TOWARDS SCALABLE DEEP SPECIES DISTRIBUTION MODELLING USING GLOBAL REMOTE SENSING

Emily F. Morris[†], **Anil Madhavapeddy**[†], **Sadiq Jaffer**[†] **and David Coomes**
Departments of Computer Science[†] and Plant Sciences
University of Cambridge, United Kingdom

ABSTRACT

Destruction of natural habitats and anthropogenic climate change are threatening biodiversity globally. Addressing this loss necessitates enhanced monitoring techniques to assess the impact of environmental shifts and to guide policy-making efforts. Species distribution models are crucial tools that predict species locations by interpolating observed field data with environmental information. We develop an improved, scalable method for species distribution modelling by proposing a dataset pipeline that incorporates global remote sensing imagery, land use classification data, environmental variables, and observation data, and utilising this with convolutional neural network (CNN) models to predict species presence at higher spatial and temporal resolutions than well-established species distribution modelling methods. We apply our approach to modelling Protea species distributions in the Cape Floristic Region of South Africa, demonstrating its performance in a region of high biodiversity. We train two CNN models and compare their performance to Maxent, a popular conventional species distribution modelling method. We find that the CNN models trained with remote sensing data outperform Maxent, underscoring the potential of our method as an effective and scalable solution for modelling species distribution.

1 INTRODUCTION

Species distribution models (SDMs) are an essential tool for biodiversity conservation due to their linkage of science to decision-making (McShea, 2014), with two prevalent approaches to constructing them. The first uses expert knowledge of a species' range and its habitat preferences within that range (Luedtke et al., 2023). The second fits SDMs using machine learning approaches that model field observations of species occurrences as non-linear functions of bioclimatic and environmental spatial layers. Maxent (Phillips et al., 2006) is the most widely adopted approach (Elith et al., 2011); a recent dataset on the global distribution of utilised plants fitted Maxent models to predict distributions of 28,235 plant species (Pironon et al., 2024). SDMs have traditionally used spatial layers with course-grained spatiotemporal resolution, making it hard to produce local fine-grained predictions for species occurrence. Furthermore, SDMs generally only use the data at a single observation point and do not leverage information about the surrounding area, which can provide valuable insights into a species' habitat such as proximity to water and neighbouring vegetation type.

We aim to provide a globally applicable method for creating SDMs, with the broader goal of providing an accurate view of where wild plant and animal species live across the planet for decision-makers to balance biodiversity preservation with human needs. Mapping at high resolution is becoming increasingly important for policymaking as climate and anthropogenic changes have local effects that are not captured by coarse environmental variables. Our approach combines convolutional neural network (CNN) models (LeCun et al., 2015) that exploit spatial information with high-resolution satellite data to produce accurate SDMs that also track local habitat changes. Using remote sensing data as predictive features for SDMs also makes the models more applicable to regions where ground-based features are unavailable due to funding constraints, political instability or lack of capacity (Cavender-Bares et al., 2022).

Previous work has created datasets to train SDMs (Joly et al., 2014; Gillespie et al., 2021) along with approaches (Botella et al., 2018; Deneu et al., 2019; 2021), but they focus on data-rich areas

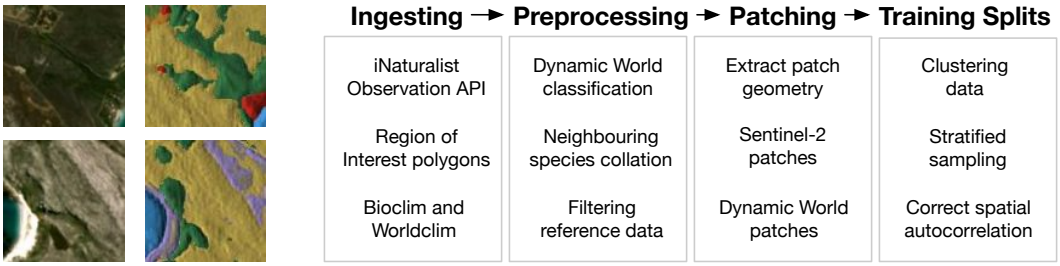

Figure 1: Sample S2-Cloudless and LULC classifications data *(left)* and the data pipeline *(right)*

such as Europe and the USA. We extend these processes to globally available data and investigate how they perform in regions of high biodiversity, specifically via a case study of modelling species under the Protea genus in the Cape Floristic Region in South Africa, one of the global biodiversity hotspots (Myers et al., 2000). We present a dataset pipeline that combines remote sensing imagery data, land use land classification data and environmental variables with species observations to a create species distribution modelling dataset, which we then use to train two different versions of a CNN architecture to perform species distribution modelling. We compare the performance of these approaches to Maxent, which we train using only environmental variables.

2 Dataset, Model and Methods

2.1 Data Sources

A key aspect of species distribution modelling is the collection of ground truth observation data which provides information on species' locations. Most models are created using historical datasets from herbariums or smaller-scale, local datasets collected by ecologists. Historical datasets are difficult to utilise with more temporally high-resolution remote sensing data. Given the rapid rate of climate change and anthropogenic habitat changes over recent years, there is no guarantee that natural habitats are still found at the site of these observations (Bracken et al., 2022). We use iNaturalist (Van Horn et al., 2018) as a crowd-sourced reference dataset for our pipeline, as it provides both recent and global species presence observation data.

We use Sentinel 2 (S2) cloudless satellite basemaps (EOX) in the RGB $10m$ resolution bands along with land-use land classification (LULC) data (see Figure 1), combined with traditional low-resolution $1km$ bioclimatic environmental variables from WorldClim V2 based on temperature and precipitation. The S2 data provides a visual representation of the environment from which key habitat features can be extracted. Recent work by Díaz et al. (2019) highlighted that anthropogenic land use changes, such as land clearing for agriculture or settlement expansion, have been the primary drivers of biodiversity loss over the last 50 years, making it an important variable to include in SDMs. We use "Dynamic World", a near real-time map (Brown et al., 2022) to leverage fine-grained $10m$ resolution LULC classifications for SDMs.

2.2 Dataset Pipeline

The dataset creation pipeline in Figure 1 *(right)* extends that proposed by Gillespie et al. (2021), incorporating additional filtering, data sources and methods for creating dataset splits. The following sections briefly detail the key elements of filtering observations and creating the training splits.

2.2.1 Filter Reference Data

We use the Global Biodiversity Information Facility, an online network that combines biodiversity data from a variety of sources, to download and preliminarily filter the iNaturalist observation data based on parameters such as location, recording date, and location uncertainty distance. Once these observations are downloaded we then further filter them via a shapefile for the region of interest – in our case, the Cape Floristic Region (Hoffman et al., 2016). To address the uncertainty in the location of the data points, as well as address any potential changes that may have occurred to vegetation coverage over time, we extract the Dynamic World land classification for each observation location and remove observations classified as "water" or "built".

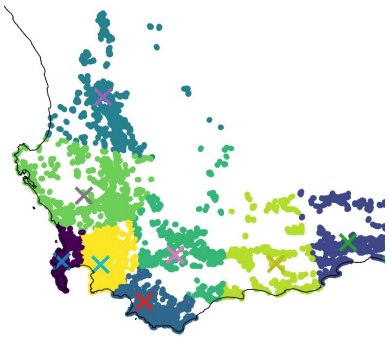

Table 1: *(left)* Example clusters for creating training splits and *(below)* statistics for test, train, and validation sets.

	Test Set	Train Set	Valid Set
Observations	7582	153703	5297
Species	2108	4019	1848
Protea Observations	192	5620	299
Protea Occurrences	1741	93576	1323

The co-occurrence of species has an important impact on the existence of many species (Wisz et al., 2013). We leverage species co-occurrence information by training a model to simultaneously predict the occurrence of all species at a given location, not just the species of interest. Furthermore, including other species should have a regularising effect on the model and stop it from overfitting when training, as well as providing "pseudo-absences" (Barbet-Massin et al., 2012) that allow the model to learn where the species do not occur. Thus although we aim to only perform species distribution modelling for Proteas, we include all species in the *Tracheophyta phylum* in our dataset. To include co-species information in our data points, we use a per-observation method (Gillespie et al., 2021). This method creates patches centered on an *observation* that represents a single iNaturalist entry. Each patch is a geographic area and is labelled with all species in observations intersecting with it. This means a single observation may *occur* in multiple neighbouring patches.

2.2.2 Create Training Splits

When performing data analysis with geospatial data, a commonly encountered phenomenon is spatial autocorrelation (Getis, 2008). Spatial autocorrelation between the training and evaluation sets can cause an overly optimistic view of model performance (Karasiak et al., 2022; Ploton et al., 2020; Kattenborn et al., 2022). Including points in the test set that are spatially close to points in the train set could mean these points share very similar features. One solution to this is to split out data into test and train sets using some form of spatial clustering, to ensure spatial separation between training and evaluation splits.

To create a test set free from spatial autocorrelation, we use the method by Gillespie et al. (2021) with a different method for creating the validation set that is more representative of the statistics of the test set. We focus our evaluation on the six Protea species (Table 2) which had more than 100 test occurrences, more than 10 test observations and more training occurrences than test occurrences.

To create the validation set, we first spatially cluster the data using K-means clustering (Arthur & Vassilvitskii, 2007) to ensure that the validation set provides good spatial coverage across the whole region of interest. We use 10 clusters in this work. For each cluster, we order the observations by the number of overlapping points and use the lowest 8% of each species of interest for the validation set. This reduces the number of observations that need to be removed from the training set due to their overlap with validation samples. To increase the number of occurrences for the species we then also include all neighbouring observations in the validation set. To account for spatial autocorrelation, we remove all the observations that overlap with any observations in the validation set from the train set.

However, constructing the split in this manner means that all samples in the validation set contained at least one Protea. This is not representative of the test set, where only about a quarter of the samples contain Protea species. Thus, we use the same method previously described to add samples that do not overlap with any Protea species to replicate the Protea presence/absence ratio. We sample $x \times 3 \times num\_protea\_examples$ observations from each cluster, where x is determined by the proportion of the total number of samples that do not contain Proteas in that cluster. General statistics about the data split can be found in Table 1 and species splits can be found in Table 2.

Table 2: Number of observations and occurrences for the six Protea species of interest across the test, train, and validation sets.

	Observations per Set			Occurrences per Set		
	Test	Train	Validation	Test	Train	Validation
Protea repens (PR)	28	662	64	416	35917	599
Protea laurifolia (PLA)	20	400	34	346	13356	228
Protea nitida (PNI)	18	500	49	479	25301	369
Protea cynaroides (PC)	13	634	59	131	34137	451
Protea neriifolia (PNE)	13	354	30	200	19998	285
Protea lorifolia (PLO)	11	108	11	170	3530	29

2.3 Models and Experimentation

2.3.1 Maxent

We use the Google Earth Engine (Crego et al., 2022) implementation of Maxent (Phillips et al.) with default settings to mirror the lack of fine-tuning in the deep learning model using only the bioclim variables. To create Maxent models for our species of interest, we use the same observations from the dataset described in the previous section. While some presence points in this dataset have been removed through the pipeline filtering, these points were spatially close enough that they would have had very similar if not duplicate environmental variable values, given the resolution of the WorldClim V2 rasters. Thus it should not affect predictive performance.

2.3.2 Deep Learning Model

We use the *Deepbiosphere* model (Gillespie et al., 2021) to perform our experimentation with our remote sensing dataset. Given an input data sample, including image data and environmental variable values, the model can be configured to predict either the families, genera, and species that occur or just the species that occur. We choose the former option, as this gives us the ability to predict where the Protea genus occurs, as well as each of our species of interest. Gillespie also presents a novel loss function, frequency-scaled binary cross-entropy loss, which proportionally weights absence and presence predictions equally. Since samples mostly consist of absence predictions, this prevents the model from learning to always predict species as absent.

For data preprocessing, we follow the TResNet (Ridnik et al., 2021) image preprocessing procedure. We do not use any augmentation strategies while training. We train two models, one using satellite images and environmental variables, which we shall refer to as *Deepbiosphere*$_{Image}$, and the other using satellite images, Dynamic World LULC images and environmental variables, which we shall refer to as *Deepbiosphere*$_{Image+DW}$. For the latter approach, we stack the satellite images and Dynamic World LULC images and pass the six-channel input to the CNN. To train the models we use the Adam optimizer with a learning rate of $1\mathrm{e}{-4}$, and a batch size of 165 and train for 100 epochs. Using our validation metrics, we choose the best-performing model checkpoint, and evaluate these models on the test set.

3 Results

Comparative Efficacy. Area Under the Curve Receiver Operating Characteristics (AUC_{ROC}) is one of the most common metrics used to measure species distribution model performance. We report this metric to compare our models across the selected Protea species (see Table 3). Both *Deepbiosphere*$_{Image}$ and *Deepbiosphere*$_{Image+DW}$ on average outperform Maxent with respect to AUC_{ROC} by 2.99 and 2.72 percent respectively. The results also suggest that this method for species distribution modelling performs well in areas of high biodiversity, which is crucial to scaling SDMs globally to include the tropical belt where an estimated two-thirds of the world's terrestrial biodiversity lives. Despite our dataset containing about double the number of total species as the

Table 3: AUC_{ROC} results for the Maxent model, *Deepbiosphere*$_{Image}$ model, and *Deepbiosphere*$_{Image+DW}$ model for the *Protea repens* (PR), *Protea laurifolia* (PLA), *Protea nitida* (PNI), *Protea cynaroides* (PC), *Protea neriifolia* (PNE), and *Protea lorifolia* (PLO) species.

	PR	PLA	PNI	PC	PNE	PLO
Maxent	0.7524	0.9282	0.8165	0.9166	0.8453	0.8709
Deepbiosphere$_{Image}$	0.8159	**0.9392**	**0.8280**	0.9143	**0.9040**	**0.9082**
Deepbiosphere$_{Image+DW}$	**0.8292**	0.9192	0.8036	**0.9324**	0.9031	0.9059

dataset created by Gillespie et al. (2021), we achieve similar performance albeit on different remote sensing data.

Presence vs Absence. The crowd-sourced dataset (iNaturalist) used for our observations does not contain true absence points which makes it difficult to compare and interpret the negative predictions of the models (Lobo et al., 2010). Thus to fully evaluate the models, a dataset comprised of true absence and presence data collected in a structured field campaign is required. Such a dataset would also allow for analysis of biases in the citizen science dataset and understanding how these affect model performance.

Land Use Datasets. While there is a difference in performance between the *Deepbiosphere* models and Maxent, there is no substantial difference between the performance of the two *Deepbiosphere* models. The *Deepbiosphere*$_{Image}$ model mostly outperforms the *Deepbiosphere*$_{Image+DW}$ model in the per-species metrics. Our hypothesis here is that the land use land cover classification (LULC) classes are too course-grained in Dynamic World and, being derived from the same Sentinel-2 imagery used as input to our model, the network may be extracting relevant information directly from the images during training. The decrease in performance of the *Deepbiosphere*$_{Image+DW}$ model versus Maxent for the *Protea Nitida* reveals the brittleness caused by this duplication, as the *Deepbiosphere*$_{Image}$ model remains an improvement over Maxent.

4 Conclusions and Future Work

In this work, we investigated the use of remote sensing data and CNN models to provide an improved method for performing scalable species distribution modelling. We found that using a deep learning approach has provided us with an exciting alternative that is at least as accurate as the prevalent Maxent method, and also one that naturally scales up with more data availability. While we used a case study of Protea species in the Cape Floristic Region in South Africa to illustrate the technique, our ambition is to extend this analysis to the full spectrum of plant and animal species worldwide to facilitate more accurate policymaking for environmental preservation. However, a major barrier is the sparseness of occurrence datasets for many species; incredibly, 30% of utilised plant species have fewer than 10 records in digital databases, which is too few to fit simple SDMs. Thus there are huge discrepancies in the amount of data available across geographies, with the tropics most poorly represented (Chapman et al., 2024). While the collection and digitalisation of large field datasets is the long-term solution, the approach presented here of combining LULC and habitat classification with sparse observation data could also work for data-deficient species if we combine expert ecological knowledge (Merow et al., 2022) in the training process.

Another advantage of Maxent which needs to be incorporated into deep learning SDMs is that approaches for addressing sample biases and other problems are well developed (Elith et al., 2010), although there is evidence that non-parametric models (especially ensembles thereof) can outperform this approach (Valavi et al., 2021). Future work should properly investigate the effect of and methods for addressing sample bias in deep learning SDMs.

Acknowledgments

EF. Morris was funded by the DeepMind Cambridge Scholarship and S. Jaffer by a donation from John Bernstein.

REFERENCES

David Arthur and Sergei Vassilvitskii. K-means++ the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1027–1035, 2007.

Morgane Barbet-Massin, Frédéric Jiguet, Cécile Hélène Albert, and Wilfried Thuiller. Selecting pseudo-absences for species distribution models: How, where and how many? *Methods in ecology and evolution*, 3(2):327–338, 2012.

Christophe Botella, Alexis Joly, Pierre Bonnet, Pascal Monestiez, and François Munoz. A deep learning approach to species distribution modelling. *Multimedia Tools and Applications for Environmental & Biodiversity Informatics*, pp. 169–199, 2018.

Jason T Bracken, Amelie Y Davis, Katherine M O'Donnell, William J Barichivich, Susan C Walls, and Tereza Jezkova. Maximizing species distribution model performance when using historical occurrences and variables of varying persistency. *Ecosphere*, 13(3):e3951, 2022.

Christopher F Brown, Steven P Brumby, Brookie Guzder-Williams, Tanya Birch, Samantha Brooks Hyde, Joseph Mazzariello, Wanda Czerwinski, Valerie J Pasquarella, Robert Haertel, Simon Ilyushchenko, et al. Dynamic world, near real-time global 10 m land use land cover mapping. *Scientific Data*, 9(1):251, 2022.

Jeannine Cavender-Bares, Fabian D Schneider, Maria João Santos, Amanda Armstrong, Ana Carnaval, Kyla M Dahlin, Lola Fatoyinbo, George C Hurtt, David Schimel, Philip A Townsend, et al. Integrating remote sensing with ecology and evolution to advance biodiversity conservation. *Nature Ecology & Evolution*, 6(5):506–519, 2022.

Melissa Chapman, Benjamin R. Goldstein, Christopher J. Schell, Justin S. Brashares, Neil H. Carter, Diego Ellis-Soto, Hilary Oliva Faxon, Jenny E. Goldstein, Benjamin S. Halpern, Joycelyn Longdon, Kari E. A. Norman, Dara O'Rourke, Caleb Scoville, Lily Xu, and Carl Boettiger. Biodiversity monitoring for a just planetary future. *Science*, 383(6678):34–36, January 2024. ISSN 1095-9203. doi: 10.1126/science.adh8874. URL http://dx.doi.org/10.1126/science.adh8874.

Ramiro D. Crego, Jared A. Stabach, and Grant Connette. Implementation of species distribution models in google earth engine. *Diversity and Distributions*, 28(5):904–916, February 2022. ISSN 1472-4642. doi: 10.1111/ddi.13491. URL http://dx.doi.org/10.1111/ddi.13491.

Benjamin Deneu, Maximilien Servajean, Christophe Botella, and Alexis Joly. Evaluation of deep species distribution models using environment and co-occurrences. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10*, pp. 213–225. Springer, 2019.

Benjamin Deneu, Maximilien Servajean, Pierre Bonnet, Christophe Botella, François Munoz, and Alexis Joly. Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment. *PLoS computational biology*, 17(4):e1008856, 2021.

Sandra Myrna Díaz, Josef Settele, Eduardo Brondízio, Hien Ngo, Maximilien Guèze, John Agard, Almut Arneth, Patricia Balvanera, Kate Brauman, Stuart Butchart, et al. The global assessment report on biodiversity and ecosystem services: Summary for policy makers. 2019.

Jane Elith, Steven J. Phillips, Trevor Hastie, Miroslav Dudík, Yung En Chee, and Colin J. Yates. A statistical explanation of maxent for ecologists: Statistical explanation of maxent. *Diversity and Distributions*, 17(1):43–57, November 2010. ISSN 1366-9516. doi: 10.1111/j.1472-4642.2010.00725.x. URL http://dx.doi.org/10.1111/j.1472-4642.2010.00725.x.

Jane Elith, Steven J Phillips, Trevor Hastie, Miroslav Dudík, Yung En Chee, and Colin J Yates. A statistical explanation of maxent for ecologists. *Diversity and distributions*, 17(1):43–57, 2011.

EOX. The global and cloudless sentinel-2 map by eox. https://s2maps.eu/#.

Arthur Getis. A history of the concept of spatial autocorrelation: A geographer's perspective. *Geographical analysis*, 40(3):297–309, 2008.

Lauren Gillespie, Megan Ruffley, and Moisés Expósito-Alonso. An image is worth a thousand species: Scaling high-resolution plant biodiversity prediction to biome-level using citizen science data and remote sensing imagery. *Biodiversity Information Science and Standards*, 5:e74052, 2021. doi: 10.3897/biss.5.74052. URL https://doi.org/10.3897/biss.5.74052.

Global Biodiversity Information Facility. https://www.gbif.org/.

Michael Hoffman, Kellee Koenig, Gill Bunting, Jennifer Costanza, and Kristen J. Williams. Biodiversity hotspots (version 2016.1), April 2016. URL https://doi.org/10.5281/zenodo.3261807.

Alexis Joly, Hervé Goëau, Hervé Glotin, Concetto Spampinato, Pierre Bonnet, Willem-Pier Vellinga, Robert Planque, Andreas Rauber, Robert Fisher, and Henning Müller. Lifeclef 2014: multimedia life species identification challenges. In *Information Access Evaluation. Multilinguality, Multimodality, and Interaction: 5th International Conference of the CLEF Initiative, CLEF 2014, Sheffield, UK, September 15-18, 2014. Proceedings 5*, pp. 229–249. Springer, 2014.

Nicolas Karasiak, J-F Dejoux, Claude Monteil, and David Sheeren. Spatial dependence between training and test sets: another pitfall of classification accuracy assessment in remote sensing. *Machine Learning*, 111(7):2715–2740, 2022.

Teja Kattenborn, Felix Schiefer, Julian Frey, Hannes Feilhauer, Miguel D Mahecha, and Carsten F Dormann. Spatially autocorrelated training and validation samples inflate performance assessment of convolutional neural networks. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 5:100018, 2022.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Jorge M. Lobo, Alberto Jiménez-Valverde, and Joaquín Hortal. The uncertain nature of absences and their importance in species distribution modelling. *Ecography*, 33(1):103–114, February 2010. ISSN 1600-0587. doi: 10.1111/j.1600-0587.2009.06039.x. URL http://dx.doi.org/10.1111/j.1600-0587.2009.06039.x.

Jennifer A. Luedtke, Janice Chanson, Kelsey Neam, Louise Hobin, Adriano O. Maciel, Alessandro Catenazzi, Amaël Borzée, Amir Hamidy, Anchalee Aowphol, Anderson Jean, Ángel Sosa-Bartuano, Ansel Fong G., Anslem de Silva, Antoine Fouquet, Ariadne Angulo, Artem A. Kidov, Arturo Muñoz Saravia, Arvin C. Diesmos, Atsushi Tominaga, Biraj Shrestha, Brian Gratwicke, Burhan Tjaturadi, Carlos C. Martínez Rivera, Carlos R. Vásquez Almazán, Celsa Señaris, S. R. Chandramouli, Christine Strüssmann, Claudia Fabiola Cortez Fernández, Claudio Azat, Conrad J. Hoskin, Craig Hilton-Taylor, Damion L. Whyte, David J. Gower, Deanna H. Olson, Diego F. Cisneros-Heredia, Diego José Santana, Elizah Nagombi, Elnaz Najafi-Majd, Evan S. H. Quah, Federico Bolaños, Feng Xie, Francisco Brusquetti, Francisco S. Álvarez, Franco Andreone, Frank Glaw, Franklin Enrique Castañeda, Fred Kraus, Gabriela Parra-Olea, Gerardo Chaves, Guido F. Medina-Rangel, Gustavo González-Durán, H. Mauricio Ortega-Andrade, Iberê F. Machado, Indraneil Das, Iuri Ribeiro Dias, J. Nicolas Urbina-Cardona, Jelka Crnobrnja-Isailović, Jian-Huan Yang, Jiang Jianping, Jigme Tshelthrim Wangyal, Jodi J. L. Rowley, John Measey, Karthikeyan Vasudevan, Kin Onn Chan, Kotambylu Vasudeva Gururaja, Kristiina Ovaska, Lauren C. Warr, Luis Canseco-Márquez, Luís Felipe Toledo, Luis M. Díaz, M. Monirul H. Khan, Madhava Meegaskumbura, Manuel E. Acevedo, Marcelo Felgueiras Napoli, Marcos A. Ponce, Marcos Vaira, Margarita Lampo, Mario H. Yánez-Muñoz, Mark D. Scherz, Mark-Oliver Rödel, Masafumi Matsui, Maxon Fildor, Mirza D. Kusrini, Mohammad Firoz Ahmed, Muhammad Rais, N'Goran G. Kouamé, Nieves García, Nono Legrand Gonwouo, Patricia A. Burrowes, Paul Y. Imbun, Philipp Wagner, Philippe J. R. Kok, Rafael L. Joglar, Renoir J. Auguste, Reuber Albuquerque Brandão, Roberto Ibáñez, Rudolf von May, S. Blair Hedges, S. D. Biju, S. R. Ganesh, Sally Wren, Sandeep Das, Sandra V. Flechas, Sara L. Ashpole, Silvia J. Robleto-Hernández, Simon P. Loader, Sixto J. Incháustegui, Sonali Garg, Somphouthone Phimmachak, Stephen J. Richards, Tahar Slimani, Tamara Osborne-Naikatini, Tatianne P. F. Abreu-Jardim, Thais H. Condez, Thiago R. De Carvalho, Timothy P. Cutajar, Todd W. Pierson, Truong Q. Nguyen, Uğur Kaya, Zhiyong Yuan, Barney Long, Penny Langhammer, and Simon N. Stuart. Ongoing declines for the world's amphibians in the face of emerging threats. *Nature*, 622

(7982):308–314, October 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06578-4. URL http://dx.doi.org/10.1038/s41586-023-06578-4.

William J. McShea. What are the roles of species distribution models in conservation planning? *Environmental Conservation*, 41(2):93–96, January 2014. ISSN 1469-4387. doi: 10.1017/s0376892913000581. URL http://dx.doi.org/10.1017/S0376892913000581.

Cory Merow, Peter J. Galante, Jamie M. Kass, Matthew E. Aiello-Lammens, Cecina Babich Morrow, Beth E. Gerstner, Valentina Grisales Betancur, Alex C. Moore, Elkin A. Noguera-Urbano, Gonzalo E. Pinilla-Buitrago, Jorge Velásquez-Tibatá, Robert P. Anderson, and Mary E. Blair. Operationalizing expert knowledge in species' range estimates using diverse data types. *Frontiers of Biogeography*, 14(2), June 2022. ISSN 1948-6596. doi: 10.21425/f5fbg53589. URL http://dx.doi.org/10.21425/F5FBG53589.

Norman Myers, Russell A Mittermeier, Cristina G Mittermeier, Gustavo AB Da Fonseca, and Jennifer Kent. Biodiversity hotspots for conservation priorities. *Nature*, 403(6772):853–858, 2000.

Steven J. Phillips, Miroslav Dudík, and Robert E. Schapire. Maxent software for modeling species niches and distributions (version 3.4.1). Software. https://biodiversityinformatics.amnh.org/open_source/maxent/.

Steven J. Phillips, Robert P. Anderson, and Robert E. Schapire. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190(3):231–259, January 2006. ISSN 0304-3800. doi: 10.1016/j.ecolmodel.2005.03.026. URL https://www.sciencedirect.com/science/article/pii/S030438000500267X.

S. Pironon, I. Ondo, M. Diazgranados, R. Allkin, A. C. Baquero, R. Cámara-Leret, C. Canteiro, Z. Dennehy-Carr, R. Govaerts, S. Hargreaves, A. J. Hudson, R. Lemmens, W. Milliken, M. Nesbitt, K. Patmore, G. Schmelzer, R. M. Turner, T. R. van Andel, T. Ulian, A. Antonelli, and K. J. Willis. The global distribution of plants used by humans. *Science*, 383(6680):293–297, January 2024. ISSN 1095-9203. doi: 10.1126/science.adg8028. URL http://dx.doi.org/10.1126/science.adg8028.

Pierre Ploton, Frédéric Mortier, Maxime Réjou-Méchain, Nicolas Barbier, Nicolas Picard, Vivien Rossi, Carsten Dormann, Guillaume Cornu, Gaëlle Viennois, Nicolas Bayol, and et al. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications*, 11(1), 2020. doi: 10.1038/s41467-020-18321-y.

Tal Ridnik, Hussam Lawen, Asaf Noy, Emanuel Ben Baruch, Gilad Sharir, and Itamar Friedman. Tresnet: High performance gpu-dedicated architecture. In *proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1400–1409, 2021.

Roozbeh Valavi, Jane Elith, José J. Lahoz-Monfort, and Gurutzeta Guillera-Arroita. Modelling species presence-only data with random forests. *Ecography*, 44(12):1731–1742, October 2021. ISSN 1600-0587. doi: 10.1111/ecog.05615. URL http://dx.doi.org/10.1111/ecog.05615.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, June 2018. doi: 10.1109/cvpr.2018.00914. URL http://dx.doi.org/10.1109/CVPR.2018.00914.

Mary Susanne Wisz, Julien Pottier, W Daniel Kissling, Loïc Pellissier, Jonathan Lenoir, Christian F Damgaard, Carsten F Dormann, Mads C Forchhammer, John-Arvid Grytnes, Antoine Guisan, et al. The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biological reviews*, 88(1):15–30, 2013.

WorldClim. Bioclimatic variables. https://www.worldclim.org/data/bioclim.html.