Multi-Scale and Multimodal Species Distribution Modeling

Nina van Tiel[1], Robin Zbinden[1], Emanuele Dalsasso[1] Benjamin Kellenberger[2], Loïc Pellissier[3,4], and Devis Tuia[1]

[1] Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
[2] University College London, London, United Kingdom
[3] ETH Zurich, Zurich, Switzerland
[4] Swiss Federal Institute for Forest, Snow and Landscape Research, WSL, Birmensdorf, Switzerland

Abstract. Species distribution models (SDMs) aim to predict the distribution of species by relating occurrence data with environmental variables. Recent applications of deep learning to SDMs have enabled new avenues, specifically the inclusion of spatial data (environmental rasters, satellite images) as model predictors, allowing the model to consider the spatial context around each species' observations. However, the appropriate spatial extent of the images is not straightforward to determine and may affect the performance of the model, as scale is recognized as an important factor in SDMs. We develop a modular structure for SDMs that allows us to test the effect of scale in both single- and multi-scale settings. Furthermore, our model enables different scales to be considered for different modalities, using a late fusion approach. Results on the GeoLifeCLEF 2023 benchmark indicate that considering multimodal data and learning multi-scale representations leads to more accurate models.

1 Introduction

In the face of the current biodiversity crisis, biodiversity models informed by ever-growing data are crucial to support conservation efforts [23]. In particular, information about the suitability of species in areas where few observations are recorded enables robust decision-making [15]. Such information is obtained from species distribution models (SDMs), which relate species occurrence data with environmental variables through statistical methods [1, 12, 26].

With the increasing availability of species occurrence data, notably through crowd-sourcing [17,28], the use of deep learning (DL) has recently been explored in SDMs [4,25,29]. DL models are of particular interest thanks to their flexibility in terms of architecture and input data types, their scalability, and their ability to model the distributions of many species with a single model [5, 7, 13, 31].

While most SDM approaches use point values of environmental variables as predictors (*i.e.* tabular data), DL facilitates the integration of geospatial information, which allows the model to consider the spatial context surrounding each species observation. Recent works have used convolutional neural net-

works (CNNs) to integrate patches extracted from rasters of environmental variables [4, 8], satellite images [10, 13], or both [25]. However, the size of the image patches considered by the model is often not justified. Studies on non-DL SDMs have shown that scale affects model performance and that the appropriate scale may depend on the species or the type of environment [14, 19, 21]. Yet, when handling tabular data, these effects are essentially linked to the resolution of the predictors. To the best of our knowledge, no study has focused on the effect of spatial extent[5] of the image patches on the performance of DL-SDMs.

Furthermore, little work has gone into considering spatial data with different resolutions in multimodal DL-SDMs. One study used the same patch size for all images, treating them as a single stack but resulting in different spatial extents for each data source [8]. Others have aligned the resolutions of the different images in pre-processing, even though this leads to unnecessarily large models for coarser grain images [25]. Finally, another study used satellite images alongside coarser bioclimatic data considered as tabular data, hence not considering the spatial context for the latter modality [10].

In this study, we analyze the effects of scale in a modular structure for SDMs based on CNNs. Inspired by works in multi-scale modeling [6, 24] and multimodal modeling with late fusion [10, 22], we design a model that can extract features at multiple scales from a single feature map and from multiple modalities with different resolutions. Our architecture enables each modality to be considered at its native resolution and at different scales. Using the GeoLife-CLEF 2023 (GLC23) benchmark [3], we investigate the effect of spatial extent on model performance in both single- and multi-scale settings, as well as uni- and bi-modal models. Our results indicate superior performance of multi-scale, multimodal approaches. Code to replicate our models can be found on GitHub.

2   Methods

Species data The GLC23 dataset [3] contains two types of georeferenced plant species observation data. It includes 5 million presence-only (PO) observations across Europe. This type of data consists of opportunistic observations in which the non-observation of a species does not confirm its absence. It is widely available but is subject to many biases [9]. Additionally, presence-absence (PA) data is available for 26k sites in France and Great Britain. PA data is more difficult to obtain as it reflects exhaustive sampling with confirmed species absences. It represents species distribution more accurately but is often not available [11]. To reflect the reality of available data for most species, we train our models on PO data and validate them with PA data. In the GLC23 challenge, all teams used PA data for training [2], thus our results are not comparable to those of the leaderboard. As only part of the PA data is openly available in GLC23, we use $7,438$ PA sites for validation. The labels for the remaining $22,404$ sites are kept secret for evaluation through the GLC23 Kaggle page. We use this second

---

[5] We use *scale* or *spatial extent* to refer to the size of image patches considered by the model (i.e. its receptive field), and *resolution* to refer to the size of the pixels.

dataset as a test set, even though it only allows us to report the evaluation metric used by the challenge. For training, we keep the PO occurrences for the $2,173$ species in the validation set and merge occurrences recorded at the same location and date, amounting to $2,856,818$ training samples.

Model predictors We use two modalities among those included in GLC23 dataset [3]: 19 bioclimatic rasters describing temperature and precipitation at a 30-arc seconds resolution (≈ 600 m at $50°$ N, the median latitude of occurrence records) [18] and images from the Sentinel-2 satellite, providing RGB and a near-infrared (NIR) bands at a 10-meter resolution. All input data is normalized by subtracting the mean and dividing by the standard deviation. From the bioclimatic rasters, patches of various sizes between 1×1 and 25×25 pixels are extracted around species occurrence records, corresponding to a ground footprint from $0.6\text{km} \times 0.6\text{km} = 0.36\text{km}^2$, up to $15\text{km} \times 15\text{km} = 225\text{km}^2$. For the satellite data, the dataset provides patches of 128×128 pixels, centered around the species occurrence records. We extract patches with sizes of 25×25, 59×59, and 115×115 pixels, which correspond to a ground footprint of $0.25\text{km} \times 0.25\text{km} = 0.06\text{km}^2$, $0.59\text{km} \times 0.59\text{km} = 0.35\text{km}^2$, and $1.15\text{km} \times 1.15\text{km} = 1.33\text{km}^2$, respectively. Our models take as input image patches with the size required for the largest spatial extent considered for each modality.

Model We propose a model structured in three parts: (1) a common encoder for all scales, (2) a spatial module that can have one or multiple branches for single- or multi-scale models, and (3) a linear classification layer with the same number of output neurons as species, $2,174$ in our case, that is applied to the concatenated outputs of the previous module (Fig. 1a,b). The sigmoid function is applied to the output to obtain predictions between 0 and 1. When considering multiple modalities, each one is encoded separately and the spatial module is adapted to the resolution and the scales to be considered for each modality. We use late-fusion and concatenate the 1024-dimensional feature vectors output by each branch of the spatial modules before the final classification layer (Fig. 1c). Late-fusion has been shown to work well in single-scale multimodal SDMs [10]. Our approach, where modality-specific streams are only fused at the end, allows us to consider each modality at its native resolution and at different scales.

The encoder for bioclimatic variables is composed of four convolutional layers with a kernel size of 1, keeping the encoder's receptive field at 1 pixel, or $0.6\text{km} \times 0.6\text{km}$. Between each convolution, batch normalization and ReLU are applied. The receptive field of 1 allows the downstream spatial module to consider a small number of pixels as spatial extent, which is pertinent with the coarse resolution of 30-arc seconds. The satellite image encoder consists of nine convolutional layers, corresponding to the first layers up to the second residual block of a ResNet model [16], resulting in a receptive field of 25×25 pixels or $0.25\text{km} \times 0.25\text{km}$.

The spatial module may consist of one or multiple branches, where the number of branches corresponds to the number of scales taken into account. Each branch consists of a series of convolutional and max pooling layers, after which
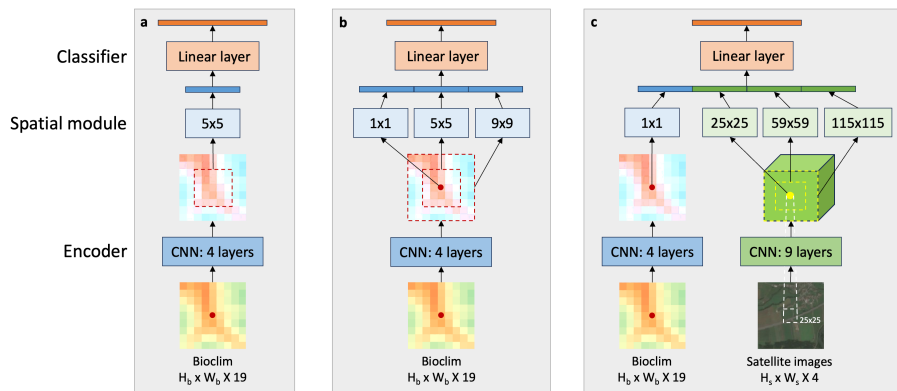
Fig. 1: Example of architectures with our modular structure for SDMs. **a.** Single-scale, unimodal model architecture for bioclimatic variables at scale (5×5). **b.** Multi-scale unimodal model architecture for bioclimatic variables at scales (1×1), (5×5) and (9×9). **c.** Multi-scale multimodal model architecture for bioclimatic variables at scale (1×1), and Sentinel-2 satellite images at scales (25×25), (59×59) and (115×115). The receptive fields after the encoders for bioclimatic variables and satellite image are 1×1 and 25×25 pixels, respectively.

the central pixel of the tensor is extracted to obtain a vector of length 512. The receptive field of the central pixel corresponds to the extent considered by that branch of the spatial module. Finally, a linear layer with $1{,}024$ outputs and ReLU are applied to each vector.

We use a weighted loss function for multi-label classification for SDMs [29], which was shown to perform well on the GLC23 dataset [30]. We consider the records of other species as negatives and do not sample additional pseudo-absences, to avoid the costly operation of downloading new satellite images on the fly. We use a stochastic gradient descent optimizer with a learning rate of 0.01 and weight decay of 0.0001. All models are trained end-to-end for 30 epochs with a batch size of 256 on NVIDIA A100 with 80GB video memory.

Evaluation We evaluate our models on the validation set with the area under the receiver operating characteristic curve (AUC). AUC is widely used in SDMs and measures how well the model discriminates presence from absence sites for each species [27]. We consider the median AUC across species. Additionally, we compute the metric used for the GLC23 challenge [2], the micro-F1 score, on the validation and test sets. This metric measures the overlap between the predicted and actual set of species, averaged over the sites. While AUC can be computed directly on the probabilistic output of our models, the F1 requires binary predictions. Although species- or model-specific binarization schemes may be used, we chose a fixed binarization threshold of 0.5 to ensure comparability among models and avoid overfitting on the validation data.

Table 1: Model performance and training time for unimodal models considering various scales in single- and multi-scale settings. The median species AUC is computed on the validation data. The micro-F1 score is computed on the validation and test data. The best and second-best scores per column are in bold and underlined, respectively. Scales are indicated in pixels and performance metrics are in %.

| Scales | Validation | | Test | |
	AUC	F1	F1	Runtime
1	**86.91**	3.05	<u>2.37</u>	1.5 hrs
5	85.69	**3.16**	<u>2.37</u>	2.6 hrs
9	83.63	2.95	2.36	2.0 hrs
17	83.59	2.30	2.08	3.6 hrs
25	83.65	1.98	1.58	9.7 hrs
1, 5	85.73	3.13	**2.39**	2.8 hrs
1, 5, 9	85.37	<u>3.14</u>	2.30	2.9 hrs
1, 5, 9, 17	<u>86.28</u>	3.05	2.35	6.8 hrs
1, 5, 9, 17, 25	85.12	3.13	2.28	17.9 hrs

(a) Bioclimatic variables

| Scales | Validation | | Test | |
	AUC	F1	F1	Runtime
25	80.41	2.76	1.69	3.9 hrs
59	<u>81.38</u>	2.88	1.98	8.0 hrs
115	80.85	2.98	2.00	28.7 hrs
25, 59	80.67	<u>3.19</u>	<u>2.15</u>	9.6 hrs
25, 59, 115	**81.80**	**3.53**	**2.25**	38.6 hrs

(b) Satellite images

3   Results

Unimodal models First, we train models with only bioclimatic variables or satellite images as predictors. Table 1 shows the performance and training time of single- and multi-scale models with different spatial extents.

Table 1a shows that, when considering only bioclimatic variables, small spatial extents obtain the best performance. While the performance decreases with increasing spatial extent for single-scale models, the different multi-scale models recover the performance of the best single-scale models, albeit with longer training time. We note that the 1×1 scale obtains the best AUC, and the 5×5 scale yields the best F1 score on the validation set. Interestingly, the combination of these scales slightly outperforms both single-scale models on the test set, indicating the marginal advantage of a multi-scale approach for this modality. The differences in performance among these models are relatively small. This may be explained by the high spatial autocorrelation in bioclimatic variables, leading to limited additional information for medium-sized spatial contexts. However, considering large spatial extents can be disadvantageous, indicating that, beyond a certain extent, spatial context is not informative and may even introduce spurious correlations for modeling the distributions of plant species.

In contrast, the results for models with satellite images in Table 1b indicate that taking multiple scales into account leads to better performance. While the F1 scores generally increase with model complexity, the AUC does not consistently follow this trend, with no clear relationship between spatial extent and performance among the single-scale models. This result suggests a larger variability of which scales are most informative, possibly due to the much higher resolution and semantic content of satellite images compared to bioclimatic vari-
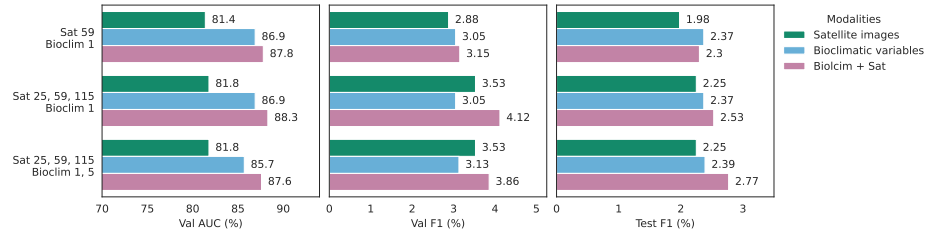
Fig. 2: Performance of bimodal and corresponding unimodal models, quantified by their validation median AUC, and test micro-F1 scores.

ables. We speculate that different scales may be informative for different species or sites and that the multi-scale architecture may learn which spatial extents are relevant, leading to its superior performance.

Multimodal models We compare the performance of bimodal models to their unimodal counterparts (Fig. 2). We consider the best scale or combination of scales for each modality. Concerning the AUC, we find that the models with bioclimatic variables outperform those with satellite images, and combining both modalities leads to further improvement. Regarding the site-wise performance, quantified by the F1 scores, we find that the unimodal models with multiple scales for satellite images outperform the models with bioclimatic variables on the validation set, but the opposite is observed on the test set. The bimodal models perform better than both of their unimodal counterparts, in particular with multi-scale feature extraction for the satellite imagery. Furthermore, the inclusion of multiple scales for both bioclimatic variables and satellite images leads to the best F1 score on the test set. These results confirm the advantage of combining modalities describing the environmental conditions with satellite images [10]. Furthermore, they indicate that multi-scale representations for both modalities lead to better species community predictions. However, such complex models require 20-fold longer training times than the simplest unimodal models. One may consider whether the performance increase is worth the carbon emissions associated with training these models, estimated at 6.91 and 0.35 kg, respectively [20].

Species- and site-level differences in performance While the differences in performance are sometimes relatively small when aggregated, some species or sites have large differences among models. To illustrate this, we compare two models with bioclimatic variables as predictors and two bimodal models (Fig. 3). These pairs of models have differences in median AUC of 1.2% and 0.7%, respectively, and, accordingly, the vast majority of species have a small difference in performance: 90% of species have a difference in AUC of less than 11% between the two pairs of models.The species with larger differences have few records in the validation data (Fig. 3a,d), but no clear trend can be found
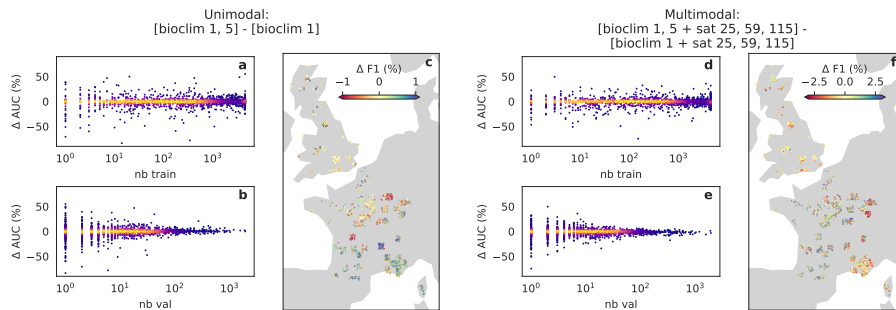
Fig. 3: Difference in median AUC (ΔAUC) and micro-F1 (ΔF1) between two unimodal and two bimodal models. Positive Δ values indicate that models [bioclim 1,5] or [bioclim 1, 5 + sat 25, 59, 115] outperform models [bioclim 1] or [bioclim 1 + sat 25, 59, 115], respectively, and vice-versa for negative values. **a, b, d, e.** ΔAUC values plotted against the number of occurrences in the training data (nb train) and the validation data (nb val) for 2,173 species. Colors indicate point density, with higher densities in yellow. **c, f.** 7,348 validation sites plotted on maps and colored by ΔF1.

with the number of records in the training data (Fig. 3b,e). While this may be explained by the sensitivity of AUC to small sample sizes [27], our results suggest that rare species may be more sensitive to the scale of the predictors. Mapping the differences in the validation F1 score per site reveals geographical clusters (Fig. 3c,f). These clusters are more locally marked for bioclimatic variables, with a clear preference for a certain model in some regions, despite a median difference in F1 of 0.03%. For the bimodal models, considering a single scale for bioclimatic variables generally leads to better site-wise performance, with a median difference in F1 of -0.23%. However, this difference is less marked in certain regions, such as the north and west of France. These qualitative results indicate that multi-scale models may be more informative for some species or in some regions, characterized by a specific type of ecosystem. We leave to future work the further investigation of these relationships.

4 Conclusion

In this study, we develop a modular structure for SDMs and explore the effect of the scale of spatial predictor variables on the GLC23 dataset for European plant species distributions. We find that small scales are most appropriate when considering bioclimatic variables. When using satellite images, our multi-scale approach showed a clear benefit in performance. Combining the best architectures for each modality with a late fusion scheme leads to further increases in performance, indicating the complementarity of both modalities. Overall, our multi-scale and multimodal model achieved the best performances for both species-wise and site-wise evaluation. Our results suggest that the most informative scales may depend on the species or site. Future work may explore these relationships further, and investigate scale-dependencies in other species groups beyond plants.

References

1. Beery, S., Cole, E., Parker, J., Perona, P., Winner, K.: Species distribution modeling for machine learning practitioners: A review. In: Proceedings of the 4th ACM SIGCAS Conference on Computing and Sustainable Societies. pp. 329–348 (2021)
2. Botella, C., Deneu, B., Gonzalez, D.M., Servajean, M., Larcher, T., Leblanc, C., Estopinan, J., Bonnet, P., Joly, A.: Overview of geolifeclef 2023: Species composition prediction with high spatial resolution at continental scale using remote sensing. In: CLEF 2023: Conference and Labs of the Evaluation Forum (2023)
3. Botella, C., Deneu, B., Marcos, D., Servajean, M., Estopinan, J., Larcher, T., Leblanc, C., Bonnet, P., Joly, A.: The geolifeclef 2023 dataset to evaluate plant species distribution models at high spatial resolution across europe (2023)
4. Botella, C., Joly, A., Bonnet, P., Monestiez, P., Munoz, F.: A deep learning approach to species distribution modelling. Multimedia Tools and Applications for Environmental & Biodiversity Informatics pp. 169–199 (2018)
5. Brun, P., Karger, D.N., Zurell, D., Descombes, P., de Witte, L.C., de Lutio, R., Wegner, J.D., Zimmermann, N.E.: Multispecies deep learning using citizen science data produces more informative plant community models. Nature Communications **15**(1), 4421 (2024)
6. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)
7. Cole, E., Van Horn, G., Lange, C., Shepard, A., Leary, P., Perona, P., Loarie, S., Mac Aodha, O.: Spatial implicit neural representations for global-scale species mapping. In: International Conference on Machine Learning. pp. 6320–6342. PMLR (2023)
8. Deneu, B., Servajean, M., Bonnet, P., Botella, C., Munoz, F., Joly, A.: Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment. PLoS computational biology **17**(4), e1008856 (2021)
9. Di Cecco, G.J., Barve, V., Belitz, M.W., Stucky, B.J., Guralnick, R.P., Hurlbert, A.H.: Observing the observers: how participants contribute data to inaturalist and implications for biodiversity science. BioScience **71**(11), 1179–1188 (2021)
10. Dollinger, J., Brun, P., Sainte Fare Garnot, V., Wegner, J.D.: Sat-sinr: High-resolution species distribution models through satellite imagery. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences **10**, 41–48 (2024)
11. Elith, J., Graham, C., Valavi, R., Abegg, M., Bruce, C., Ferrier, S., Ford, A., Guisan, A., Hijmans, R.J., Huettmann, F., et al.: Presence-only and presence-absence data for comparing species distribution modeling methods. Biodiversity informatics **15**(2), 69–80 (2020)
12. Elith, J., Leathwick, J.R.: Species distribution models: ecological explanation and prediction across space and time. Annual review of ecology, evolution, and systematics **40**(1), 677–697 (2009)
13. Estopinan, J., Servajean, M., Bonnet, P., Munoz, F., Joly, A.: Deep species distribution modeling from sentinel-2 image time-series: a global scale analysis on the orchid family. Frontiers in Plant Science **13**, 839327 (2022)
14. Guisan, A., Thuiller, W.: Predicting species distribution: offering more than simple habitat models. Ecology letters **8**(9), 993–1009 (2005)

15. Guisan, A., Tingley, R., Baumgartner, J.B., Naujokaitis-Lewis, I., Sutcliffe, P.R., Tulloch, A.I., Regan, T.J., Brotons, L., McDonald-Madden, E., Mantyka-Pringle, C., et al.: Predicting species distributions for conservation decisions. Ecology letters **16**(12), 1424–1435 (2013)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
17. Joly, A., Goëau, H., Champ, J., Dufour-Kowalski, S., Müller, H., Bonnet, P.: Crowdsourcing biodiversity monitoring: how sharing your photo stream can sustain our planet. In: Proceedings of the 24th ACM international conference on Multimedia. pp. 958–967 (2016)
18. Karger, D.N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R.W., Zimmermann, N.E., Linder, H.P., Kessler, M.: Climatologies at high resolution for the earth's land surface areas. Scientific data **4**(1), 1–20 (2017)
19. König, C., Wüest, R.O., Graham, C.H., Karger, D.N., Sattler, T., Zimmermann, N.E., Zurell, D.: Scale dependency of joint species distribution models challenges interpretation of biotic interactions. Journal of Biogeography **48**(7), 1541–1551 (2021)
20. Lacoste, A., Luccioni, A., Schmidt, V., Dandres, T.: Quantifying the carbon emissions of machine learning. arXiv preprint arXiv:1910.09700 (2019)
21. Lu, M., Jetz, W.: Scale-sensitivity in the measurement and interpretation of environmental niches. Trends in Ecology & Evolution **38**(6), 554–567 (2023)
22. Mac Aodha, O., Cole, E., Perona, P.: Presence-only geographical priors for fine-grained image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9596–9606 (2019)
23. Pollock, L.J., O'connor, L.M., Mokany, K., Rosauer, D.F., Talluto, M.V., Thuiller, W.: Protecting biodiversity (in all its complexity): new models and methods. Trends in Ecology & Evolution **35**(12), 1119–1128 (2020)
24. Reed, C.J., Gupta, R., Li, S., Brockman, S., Funk, C., Clipp, B., Keutzer, K., Candido, S., Uyttendaele, M., Darrell, T.: Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4088–4099 (2023)
25. Teng, M., Elmustafa, A., Akera, B., Bengio, Y., Radi, H., Larochelle, H., Rolnick, D.: Satbird: a dataset for bird species distribution modeling using remote sensing and citizen science data. Advances in Neural Information Processing Systems **36** (2024)
26. van Tiel, N., Lyu, L., Fopp, F., Brun, P., van der Hoogen, J., Karger, D.N., Casadei, C.M., Tuia., D., Zimmermann, N.E., Crowther, T., Pellissier, L.: Regional uniqueness of tree species composition and response to forest loss and climate change. Nature Communications **15**(4375) (2024)
27. Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J.J., Elith, J.: Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. Ecological Monographs **92**(1), e01486 (2022)
28. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8769–8778 (2018)
29. Zbinden, R., van Tiel, N., Kellenberger, B., Hughes, L., Tuia, D.: On the selection and effectiveness of pseudo-absences for species distribution modeling with deep learning. arXiv preprint arXiv:2401.02989 (2024)

30. Zbinden, R., Van Tiel, N., Rußwurm, M., Tuia, D.: Imbalance-aware presence-only loss function for species distribution modeling. arXiv preprint arXiv:2403.07472 (2024)
31. Zbinden, R., Van Tiel, N.M.A., Kellenberger, B.A., Hughes, L., Tuia, D.: Exploring neural networks and their potential for species distribution modeling. In: 11th International Conference on Learning Representations (ICLR) Workshops (2023)