batch Stochastic Gradient Descent; Bottou, 1991), with the specificity that we approximate the normalization within mini-batches. Specifically, we compute the loss function on a small random subset of sites $B \subset \{1, \ldots, K\}$, called mini-batch, hence normalizing the intensities within the mini-batch. The mini-batch-wise loss is then written as follows:

$$\mathscr{L}\left(\widetilde{\mathbf{\Lambda}}_{i\in B}, \widetilde{\mathbf{Y}}_{i\in B}\right) = -\frac{1}{|B|N}\sum_{i\in B}\sum_{j=1}^{N}\frac{y_{ij}}{\sum_{i\in B} y_{ij}+\epsilon}\log\left(\frac{\lambda_j(\mathbf{x}_i)}{\sum_{i\in B}\lambda_j(\mathbf{x}_i)}\right). \tag{4}$$

Note that the denominator $\sum_{i\in B}\lambda_j(\mathbf{x}_i)$ is strictly positive under typical model assumptions where the Poisson intensity functions $\lambda_j(\mathbf{x}_i)$ are positive. However, the denominator $\sum_{i\in B} y_{ij}$, representing the count sum for class j in the mini-batch, may be zero if no samples of class j are present in the mini-batch. In practice, this requires either ensuring that each mini-batch contains at least one example of every class to avoid division by zero, or the addition of a small, but positive, ϵ. This mini-batch-wise loss makes the model optimization computationally feasible by computing it one mini-batch at a time and avoiding computing the full partition functions for every iteration, and thus allows it to be scalable to large domains and trained efficiently with any optimiser based on random mini-batches, such as the commonly used Adam (Kingma, 2014). However, it must be noted that the batch-wise loss is not a simple approximation of the full loss as, for instance, the normalized occurrences tend to have a larger value for smaller mini-batches. Nevertheless, we provide the mathematical guarantee that, for any mini-batch size n ($1 < n < K$), minimizing the model loss on all mini-batches also minimizes the full loss (see Appendix A.2). This suggests that our final estimator should be close to the global minimizer of the full loss, even though there is not guarantee to obtain the latter due to the non-convexity induced by the non-linear feature extractor. Similarly to supervised contrastive methods (Khosla et al., 2020), the intensity predictions could become more specific, and more concentrated around the occurrences, as the mini-batch size increases. A small mini-batch size could, on the other hand, result in smoother species intensities. Therefore, we tested the impact of the mini-batch size during training on the final predictions of DeepMaxent. Note that, given the expression of $\lambda_j(\mathbf{x}_i)$ in Equation (3), the normalized intensities across the mini-batch B used in Equation (4) are given by $\exp\left(\gamma_j^\top g_\theta(\mathbf{x}_i)\right) / \sum_{k\in B}\exp\left(\gamma_j^\top g_\theta(\mathbf{x}_k)\right)$. This is a particular case of the normalization function commonly referred to as softmax, applied to the logits $\gamma_j^\top g_\theta(\mathbf{x}_k)$ over the mini-batch B.

2.1.4 | Spatial sampling bias correction with TGB correction

When occurrence concentration is biased by spatial variations in sampling effort, a popular SDM correction approach is the TGB method (Phillips et al., 2009), which was initially proposed to correct sampling bias in Maxent. The method basically approximates the spatial sampling effort through the distribution of occurrences of a TG of species, providing background points to Maxent for each site where TG species were reported. In other words, TGB restricts the study domain to the sites with at least an evidence of sampling effort (one observation), which reduces the problem of false absences. The strategy is expected to work when TG species are reported jointly with the focal species (e.g. the TG is a biological group targeted by the same citizen science programme).

In DeepMaxent, which models multiple species simultaneously, the TGB strategy emerges implicitly by considering all samples within a mini-batch in the computation of the partition function. This makes the implementation of TGB particularly efficient, as the calculation of the intensity prediction for each species is recycled to be used for the TGB-enhanced partition function.

2.1.5 | L2-regularization implementation in DeepMaxent

Maxent makes use of L1 penalization, on the species weights γ_j that model the relation between the features and the density prediction. The L1 term, known as LASSO penalty, encourages γ_j to become

sparse, thus selecting a subset of features. The L1 regularization is important in Maxent due to a number of features that grows more than quadratically with the number of environmental variables. In DeepMaxent, the latent features are learnt to maximize prediction performances and can be kept to a fixed dimensionality, removing the need for feature selection (Goodfellow et al., 2016). For DeepMaxent, we employ L2 regularization, that is, a penalty on the Euclidean norm of the γ_j and the rest of the model weights, which encourages small but non-zero weights, which tends to induce a smoothing of the estimated species intensities. The intercept term **b**, on the other hand, is not penalized, since it controls only the baseline level of the intensity function per species and does not affect the effective capacity of the model. The total loss function thus becomes:

$$\mathscr{L}_{\text{total}}\left(\tilde{\lambda}, y; \theta, \gamma\right) = \mathscr{L}_{\mathscr{H}}\left(\tilde{\lambda}, y; \theta, \gamma\right) + \frac{\tau}{2}\left(\| \theta \|_2^2 + \| \gamma \|_2^2\right) \quad (5)$$

where τ is the weight decay coefficient. This term penalizes large weight values, encouraging the model to learn smaller weights without enforcing sparsity.

2.2 | Evaluation of model performance

2.2.1 | Datasets

For our experiments, we used two openly available datasets: (i) one from the NCEAS (Elith et al., 2020) and (ii) GeoPlant (Picek et al., 2025).

The NCEAS dataset

This dataset includes 52,605 PO records (for SDM training) and PA (for SDM evaluation) data from six global regions: Australian Wet Tropics (AWT), Canada (CAN), New South Wales (NSW), New Zealand (NZ), South America (SA) and Switzerland (SWI) (Elith et al., 2020). Each regions is associated with a specific set of species, and sometimes from several biological groups (see Table 1), with a total of 226 anonymous species. The dataset provides specific environmental variables with specific spatial resolution for each region, including climatic, soil or location variables (see more details in Elith et al., 2020). The total area of all regions is of 13,607,500 km^2 (Table 1), so that there is 0.004 PO records per km^2 on average when pooling all species and regions. However, the spatial concentration of records is extremely heterogeneous across regions and taxonomic groups, varying from 0.0002 records/km^2 in SA to 0.887 records/km^2 in SWI, as computed from Table 1.

Various SDM methods have been evaluated using this dataset (Elith et al., 2006; Phillips et al., 2009; Valavi et al., 2022; Zbinden et al., 2024), which allows for a direct comparison of DeepMaxent's performance to many state-of-the-art SDM methods. Phillips et al. (2009) studied spatial sampling biases and found that the PO data in certain regions (AWT, CAN and SWI) contained high levels of such biases, making this a good benchmark to assess model robustness.

The GeoPlant dataset

Described by Picek et al. (2025), we use this dataset to evaluate DeepMaxent under a different data regime, with 100 times more PO records, much more concentrated in space, and 40 times more species covered, than NCEAS. Indeed, GeoPlant contains 5,079,797 PO observations of 9709 plant species (i.e. almost half of Europe's flora) from 13 selected datasets of the Global Biodiversity Information Facility (www.gbif.org), to be used for model training. These records cover 38 European countries spanning a total area of about 5,914,500 km^2, with 0.859 records per km^2 on average. This spatial concentration in Geoplant is thus 200 times higher than the average of NCEAS, and comparable to the one of the SWI region, but for an area 150 times larger. Additionally, 88,987 PA records from the European Vegetation Archive were used for model evaluation.

Within GeoPlant, two types of input data were considered in this work for training SDMs. The first configuration (Bioclim-GeoPlant) used bioclimatic variables aggregated to a 10 km resolution, enabling direct comparison with MaxEnt. The second configuration

TABLE 1 The total number of species, the occurrence number in presence-only (PO) data and the total number of species presence in presence–absence (PA) data for each region and biological group.

				Occurrences number		
Code	Location	Biological group	Species number	PO	PA	Area ('000' km^2)
AWT	Australian wet tropics	Bird	20	3105	340	24
AWT	Australian wet tropics	Plant	20	701	102	24
CAN	Ontario, Canada	Bird	20	5063	14,571	979.3
NSW	New South Wales	Bates	10	187	570	76.2
NSW	New South Wales	Birds	7	1781	1839	76.2
NSW	New South Wales	Plants	29	680	5329	76.2
NSW	New South Wales	Reptile	8	675	1008	76.2
NZ	New Zealand	Plant	52	3088	19,120	265.4
SA	South America	Plant	30	2220	152	12,223.2
SWI	Switzerland	Tree	30	35,105	10,013	39.6

(LandSat-GeoPlant) focused on applying deep learning-based approaches to multi-band time series of satellite data. Only methods that include a neural network feature extractor were considered here. We used Landsat-based covariates at 30 m resolution, derived from the seasonally aggregated and gap-filled GLAD analysis-ready dataset (Potapov et al., 2020) and accessed via the EcoDataCube platform (Witjes et al., 2022). These data covered six spectral bands: Red (R), Green (G), magenta (B), Near Infrared (NIR), Shortwave Infrared 1 (SWIR1) and Shortwave Infrared 2 (SWIR2).

2.2.2 | Evaluation metrics

To directly compare our results to Phillips et al. (2009), Zbinden et al. (2024) and Valavi et al. (2022), we evaluated our method performances with the area under the ROC curve (AUC) computed for each species across the PA plots, none of which were used in model training. The AUC is the empirical probability that a presence site has a higher model-predicted value than an absence site. In other words, it measures the model's ability to distinguish between presence and absence classes based on its predicted scores. The NCEAS dataset being decomposed into regions and biological groups, we first averaged each species-wise AUC per biological group and then per region (across groups), and then took the average over regions as our general performance metric.

2.2.3 | Implementation details

NCEAS dataset

For the NCEAS dataset, the feature extractor g_θ was implemented as a multilayer perceptron (MLP) with rectifier linear unit non-linearities and skip connection between hidden layers (see Figure B1) in order to mitigate potential vanishing gradient issues when exploring deeper architectures (He et al., 2016). The use of an MLP as a feature extractor was justified both by the structure of the input data (vectors of environmental variables for each site) and as a standard reference in deep learning-based multi-species SDMs (Hu et al., 2025; Kellenberger et al., 2024; Zbinden et al., 2024).

In this study, the parameters related to the neural network architecture and the optimizer, referred to as hyper-parameters, were selected through a cross-validation procedure. The search included the number of hidden layers, learning rate, mini-batch size and weight decay, and was performed using spatially blocked folds based on geographical data (Roberts et al., 2017; Valavi et al., 2019). As suggested in Zbinden et al. (2024), the cross-validation was performed using PO data (see details in Appendix C). All models were trained for 100 epochs. Once cross-validation has been performed, the hyper-parameter values were chosen to be the same for all regions and biological-groups. These hyper-parameters were: two hidden layers, Adam as optimizer, a learning rate of 0.0002 and a both the mini-batch size and hidden layer size of 250. Each loss function was evaluated both with and without TGB correction. Among the hyper-parameters tested, these values consistently yielded the best results for all loss functions (see Annex C). Regarding weight decay τ, we did not observe any performance improvement across the tested loss functions, except in the case of DeepMaxent, where it led to better results. The final model was then calibrated with these hyper-parameter values on the whole PO data and applied to the PA data. To account for variability arising from model initialization, each model was trained and evaluated across 10 different random seeds, following the same procedure as used in Zbinden et al. (2024).

For the NCEAS dataset, although DeepMaxent implicitly implements TGB correction, we additionally added random background points in locations with no species observations to verify that adding such points would not contribute to improve model performance. Background points were sampled uniformly from the raster data, producing a dataset 10 times larger than the PO occurrences. This procedure ensured a consistent representation of environmental conditions across the study area and maintained a percentage-based approach to guarantee comparability across datasets (e.g. SWI or CAN).

GeoPlant datasets

For the Bioclim-GeoPlant dataset, we used the same MLP architecture as for NCEAS. This choice was again driven by the structure of the data, consisting of 19 bioclimatic variables. For the LandSat-GeoPlant dataset, each occurrence in our dataset was represented as a multidimensional data cube of 6 spectral bands × 4 seasons × 21 years, which serves as input to the model (see Figure I20). As a feature extractor, we used the adapted ResNet-18 model proposed in Picek et al. (2025). The output features from the ResNet-18 were processed through two fully connected layers, resulting in the final predictions. The number of epochs was fixed at 20 throughout the training process and each hidden layer was composed of 250 neurons.

For both GeoPlant datasets, cross-validation was not performed due to computational constraints. Instead, a validation set was created by randomly splitting the PO data to select the best model. This approach ensured that model performance was assessed without excessive computational overhead. The hyper-parameters used for training were adapted from Picek et al. (2025), maintaining consistency with previously established configurations.

2.2.4 | Baseline losses

We implemented the Poisson regression loss (see Equation 1), using the same neural network architecture as in DeepMaxent to model λ_{ij}, to test the effect of the density normalization in DeepMaxent on the estimator quality. Other commonly used loss functions in deep learning, and notably for SDMs, namely CE over species (Brun et al., 2024; Deneu et al., 2021) and BCE (Benkendorf & Hawkins, 2020; Zbinden et al., 2024) were implemented. The CE loss over species, $\mathscr{L}_{\mathrm{CE}}(\mathbf{\Lambda}, \mathbf{Y})$ (Equation 6), measures for each site the deviation between a predicted probability distribution across species and the associated

empirical distribution based on the species observations in that site. In this case, the predicted probabilities are obtained by normalizing the intensity values over the species $\mathbf{\Lambda} := \{\lambda_{ij}\}_{i=1,j=1}^{K,N}$, at each site separately, implemented using the softmax function:

$$\mathscr{L}_{\mathrm{CE}}(\mathbf{\Lambda}, \mathbf{Y}) = -\frac{1}{K}\sum_{i=1}^{K}\sum_{j=1}^{N}\frac{y_{ij}}{\sum_{k=1}^{N} y_{ik}}\log\left(\frac{\lambda_{ij}}{\sum_{k=1}^{N}\lambda_{ik}}\right) \quad (6)$$

The BCE loss, $\mathscr{L}_{\mathrm{BCE}}(\mathbf{\Lambda}, \mathbf{Y}_{\mathrm{b}})$, was implemented for the case where $\mathbf{Y}_{\mathrm{b}} \in \{0, 1\}$ is treated binary variable of $\mathbf{Y}$, taking the value 1 if the species was observed at least once in the pixel, and 0 otherwise. Unlike the CE case, where the probabilities across species are required to sum to 1 in each site, the predicted probability of a species learned using BCE does not directly restrict the probabilities of other species. In this setting, the softmax function reduces to a sigmoid function. The BCE loss is defined as:

$$\mathscr{L}_{\mathrm{BCE}}(\mathbf{\Lambda}, \mathbf{Y}_{\mathrm{b}}) = -\frac{1}{KN}\sum_{i=1}^{K}\sum_{j=1}^{N}\left(y_{\mathrm{b},ij}\log\sigma(\log(\lambda_{ij})) + (1-y_{\mathrm{b},ij})\log(1-\sigma(\log(\lambda_{ij})))\right), \quad (7)$$

where σ is the logistic sigmoid function here applied to the linear predictor, or 'logit', of each species $\log(\lambda_{ij}) = \gamma_j^{\top} g_{\theta}(\mathbf{x}_i) + b_j$.

3 | RESULTS

3.1 | Comparative analysis of SDM methods

Table 2 shows the performances of various traditional SDM methods, including Maxent, BRT with or without TGB correction, and the recent neural network model for multi-species proposed by Zbinden et al. (2024), all evaluated with the average AUC per region, along with overall average (Phillips et al., 2006; Valavi et al., 2022; Zbinden et al., 2024). It also contains the performance of our main DeepMaxent implementation and the baseline deep learning losses, with and without the TGB correction.

Without TGB sampling bias correction, performances are overall lower and close among methods, ranging from 0.716 to 0.723 in overall average AUC (Table 2), except for the CE loss which achieves 0.731. Except for the latter, we observe no general performance gain for the tested deep learning losses (BCE, Poisson, DeepMaxent, ranging from 0.719 to 0.720) compared to the literature methods, for example, Maxent (0.721) or the best SDM Ensemble of Valavi et al. (2022) (0.723).

The TGB bias correction brings a consistent performance improvement for all methods, including the ones in the literature and our implementations. However, not all approaches respond equally strongly to TGB. For instance, Maxent gains 0.039 in overall averaged AUC by using TGB, and it is the same for BRT. Regarding our implemented baseline losses, TGB induces an AUC gain of 0.011 for the CE loss, 0.045 for the BCE loss and 0.040 for the Poisson loss. Finally, DeepMaxent achieves an improvement of 0.048 with TGB, resulting in the highest overall AUC (0.768). These results show that the proposed DeepMaxent is well adapted to this bias correction technique while it enables leveraging the predictive potential of multi-species neural networks for spatial density estimation. The largest region average AUC gains were mostly seen in regions CAN and AWT, where spatial sampling bias is the strongest according to Phillips et al. (2009). Note that the best method of Zbinden et al. (2024), achieving an overall averaged AUC of 0.755, incorporated both random and TGB points as absences in their BCE loss and their results specifically showed the key role of the TGB points in this performance. DeepMaxent with TGB also had the best AUC in four of the six regions (CAN, NSW, NZ, SWI), showing that it is robust across regions and biological groups (NSW includes four biological groups, see Table 1). BCE with TGB is the second-best method in overall AUC (0.764). In contrast, using TGB, CE and Poisson yield poorer overall AUC (0.745 and 0.759) than Maxent (0.760) or BRT (0.759).

Indeed, although we would expect similar results for DeepMaxent and Poisson when using the same TGB strategy, given the equivalence of their global minimizer (see Appendix A.1), DeepMaxent resulted in a 0.023 gain in AUC, closing the gap towards a perfect score by close to 10% (i.e. 10% less mis-ordered pairs of presence and absence sites) compared to Poisson, a significant gain according to statistical tests (Appendix E).

Figure 2 shows AUC scores by observation abundance class (rare, common and abundant) for each loss function, averaged over all regions. Averaging takes into account region-specific abundance class distributions (see Appendix I17).

DeepMaxent consistently outperforms the other losses across all abundance classes, while BCE ranks second in every class. Poisson loss performs moderately well for abundant and common species. For rare species, the Poisson loss is dominated by the many zero-observation terms, which directly penalize the predicted intensities via the λ_{ij} component. This tends to enforce uniformly low intensity values, potentially leading to underfitting beyond what data scarcity alone would justify. In contrast, CE shows the opposite trend: It achieves its best results for rare species, but lower AUC scores for abundant and common species, suggesting that its sensitivity to class imbalance limits its overall effectiveness in these abundance regimes. These results suggest that DeepMaxent is particularly well suited for modelling species with heterogeneous distributions and low occurrence counts, where classical Poisson-based methods may underperform due to their tendency to over-penalize predictions in data-sparse regions.

Estimated probabilities for can01, a rare species, differ widely depending on the loss function applied (see Figure 3). For this species, PO data are rare, while PA data are more abundant. In contrast, the maps generated for can02, a common species, give much more consistent estimates between the different methods (see Appendix G8).

Table 3 reports the general averaged AUC values for different loss functions evaluated on two datasets, Bioclim-GeoPlant and Landsat-GeoPlant, using TGB sampling bias correction.

On the Bioclim-GeoPlant dataset, the DeepMaxent model achieves the highest average AUC (0.860), outperforming all losses, including BCE (0.839), Poisson (0.837) and CE (0.830), as well as the Maxent model (0.823). These results highlight the improved predictive

TABLE 2 Comparison of method performance by region-averaged area under the ROC curve (AUC) and general averaged AUC over all regions.

	Regions						
	AWT	CAN	NSW	NZ	SA	SWI	avg
Results from the literature							
Single-species models							
Maxent [1]	0.686	0.587	0.700	0.738	0.804	0.809	0.721
BRT [1]	0.681	0.577	0.701	0.735	0.795	0.816	0.718
RF down-sampled [1]	0.675	0.572	0.715	0.746	*0.813*	0.818	0.723
Ensemble [1]	0.683	0.580	0.710	*0.749*	0.806	0.812	0.723
IWLR-GAM [1]	0.674	0.595	0.689	0.747	0.796	0.798	0.716
Maxent (using TGB) [2]	**0.732**	0.716	0.741	0.738	0.798	0.837	0.760
BRT (using TGB) [2]	0.700	0.728	0.738	0.740	0.792	0.842	0.757
Multi-species models							
Zbinden et al. [3]	0.704	0.714	0.719	0.741	**0.815**	0.838	0.755
Results from our implementations							
Baseline losses							
CE	0.701	0.661	0.732	0.724	0.772	0.793	0.731 ± 0.001
CE (using TGB)	*0.727*	0.708	0.739	0.732	0.771	0.792	0.745 ± 0.001
BCE	0.656	0.600	0.718	0.736	0.804	0.799	0.719 ± 0.002
BCE (using TGB)	0.722	*0.730*	*0.743*	0.738	0.804	*0.849*	*0.764* ± 0.002
Poisson loss	0.658	0.599	0.714	0.737	0.804	0.799	0.719 ± 0.002
Poisson loss (using TGB)	0.712	*0.730*	0.732	0.729	0.801	*0.849*	0.759 ± 0.002
Proposed loss							
DeepMaxent	0.654	0.593	0.718	0.744	0.803	0.810	0.720 ± 0.001
DeepMaxent (using TGB)	0.712	**0.732**	**0.752**	**0.753**	0.806	**0.850**	**0.768** ± 0.001

Note: The best average AUC for each column is highlighted in bold, while the second-best averaged AUC is italicized. The references correspond to results from the following articles: [1] Valavi et al. (2022), [2] Phillips et al. (2009) and [3] Zbinden et al. (2024).

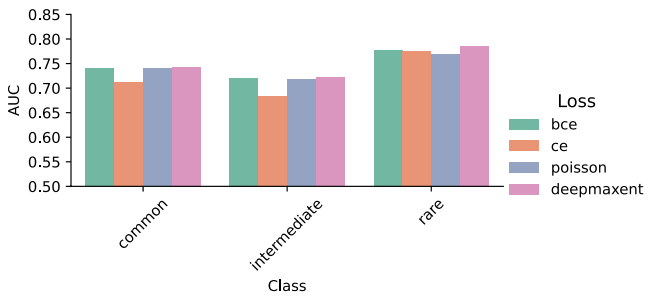

FIGURE 2 Comparison of average area under the ROC curve (AUC) values across all regions by loss and abundance classes on National Centre for Ecological Analysis and Synthesis (NCEAS) dataset.

capacity of DeepMaxent when using environmental variables from Bioclim. A similar result is observed on the Landsat-GeoPlant dataset, where DeepMaxent also performs best with a general averaged AUC of 0.887, closely followed by BCE (0.885), while CE and Poisson yield lower performances (0.829 and 0.862, respectively). DeepMaxent performs well on satellite time series data. BCE Loss also performs well in this context, especially for a large dataset. Notably, the standard deviations across random initializations are very small for all models (≤0.002), indicating high consistency across runs.

3.2 | Sensitivity study

Table 4 shows the average AUCs calculated for all regions, according to six different values for each hyper-parameter: mini-batch size, number of hidden layers and weight decay. A detailed analysis of AUC values for each region is provided in Appendix D. The general performance of DeepMaxent-TGB was quite robust to hyper-parameter choices, with the largest difference in average AUC across all tested values being only 0.010, illustrating the model's stability with respect to mini-batch size, number of hidden layers and weight decay. In particular, DeepMaxent with TGB kept a general average AUC above 0.764, that is, above the best results using all other methods, for all tested mini-batch sizes, ranging from 10 to 2500. Qualitatively, a smaller mini-batch size induces smoother species intensity maps, while larger mini-batch size tends to concentrate the intensity in higher abundance areas, as illustrated for one species

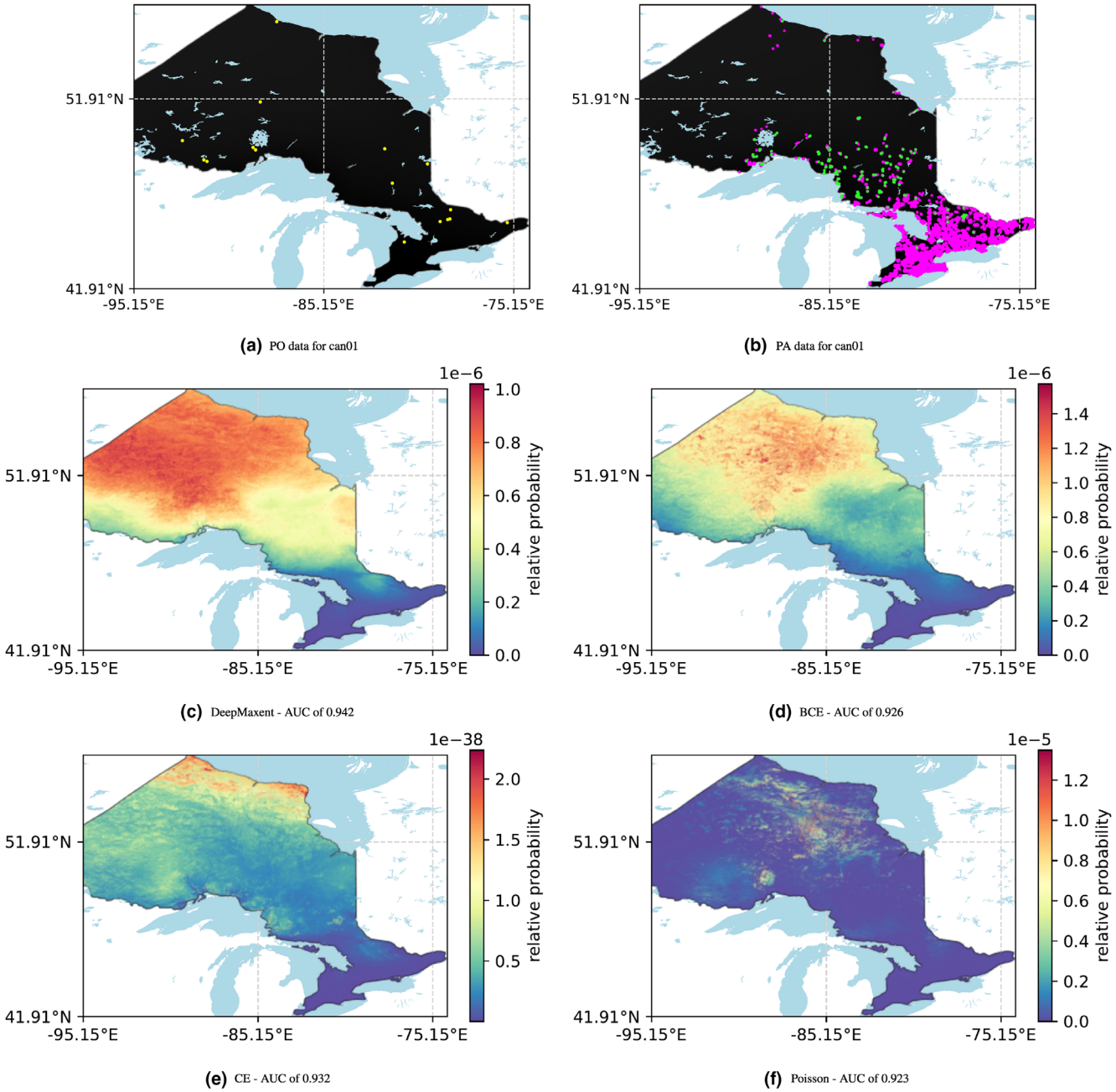

FIGURE 3 Estimated relative probabilities for the species can01 (a rare species with 16 PO points): (a) presence-only (PO) data, yellow points; (b) presence–absence (PA) data where green corresponds to presences, and magenta to absences. (c–f) Estimated from different loss functions: (c) DeepMaxent; (d) binary cross-entropy (BCE); (e) cross-entropy (CE); (f) Poisson loss.

in the region CAN in Figure 4. The L2 regularization (weight decay) has an important impact on DeepMaxent performance. It has a small but consistently positive impact on the performance up to a value of 3×10^{-4} (see Table 4), while further increasing the weight decay value results on a performance degradation due to oversmoothing (see Figure 4). Varying the number of hidden layers in the neural network architecture of DeepMaxent from one to two had almost no effect, with a same general averaged AUC of 0.767 (Table 4), and the AUC softly and progressively decreased for three (0.766), four (0.764), five (0.762) and six layers (0.759).

4 | DISCUSSION

In this study, we propose DeepMaxent, a new SDM method for PO data that generalizes Maxent to multi-species deep neural networks with a direct relationship to a Poisson count loss. DeepMaxent can be trained with a scalable batched algorithm, and implicitly implements the TGB sampling bias correction initially proposed for Maxent (Phillips et al., 2009). We conduct an extensive evaluation of the method on two SDM benchmark datasets: NCEAS (Elith et al., 2020) and GeoPlant (Picek et al., 2025), which together span a wide range

of species, biological groups, regions and data regimes. The NCEAS dataset enables comparison with a broad set of state-of-the-art SDM methods previously evaluated on this dataset (Phillips et al., 2009; Valavi et al., 2022; Zbinden et al., 2024), such as Maxent and BRT. We also use NCEAS to compare the DeepMaxent loss against batched implementations of alternative losses (Poisson, BCE, CE), and to assess the efficiency of the implicit TGB bias correction compared to uniform random background for each loss. Consistent with previous findings on traditional SDM methods (Barber et al., 2022; Phillips et al., 2009; Ranc et al., 2017), deepSDMs benefit strongly and systematically from TGB regardless of the loss function. DeepMaxent is well adapted to the TGB correction as it outperforms the same model architecture trained using the three alternative losses: the Poisson loss from which it derives, the CE over species (Brun et al., 2024; Deneu et al., 2021) and the BCE loss (encoding occurrences as presence and pseudo-absences; Benkendorf & Hawkins, 2020; Zbinden et al., 2024). DeepMaxent with TGB achieves the highest average AUC across all NCEAS regions and outperforms all other methods in four of the six regions (see Table 2). It notably surpasses the single-species Maxent and BRT methods with TGB correction.

More broadly, our results illustrate that, like other SDM methods, deepSDMs must properly account for spatial sampling biases to reveal their performance potential. This can be achieved through the compatibility of DeepMaxent and TGB, but many other proposed bias correction methods could bring further improvements to DeepMaxent (e.g. Boria et al., 2014; Botella et al., 2021). Future efforts could consider more sophisticated bias correction strategies, such as adapting background sampling strategies to biases specific to each biological group and species, as these biases are often a major source of variation in the performance of SDMs (Schartel & Cao, 2024), or explicitly model spatial sampling effort (Botella et al., 2021; Warton et al., 2013).

For feature extraction on the NCEAS dataset, DeepMaxent with a fully connected neural network with just two hidden layers is found to be optimal across most regions. This simple MLP architecture performs well on tasks involving low-dimensional and unstructured tabular data. While deep learning is often associated with large models, simpler architectures can be more efficient and beneficial for such tasks. For more complex tasks such as the GeoPlant dataset, DeepMaxent is compatible with any type of neural network architecture, allowing the model to obtain the best overall results on Landsat-GeoPlant using a convolutional architecture for time series.

In multi-species settings, DeepMaxent, like other deepSDMs (and some specific approaches such as concurrent ordination; van der Veen et al., 2023), reduces overall computational cost by learning a single feature extractor shared across species (Ba & Caruana, 2014; Raghu et al., 2017), in contrast to most methods that rely on more resource-intensive, per-species computations (Merow et al., 2014). Similar ideas of reducing multi-species complexity into shared latent dimensions exist in statistical ecology (e.g. concurrent ordination [van der Veen et al., 2023] and community-level driver models [Ovaskainen, Tikhonov, Dunson, et al., 2017]), but DeepMaxent implements it through deep neural feature learning rather than explicit latent-variable modelling. It is important to note that each layer added increases the number of model parameters. Although these parameters are shared between the different species, this increase in complexity can lead to overfitting, particularly when data are limited or noisy. A prudent approach, therefore, is to start with simple architectures and then gradually make the model more complex, by adding hidden layers, as long as validation performance continues to improve. A deterioration in validation performance then serves as a warning signal that the model is becoming too complex in relation to the amount of information available in the training data.

TABLE 3 Comparison of method performance by general averaged area under the ROC curve (AUC).

Loss	AUC
Results for Bioclim-GeoPlant	
Maxent	0.823
CE	0.830 ± 0.001
BCE	*0.839* ± 0.001
Poisson loss	0.837 ± 0.002
DeepMaxent	**0.860** ± 0.001
Results for Landsat-GeoPlant	
CE	0.829 ± 0.001
BCE	*0.885* ± 0.002
Poisson loss	0.862 ± 0.002
DeepMaxent	**0.887** ± 0.001

Note: The best average AUC for each column is highlighted in bold, while the second-best averaged AUC is italicized. All methods use TGB for the background samples.

TABLE 4 Average area under the ROC curve (AUC) values for DeepMaxent-Target-Group Background (TGB) across all regions, calculated for six different values of each hyper-parameter: mini-batch size, number of hidden layers and weight decay. The default values used are a mini-batch size of 250, two hidden layers and a weight decay of 3e-4.

Mini-batch size	AUC	Hidden layers	AUC	Weight decay	AUC
10	0.765	1	0.767	0	0.762
25	0.765	2	0.767	3e-5	0.763
100	0.767	3	0.766	1e-4	0.765
250	0.767	4	0.764	3e-4	0.767
1000	0.766	5	0.762	1e-3	0.765
2500	0.764	6	0.759	3e-3	0.757

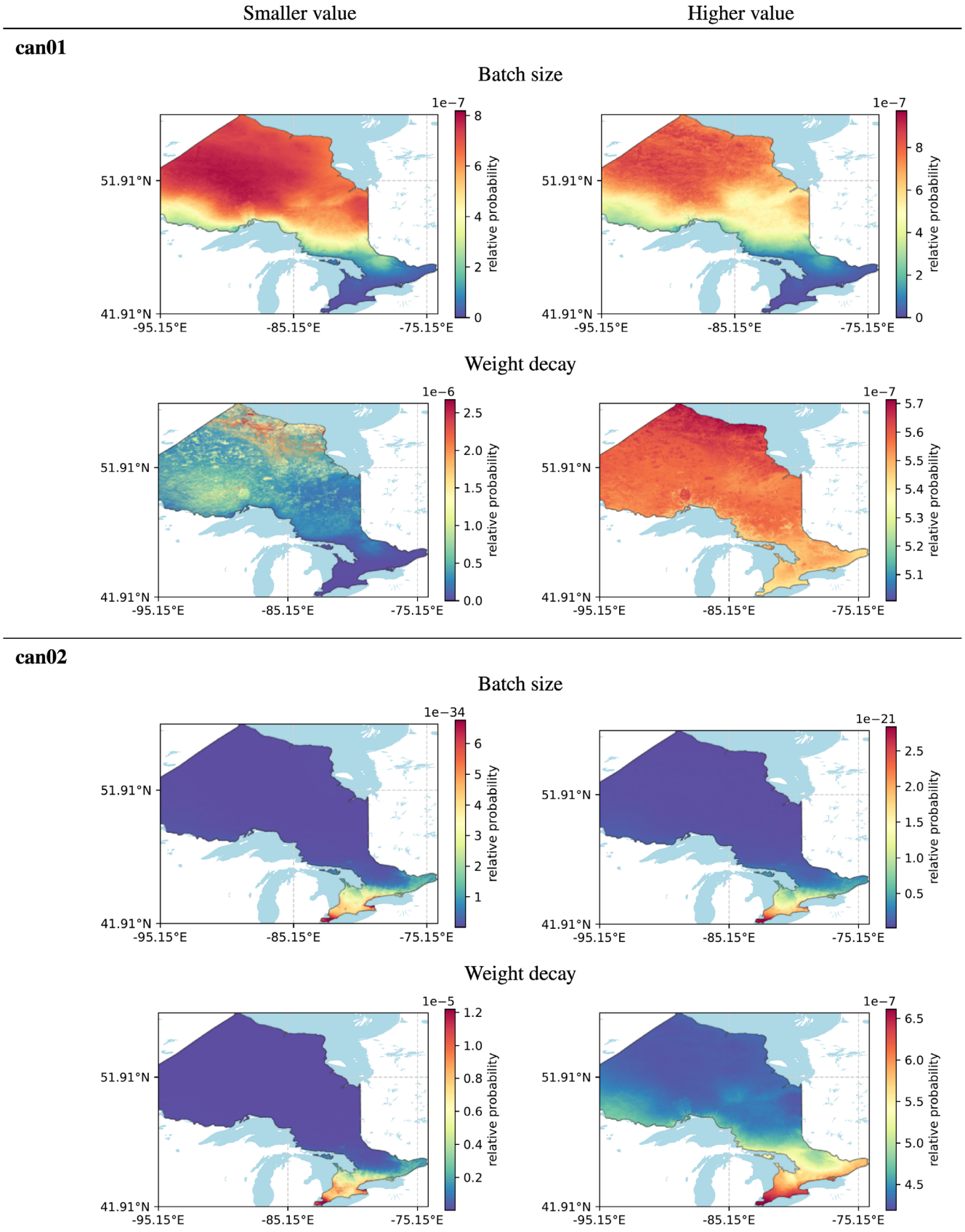

FIGURE 4 Estimated relative probabilities for the species can01 and can02 (CAN) by varying mini-batch size and weight decay, while keeping other hyper-parameters at their default values (mini-batch size = 250, hidden layers = 2, weight decay = 3×10^{-4}). Smaller mini-batch sizes correspond to 10, and larger sizes to 3000. For weight decay, smaller values correspond to 3×10^{-6}, and larger values to 0.1.

More broadly, the proposed approach is flexible regarding the type of input and species observation data and should facilitate data integration approaches in the future by using neural networks. In the case of more structured input data, the neural network architecture could also be adapted to ingest other types of inputs (e.g. spatial remote sensing imagery), which might lead to capturing complementary spatio-temporal environmental patterns (Deneu et al., 2021; Estopinan et al., 2022). Importantly, DeepMaxent is the first method to bridge the gap between deep learning and point process-based SDM (Renner et al., 2015). Using a point-process formulation of species distributions within our loss framework, the present work focuses on PO data through a Poisson likelihood. However, this formulation is general and could, in principle, be extended to incorporate other types of ecological observations, such as PA surveys (Fithian et al., 2015), detection/non-detection histories (Koshkina et al., 2017), or abundance and imperfect count data (Dorazio, 2014). In such extensions, all observation types would remain linked to the same underlying predicted ecological intensity, while each type would be associated with its own appropriate observation likelihood. When these likelihoods involve additional parameters (e.g. detection probabilities), they could be jointly estimated within the same optimization framework or through hierarchical extensions under suitable independence assumptions (see, e.g. Isaac et al., 2020).

These approaches for combining various observation types, called integrated SDMs, have recently been highlighted as a promising avenue to enhance the reliability of SDMs (Isaac et al., 2020; Miller et al., 2019; Mostert & O'Hara, 2023). Extending DeepMaxent with this approach could use standardized datasets to disentangle the real relative abundance of each species from detection biases, while harnessing the extensive geographical coverage of opportunistic PO data. To further increase ecological realism, the DeepMaxent framework could be extended to incorporate species co-occurrence patterns, following the principles of Joint SDMs (Pollock et al., 2014). This integration, which involves modeling the joint probability of species occurrences given the environment, has already proven feasible within deepSDM architectures (Chen et al., 2017).

Maxent and log-linear Poisson regression have been shown to be equivalent in terms of the estimated probability across sites (Renner & Warton, 2013), and we found a similar equivalence between the more general DeepMaxent and Poisson losses, in that they share the same global minimizer when applied to a single species (Appendix A.1). This would lead us to expect similar results for DeepMaxent and Poisson, but DeepMaxent consistently outperformed Poisson in our experiments. One key difference that could explain the discrepancy lies in the treatment of the absolute intensities in the multi-species setting: In the un-normalized Poisson loss, the model will try to capture differences in observation counts between species, leading to more importance in the loss for frequently observed species. In contrast, the DeepMaxent loss normalizes intensities across sites for each species, which removes the effect of the number of observations in the per-species loss magnitude and decouples parameter updates from absolute intensity. Additionally, the stochastic mini-batch optimization interacts with this normalization, further differentiating the parameter estimates between the two losses.

Even though the CE loss has the best results without TGB correction, and thus appears natively less sensitive to spatial sampling bias, its performances with TGB remain below the rest of the methods, which benefit more from this correction (DeepMaxent, BCE, Poisson, Maxent). This limited performance may be due to the loss of information on the spatial variations of the intensity for each species when normalizing the intensity across species for each site. More broadly, regarding the estimation of multiple species spatial intensities from biased PO data, these results suggest that learning to classify the most likely observed species (Brun et al., 2024; Deneu et al., 2021; Estopinan et al., 2022) per site leads to sub-optimal results.

In addition to the smoothing of species spatial densities achieved by the DeepMaxent loss function, which induces entropy maximization, this study highlights that L2 regularization can be used to further encourage smoothing of predicted probabilities (see Figure 4). By penalizing large model parameters, L2 reduces overfitting and produces smoother predictions, resulting in more gradual changes in species probability across space and improved performance in PO settings, potentially alleviating overfitting in the low-data regime.

Concerning the optimization process of DeepMaxent, we provided a mathematical guarantee to justify the use of a stochastic batched gradient descent algorithm and we further showed that varying the batch size (from 10 to 2500) had little impact on general performances (see Table 4). From a computational perspective, this algorithm is highly scalable for large datasets, as it drastically reduces memory and computational requirements by processing only one mini-batch at a time. Moreover, it can take full advantage of GPU parallelization, further accelerating training and enabling efficient handling of high-dimensional or large-scale data (see Appendix H). Yet, the mini-batch size may affect the learning trajectory due to the approximation of the partition function, and we noticed its influence on the final model behaviour. Similar to increasing the weight decay, we observed that decreasing the mini-batch size may smooth species spatial densities. Although mini-batch size is one of the least sensitive hyper-parameters, identifying an optimal mini-batch size remains important, and it may interact with other hyper-parameters such as the learning rate and number of epochs.

One classical limitation of PO-based SDMs is that PA data are often unavailable to validate model hyper-parameters. To address this issue, we evaluated hyper-parameter tuning directly on PO data and found that it reliably identifies suitable values, thereby removing the need for PA data during model training. The corresponding analyses are provided in Appendix C.

Overall, DeepMaxent provides improvements over both traditional PO SDMs and previous deep learning-based approaches on a variety of case studies. It leverages neural networks to learn complex, high-order non-linear relationships directly from data, enabling flexible and expressive modelling, while benefiting from the principles behind the success of the original Maxent. However, several methodological directions remain open. Future extensions

could integrate additional ecological information, such as species traits or phylogenetic relationships, to model inter-species dependencies more explicitly, an approach explored in models like HMSC (Ovaskainen, Tikhonov, Dunson, et al., 2017). Additionally, DeepMaxent could offer a framework for developing and testing bias-reduction strategies in deepSDMs, such as improved background sampling or observation models, to enhance robustness in heterogeneous and biased datasets.

AUTHOR CONTRIBUTIONS

Maxime Ryckewaert, Diego Marcos, Christophe Botella, Maximilien Servajean, Pierre Bonnet and Alexis Joly conceived the ideas and designed methodology; Maxime Ryckewaert analysed the data and led the writing of the manuscript. Maxime Ryckewaert, Diego Marcos, Christophe Botella, Maximilien Servajean, Pierre Bonnet and Alexis Joly contributed critically to the drafts and gave final approval for publication.

ACKNOWLEDGEMENTS

This work was funded by the European Union under the Horizon Europe Research and Innovation Programme through the projects B3 (Biodiversity Building Blocks for Policy; ID No. 101059592) and MAMBO (Modern Approaches to the Monitoring of Biodiversity; ID No. 101060639).

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210x.70262.

DATA AVAILABILITY STATEMENT

The datasets used in this study are derived from previously published sources. The first dataset is available at https://doi.org/10.17161/bi.v15i2.133844 (Elith et al., 2020). The second dataset is available at https://doi.org/10.52202/079017-4023 (Picek et al., 2025). The code used in this study is publicly available on GitHub at https://github.com/RYCKEWAERT/deepmaxent and has been archived and versioned on Zenodo: https://doi.org/10.5281/zenodo.18377697.

ORCID

Maxime Ryckewaert https://orcid.org/0000-0002-9494-797X
Christophe Botella https://orcid.org/0000-0002-5249-911X
Pierre Bonnet https://orcid.org/0000-0002-2828-4389

REFERENCES

Ba, J., & Caruana, R. (2014). Do deep nets really need to be deep? *Advances in Neural Information Processing Systems*, *27*, 2654–2662. https://proceedings.neurips.cc/paper/2014/hash/ea8fcd92d59581717e06eb187f10666d-Abstract.html

Barber, R. A., Ball, S. G., Morris, R. K. A., & Gilbert, F. (2022). Target-group backgrounds prove effective at correcting sampling bias in Maxent models. *Diversity and Distributions*, *28*(1), 128–141.

Benkendorf, D. J., & Hawkins, C. P. (2020). Effects of sample size and network depth on a deep learning approach to species distribution modeling. *Ecological Informatics*, *60*, 101137.

Bonnet, P., Joly, A., Faton, J.-M., Brown, S., Kimiti, D., Deneu, B., Servajean, M., Affouard, A., Lombardo, J.-C., 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), e12023. https://doi.org/10.1002/2688-8319.12023

Boria, R. A., Olson, L. E., Goodman, S. M., & Anderson, R. P. (2014). Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecological Modelling*, *275*, 73–77. https://www.sciencedirect.com/science/article/pii/S0304380013005917

Botella, C., Joly, A., Bonnet, P., Monestiez, P., & Munoz, F. (2018). A deep learning approach to species distribution modelling. In *Multimedia tools and applications for environmental & biodiversity informatics* (pp. 169–199). Springer International Publishing.

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), 933–945.

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), e0232078. https://doi.org/10.1371/journal.pone.0232078

Bottou, L. (1991). Stochastic gradient learning in neural networks. In *Proceedings of Neuro-Nîmes 91*. EC2.

Brun, P., Karger, D. N., Zurell, D., Descombes, P., de Witte, L. C., Lutio, R., Wegner, J. D., & Zimmermann, N. E. (2024). Multispecies deep learning using citizen science data produces more informative plant community models. *Nature Communications*, *15*(1), 4421.

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., & Dhiya'ulhaq, N. U. (2022). The benefits of contributing to the citizen science platform iNaturalist as an identifier. *PLoS Biology*, *20*(11), e3001843. https://doi.org/10.1371/journal.pbio.3001843

Carvalho, S. B., Brito, J. C., Crespo, E. G., Watts, M. E., & Possingham, H. P. (2011). Conservation planning under climate change: Toward accounting for uncertainty in predicted species distributions to increase confidence in conservation investments in space and time. *Biological Conservation*, *144*(7), 2020–2030. https://www.sciencedirect.com/science/article/pii/S0006320711001649

Chen, D., Xue, Y., Fink, D., Chen, S., & Gomes, C. P. (2017). Deep multispecies embedding. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (pp. 3639–3646).

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), e1008856. https://doi.org/10.1371/journal.pcbi.1008856

Dorazio, R. M. (2014). Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Global Ecology and Biogeography*, *23*(12), 1472–1484.

Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M. C. M., Peterson, A. T., … Zimmermann, N. E. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*(2), 129–151. https://doi.org/10.1111/j.2006.0906-7590.04596.x

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), 69–80. https://journals.ku.edu/jbi/article/view/13384

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*. arXiv:2401.05470. http://arxiv.org/abs/2401.05470

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

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), 424–438. https://doi.org/10.1111/2041-210X.12242

Fourcade, Y., Engler, J. O., Rödder, D., & Secondi, J. (2014). Mapping species distributions with MAXENT using a geographically biased sample of presence data: A performance assessment of methods for correcting sampling bias. *PLoS One*, *9*(5), e97122. https://doi.org/10.1371/journal.pone.0097122

Gillespie, L. E., Ruffley, M., & Exposito-Alonso, M. (2024). Deep learning models map rapid plant species changes from citizen science and remote sensing data. *Proceedings of the National Academy of Sciences*, *121*(37), e2318296121. https://doi.org/10.1073/pnas.2318296121

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Guisan, A., Tingley, R., Baumgartner, J. B., Naujokaitis-Lewis, I., Sutcliffe, P. R., Tulloch, A. I. T., Regan, T. J., Brotons, L., McDonald-Madden, E., Mantyka-Pringle, C., Martin, T. G., Rhodes, J. R., Maggini, R., Setterfield, S. A., Elith, J., Schwartz, M. W., Wintle, B. A., Broennimann, O., Austin, M., … Buckley, Y. M. (2013). Predicting species distributions for conservation decisions. *Ecology Letters*, *16*(12), 1424–1435. https://doi.org/10.1111/ele.12189

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778). IEEE. http://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366. https://www.sciencedirect.com/science/article/pii/0893608089900208

Hu, Y., Si-Moussi, S., & Thuiller, W. (2025). Introduction to deep learning methods for multi-species predictions. *Methods in Ecology and Evolution*, *16*(1), 228–246. https://doi.org/10.1111/2041-210X.14466

Isaac, N. J. B., Jarzyna, M. A., Keil, P., Dambly, L. I., Boersch-Supan, P. H., Browning, E., Freeman, S. N., Golding, N., Guillera-Arroita, G., Henrys, P. A., Jarvis, S., Lahoz-Monfort, J., Pagel, J., Pescott, O. L., Schmucki, R., Simmonds, E. G., & O'Hara, R. B. (2020). Data integration for large-scale models of species distributions. *Trends in Ecology & Evolution*, *35*(1), 56–67.

Kellenberger, B., Winner, K., & Jetz, W. (2024). The performance and potential of deep learning for predicting species distributions. *Global Ecology and Biogeography*, *35*, e70184. https://doi.org/10.1101/2024.08.09.607358

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., & Krishnan, D. (2020). Supervised contrastive learning. *Advances in Neural Information Processing Systems*, *33*, 18661–18673.

Kingma, D. P. (2014). *Adam: A method for stochastic optimization*. arXiv:1412.6980.

Komori, O., Saigusa, Y., Eguchi, S., & Kubota, Y. (2024). *Cumulant-based approximation for fast and efficient prediction for species distribution*. arXiv:2405.14456. http://arxiv.org/abs/2405.14456

Koshkina, V., Wang, Y., Gordon, A., Dorazio, R. M., White, M., & Stone, L. (2017). Integrated species distribution models: Combining presence-background data and site-occupancy data with imperfect detection. *Methods in Ecology and Evolution*, *8*(4), 420–430.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://www.nature.com/articles/nature14539

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*(4), 541–551. https://ieeexplore.ieee.org/abstract/document/6795724/

Merow, C., Smith, M. J., Edwards, T. C., Jr., Guisan, A., McMahon, S. M., Normand, S., Thuiller, W., Wüest, R. O., Zimmermann, N. E., & Elith, J. (2014). What do we gain from simplicity versus complexity in species distribution models? *Ecography*, *37*(12), 1267–1281. https://doi.org/10.1111/ecog.00845

Miller, D. A. W., Pacifici, K., Sanderlin, J. S., & Reich, B. J. (2019). The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution*, *10*(1), 22–37.

Mostert, P. S., & O'Hara, R. B. (2023). Pointedsdms: An r package to help facilitate the construction of integrated species distribution models. *Methods in Ecology and Evolution*, *14*(5), 1200–1207.

Ovaskainen, O., Tikhonov, G., Dunson, D., Grøtan, V., Engen, S., Sæther, B.-E., & Abrego, N. (2017). How are species interactions structured in species-rich communities? A new method for analysing time-series data. *Proceedings of the Royal Society B: Biological Sciences*, *284*(1855), 20170768.

Ovaskainen, O., Tikhonov, G., Norberg, A., Blanchet, F. G., Duan, L., Dunson, D., Roslin, T., & Abrego, N. (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters*, *20*(5), 561–576.

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*(3), 231–259. https://www.sciencedirect.com/science/article/pii/S030438000500267X

Phillips, S. J., & Dudík, M. (2008). Modeling of species distributions with Maxent: New extensions and a comprehensive evaluation. *Ecography*, *31*(2), 161–175.

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., & Ferrier, S. (2009). 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

Picek, L., Botella, C., Servajean, M., Leblanc, C., Palard, R., Larcher, T., Deneu, B., Marcos, D., Bonnet, P., & Joly, A. (2025). GeoPlant: Spatial plant species prediction dataset. In *The Thirty-Eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://openreview.net/forum?id=GHIJM45fWY

Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R. B., Parris, K. M., Vesk, P. A., & McCarthy, M. A. (2014). Understanding co-occurrence by modelling species simultaneously with a joint species distribution model (JSDM). *Methods in Ecology and Evolution*, *5*(5), 397–406.

Potapov, P., Hansen, M. C., Kommareddy, I., Kommareddy, A., Turubanova, S., Pickens, A., Adusei, B., Tyukavina, A., & Ying, Q. (2020). Landsat analysis ready data for global land cover and land cover change mapping. *Remote Sensing*, *12*(3), 426. https://www.mdpi.com/2072-4292/12/3/426

Raghu, M., Poole, B., Kleinberg, J., Ganguli, S., & Sohl-Dickstein, J. (2017). On the expressive power of deep neural networks. In *International conference on machine learning* (pp. 2847–2854). PMLR. https://proceedings.mlr.press/v70/raghu17a.html

Ranc, N., Santini, L., Rondinini, C., Boitani, L., Poitevin, F., Angerbjörn, A., & Maiorano, L. (2017). Performance tradeoffs in target-group bias correction for species distribution models. *Ecography*, *40*(9), 1076–1087.

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), 366–379. 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.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913–929. https://doi.org/10.1111/ecog.02881

Schartel, T. E., & Cao, Y. (2024). Background selection complexity influences Maxent predictive performance in freshwater systems. *Ecological Modelling*, *488*, 110592. https://www.sciencedirect.com/science/article/pii/S0304380023003228

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. https://www.sciencedirect.com/science/article/pii/S0893608014002135

Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2019). blockCV: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods in Ecology and Evolution*, *10*(2), 225–232. https://doi.org/10.1111/2041-210X.13107

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), e01486. https://doi.org/10.1002/ecm.1486

van der Veen, B., Hui, F. K. C., Hovstad, K. A., & O'Hara, R. B. (2023). Concurrent ordination: Simultaneous unconstrained and constrained latent variable modelling. *Methods in Ecology and Evolution*, *14*(2), 683–695.

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.

Witjes, M., Parente, L. L., Krizan, J., Antonic, L., & Hengl, T. (2022). *Ecodatacube.eu: Analysis-ready open environmental data cube for Europe*. https://www.researchsquare.com/article/rs-2277090/v3

Yackulic, C. B., Chandler, R., Zipkin, E. F., Royle, J. A., Nichols, J. D., Grant, E. H. C., & 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*. https://www.ssrn.com/abstract=4684222

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Appendix S1. Supplementary theoretical aspects.

Appendix S2. DeepMaxent architecture.

Appendix S3. Cross-validating DeepMaxent using PO data.

Appendix S4. Sensitivity analyses.

Appendix S5. Statistical significance of performance differences.

Appendix S6. Comparison of performances across abundance classes on the NCEAS dataset.

Appendix S7. Estimated probabilities maps by species with corresponding PA data, for CAN region.

Appendix S8. Computational time.

Appendix S9. Dataset description.

How to cite this article: Ryckewaert, M., Marcos, D., Botella, C., Servajean, M., Bonnet, P., & Joly, A. (2026). Applying the maximum entropy principle to neural networks enhances multi-species distribution models. *Methods in Ecology and Evolution*, *00*, 1–16. https://doi.org/10.1111/2041-210x.70262