THESIS TO OBTAIN THE DEGREE OF DOCTOR FROM THE UNIVERSITY OF MONTPELLIER

In Computer Science

Doctoral School : Information, Structures and Systems

Research units : ZENITH Inria, LIRMM, UMR AMAP, Montpellier

A predictive approach to determining the joint conservation status of species

Presented by Joaquim ESTOPINAN
On 28 November 2023

Under the direction of Alexis JOLY and François MUNOZ

In front of a jury composed of

Catherine H. GRAHAM, Professor, Swiss Federal Research Institute WSL	Reporter
Daniele SILVESTRO, Assistant Professor, University of Fribourg	Reporter
David MOUILLOT, Professor HDR, University of Montpellier	President
Jesper ERENSKJOLD MOESLUND, Senior Researcher, Aarhus University	Examiner
Alexis JOLY, Research Director, Université of Montpellier	Director
François MUNOZ, Professor HDR, Grenoble Alpes University	Director
Pierre BONNET, Research Fellow, CIRAD	Supervisor (guest)
Maximilien SERVAJEAN, Associate Professor, Paul Valéry University Montpellier 3	Supervisor (guest)

UNIVERSITÉ DE MONTPELLIER

Inria

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Informatique

École doctorale : Information, Structures, Systèmes

Unités de recherche : ZENITH Inria, LIRMM, UMR AMAP, Montpellier

Une approche prédictive de la détermination du statut de conservation conjoint des espèces

Présentée par Joaquim ESTOPINAN
Le 28 Novembre 2023

Sous la direction de Alexis JOLY
et François MUNOZ

Devant le jury composé de

Catherine H. GRAHAM, Professeur, Institut fédéral de recherches WSL	Rapportrice
Daniele SILVESTRO, Professeur assistant, University de Fribourg	Rapporteur
David MOUILLOT, Professeur HDR, Université de Montpellier	Président
Jesper ERENSKJOLD MOESLUND, Chercheur senior, Université d'Aarhus	Examinateur
Alexis JOLY, Directeur de recherche, Université de Montpellier	Directeur
François MUNOZ, Professeur HDR, Université Grenoble Alpes	Directeur
Pierre BONNET, Chargé de recherche, CIRAD	Encadrant (invité)
Maximilien SERVAJEAN, Maître de conférences, Université Paul-Valéry Montpellier 3	Encadrant (invité)

UNIVERSITÉ DE MONTPELLIER

Inria

Acknowledgements

Je vais essayer de m'en tenir à une page bien que ça risque d'être compliqué. Ces trois années de doctorat à Montpellier ont été intenses, mais menées dans un cadre privilégié tant sur le plan professionel que personnel. Commencer une thèse un 1er novembre 2020 de confinement peut paraître loin d'être idéal sur le papier. Pourtant, dès le début de cette aventure, j'ai eu la chance d'être extrêment bien encadré par mes directeurs Alexis Joly et François Munoz, ainsi que mes encadrants Maximilien Servajean et Pierre Bonnet. Je vous remercie sincèrement pour votre soutien continu. Avoir un créneau hebdomadaire pour discuter de nos avancées peut sembler anodin, mais s'est en fait avéré précieux et motivant pour garder le bon cap. Plus particulièrement, merci Alexis pour ta confiance, ta curiosité et perspicacité scientifique qui m'ont aiguillé tout au long de mes recherches. Merci François pour tes analyses qui ont toujours su apporter de la valeur à nos travaux. Tes venues à Montpellier ont été riches en apprentissages. Merci Maximilien pour ta réactivité supersonique, ton soutien scientifique mais pas que: finalement, devant n'importe quel point bloquant, tu as toujours su me rédiriger vers les ressources adaptées. Enfin, merci Pierre pour ton esprit critique, tes interprétations scientifiques et ton attention à tous les niveaux. Ta générosité pour s'assurer que les conditions et directions de travail soient idéales est inégalée.

Vient maintenant le moment de remercier toutes les personnes avec qui nous avons collaborer pour mener à bien cette thèse. Pour commencer, merci à Cathy Desseaux, Nathalie Brillouet et Nathalie Hodebert pour votre soutien administratif. Vous m'avez grandement facilité la vie de doctorant, notamment lors de mobilités, milles mercis. Merci à Alexander Zizka pour nous avoir donné accès à ses données d'occurrences d'orchidées. Merci à toi Laura Pollock pour m'avoir chaleureusement accueilli dans ton équipe à Montréal. Merci pour ton regard avisé sur nos travaux. J'ai ainsi eu l'opportunité de continuer mes recherches deux mois en immersion dans une culture nouvelle et j'en suis extrêmement reconnaissant. Je tiens également à remercier Christophe Fiorio, Ana Rodrigues et Dino Ienco pour votre écoute et vos suggestions lors des comités de suivi annuels. Merci à Patrick Valduriez, Ghislain Vieilledent, Grégoire Vincent et Joseph Salmon pour vos conseils. Merci à l'Inria pour le financement de cette thèse ainsi qu'aux bailleurs de projets européens GUARDEN et MAMBO. Enfin, merci aux membres de mon jury de thèse Catherine Graham, Daniele Silvestro, David Mouillot et Jesper Erenskjold Moeslund. C'est grâce à votre disponibilité que je pourrai défendre ma thèse dans un cadre scientifique excellent.

Bon, ça va dépasser une page c'est sûr, mais pas si grave. Évidemment merci à mes collègues rapidement devenus amis de l'Inria. Merci pour votre aide, pour nos pauses caféinées et pour nos sorties sportives en tout genre. Titouan, Benjamin D., Camille, Maximilien, César, Pierre L., Shamprikta, Théo, Diego, Benjamin B., Christophe, Antoine L., Raphael, Aïmi: on aura bien rigolé, et j'ai bien peur qu'on ait d'autres occasions à l'avenir. Merci à l'équipe PlantNet bien sûr! Matthias, Hugo, Antoine A., Hervé, Jean-Christophe, Maxime, Rémi, Vincent, Thomas: ça a été un plaisir de partager les couloirs avec vous. A mi-chemin j'ai rejoins l'UMR AMAP, merci encore Pierre pour cette opportunité. J'ai alors eu la chance de croiser la route de nombreux nouveaux camarades de boisson amère: Houssem, Laetitia, Paul, Dim & Dom, Camille S. et Camille

GT., Aurélien, Bruno, Claudia, Colin, Fred, Vanessa, François, Lily, Pablo, et j'en oublie. Merci à vous tous. Continuer ma thèse aussi à vos côtés m'a donné un nouvel élan. Merci pour votre bienveillance. Amaury, Julie, Elie, Fanny, Mathilde, Justine, j'espère avoir été un bon coloc avenue Henri Marès, moi j'ai adoré! Juan, Margot, Lorenzo, Santiago, Francisco, merci pour tout les amis. Une pensée sincère à mes amis de longue date.
Pour finir, merci infiniment à mes parents, ma soeur, et toi Belem, évidemment. Dur de mettre sur papier ma gratitude pour votre soutien inconditionnel. Cette thèse c'est aussi la vôtre. Naturellement, je vous la dédie.

Résumé

Les modèles de distribution d'espèces (SDMs) ont pour but d'apprendre les préférences environnementales des espèces et de projeter leur distribution géographique. Les récentes percées dans le domaine de l'apprentissage profond, associées à l'explosion des données sur la biodiversité, ont conduit au développement d'une nouvelle génération de modèles appelés deep-SDMs. Dans cette thèse, nous explorons leur application pour la conservation de la biodiversité.

Tout d'abord, nous évaluons la contribution des séries temporelles d'images satellites en tant que covariables environnementales. La capture de la phénologie des habitats d'espèces s'est avérée précieuse, en particulier pour les espèces rares et dans les régions riches en espèces. Deuxièmement, nous entraînons un deep-SDM pour déduire les assemblages mondiaux d'espèces d'orchidées à l'échelle du kilomètre. Des indicateurs spatiaux de leur risque d'extinction sont ensuite cartographiés à l'aide de la liste rouge des espèces menacées de l'UICN (Union Internationale pour la Conservation de la Nature). En mettant en évidence des motifs spatiaux du risque d'extinction de taxons sous-évalués, ces indicateurs multi-échelles et basés sur de grands volumes de données peuvent informer la planification de la conservation. Troisièmement, nous utilisons des représentations d'espèces issues d'un deep-SDM pour prédire avec succès les statuts UICN des espèces, tout en permettant de projeter l'étude dans des conditions bioclimatiques futures. En effet, l'évaluation automatisée du risque d'extinction est un domaine de recherche actif pour compléter les évaluations manuelles. Notre méthode de classification bénéficie du pouvoir de généralisation des deep-SDMs. Il vise à atténuer la dépendance à l'information géographique dans les évaluations du risque d'extinction de la flore et ainsi pouvoir prédire l'évolution future de ces risques.

La modélisation de la distribution des espèces est une tâche extrêmement difficile en raison de leurs dépendances biotiques et abiotiques complexes. Les modèles d'apprentissage profond peuvent s'appuyer sur les informations clefs qui sont en corrélation avec les espèces observés, et ce même lorsque des covariables environnementales riches et de très grande dimension leur sont fournies. En outre, les données sur la biodiversité sont entravées par des biais (taxonomique, géographique, temporel, etc.) que les techniques d'apprentissage automatique peuvent aider à compenser. En conclusion, nous avons étudié trois directions dans lesquelles les deep-SDM peuvent contribuer à la production de supports d'aide à la décision pour la conservation. 1) Tirer parti des données satellitaires de grande dimension pour modéliser les distributions d'espèces, 2) utiliser le pouvoir de généralisation et l'inférence des deep-SDMs pour cartographier mondialement le risque d'extinction d'assemblages d'espèces, et enfin 3) encoder des variables de grande dimension et leurs interactions reflétant les préférences environnementales des espèces pour des tâches connexes telles que la classification des statuts UICN.

Mots-clés – Modélisation de la distribution des espèces, science des données, apprentissage profond, réseaux de neurones convolutifs, réduction de dimension, télédétection, sciences de la conservation, assemblages d'espèces, risque d'extinction, orchidées

Abstract

Species distribution models (SDMs) aim to learn the environmental preferences of species and ultimately project their geographic distributions. Recent breakthroughs in deep learning, coupled with the explosion of biodiversity data, have led to the development of a new generation of models, called deep-SDMs. In this thesis, we explore their application in biodiversity conservation.

First, we evaluate the contribution of satellite image time-series as environmental covariates. Capturing the phenology of species' habitats was found to be valuable, especially for rare species and in species-rich regions. Second, we train a deep-SDM to infer global orchid species assemblages at the kilometre scale. Spatial indicators of their extinction risk are then mapped using the IUCN Red List of threatened species. By highlighting spatial patterns of extinction risk for under-assessed taxa, such scalable and data-intensive indicators can inform conservation planning. Third, we use SDM-based species features to successfully predict the IUCN extinction risk status of species while being flexible enough to project the study into future bioclimatic conditions. Indeed, automated extinction risk assessment is an active research avenue to complement manual assessments. Our classification scheme benefits from the generalisation power of deep-SDMs. It aims to mitigate the over-reliance on geographic information in flora extinction risk assessments, thus allowing prediction of future extinction risk patterns.

Modelling species distributions is an incredibly difficult task due to complex biotic and abiotic dependencies. Deep learning models can rely on the critical information that correlates with observed species patterns when provided with rich, high-dimensional environmental covariates. Furthermore, biodiversity data are hampered by biases (taxonomic, geographic, temporal, etc.) that machine learning techniques can help to compensate for. Ultimately, in our work we have investigated three directions in which deep-SDMs can contribute to the production of decision support for conservation: 1) taking advantage of high-dimensional satellite data to model species distributions, 2) using their generalisation and inference power to map the extinction risk of global species assemblages, and finally 3) encoding high-dimensional covariate information for downstream tasks such as flexible IUCN status classification.

Keywords – Species distribution modelling, data science, deep learning, convolutional neural networks, dimensionality reduction, remote sensing, conservation science, species assemblages, extinction risk, orchids

Publications

This thesis is based on three articles. These are listed below as primary publications. One of these articles (Chapter 3) has been published in the research topic *Plant Biodiversity Science in the Era of Artificial Intelligence* of *Frontiers in Plant Science*. A second one (Chapter 4) is under review in *Ecological Informatics*. The last article (Chapter 5) is being finalised for the special issue *Predictive Biogeography* of *Ecography*, where we have been invited to submit a full manuscript.

The articles listed in the secondary publications are related to, but not included in, this thesis. The first will be part of the book *On the edge of sixth extinction in biodiversity hotspots : Facts, needs, solutions and opportunities in Thailand and adjacent countries*. The second, the GeoLifeCLEF 2023 Dataset paper, is under review in *Ecology*.

Primary publications

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. (Estopinan et al., 2022)
Author contribution: data collection, modelling, analysis, writing

Estopinan, J., Servajean, M., Bonnet, P., Joly, A., Munoz, F. (2023). AI-based mapping of the conservation status of orchid assemblages at global scale. Under review in *Ecological Informatics*.
Author contribution: data collection, modelling, analysis, writing

Estopinan, J., Servajean, M., Bonnet, P., Munoz, F., Joly, A. (2023). Exploiting deep-SDMs to predict plant extinction risk and test climate change influence. Under review in *Ecography*.
Author contribution: data collection, modelling, analysis, writing

Secondary publications

Munoz, F., **Estopinan, J.**, Bose, R., Pélissier, R., Vieilledent, G. (2021). Future impacts of climate change and deforestation on endemic trees of Western Ghats, South India. In preparation to be published in a book. (Munoz et al., 2021)
Author contribution: modelling, and results description

Botella, C., Deneu, B., Marcos, D., Servajean, M., **Estopinan, J.**, Larcher, T., Leblanc, C., Bonnet, P., Joly, A. (2023). The GeoLifeCLEF 2023 Dataset to evaluate plant species distribution models at high spatial resolution across Europe. Submitted to *Ecology*. (Botella et al., 2023)
Author contribution: Assisted in parts of the data collection, pre-processing and description

Oral communications

I have had the opportunity to present my work at various occasions, including a national and international conference:

- **Conference of the French statistical ecology research group**, 4-5 April 2022, Montpellier, France
- **ESA & CSEE joint meeting**, 14-19 August 2022, Montreal, Canada

Software

Our work resulted in several online productions:

- A Zenodo repository of the *DeepOrchidSeries dataset* described in section 3.2.1 is available at https://zenodo.org/.
- The *sen2patch* GitLab repository gathers Python code to collect Sentinel-2 image time-series from a list of geo-localised occurrences. Available at https://gitlab.inria.fr/.
- The https://mapviewer.plantnet.org/ platform provides access to the AI-based global maps of the IUCN status of orchid species assemblages produced in chapter 4.
- The GeoLifeCLEF 2023 dataset is available via the kaggle challenge page at https://www.kaggle.com/ and the GitHub https://github.com/ provides useful codes to facilitate data manipulation.

Table of contents

1 Introduction **1**
- 1.1 Context . . . 2
- 1.2 Thesis motivation and contributions . . . 5
  - 1.2.1 Central research questions and organisation . . . 5
  - 1.2.2 Modelling species distribution from satellite time-series - Ch. 3 . . . 6
  - 1.2.3 Mapping the extinction risk of species assemblages - Ch. 4 . . . 8
  - 1.2.4 Extinction risk prediction under climate change - Ch. 5 . . . 9
- 1.3 A suited case study family: *Orchidaceae* . . . 11

2 State of the art **13**
- 2.1 Introduction . . . 15
- 2.2 Indicators for biodiversity conservation . . . 15
  - 2.2.1 Indicators at species level . . . 16
    - 2.2.1.1 The IUCN Red List of Threatened Species . . . 16
    - 2.2.1.2 Ecological and evolutionary originality . . . 20
    - 2.2.1.3 International agreements . . . 21
    - 2.2.1.4 Species population level: The Living Planet Index . . . 21
  - 2.2.2 Community and habitat indicators . . . 24
    - 2.2.2.1 Biodiversity: components and measures . . . 24
    - 2.2.2.2 Habitat classifications . . . 25
    - 2.2.2.3 Essential biodiversity variables . . . 27
  - 2.2.3 Spatial indicators of threat . . . 28
    - 2.2.3.1 Threatened species patterns . . . 28
    - 2.2.3.2 Mapping sources and mechanisms of threat . . . 31
    - 2.2.3.3 Threat integration in conservation planning . . . 32
- 2.3 Modelling species distribution . . . 36
  - 2.3.1 Species distributions . . . 36
    - 2.3.1.1 Mapping species geographic range . . . 36
    - 2.3.1.2 Species observation data . . . 37
    - 2.3.1.3 Models . . . 40
    - 2.3.1.4 Covariates relevance . . . 44
    - 2.3.1.5 SDMs to inform conservation . . . 47
  - 2.3.2 Deep-SDMs . . . 48
    - 2.3.2.1 Motivations . . . 48
    - 2.3.2.2 Convolutional neural networks . . . 54
    - 2.3.2.3 Resources and inference . . . 56
- 2.4 Insights for conservation planning . . . 59

2.4.1 AI for ecology and conservation . . . 59
2.4.2 Predicting the missing conservation status of species . . . 62
2.4.3 Future trajectories of conservation indicators . . . 64
2.4.4 Optimising spatial conservation . . . 66

3 Deep Species Distribution Modeling From Sentinel-2 Image Time-Series: A Global Scale Analysis on the Orchid Family 69
3.1 Introduction . . . 71
3.1.1 Context . . . 71
3.1.2 Contributions . . . 73
3.2 Materials and Methods . . . 74
3.2.1 *DeepOrchidSeries* dataset . . . 74
3.2.1.1 Raw input data description . . . 74
3.2.1.2 Dataset construction . . . 76
3.2.2 Species Distribution Models trained with satellite image time-series 80
3.2.2.1 Model definition and training procedure . . . 80
3.2.2.2 Performance evaluation of the model . . . 81
3.2.2.3 Interpretability experiments: quantifying the contribution of temporal information . . . 84
3.2.2.4 Modality contribution on a global scale . . . 86
3.3 Results . . . 87
3.3.1 Model validation and performance . . . 87
3.3.2 Results by number of species occurrences . . . 88
3.3.3 Results by region and regional diversity index . . . 89
3.3.4 Statistical tests . . . 91
3.3.5 Model evaluation regarding time and spatial data mismatches . . . 92
3.3.6 Modality contribution on a global scale . . . 93
3.4 Discussion . . . 93
3.5 Conclusion . . . 98

4 AI-based mapping of the conservation status of orchid assemblages at global scale 101
4.1 Introduction . . . 103
4.2 Materials and Methods . . . 105
4.2.1 Taxonomic focus: the *Orchidaceae* family . . . 105
4.2.2 Species assemblage prediction model . . . 106
4.2.2.1 Definition . . . 106
4.2.3 Validation . . . 107
4.2.4 Conservation indices for species assemblages . . . 108
4.2.4.1 Indices definition . . . 108
4.2.4.2 Missing status completion . . . 109
4.2.5 High-resolution maps construction . . . 110
4.2.5.1 Global grid design . . . 110
4.2.5.2 Maps definition and construction . . . 110
4.2.6 Zonal statistics . . . 110
4.2.6.1 Region spatial coverage of the most critical IUCN status 110
4.2.6.2 Region average proportions . . . 111

4.2.7 Data . . . 111
4.2.7.1 Orchid occurrences . . . 111
4.2.7.2 Predictive features . . . 111
4.3 Results . . . 113
4.3.1 $\mathcal{I}_{\mathcal{O}}$ indicator: most critical status of the species in the assemblage 113
4.3.1.1 Global patterns . . . 113
4.3.1.2 Country-level analysis . . . 113
4.3.2 $\mathcal{I}_c$ indicator: proportion of species in the assemblage with a given status . . . 114
4.3.2.1 Global patterns . . . 114
4.3.2.2 Country-level analysis . . . 116
4.4 Discussion . . . 119
4.4.1 Modelling choices . . . 119
4.4.2 Considerations on covariates . . . 120
4.4.3 Our indicators originality . . . 120
4.4.4 Orchids conservation . . . 122
4.4.5 Conclusions . . . 122

5 Exploiting deep-SDMs to predict plant extinction risk and test climate change influence 125
5.1 Introduction . . . 126
5.2 Materials and methods . . . 129
5.2.1 Method motivation . . . 129
5.2.2 Data on the *Orchidaceae* family . . . 130
5.2.2.1 Species observations . . . 130
5.2.2.2 IUCN Red List status . . . 130
5.2.2.3 Predictive features . . . 130
5.2.3 Method definition . . . 131
5.2.3.1 Deep species distribution modelling . . . 131
5.2.3.2 Dispersal scenarios . . . 132
5.2.3.3 Species niche features for IUCN classifcation . . . 132
5.2.3.4 Classifying extinction risk status from species niche features 134
5.2.3.5 Projections within the CMIP6 SSP5-8.5 scenario . . . 135
5.2.4 Model validation . . . 136
5.3 Results . . . 136
5.3.1 Continents . . . 137
5.3.2 Latitude . . . 138
5.3.3 Altitude . . . 138
5.4 Discussion . . . 139
5.4.1 Interpretations . . . 139
5.4.2 Limitations . . . 140
5.4.3 Perspectives . . . 141
5.4.4 Conclusion . . . 142

6 Conclusion and perspectives 143
6.1 Results synthesis . . . 144
6.2 Limits and perspectives . . . 145

6.2.1 Contribution of our work to conservation . . . 145
6.2.2 Uncertainty and confidence in predictions for conservation . . . 147
6.2.3 Research directions . . . 148
6.2.3.1 Species distribution modelling with deep learning . . . 148
6.2.3.2 IUCN status prediction . . . 149
6.2.4 IA, Ecology & Conservation . . . 150
6.3 Conclusion . . . 151

Bibliography **153**

Supplementary information **189**

Acronyms

AI Artificial Intelligence. 58, 120
AOH Area of Hability. 19, 29
AOO Area of Occupancy. 17, 47, 128, 129, 147
ASTER Advanced Spaceborne Thermal Emission and Reflection Radiometer. 87

BOA Bottom-of-Atmosphere. 75, 76
BRT Boosted Regression Tree. 43

CAPTAIN Conservation Area Prioritisation Through Artificial INtelligence. 66
CBD Convention on Biological Diversity. 15, 26
CITES Convention on International Trade in Endangered Species of Wild Fauna and Flora. 21
CMIP6 Coupled Model Intercomparison Project Phase 6. 128, 131
CNN Convolutional Neural Network. 53, 71, 73, 80, 129, 148
CR Critically endangered. 17, 29, 108, 127
CS Citizen Science. 37, 45, 57, 61

DD Data Deficient. 17
deep-SDM Species Distribution Model based on a deep learning architecture. 4, 36, 71, 105, 120, 126
DL Deep Learning. 4, 104, 129
DRW Deferred Re-Weighting. 81, 88

EBV Essential Biodiversity Variable. 27, 146
ED Evolutionarily Distinctiveness. 20, 30
EDGE Evolutionarily Distinct and Globally Endangered. 20, 21, 30, 199
EEA European Environment Agency. 26
EN Endangered. 17, 108, 127
EOO Extent of Occurrence. 17, 36, 47, 128, 129, 147
ESA European Space Agency. 77
EUNIS European Nature Information System. 25, 26
EVA European Vegetation Archive. 26
EVI Enhanced Vegetation Index. 93
EW Extinct in the Wild. 17
EX Extinct. 17, 20, 127

FC Fully Connected. 54, 131
FUSE Functionally Unique, Specialized, and Endangered. 21

GAM Generalised Additive Model. 43, 53
GBIF Global Biodiversity Information Facility. 36, 37, 74, 105, 111, 129, 150
GDAL Geospatial Data Abstraction Library. 135
GEO BON Group on Earth Observations Biodiversity Observation Network. 27
GEP Global Extinction Probability. 29

GLM Generalised Linear Model. 43, 62, 63
GPU Graphical Processing Unit. 50, 56, 57, 81, 96, 131, 192

IPBES Intergov. Science-Policy Platform on Biodiversity and Ecosystem Services. 2, 15, 28
IUCN International Union for Conservation of Nature. 8, 16, 74, 103, 126, 127

JSDM Joint SDM. 42, 43

LC Least Concern. 17, 20, 29, 108, 127, 139
LDAM Label-Distribution-Aware Margin. 81, 88, 131, 148, 191, 194
LPI Living Planet Index. 21

MaxEnt Maximum Entropy. 43
ML Machine Learning. 4, 43

NDVI Normalised Difference Vegetation Index. 45, 93
NE Not Evaluated. 16
NN Neural Network. 49, 63
NT Near Threatened. 17, 108, 127

PA Protected Area. 33, 48, 66, 105, 119, 122
PINN Physics-Informed Neural Network. 150

RCP Representative Concentration Pathway. 10, 127, 128
RF Random Forest. 29, 52, 62, 63, 127
RL IUCN Red List of Threatened Species. 16, 20, 37, 127
RLE Red List of Ecosystems. 20, 26, 199
RLI Red List Index. 20
RNN Recurrent Neural Network. 98
RS Remote Sensing. 3, 45

S2 Sentinel-2. 6, 73, 76, 79, 96
SAR Synthetic Aperture Radar. 62
SDG Sustainable Development Goal. 15
SDM Species Distribution Model. 4, 36, 71, 104, 126, 127
SEM Spatially Explicit Model. 105
SGD Stochastic Gradient Descent. 50, 51, 55, 81
SHAP SHapley Additive exPlanations. 52, 147
SSDM Stacked SDM. 42
SSP Shared Socioeconomic Pathway. 128, 131, 135
STAR Species Threat Abatement and Restoration. 29
SVM Support Vector Machine. 43

TOA Top-of-Atmosphere. 75, 76

VU Vulnerable. 17, 108, 127

WCSP World Checklist of Vascular Plants. 129, 199
WGSRPD World Geographical Scheme for Recording Plant Distributions. 82, 83, 107, 110, 114, 117, 135, 191, 192, 202

xAI explainable AI. 44, 52, 147

List of Figures

1.1 Big data dimensions in ecology 5

2.1 Generalised Red List assessment workflow 18
2.2 Two examples of EOO and AOO measures 19
2.3 Two contrasting views of the Living Planet Index 23
2.4 From observations to the production of EBVs and indicators 27
2.5 Protection-weighted range-size rarity of imperiled species 30
2.6 Patterns of *Cactaceae* biodiversity 31
2.7 Distribution of the predominant threats to biodiversity across Australia . 32
2.8 Example of map combination for conservation planning 33
2.9 Alexander von Humboldt's *Tableau Physique* 35
2.10 GBIF cumulative number of plant records over time 39
2.11 Timeline of the three eras of statistical learning 49
2.12 Bias-variance trade-off in machine learning 51
2.13 Trade-off between accuracy and interpretability 52
2.14 Convolution example 55
2.15 Promises, pitfalls and priorities for deep learning in conservation 59
2.16 Evolution of article citations with 'prediction' and 'ecology' as keywords . 60

3.1 Visual abstract of the method 72
3.2 Orchid observation distributions 75
3.3 Creation workflow of the *DeepOrchidSeries* dataset 77
3.4 Number of occurrences per tile and patch size 78
3.5 Image time-series associated to three orchid observations 79
3.6 Cloud cover percentages of the Sentinel-2 tested products 80
3.7 Scheme of the temporal information contribution experiment 84
3.8 Top-30 accuracy for model validation and test sets 87
3.9 Performance analysis in function of species number of training observations 89
3.10 Graphs and global maps illustrating the performance per region 90
3.11 Performance analysis in function of observation year and location uncertainty 92
3.12 Global comparison of SDM modality contribution 93

4.1 Global maps of the most critical IUCN status with three methods 112
4.2 Four indicators based on orchid assemblage predictions 115
4.3 Country proportion of threatened species *versus* Shannon index 117
4.4 Five indicators of orchid assemblage extinction risk in Sumatra 118
4.5 Screenshot example of the website tools 121

5.1 Scheme of our extinction risk classifier fed with flexible SDM-based features 133

5.2 Dataset IUCN status distribution and classification performance 136
5.3 Status proportions per continent and time period 137
5.4 Status predictions in function of species average latitude 138
5.5 Status predictions in function of species average elevation 139

S1 Histogram of orchid observation geolocation uncertainty 190
S2 Performance analysis in function of the number of observations per region 190
S6 Status distributions of IUCN-assessed orchids and *IUCNN* predictions . . 196
S7 Correlation between orchid diversity and threat levels 198
S9 2D input data associated to three observations 206
S10 Confusion matrices of the extinction risk classifier 209
S11 Global increase of species predicted to be threatened over time 210
S12 Status proportions per continent and time period only for IUCN-assessed species 210
S13 Status predictions in function of species average latitude only for IUCN-assessed species 211
S14 Status predictions in function of species average elevation only for IUCN-assessed species 212
S15 Cumulative histogram of threatened species per average latitude over time 213
S16 Histograms of the number of species per 2009 human footprint bin 214
S17 Predicted status proportions per human footprint bin across time periods 215

List of Tables

2.1 Sources of species observation data 38

3.1 Number of occurrences and species in training, validation and test sets . 82

4.1 Top-15 countries with the largest area share of CR or EN as most critical IUCN status 114

4.2 Top-15 average status proportions per country: threatened, CR and EN . 116

S2 List of covariates used as deep-SDM input 207

CHAPTER 1

INTRODUCTION

Table of contents

1.1 Context 2

1.2 Thesis motivation and contributions 5

1.2.1 Central research questions and organisation 5

1.2.2 Modelling species distribution from satellite time-series - Ch. 3 6

1.2.3 Mapping the extinction risk of species assemblages - Ch. 4 . . . 8

1.2.4 Extinction risk prediction under climate change - Ch. 5 9

1.3 A suited case study family: *Orchidaceae* 11

1.1 Context

Changes and challenges. Climate change is one of the proven drivers of the ongoing sixth mass extinction (Barnosky et al., 2011; Pacifici et al., 2015). As a reminder, July 2023 was the hottest month ever recorded on Earth[1]. Accompanied by a series of extreme weather events (heatwaves, wildfires, floods) and record sea surface temperatures, the weather of July 2023 has had catastrophic impacts[2]. Against this background, it is fair to ask whether the manifestations of climate disruption are likely to provoke a strong political and social response.

Many anthropogenic drivers explain the increasing extinction risk of species (Sala et al., 2000). Indeed, extinction rate is estimated to be up to a thousand times higher than historical background levels (Pimm et al., 2014). If species threatened with extinction collapse under the pressures they face, the rate could rise to 10,000 times the background rate. Besides climate, land-use change (Foley et al., 2005), chemical and industrial pollution, invasive species (Gurevitch & Padilla, 2004) and international trade (Lenzen et al., 2012) are at the forefront. In addition to being taxon-specific, the causes are interrelated, difficult to measure in isolation and therefore difficult to understand.

Measuring the multiple facets of biodiversity (taxonomic, genomic, phylogenetic, functional, ecosystemic, etc.) and its vulnerability is a challenging task (Purvis & Hector, 2000). At the taxon level, the research community considers that some species may go extinct before they have even been described (Costello et al., 2013). Biodiversity scales are nested. Species loss is a major driver of ecosystem change, compromising the goods and services that directly and indirectly benefit humans (Hooper et al., 2012).

A plural response. In the face of this incredibly complex biodiversity crisis, the response has no choice but to be multifaceted and collaborative (Wheeler et al., 2012). To be successful, the collaboration should be i) societal to raise awareness and embrace change, ii) political to lead and fund change at scale, and iii) scientific to advance understanding and levers.

The Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES) delivered an impactful Global Assessment Report on Biodiversity and Ecosystem Services in 2019 (IPBES, 2019). This structure aims to improve the interface between science and policy on biodiversity and ecosystem services issues. While the quality and need for such synthesis work is manifest, uptake by policymakers still appears to be low (Garcia et al., 2022). The most recent example occurred in the European Parliament, where the highly anticipated Nature Restoration Law was voted in favour, but only in a heavily watered-down version[3]. Besides, science-based solutions do exist. Spatial ecology science is for instance producing new indicators on the health of biodiversity (see Section 2.2) that could help reporting progresses when integrated in international agreements like the post-2020 global biodiversity framework (Jetz et al., 2022).

The weak collective and political response to increasingly well-documented challenges can lead to frustration, eco-anxiety or eco-anger (Stanley et al., 2021). While this is an alarming mental health issue, especially among the youngest (Benoit et al., 2022),

[1] https://climate.copernicus.eu/
[2] https://www.nasa.gov/
[3] https://www.birdlife.org/

the positive management of these strong emotional drivers to engage in action is a key process to restore wellbeing and nourish transformative change (Díaz et al., 2019). Action is meant here in a large and personal sense. It includes discussing these concerns with one's surroundings, vulgarising and disseminating current knowledge, producing scientific research, or engaging in politics and activism.

The role of research. In this context, research clearly has a crucial role to play. Understanding and mapping biodiversity and its threats across scales, but also providing answers (restoration and protection, environmental justice, responsible farming, mitigation of extreme events). Interdisciplinary science and nexus approaches are key to addressing the multidimensional challenges of the 21st century (Allan et al., 2018; Haider et al., 2018; Liu et al., 2018; Zarnetske et al., 2019). For example, promoting integrative studies from masters and graduate level is a promising direction to train qualified professionals[4]. Adopting novel success metrics to recognise and promote interdisciplinary projects is another avenue to consider (Goring et al., 2014).
A major technical barrier to the study of biodiversity - and ecology more broadly - is scale. Studies carefully focused on a particular taxon and/or spatio-temporal context do indeed shape our knowledge of biodiversity, and will always be much needed. However, scaling up biodiversity assessment using the same methods is impossible in practice (time-consuming, expensive manual data collection and analysis using tailored methods). Current knowledge of biodiversity is biased i) geographically towards the poles, although species richness is concentrated in the tropics, ii) taxonomically towards well-documented, common or charismatic taxa (mammals, birds), iii) temporally towards recent years due to increasing resources and citizen data, and iv) functionally towards species with large spatial distributions, although most species have small ranges and are known from only a few localities (Collen et al., 2008; Pimm et al., 2014). A biased assessment of biodiversity would lead to unrepresentative - or at least equally biased - conservation, leaving facets of biodiversity vulnerable to pressures.

Data and biodiversity science. Advances in technology, data science and ecoinformatics are thankfully bringing new ways of collecting, pre-processing and analysing information (Farley et al., 2018; Hampton et al., 2013). Using the common *Four Vs* definition of big data, it appears clear that ecology has become a data-intensive science (Michener & Jones, 2012), see Figure 1.1. i) *Volume* can be easily illustrated by Remote Sensing (RS). For example, the two satellites of the Sentinel-2 imaging mission generate 1.6 Tbytes of raw imagery per day[5]. ii) *Variety* can be represented by the wide range of scientific projects across temporal, spatial and taxonomic scales. The multiple facets of biodiversity give a sense of how diverse ecological data can be. iii) *Veracity* includes the different levels of expertise in species identification, but also taxonomic and distributional inaccuracies, among others. iv) *Velocity* is probably the weakest dimension represented in ecology, but improvements are needed and promising, with two examples in iEcology (Jarić et al., 2020) and camera traps. However, the collection of vast amounts of data alone is not sufficient to address current conservation challenges (Pimm et al., 2015).

[4] https://idil.edu.umontpellier.fr/en/
[5] https://www.esa.int/

Effective use of this wealth requires the successful extraction of valuable information, which is likely to be hampered by multiple biases (sampling, detection, taxonomy). This has the potential to reduce current gaps in biodiversity knowledge (Hortal et al., 2015). In addition, efficient biodiversity conservation requires the integration of biodiversity modelling and conservation, two fields that are still largely independent (Guisan et al., 2013; Pollock et al., 2020).
Extracting critical information is a key challenge for scientific progress in the age of big data. Distilling rich information and capturing relevant features for a given optimisation problem is a machine learning task. There is a clear need to bring data science and biodiversity closer together. Machine Learning (ML) and Deep Learning (DL) are increasingly being used to solve applications in ecology (Christin et al., 2019), see Section 2.4 for a review in conservation. They provide the ability to handle the high-dimensional data that characterises biodiversity and to capture complex relationships between predictors and the desired response function (Maxwell et al., 2018). Ecology produces global and ever-increasing amounts of data at ever finer resolutions (spatial, temporal and spectral). Unlike traditional models, deep learning can make the most of this opportunity by analysing these data jointly and across scales (Zhu et al., 2017).
Exploiting machine learning models in frameworks that are flexible enough to deal with heterogeneous biodiversity data and uncertainties is a major challenge for computer scientists (Farley et al., 2018). With the increasing sharing of data in open repositories, there is an unprecedented opportunity to co-evolve data science and biodiversity knowledge (Soranno et al., 2015). Finally, in a context where misinformation about climate change is on the rise (Linden et al., 2017), the question of how information is presented takes on a particular importance. We will therefore pay attention to the appropriate representation of our results (maps, figures, scenarios, online tools, etc.).

Modelling species distributions and remote sensing. A common modelling framework has been used across our studies: Species Distribution Models (SDM), and more specifically models based on deep learning architectures, called deep Species Distribution Models (deep-SDM), Botella et al. (2018a). SDMs are statistical tools designed to shape the distribution of species from the association of species observations with environmental descriptors under modelling and ecological assumptions. The properties and challenges posed by this class of models are detailed in the state of the art section 2.3. The use of SDMs to guide conservation decisions has already led to successful actions. However, more dialogue between modellers and decision-makers is needed in the design of SDMs to make them fit for purpose (Guisan et al., 2013).
Remote sensing has revolutionised the monitoring of life on Earth, revealing the spatial and temporal aspects of biodiversity through assessments of ecosystem structure, composition and function (Cavender-Bares et al., 2022). To understand and safeguard biodiversity, it is essential to embrace an integrative approach to biodiversity science (König et al., 2019). This involves extending our understanding of local-scale processes to much larger scales through the use of remote sensing. For example, plant spectral diversity can be monitored at large scales through remote sensing and has been shown to integrate functional and phylogenetic components of biodiversity (Schweiger et al., 2018).
Remote sensing products provide globally consistent and continuous spatio-temporal data on factors influencing the distribution of organisms. This opens up the possibility of

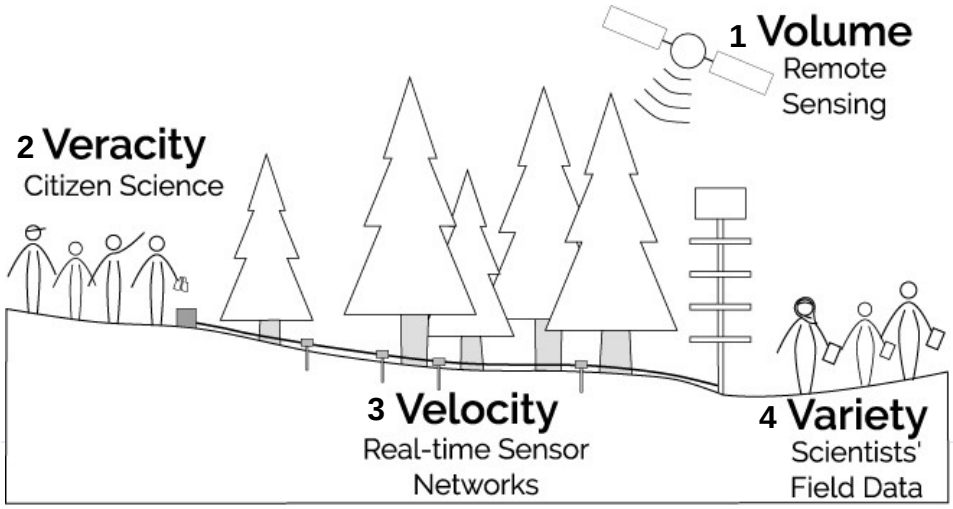


Figure 1.1: *Big data dimensions (four Vs) in ecology: **(1)** Remote sensors of the earth system, mounted on a variety of platforms, which generate large data volumes. **(2)** Citizen-science efforts that collect data at volumes far beyond the capacity of scientific experts, by individuals with varying degrees of expertise. **(3)** Near real-time sensor networks that can deliver on-going data feeds at low latency and high velocity. **(4)** Field observations and experiments by scientists, across a wide variety of measurements, systems, and scales. This mapping of the four Vs to data types is illustrative; all four dimensions are present in all data types. Reproduced from (Farley et al., 2018) with permission from Oxford University Press.*

improving the accuracy of SDMs for conservation purposes (Randin et al., 2020). Indeed, satellite missions deliver spatially-explicit environmental predictors that can help models accurately capture species' habitat characteristics and preferences (He et al., 2015). In addition, the increasing spatial resolution and temporal revisit frequency of satellites allows for the development of rapidly repeatable and scalable biodiversity indicators. SDMs based on deep learning can produce high resolution maps globally (1 km, see Chapter 4), or even very high resolution (50 m, Deneu et al., 2022) when provided with appropriate remote sensing resources. Achieving such spatial resolution would ultimately allow conservation decisions to be informed by consistent and scale-appropriate indicators.

It is in this challenging societal and scientific context, at the intersection of data and biodiversity science, that my PhD project takes place. After presenting the overall organisation of the manuscript, we will now focus more specifically on the motivations behind my doctoral research.

1.2 Thesis motivation and contributions

1.2.1 Central research questions and organisation

Our overarching research objective is to take advantage of recent advances in deep learning to model plant species distributions and provide new decision support for conservation practitioners. With species distribution models based on deep learning at the heart of our project, we can identify three main research questions that have guided our work:

1. Leveraging the temporal revisit of recent satellite imagery missions, does characterising species habitat in time improve species distribution modelling? If so, where are the greatest gains?
2. Can deep learning help map the extinction risk of species at both high resolution and global scale? If so, how are threatened orchids distributed across scales?
3. Can deep-SDMs produce species features that are informative of the extinction risk status, and how do these predictions respond to a business-as-usual climate change scenario?

Manuscript organisation
This thesis is structured in six chapters. Chapter 1 is the current introduction, setting out the general context of this doctoral thesis, and continues below with our specific motivations and contributions. Chapter 2 provides an overview of the state of the art research that underpins our work: indicators for biodiversity conservation, species distribution models (especially those using deep learning), and insights for conservation planning based on predictive approaches. Then chapters 3-5 are our three original research papers. Chapter 3 has been published in a peer-reviewed journal, chapter 4 is under review and chapter 5 is being finalised prior to submission. Finally, chapter 6 discusses our scientific findings and their limitations, concludes on our research contribution and suggests new directions.

1.2.2 Modelling species distribution from satellite time-series - Ch. 3

Describing species environment. In order to model accurately the distribution of species, it seems necessary to provide the best possible descriptors of their environment. While knowledge of the proximal features behind species' environmental preferences is still very limited, a legitimate approach is to describe the observed species' habitat as precisely as possible (see Section 2.3.1.4 for a review on SDM covariates). The Sentinel-2 (S2) mission now provides access to RGB+IR imagery with a spatial resolution of ten metres. This makes it one of the best sources of information, both free and available on a global scale, for describing species habitats. Such a data-intensive method is enabled by the ability of deep learning to generalise the potential niches of species without being over-parameterised and then over-reliant on the very high dimensional training data (Poggio et al., 2017).
Capturing habitat phenology. Modern remote sensing missions represent an unprecedented opportunity to collect images of species habitat at high spatial resolution, but also with high temporal revisit frequency. Providing not only the central value of environmental predictors as input to deep-SDMs, but also the two-dimensional spatial context around species observations has been proven valuable for shaping potential niches (Deneu et al., 2021b). What about the temporal dynamics of remotely sensed habitats? We anticipate that the temporal dimension - still largely under-exploited by species distribution modellers - would contribute significantly to habitat qualification. Better habitat characterisation allows models to better explore and shape the environmental

conditions that have allowed plant species to flourish. Ultimately, we expect that describing the phenology of habitats would lead to a thinner capture of species' environmental sensitivity.

Comparison and hypothesis. An analogous case is photogrammetry, where reliable information about physical objects is obtained from multiple images taken from different angles. In our case, we do not multiply the viewing angles, but rather the shots along a year. In both cases, multiplying the inputs along one of the dimensions of the measurement allows a richer description of the target. A first goal is then to comprehensively test this hypothesis i) with remote sensing data and ii) on a global scale and iii) on a particularly rich, diverse and sensitive taxon.

Objectives. The main objective of this first study is to test the global contribution of habitat phenology to the performance of deep-SDMs. This will require the collection and pre-processing of a remote sensing dataset with a spatial coverage and temporal revisit that, to our knowledge, has never been mobilised for modelling species distributions. We will also investigate for which species and where the addition of the temporal dimension is most beneficial for modelling distributions. On a global scale, an additional study will test the contributions of three SDM modalities: remote sensing imagery, bioclimatic variables and static variables (elevation, position, human footprint and ecoregions). The motivation behind this work is to assess, at a global scale and for a given study taxon, the trade-off between the processing cost of remote sensing and its predictive power compared to other readily available variables.

Contributions

1. A remote sensing dataset called *DeepOrchidSeries* is generated from a set of 1 million orchid observations for 14 thousand species distributed worldwide. Twelve-month image time-series from the Sentinel-2 mission are associated with each sighting. The four 10-metre resolution channels (RGB+IR) are sampled with an extent of 640 x 640 m centred on the observation.
2. An SDM based on a convolutional neural network architecture (Inception V3) is adapted to learn potential species niches from the satellite image time-series and with a tailored training procedure.
3. An ablation study tests the contribution of the temporal dimension of the input (capturing habitat phenology) to species distribution modelling.
4. As expected, the temporal dimension helps in modelling species distributions by allowing a better characterisation and differentiation of habitats. The macro-average performance with a time-series of satellite images as input is almost two times higher than with a random single view. Interestingly, hard predictions of rare species and in species-rich regions particularly benefit from the additional temporal context.
5. The contributions of three SDM modalities are compared on a global scale: bioclimatic variables, satellite image time-series and static variables (including altitude, position, human footprint and ecoregions).

1.2.3 Mapping the extinction risk of species assemblages - Ch. 4

Prioritise conservation. Species around the world are increasingly threatened by extinction due to human activities. Conservation science is a mission-driven discipline that aims to prevent species from going extinct. All facets of biodiversity have intrinsic and potential value, and in a utopian world the entire tree of life would be protected or at least monitored (Pollock et al., 2020). However, the costs of conservation and the socio-economic context leave no choice but to prioritise action on certain aspects of biodiversity (Juffe-Bignoli et al., 2016).

The dimensions of threat. Threat is a multidimensional concept that cannot be represented in a single way. The section 2.2.3 will focus on describing the literature on threat indicators. Existing threat maps mostly focus on pressures *sources* and *processes* with remotely sensed data, such as land use change, urbanisation or trade (Harfoot et al., 2021). However, i) threat sources and mechanisms do not necessarily translate into threatened *states* and ii) these threatened states or *stresses* of biodiversity, such as population decline or niche shrinkage, are rarely mapped (Balmford et al., 2009). It is precisely these symptoms, these causes of concern, that the International Union for Conservation of Nature (IUCN) criteria are intended to measure for species and, more recently, for ecosystems.

Species range knowledge. Mapping the concentration of threatened species therefore appears to be a useful input for decision making. According to (Di Marco et al., 2017), the alignment between global conservation science and the distribution of biodiversity remains poor, especially for threatened species. Species assessments continue to show a clear geographical bias. Furthermore, spatial data to support extinction risk assessment with known ranges are extremely scarce for the *plantae* kingdom. Of the 150,300 species assessed by the IUCN, 82% have spatial data. However, this is mostly the case for comprehensively assessed taxonomic groups such as vertebrates or amphibians. In fact, the proportion of IUCN assessments supported by spatial data drops to one sixth for plant species[6]. Spatial data may also be incomplete, as IUCN acknowledges. Generalisation of species distributions through modelling can help to overcome the scarcity and bias of IUCN plant range data. We believe that this is a key modelling step towards the design of indicators that are truly representative of biodiversity patterns.

Biodiversity indicators. There are three basic functions of indicators: simplification, quantification and communication (Pinborg, 2002). Typically, indicators simplify complex phenomena to make them measurable and enable information to be communicated. Spatial indicators are naturally based on geospatial data, and this additional output dimension enables spatial analysis. Biodiversity indicators play a crucial role in facilitating communication about the current status and trends of biodiversity and the underlying relationships driving these changes.

Our proposal. We believe that scalable spatial indicators of the extinction risk of plant assemblages are needed to inform conservation decisions. What we map is i) the spatial concentration of threatened species as defined by IUCN and ii) the most critical status of species that may be present at a given location. The development of these indicators implies to fill the current spatial and taxonomic gaps in extinction risk assessments.

Objectives. Questions addressed in our study relate to the spatial distribution of plant

[6] https://www.iucnredlist.org/, accessed on 09/08/23.

extinction risk. How can we complement IUCN information to fill assessment gaps and provide global estimates at high resolution? Can we identify local, regional or global patterns in the designed indicators and where are the most affected countries?
Another motivation for our work is to stimulate methodological research. Indeed, addressing such a global issue at high resolution requires technical expertise and the formulation of appropriate hypotheses that need to be carefully tested and challenged by the community. This is an illustration of the needed interdisciplinary context that we advocate in the introduction.
Summary. To conclude, we believe that there is a strong need to map the concentration of potentially threatened species in under-assessed taxonomic groups. This fills a taxonomic, spatial, but also a target gap in the representation of threats. Such indicators would ultimately benefit conservation decision making. This study will hopefully contribute to and motivate the development of methodological research by interdisciplinary teams. We are confident that challenging and complementing these results would ultimately benefit conservation science.

Contributions

1. A deep-SDM is calibrated to predict species assemblages worldwide at kilometre resolution. Based on globally available information (bioclimatic, pedological, human footprint, ecoregions and position), the model is guaranteed to return the true species within its predicted assemblage with 97% confidence (conformal prediction).
2. Spatial indicators are developed from these predictions and from the IUCN Red List completed by a reference predictive model. They are defined as the proportion of threatened species (and each IUCN extinction risk category), the most critical IUCN status and the Shannon index. The interactive maps are available online at https://mapviewer.plantnet.org/?config=apps/store/orchid-status.xml.
3. Summary spatial statistics are calculated at country level. The most threatened countries are compared with the most species-rich countries. Indicators are compared with current protected areas on the island of Sumatra.
4. The highest level of threat is observed in Madagascar and the surrounding islands. In Sumatra, we found a favourable alignment between protected areas and our indicators. However, when we enhance the existing IUCN assessments with predictions of species status, we find worrying levels of threatened species across the island and throughout the world.

1.2.4 Extinction risk prediction under climate change - Ch. 5

Coverage biases in the IUCN Red List. As of 2022, only 15% of the world's known plant species are assigned an IUCN status[7]. This is a consequence of the aforementioned taxonomic bias in Red List assessments, with some taxa being comprehensively

[7] https://www.iucnredlist.org/, accessed on 09/08/23.

assessed, such as mammals, birds or amphibians, and others barely assessed, such as invertebrates (2% assessed) or marine species (15% assessed). Overall, 7% of the world's described species have been assessed by IUCN. Despite important efforts to speed up manual assessments, the task is Herculean (e.g. 1,305,250 estimated number of described invertebrates). Limitations in Red List coverage remain numerous (taxonomic, but also spatial, temporal and thematic biases, Bachman et al., 2019).

Compensatory automatic methods. In addition, threatened but unassessed species are excluded from conservation efforts worldwide, but also from private sector protection and specific funding. In response to these massive concerns, research has developed compensatory automated assessment methods, see the state of the art section 2.4.2. The motivations are numerous, with the need to update the Red List and the cost of assessment also coming into play (Rondinini et al., 2014).

Climate change and extinction risk. Climate change is expected to threaten up to one in six species with extinction under current atmospheric emissions trajectories (Urban, 2015). Overall, red listing has been shown to be a relevant warning to identify and protect species threatened by climate change (Stanton et al., 2015). However, the consideration of climate change with appropriate hypotheses in red list assessment is still the fruit of active research and can massively influence assessment results (Moat et al., 2019), see section 2.4.3.

Objectives. In our third study, our motivation lies in the intersection of these two research areas. We investigate the influence of climate change projections on automated Red List assessments based on SDM features. Our classification scheme benefits from the generalisation power of deep-SDMs. It aims to mitigate the over-reliance on geographic information in flora extinction risk assessments, thus allowing prediction of future extinction risk patterns. We test if a deep-SDM used as a dimension reduction algorithm can provide species features predictive of IUCN status. Our use of SDM to support red listing is novel and conditioned by species dispersal scenarios. In particular, we are interested in projections of threatened species across continents, across latitudinal gradients and across altitudes.

Contributions

1. A trained deep-SDM is used as a dimension reduction algorithm to encode species features that are predictive of IUCN extinction risk status and benefit from the niche generalisation power of SDMs. The method validation demonstrates competitive performances.
2. The business-as-usual emission rate scenario (RCP 8.5) along with two extreme dispersal scenario (null and unlimited) are adopted to model climate change until 2100. The response in threatened species predictions is declined by continents, latitude and elevation.
3. The proportion of threatened species is globally increasing, with alarming levels in Africa, Asia and South America. It is also predicted to peak around both Tropics and the Equator, in the lowlands and in the 800-1,500 m altitudinal range.

1.3 A suited case study family: *Orchidaceae*

An hyperdiverse and threatened family. With c. 31,000 species[8] worldwide, orchids are one of the largest families of flowering plants. This hyperdiverse family is also one of the most threatened, in part due to their complex life history strategies (Cozzolino & Widmer, 2005; Fay, 2018). This, together with complex biotic interdependencies including with mycorrhizal fungi (McCormick et al., 2018) and fire exposure, is considered to make their distributions particularly sensitive to climate change. Other major threats they face include habitat conversion and harvesting (horticulture, illegal trade). This family is therefore experiencing unprecedented challenges, surpassing those of many other plant groups.

A flagship ecological indicator... The specificity of orchids provide them good indicator properties of the health of their ecosystem (Newman, 2009). In addition to their sensitivity to climate change, orchids have been shown to respond measurably to past and present environmental disturbances (Kull & Hutchings, 2006). The family is also involved in ecosystem functioning through their highly specific to generalist interactions with pollinators (sometimes resulting in high levels of co-evolution) and mycorrhizal associations (Swarts & Dixon, 2009). Finally, orchids are easy to monitor in the sense that once populations are established, they are easy to find every year. Orchids can therefore be considered as a suitable ecological indicator of the health of their ecosystem, as defined by (Jørgensen et al., 2016). Indeed, the family is i) easy to monitor, ii) sensitive to small-scale environmental changes with quantifiable and predictable responses, and iii) globally distributed. Moreover, the effectiveness of surrogate species for biodiversity planning has been tested and validated, at least for taxa within the same realm (Rodrigues & Brooks, 2007). Through the prism of our orchid-based indicators, we do not pretend to accurately represent ecosystem health. The orchid family has some characteristics that are associated with ecosystem health and services, but naturally cannot capture all facets of ecosystem biodiversity.

But with scarce knowledge on spatial range and extinction risk. As of July 2023, the Red List has evaluated 1,970 orchids, representing 6.3% of the estimated 31,000 species. Although this is low, orchids are the vascular plant family with the third highest increase in evaluation effort over the last decade, behind *Fabaceae* and *Cactaceae* (Bachman et al., 2019). In terms of the availability of spatial data, less than 11% of the species assessed by IUCN (1,970) have recorded spatial ranges (210). In other words, less than 1% of described orchids have both an IUCN assessment and a described spatial range. This highlights the need to fill the gaps in assessment and range data before developing scalable biodiversity indicators to guide conservation.

Orchid conservation at all scales. As a beloved and charismatic family with a profile that can benefit plants on a large scale, orchid conservation receives attention from a wide range of actors. This has resulted in large orchid networks (Orchid Specialist Group, European Orchid Council, American Orchid Society, etc.) and numerous academic studies at global (Cribb et al., 2003; Vitt et al., 2023) and national levels. Here are examples from Costa Rica, Greece, France, Australia and China (Crain & Fernández, 2020; Tsiftsis

[8] registered on the Kew platform *Plants of the World Online* https://powo.science.kew.org/, accessed on 09/08/23, accepted names only.

& Tsiripidis, 2020; Vogt-Schilb et al., 2015; Wraith & Pickering, 2019; Zhang et al., 2015). Moreover, Gale et al. (2018) formulate eight recommendations to guide global orchid conservation, including the creation of orchid reserves that will benefit a wide range of other species, and more monitoring of all forms. Citizen science observations are also improving our understanding of the family distribution. However, knowledge gaps are still very important and lead to a difficult prioritisation work before implementing effective conservation (Gale et al., 2018; Wraith et al., 2020). Here is the conclusion of Michael F. Fay from his study entitled *"Orchid conservation: how can we meet the challenges in the twenty-first century?"*:

> *"The level of these threats [habitat destruction, climate change, harvest] now outstrips our abilities to combat them at a species-by-species basis for all species in such a large group as Orchidaceae; if we are to be successful in conserving orchids for the future, we will need to develop approaches that allow us to address the threats on a broader scale to complement focused approaches for the species that are identified as being at the highest risk."*
>
> (Michael F. Fay, 2018)

This author emphasises that species-specific research remains vital to understanding orchid biology, pollination, mycorrhizal associations, population genetics and demography. However, the threats are too severe to allow only this type of focused study if we aspire to conserve the incredible diversity of orchids.

Our contribution to orchid conservation. Our research feeds into orchid conservation at several levels. We are helping to map their global diversity at the kilometre scale with the Shannon Index. With this unprecedented spatial resolution, we are also revealing global patterns of extinction risk for the family. In addition, we show how supplementing the IUCN Red List with automated assessments leads to drastically different risk levels and patterns. An interactive map allows users to explore regional or local gradients. Spatial statistics provide insights and priorities at the country level. Finally, we complement the IUCN Red List with our own predictive method that benefits from the generalisation power of deep-SDMs. Thanks to bioclimatic projections, we provide automated extinction risk assessments for 13,240 orchids, not only for the present, but also for four twenty-year periods up to 2100. We discuss projections of extinction risk levels per continent, as a function of latitude and as a function of altitude.

CHAPTER 2

STATE OF THE ART

Table of contents

2.1 Introduction 15
2.2 Indicators for biodiversity conservation 15
2.2.1 Indicators at species level 16
2.2.1.1 The IUCN Red List of Threatened Species 16
2.2.1.2 Ecological and evolutionary originality 20
2.2.1.3 International agreements 21
2.2.1.4 Species population level: The Living Planet Index . . 21
2.2.2 Community and habitat indicators 24
2.2.2.1 Biodiversity: components and measures 24
2.2.2.2 Habitat classifications 25
2.2.2.3 Essential biodiversity variables 27
2.2.3 Spatial indicators of threat 28
2.2.3.1 Threatened species patterns 28
2.2.3.2 Mapping sources and mechanisms of threat 31
2.2.3.3 Threat integration in conservation planning 32
2.3 Modelling species distribution 36
2.3.1 Species distributions 36
2.3.1.1 Mapping species geographic range 36
2.3.1.2 Species observation data 37
2.3.1.3 Models 40
2.3.1.4 Covariates relevance 44
2.3.1.5 SDMs to inform conservation 47
2.3.2 Deep-SDMs 48
2.3.2.1 Motivations 48
2.3.2.2 Convolutional neural networks 54
2.3.2.3 Resources and inference 56
2.4 Insights for conservation planning 59

2.4.1 AI for ecology and conservation 59
2.4.2 Predicting the missing conservation status of species 62
2.4.3 Future trajectories of conservation indicators 64
2.4.4 Optimising spatial conservation 66

2.1 Introduction

The background information that underpins our work is largely interdisciplinary. We decided to start this review of the state of the art by introducing the application targets behind our deep learning models in a first section 2.2 related to indicators for biodiversity conservation. This first section organises classical metrics and new approaches to quantify biodiversity. It is not intended to be exhaustive, but rather to provide an organised overview of i) the commonly used metrics that guide conservation measures, and ii) the original and ongoing research that attempts to fill current knowledge gaps. Whole facets of biodiversity such as ecosystem services and functions (Burkhard & Maes, 2017; Oliver et al., 2015) or genetic diversity are not covered in this review, but should also be measured to guide conservation.
In a second section 2.3, we will consider the state of the art in species distribution modelling. After an overview of SDMs, we will present recent models based on deep learning architectures. Indeed, deep-SDMs are the common building block of our three scientific contributions. Special attention will be given to the motivation for adopting deep learning successes in ecology and SDMs. The deep learning principles relevant to our work are also introduced and illustrated.
Finally, the third section 2.4 provides insights for conservation planning, especially when they are based on predictive approaches. While such applications are now many and varied, we will focus on automated assessment of species extinction risk, prediction of climate change impacts on biodiversity, and optimisation of conservation resources.

2.2 Indicators for biodiversity conservation

Indicators simplify, quantify and communicate complex concepts. Spatial indicators exploit geospatial data and allow spatial analysis. Biodiversity indicators are essential for reporting the status, trends and drivers of biodiversity change. They allow conservation targets to be set and progress to be tracked against countries' commitments under international agreements. A wide range of conservation-oriented indicators are used to try to cover as many facets of biodiversity as possible. Multiplying indicators may seem contrary to their primary function of simplifying concepts. It remains necessary, however, as biodiversity cannot be reduced to a single measure of species, ecosystems or functions. The Biodiversity Indicators Partnership[1] is a global initiative to coordinate the development and use of indicators for biodiversity-related conventions: IPBES, the Convention on Biological Diversity (CBD), the Sustainable Development Goals (SDG), national and regional agencies.
Here we present an overview of indicators for biodiversity conservation that we consider to be key reference points for situating our work. Our first focus is on indicators at the central and base level of species in section 2.2.1. Next, we will focus on higher level indicators of species communities and habitats in Section 2.2.2. Finally, we will present the active literature on the spatial indicators of biodiversity threats Section 2.2.3.

[1] https://www.bipindicators.net/

Indicator species of ecosystem health
One way of assessing the health of an ecosystem is to measure species or responses that are representative of its state or vigour (Newman, 2009). These measures, or indicators, can be the presence or absence of species, biomass or species metabolism. Examples of indicators are the use of *Posidonia oceanica (L.) Delile* meadows to assess water quality (Pergent-Martini et al., 2005), the presence of freshwater fish and common birds in Europe to represent the quality of their respective habitats (Vallecillo et al., 2016; Yousefi et al., 2020) and finally butterfly assemblages to capture environmental heterogeneity (topography and moisture) (Kremen, 1992). Recent research suggests that, although work is needed to refine understanding of their response to specific disturbances, orchids have good indicator properties (presence and abundance, growth and symbionts) of ecosystem health (Cribb & Hermans, 2007; Newman, 2009). Taking such properties into account, orchid-based indicators can be considered to have a wider scope than just qualifying their family, but also a degree of habitat quality. Again, however, we do not pretend to be able to fully capture ecosystem health through a single family of indicators. In practice, achieving this goal would require a large number of indicators and measurements.

2.2.1 Indicators at species level

2.2.1.1 The IUCN Red List of Threatened Species

Introduction and relevance
The International Union for Conservation of Nature (IUCN) maintains the most widespread and authoritative information on the extinction risk of species in the IUCN Red List of Threatened Species (RL), see (Mace et al., 2008). Its strength lies in the unified assessment process based on five criteria that include population size and dynamics, geographic range, and direct quantitative estimates of species' probability of extinction. For a comprehensive summary on IUCN criteria, the interested reader is invited to consult the *Red List Criteria Summary Sheet* available online[2] and the official guidelines for more details (Commission et al., 2001). These criteria are the result of an iterative refinement process that began in the 1960s with emblematic species and converged on the current version, established in 2001 and revised in 2012. The Red List has gained political weight over the years and its conservation value is recognised by the community (Rodrigues et al., 2006). Betts et al. (2020) proposed a framework for measuring the impacts of red listing species. It allows conservation funding to be directed (Bachman et al., 2019), but also restricts trade in species (Challender et al., 2023) and sets limits on the private sector (Bennun et al., 2018). Many extinction risk assessment are led at the regional or national level.

Assessment process and categories
The IUCN Red List divides species into nine categories of extinction risk. By default, an accepted species that has never been evaluated falls into the Not Evaluated (NE) category. These species are not published on the IUCN Red List, but represent 93% of the world's described species. When an assessment is undertaken, the first step is to gather the information that supports the various criteria. Ideally, all criteria are assessed against official thresholds and the most critical status is retained (precautionary

[2] https://www.iucnredlist.org/

approach). In practice, information must be available to assess the risk of extinction of species against at least one of the five criteria. Otherwise, species inherit the status of Data Deficient (DD). With sufficient information, the assessment results in:

- Least Concern (LC) if the species is far from qualifying as threatened
- Near Threatened (NT) if a threshold is close or likely to be reached in the near future
- Vulnerable (VU), Endangered (EN), Critically endangered (CR) when a corresponding threshold is reached, resulting in a high, very high or extremely high risk of extinction in the wild, respectively.

Finally, a taxon may be Extinct in the Wild (EW) if it is known to occur only in a controlled environment, or Extinct (EX) if there is no reasonable doubt that the last individual has died. Classifying a species as extinct is i) a difficult task, as it requires an exhaustive survey of the species niche or habitat, and ii) a weighty task, as it halts any potential conservation and recovery of the species. The generalised Red List assessment process is outlined in Figure 2.1.

There are specific guidelines for adapting species assessments at regional and national levels (IUCN, 2012a). It is indeed important to complement global assessments at lower levels, where conservation decisions are often made and implemented. When regional assessments concern endemic species, they are equivalent to global assessments. It is estimated that 60% of plant species are endemic to a single region (Bachman et al., 2018). As regional assessments are abundant, they represent an important potential to extend the global coverage of the IUCN Red List (Bachman et al., 2019). Limitations to their integration, such as translation needs or manual entry, are being addressed by the community.

Area of Occupancy and Extent of Occurrence

Criterion B focuses mainly on the geographic range of species and is commonly used to assess the extinction risk of plants. It measures firstly two geographical descriptors, Area of Occupancy (AOO) and Extent of Occurrence (EOO). EOO is defined as "the area contained within the shortest continuous imaginary boundary which can be drawn to encompass all known, inferred or projected sites of present occurrence of a taxon, excluding cases of vagrancy". (Commission et al., 2001). It is not a measure of the size of a species' range, as already assumed (Jetz et al., 2008), but a measure of the species' spatial spread. AOO is defined as "the area within its EOO occupied by a taxon, excluding cases of vagrancy" (Commission et al., 2001). The cell size should be "appropriate to the relevant biological aspects of the taxon, the nature of threats and the available data". In practice, 2 x 2 km^2 cells are widely used. Two examples are shown in Figure 2.2. In addition to meeting the official thresholds for EOO and/or AOO, Criterion B requires two subcriteria out of three possible to be fully assessed. These are whether the species is (a) highly fragmented or has a limited number of locations, (b) in continuing decline, and (c) experiencing extreme fluctuations. An IUCN location is defined as "a geographically or ecologically distinct area in which a single threatening event can rapidly affect all individuals of the taxon present". While continuing decline (b) can usually be observed or inferred, a second subcriterion is often difficult to demonstrate. Indeed, subcriterion (c) on extreme fluctuations only concerns a limited range of species

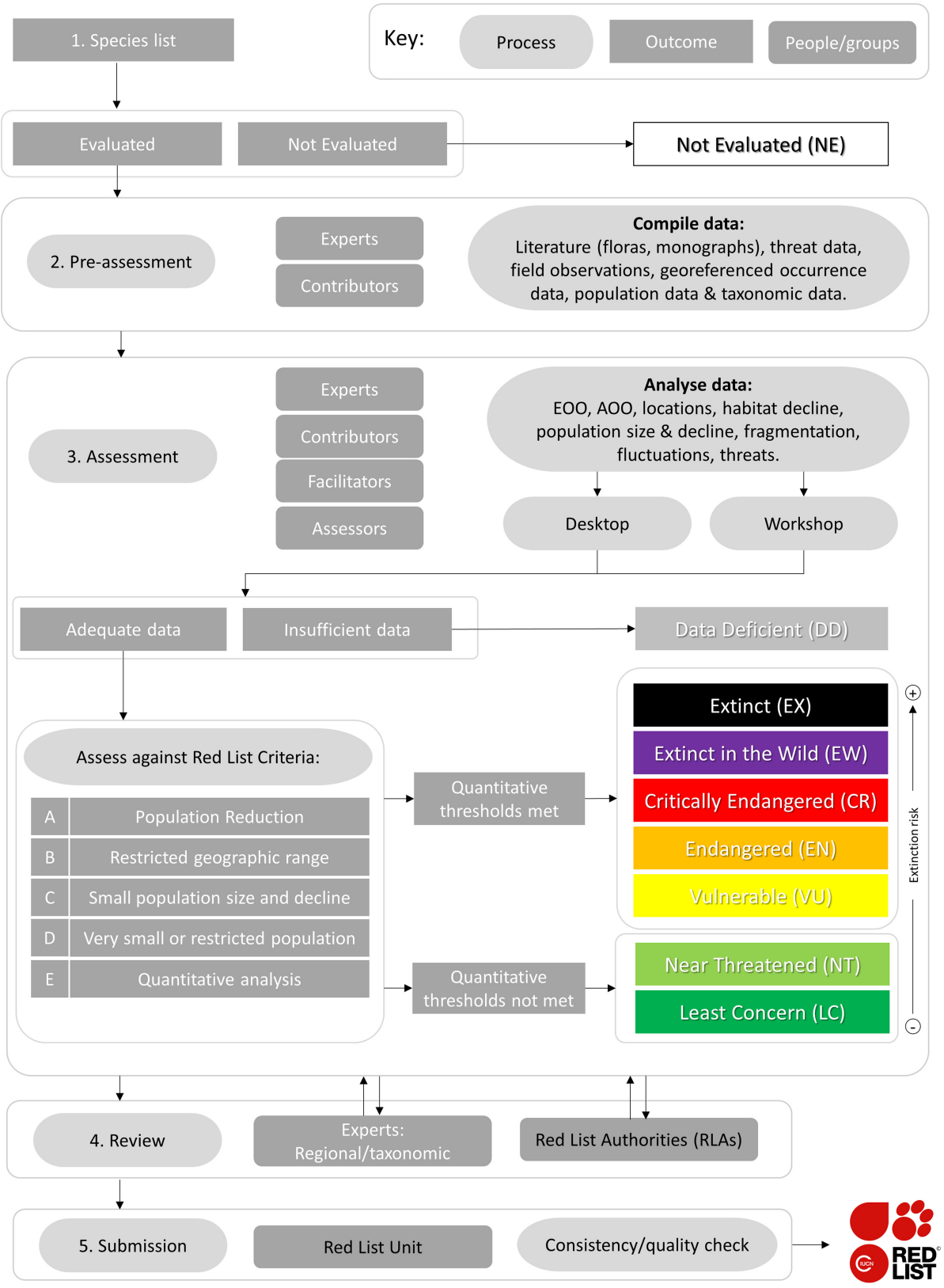


Figure 2.1: *Generalised Red List assessment workflow from species list to publication on the Red List. Ovals represent processes, grey and coloured rectangles are outcomes and curved rectangles are people or groups. EOO = Extent of occurrence, AOO = Area of occupancy. Arrows indicate direction of flow through different stages, including feedback. Reproduced from (Bachman et al., 2019), CC-BY license.*

such as migratory birds, often leaving subcriterion (a) as the only and difficult remaining option. As the calculation of EOO and AOO can be easily coded, most attempts to automate the IUCN Red List assessment while respecting the official thresholds focus on this B criterion (see section 2.4.2).

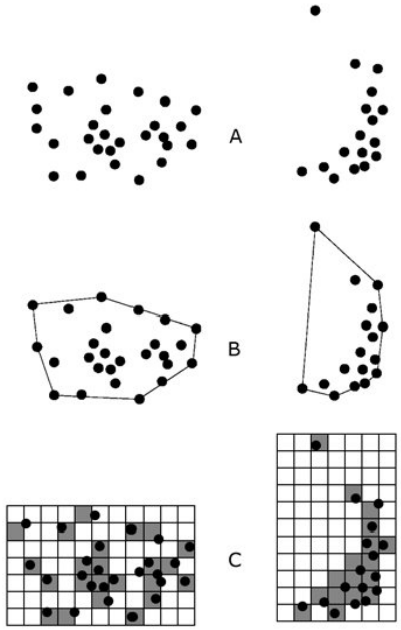


Figure 2.2: *Two examples of EOO and AOO measures.* ***(A)*** *distribution of records of occurrence.* ***(B)*** *minimum convex hulls around records of occurrence to measure EOO.* ***(C)*** *shows two measures of AOO with the sum of the occupied grid squares. Reproduced from (Commission et al., 2001).*

Limitations and scientific debate

As an influential species conservation indicator, the RL and its assessment process are a source of scientific debate and research. Major criticisms have focused on the lack of flexibility or representativeness of the assessment process (Akcakaya et al., 2000; Bachman et al., 2019; Jarić et al., 2016). In an attempt to improve the evaluation process, new or modified criteria are regularly proposed. For example, they may question the calculation of EOO and AOO (Breiner & Bergamini, 2018; Joppa et al., 2016; Marsh et al., 2019). Studies also promote the inclusion of additional information in the assessment, be it species Area of Hability (AOH) Brooks et al. (2019), environmental niche (Breiner et al., 2017), ecological traits (Mattila et al., 2008) or genetic diversity (Rivers et al., 2014).
Biases in RL coverage are well known: taxonomic, but also spatial towards the poles, temporal with 83% of assessments being outdated by 2025 (Rondinini et al., 2014) and functional: species with restricted distributions, small body sizes and low dispersal abilities are underrepresented (Pimm et al., 2014). In addition, assessing biodiversity through the RL is costly (Juffe-Bignoli et al., 2016).
These concerns are legitimate. There is considerable value in identifying and attempting to tackle areas for improvement, even if the RL criteria do not set new standards. IUCN guidelines are regularly updated to incorporate new approaches. Misunderstandings of IUCN concepts can also occur (Collen et al., 2016). Although it can be frustrating,

changing the RL assessment process could compromise its objectivity and authority (Breiner, 2016). Finally, new approaches are being developed to complement the RL, such as the Red List of Ecosystems (RLE), which will be introduced in section 2.2.2, and the Red List Index (RLI) introduced below. Short-term information is provided by other large-scale measures of changing nature, such as population size, number, and habitat extent. They can be easily linked to economic values and public concerns (Balmford et al., 2003).

Red List Indices

As the human impact on biodiversity was increasingly recognised, the need to quantify overall biodiversity loss in order to assess countries' commitments arose at the beginning of the century (Butchart et al., 2010). In response, IUCN and its partners developed the Red List Index (Butchart et al., 2007, 2004). The RLI aggregates IUCN Red List information to provide information on trends in the status of biodiversity for sets of species. The IUCN categories are quantified from Least Concern (LC) (0) to Extinct (EX) (5). An RLI of 1 indicates that all species in the set are LC and an RLI of 0 indicates that all species are extinct. It was originally designed for groups that are extensively and repeatedly assessed, such as birds and mammals. When two different RL assessments are available in time, biodiversity loss can be quantified. The RLI targets genuine changes in threat status, not reassessments due to improved knowledge or changes in criteria. The RLI can be calculated for any subset of species. It can therefore be calculated for countries (Rodrigues et al., 2014; Saiz et al., 2015) or thematic groups of species such as bird and mammal pollinators (Regan et al., 2015). As an exhaustive RL assessment is far from being achieved for many taxonomic groups such as plants, a sampled RLI has also been developed (Brummitt et al., 2015). It is based on a representative sample of 1,500 species and can provide a representative picture of biodiversity change for under-assessed but threatened groups.

2.2.1.2 Ecological and evolutionary originality

Phylogenetic diversity

To protect the potential and inherent value of phylogenetic (evolutionary) diversity, new approaches have been praised (Mace et al., 2003; Pollock et al., 2020; Vitt et al., 2023). Phylogenetic diversity is defined as "the evolutionary diversity represented by sets of taxa, where the most common metric (Faith's phylogenetic diversity) is the branch length of the minimum spanning tree connecting a set of species in a phylogeny" (Pollock et al., 2020). The Evolutionarily Distinct and Globally Endangered (EDGE) metric contributes to identify and conserve such threatened species (Gumbs et al., 2023; Isaac et al., 2007). Evolutionarily Distinctiveness (ED) scores are calculated by dividing the phylogenetic diversity of a clade among its members. Extinction risk is considered using numerical RL categories ranging from $LC = 0$ to $CR = 4$. By combining these two metrics, EDGE places emphasis on species with high evolutionary heritage but which are threatened with extinction. However, the approach requires that clade phylogenies have been established. It has been successfully implemented for mammals, amphibians, corals, birds, reptiles, sharks and rays, but many taxa remain to be considered.

Functional diversity
Another facet of biodiversity of great concern, but overlooked in current RL assessments, is functional diversity. It is defined as "the diversity of functional forms in a species set (or community) as measured by a variety of metrics using dendrograms or representations in multidimensional space" (Pollock et al., 2020). Similar to the EDGE index, new metrics attempt to quantify the functional diversity of species and cross the result with the RL to give a sense of functional extinction risk priority. Recent examples are the Functionally Unique Specialized and Endangered (FUSE) index and the Ecologically Distinct and Globally Endangered (EcoDGE) index (Griffin et al., 2020; Hidasi-Neto et al., 2015). Finally, the latter authors also propose to combine both phylogenetic and functional diversity with extinction risk in a general index called EcoEDGE.

2.2.1.3 International agreements

At the international level, one of the most influential tools for protecting endangered species from excessive trade is the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES)[3]. The Convention was originally drafted by IUCN members in the 1960s and was adopted by 80 countries in 1975. It aims to limit trade in charismatic animals and plants that are exploited alive and for derived products such as food, exotic leather goods, wooden musical instruments, timber, tourism and medicines. In addition, a significant number of species are threatened by complex trade systems where consumers in developed countries cause species to be threatened by goods produced in developing countries (Lenzen et al., 2012).

The European Union has also provided a framework for protection through the Birds and Habitats Directives (Commission et al., 2015). Thirteen years after the Birds Directive, the Habitats Directive was passed in 1992. It aims to protect over a thousand species, with the overall aim of "ensuring that these species and habitat types are maintained at or restored to a favourable conservation status within the EU" (Commission et al., 2015). A new impulse in 2017 aims to accelerate progress towards halting and reversing the loss of biodiversity and ecosystem services (Commission et al., 2017).

2.2.1.4 Species population level: The Living Planet Index

Definition
The Living Planet Index (LPI) is a highly visible biodiversity indicator whose alarming rates are often reported in the media. However, it is subject to many misinterpretations, probably because of its apparent simplicity, which contrasts with the highly complex nature of biodiversity. A misleading example from the Guardian: *"Humanity has wiped out 60% of animal populations since 1970, report finds"*[4]. What the LPI measures is *the average change in the number of individuals in the animal populations surveyed*. An accurate formulation of the global trend is *between 1970 and 2018, there was an average*

[3] https://cites.org/
[4] https://www.theguardian.com/

decline in population size of 69% across the 31,821 populations studied. A population is defined as "a group of individuals of the same species living in the same geographical area. A species often has several or many populations, each living in a different area". This results in two populations of the same species, but with very different numbers of individuals, having the same contribution. Also, a common species with dozens of populations will be strongly represented, as opposed to a rare species with only one population. For more details and a clear example of the LPI calculation, the interested reader is invited to consult (Loh et al., 2005; Ritchie et al., 2022).

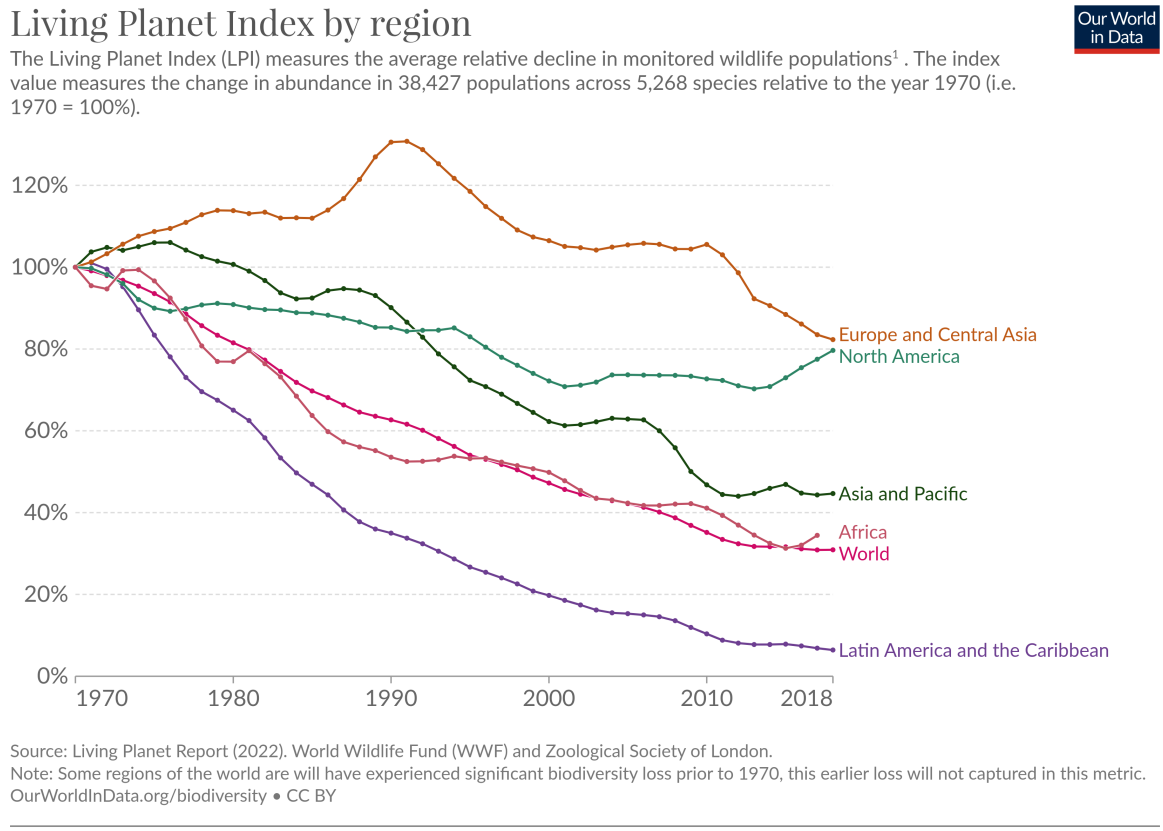
A biased species sample

The LPI is based on 31,821 populations of 5,230 species. This may seem a large sample, but in reality it is limited and highly biased to represent global biodiversity trends. First, the taxonomic bias: only vertebrates are included, with 16% of known bird species, 11% of mammals, 6% of fish and 3% of amphibians and reptiles. Entire taxonomic groups are overlooked: insects, fungi, corals or plants. Moreover, even represented groups are not evenly sampled in terms of IUCN extinction risk: threatened mammals and charismatic species are over-represented (Collen et al., 2009). Second, the spatial bias: the tropics are under-represented, even though they are both a biodiversity hotspot and a threat hotspot.

On the importance of results visualisation

Figure 2.3 provides two contrasting views of the same information, the LPI. The top figure shows the averages per region over time and the bottom figure shows the overall trends per taxonomic group. Let's first focus on the two regions with the largest and smallest decreases in LPI. Latin America and the Caribbean experienced an average decline of 94% in its *studied* populations between 1970 and 2018. The main drivers are intense deforestation and the expansion of agricultural land. Europe and Central Asia have seen the smallest decrease in LPI, with an average decline of 18% since 1970. However, this low figure masks an additional bias in the data: in Europe, land-use change occurred well before 1970. Therefore, the 1970 reference represents already reduced populations and distorts the comparison with other regions. Details of the dynamics of each region are given in (Ritchie et al., 2022).

Plots of overall LPI trends by taxonomic group (Figure 2.3 below) show a very different and more nuanced picture. Here, across all groups, we observe a roughly 50-50 split between declining and increasing populations. This apparent status quo is quite surprising compared to regional and world trends. It can be explained by small population increases and overwhelming decreases, which are not shown here, but may influence the calculation of the LPI geometric average. This more balanced view i) highlights the need to prioritise conservation action where it is most needed and ii) offers hope in the sense that populations are not doomed to collapse everywhere on Earth, as the LPI headline figures may suggest. Although it may seem secondary, the latter point is in fact essential if positive conservation action is to be taken. For example, some wild mammals are making a comeback in Europe (Ritchie et al., 2022). Finally, the Living Planet Report 2022, entitled *Building a nature-positive society*, provides an accurate global picture of the biodiversity crisis while clearly remaining solution-oriented.



1. **Population**: A population is a group of individuals of the same species that live in the same geographic area. A species will often have multiple or many populations, each living in a different area.

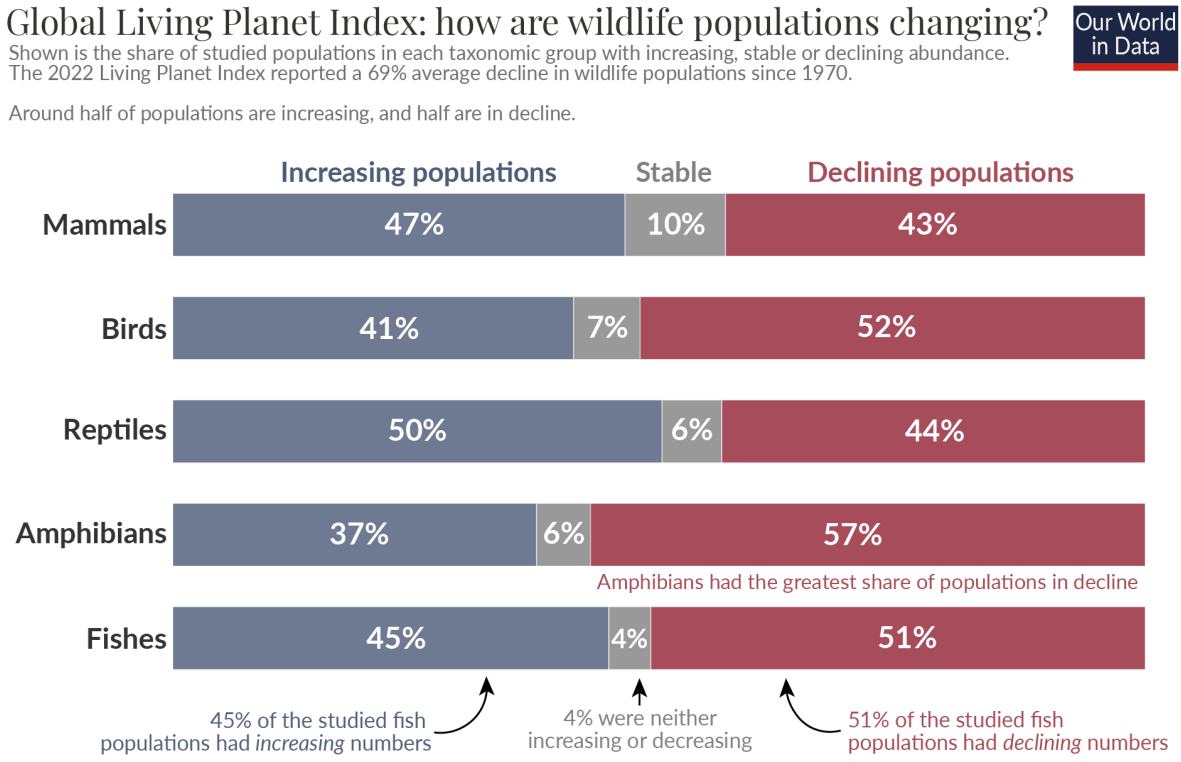


Figure 2.3: *Two contrasting views of the Living Planet Index:* ***(top)*** *by region and* ***(bottom)*** *trends by taxonomic group. Reproduced from (Ritchie et al., 2022), CC-BY license.*

These two contrasting views of the LPI illustrate the importance of data representation. Indeed, the aim of such an indicator is to guide and fund conservation where it is most needed. Result representations should therefore be varied and declined along different explanatory variables to accurately account for the multiple dimensions of the problem.

2.2.2 Community and habitat indicators

2.2.2.1 Biodiversity: components and measures

Biodiversity definition and diversity components

Michel Loreau provided a coherent definition of biodiversity in 2005:

> *"The Earth is home to an extraordinary biological diversity, which includes not only the species that inhabit our planet, but also the diversity of their genes, the multitude of ecological interactions among them and with their physical environment, and the variety of complex ecosystems that they form. This biodiversity, the product of more than 3 billion years of evolution, is a natural heritage and a vital resource on which humanity depends in many ways."*
>
> (Barbault and Loreau, 2005)

Biodiversity is therefore not limited to species richness. However, here and throughout this thesis, species will be the level of study chosen.

Given a pool of species, two key components influence perceived species diversity: species *richness* and species *evenness* (or *equitability*, Marcon, 2015). Richness is the number of different species, while evenness is their proportion. Consider two sets of species, A and B, with the same species richness. If set A contains a largely dominant species that leaves little room for other species, while set B contains equally balanced species, set B will appear more diverse. Most indices, such as Simpson's or Shannon's, assess both richness and evenness, see (Marcon, 2015) for a comprehensive review. These *species-neutral* diversity measures do not take into account the distance between classes. However, two species from the same genus are obviously closer than two species from different families. This idea is reflected in measures of phylogenetic and functional diversity. *Disparity* is the third component of diversity, measuring the degree of difference between species. Based on a broad literature review in disciplines concerned with diversity (beyond biodiversity), A. Stirling asserts that these three components (richness, evenness and disparity) cover all aspects of diversity (Stirling, 2007).

A diversity measure: the Shannon index

The Shannon index is a classic measure of diversity. It is species-neutral in the sense that no specific species trait is taken into account in its calculation (Marcon, 2015). Derived from information theory (Shannon, 1948), the Shannon index is also known as the Shannon-Weaver or Shannon-Wiener index, or simply entropy:

$$H = -\sum_{s=1}^{S} p_s \ln(p_s) \tag{2.1}$$

where S is the number of different species in the system under study and p_s is the probability that a random species from the system belongs to species s (given by the species prevalence, assuming uniform detection). The Shannon index provides a measure of biodiversity as an information quantity.

Biodiversity types: α, β, γ
Diversity is classically estimated at several nested levels. Whittaker calls them α, β and γ biodiversity types (Whittaker, 1960):

- α diversity is the local diversity measured within a bounded system. More precisely, it is the diversity within a uniform habitat of fixed size.
- β diversity measures how different local systems are. This definition is still under discussion (Moreno & Rodriguez, 2010).
- γ diversity is similar to alpha diversity, but this time the measure is over all the systems in the study.

The measures of α and β diversity depend on how finely the habitat is defined. Distinguishing many habitats reduces α diversity in favour of β diversity. It is therefore important to define an indicator that does not depend on this subdivision, hence γ diversity. Recent research suggests using deep learning to estimate species richness, α, β and γ diversity from environmental covariates, see (Andermann et al., 2022).

2.2.2.2 Habitat classifications

Compared to the authoritative IUCN Red List of Threatened Species, biodiversity indicators at higher levels, such as habitat or ecosystem, have a long way to go in terms of visibility and impact. Indeed, conservation science has long focused on the species level (considered more tangible), but biodiversity loss occurs at all levels (Keith et al., 2013). Advancing both knowledge and conservation at the ecosystem level is therefore a top priority. New initiatives in this direction are numerous and can benefit from remote sensing contributions.
Ecosystem is defined as "the living components (biotic complexes and assemblages of species), the non-living components (abiotic environment), the processes and interactions within and between the biotic and abiotic, and the physical space in which they operate" (Nicholson et al., 2021). Habitats are inherently defined from a species perspective and, in simple terms, correspond to the ecosystem(s) in which species can subsist. The European Nature Information System (EUNIS) defines habitats more pragmatically as "a place where plants or animals normally live, characterised primarily by its physical features (topography, species morphology, soil characteristics, climate, water quality, etc.) and secondarily by the species that live there"[5].

The IUCN Habitats Classification Scheme and the Red List of Ecosystems
When assessing the risk of species extinction, the IUCN requires species habitats to be reported according to a classification scheme (IUCN, 2012b). Crossed with global data

[5] https://www.eea.europa.eu/

on land cover (remotely sensed), climate and land use, this classification has allowed the creation of a global map of terrestrial habitat types (Jung et al., 2020). However, the IUCN itself acknowledges that this classification scheme is not entirely satisfactory and that a review is needed [6]. A spatial description of nature in terms of ecosystems rather than species habitats might have been preferable and was the origin of the Red List of Ecosystems (RLE).

The foundations of the RLE begin with the description of ecosystem risk assessments and the concept of ecosystem collapse (Keith et al., 2013). Based on five criteria, including the rate of ecosystem decline, distribution and degradation, the assessment process and threat categories mirror those of the IUCN Red List of threatened species. Improving the assessment of ecosystems will hopefully make it possible to prevent their collapse through restoration (Valderrábano et al., 2021). The RLE also relies on a comprehensive and globally consistent typology of ecosystems, launched in 2020 and recently updated (IUCN, 2020; Keith et al., 2022). The new typology is described as conceptually robust, scalable, spatially explicit and adapted to reflect functional responses to change and management. Research is underway to develop biodiversity indicators based on the RLE (Rowland et al., 2020).

The EUNIS Habitat Classification

The European Nature Information System (EUNIS) was developed in the 1990s and early 2000s for the European Environment Agency (EEA, Davies et al., 2004). It covers both marine and terrestrial habitats and is the largest comprehensive pan-European hierarchical habitat classification. The classification of terrestrial habitats is based on species composition and vegetation structure (phytosociological vegetation types), but also on the abiotic environment and geographical location. The classification is widely recognised as a key tool for monitoring progress towards the EU biodiversity targets. It is a common reference for habitat characterisation by vegetation science or satellite imagery and feeds both research and EU policy. Classification today relies on an expert-based system, mostly based on vegetation plots such as the European Vegetation Archive (EVA) (Chytrý et al., 2016, 2020), and new automated classification methods are under development. At the European level, the EEA is also supporting the development of (i) a set of thematic biodiversity indicators [7] (one of which is *Biodiversity - Ecosystems*) and ii) the Mapping and Assessment of Ecosystems and their Services (Maes et al., 2013).

Natura 2000

Natura 2000 is a network of natural and semi-natural sites in Europe, targeting breeding or resting sites for rare, threatened species and rare natural habitats. The sites are spread across the 27 EU countries (both on land and at sea) and the network is governed by the Birds and Habitats Directives (Commission et al., 2015, 2017). It is the EU's main response to the recommendations of the CBD. It currently covers 18% of the EU's land area and 8% of its marine territory[8]. It is important to note that Natura 2000 is not a strict network of protected areas where human activities are excluded (indeed, most of the

[6] https://www.iucnredlist.org/
[7] https://www.eea.europa.eu/ims
[8] Interactive map available at https://natura2000.eea.europa.eu/

sites are privately owned). Instead, EU countries must ensure that the sites are managed in a sustainable way, both ecologically and economically. Details and illustrations of the sites' objectives can be found in the book (Sundseth & Creed, 2008).

2.2.2.3 Essential biodiversity variables

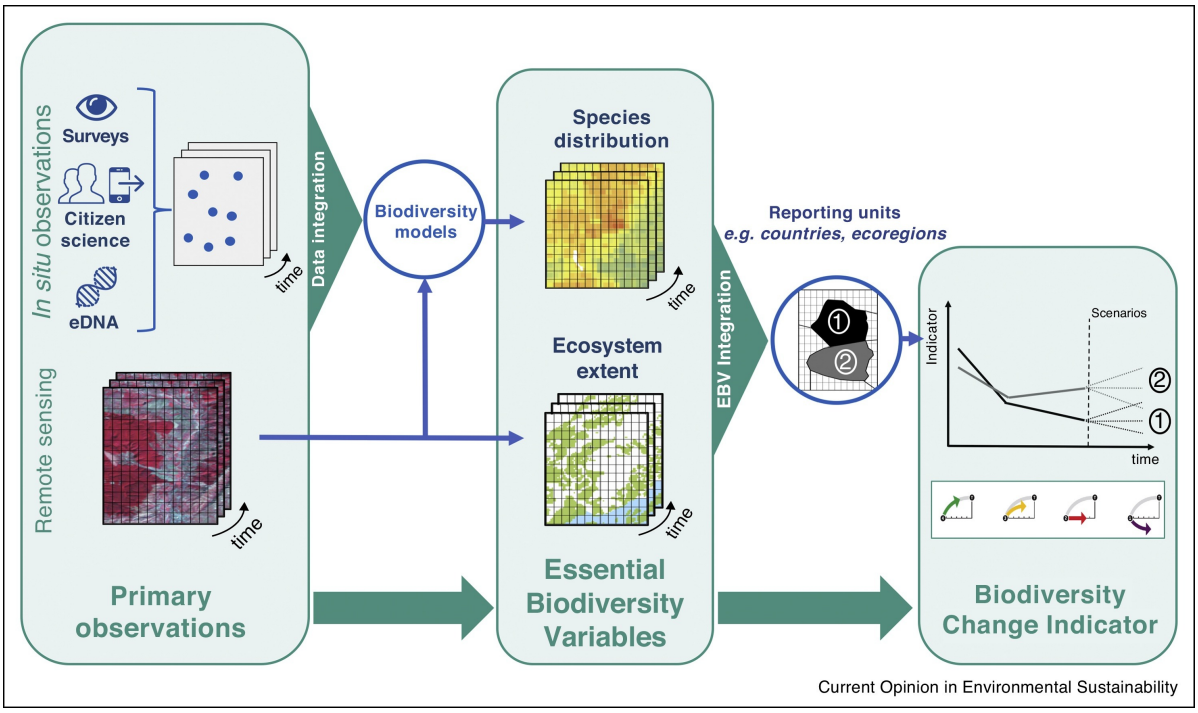


Figure 2.4: *From observations to the production of EBVs and indicators. In this example, integrated data from different primary sources of observations (e.g. in situ, remote sensing) are combined within biodiversity models to produce layers of spatial and temporal variation in ecosystem extent and species distribution EBVs. This information is then integrated and summarised within reporting units to calculate an indicator of biodiversity change, used, for instance, for reporting progress towards an Aichi conservation target. EBVs and models can also be used to project changes in the indicator using scenarios. Reproduced from (Navarro et al., 2017), CC-BY license.*

Essential Biodiversity Variables (EBV) have been introduced to promote the collection, sharing and use of biodiversity information (Pereira et al., 2013). EBVs are biological state variables. They provide a way to aggregate biodiversity observations from different methods, including citizen science and remote sensing (Proença et al., 2017). EBVs have been developed by the Group on Earth Observations Biodiversity Observation Network (GEO BON). Their aim is to support decision-makers by providing a general framework for biodiversity monitoring.https://geobon.org/ebvs/what-are-ebvs/ EBVs must be consistent over time (Mihoub et al., 2017) and space, similar to Essential Climate Variables and Essential Ocean Variables. There are currently 6 EBV classes and 21 EBV names[9]. In the *species populations* class, the EBVs are *species distributions* and *species abundances*. *Ecosystem phenology* is an EBV from the class *ecosystem functioning*. EBVs

[9] https://geobon.org/

and BONs (a system of coordinated Biodiversity Observation Networks) are the two core components developed by GEO BON.

EBVs themselves can either be directly linked to the Aichi Targets and Sustainable Development Goals, or used as inputs to biodiversity models to derive indicators, see Figure 2.4. GEO BON is therefore developing a set of global biodiversity change indicators (BON, 2015) to inform international agreements and reports such as the IPBES assessments. EBVs provide a common formalisation framework to qualify heterogeneous data, for example on species populations (Jetz et al., 2019).

Community and habitat indicators have a spatial extent and can easily be mapped. Our focus now will be on mapping the spatial dimensions of threats and how they can be integrated into conservation planning.

2.2.3 Spatial indicators of threat

Conservation planning is an inherently spatial process (Evans et al., 2011). Community and habitat indicators can be readily mapped and integrated into conservation planning. However, this is not the case for the species level indicators introduced in the previous section 2.2.1. These indicators must first be projected into geographic space.
Following the dimensions of threat defined in Balmford et al. (2009), mapping species at risk of extinction amounts to representing stresses or *unfavourable states* and is the topic of the first section 2.2.3.1. Next, we are interested in mapping sources and mechanisms of threat (Section 2.2.3.2) and finally in integrating threat into conservation planning (Section 2.2.3.3).

2.2.3.1 Threatened species patterns

Spatial range of species
Spatial ranges of species are often used to project species-level indicators into geographic space. Such ranges may be expert-based, simply inferred from species sightings, or modelled (see section 2.3). Only birds, mammals and amphibians have both comprehensive range maps and conservation status. Many studies have therefore focused on these taxonomic groups. However, these well-assessed groups are not always good surrogates for vulnerable species (Hamilton et al., 2022). As spatial knowledge is still very scarce and poorly integrated (Jetz et al., 2012), range generalisation is often necessary:

> *"Given this pivotal role of species distribution information, it might be surprising to realize how poorly documented the geography of life on earth is, an impediment termed the Wallacean shortfall."*
>
> (Jetz et al., 2012)

The Wallacean shortfall was named after Alfred Russel Wallace (1823–1913) by Lomolino et al. (2004) and refers to *the paucity of information on the geography of nature.*

In the following paragraphs, examples of studies are given where threatened species patterns have been mapped using known species ranges or SDM results. The aim is to illustrate a range of work comparable to the maps in our chapter 4 (maps that are not not reproduced here are linked to their online versions).

Global extinction probability
The Global Extinction Probability (GEP) is an approach that quantifies the contribution of potential local biodiversity loss events to global biodiversity loss (Kuipers et al., 2019). It is based on species range, IUCN status and species richness. The IUCN extinction risk categories are quantified according to a given scheme, e.g. LC= 0 to Critically endangered (CR)= 4. GEP is a relative measure that indicates, for a given area, whether many species are threatened with extinction and how dependent they are on that area. If there are many threatened species with very small ranges, or if there are endemic species, the GEP is likely to be higher than in other areas. Verones et al. (2022) calculated GEPs for 98,000 species in 20 species groups across marine, terrestrial and freshwater ecosystems (see their Figure 1). For taxonomic groups for which no IUCN range was available, polygons based on occurrence records (freshwater fishes) or SDM-based ranges (terrestrial vascular plants) were used (Borgelt et al., 2022b). The GEP calculation can also inform biodiversity life cycle assessments. This is a method of estimating the potential environmental impacts of a product (or in this case, threat sources) over its entire life cycle (or over time, Winter et al., 2017).

Species threat abatement and restoration (STAR) metric
The STAR quantifies the extent to which mitigating threats and restoring habitats locally contributes to reducing extinction risk (Mair et al., 2021). The metric is scalable across species, threats and geographies, and is mapped at a grid cell resolution of 5 km (see their Figure 2). It consists of two complementary measures: STAR_T, i.e. how much the reduction of a specific threat in a given location contributes to the reduction of the extinction risk of all studied species, and STAR_R, which assesses the habitat restoration potential for a given species in a given location relative to its current habitat. The STAR calculation requires the species' IUCN extinction risk, threats, spatial range, current and restorable AOH. 5,359 species (2,055 amphibians, 1,957 birds and 1,347 mammals) were included in the analysis. From now on, only extensively assessed taxonomic groups can be included. Current and original AOH were calculated by matching species' IUCN ranges to synthetic land use and land cover maps from 2015 and 1992, respectively. It was also constrained to digital elevation maps with known IUCN elevation ranges. Finally, the recoverable AOH is calculated as the difference between the original and current AOH.

An SDM-based approach to inform under-assessed taxa in the United States
Hamilton et al. (2022) developed a novel spatial index at fine resolution (990 m) in the United States that highlights species with i) small modelled habitat ii) largely unprotected, i.e. with a large proportion of modelled habitat outside current protected areas. To do this, they used Random Forest (RF) models to shape suitable habitats for common, but under-assessed, taxa: vascular plants, terrestrial vertebrates, freshwater animals and pollinators. The suitability maps were then thresholded and validated by experts. The authors note that the spatial patterns of invertebrates and vascular plants are not well

represented by the groups commonly used in biodiversity indicators (birds, mammals, amphibians), see Figure 2.5. These understudied groups are yet important contributors to ecosystem function, providing habitat, prey, pollination and nutrient cycling (Hamilton et al., 2022). They also compared the spatial patterns of their indicator based on either modelled habitats or species range maps for vertebrates. Using modelled habitats resulted in much more nuanced conservation opportunities. The scientific approach taken by these authors is probably the closest to our work developed in chapter 4 on the extinction risk of orchid assemblages.

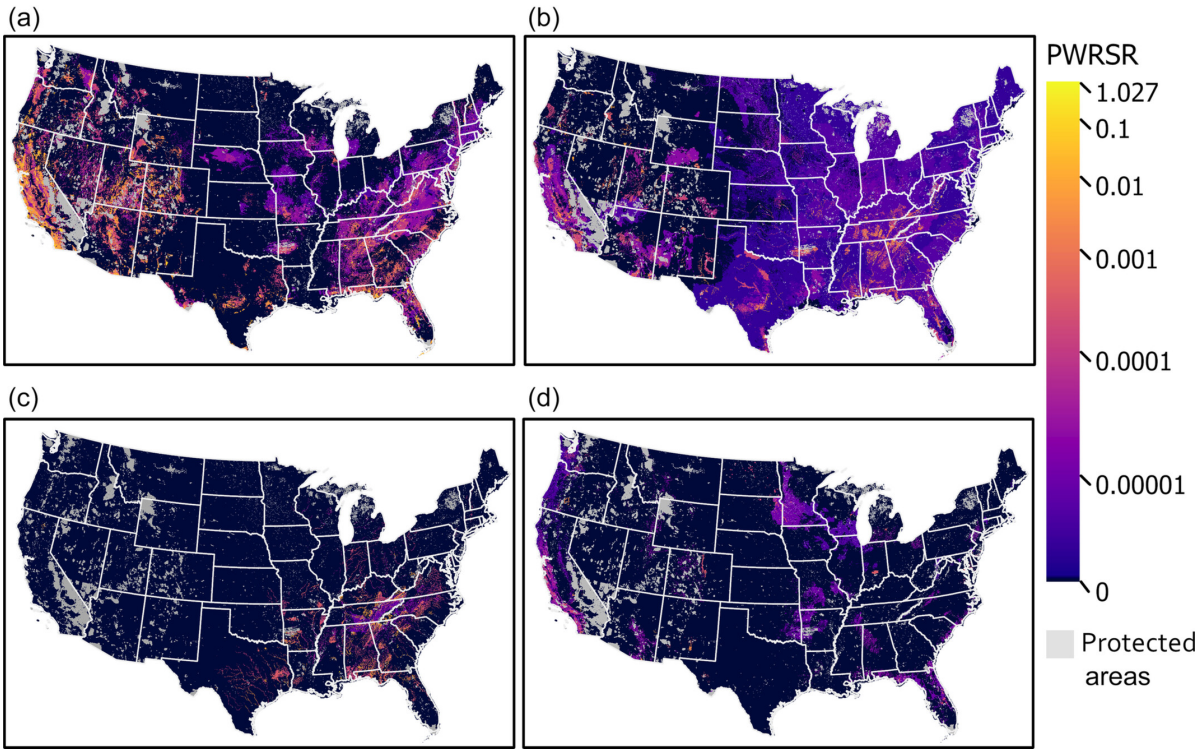


Figure 2.5: *Protection-weighted range-size rarity of imperiled species by taxonomic group:* ***(a)*** *vascular pants,* ***(b)*** *vertebrates,* ***(c)*** *freshwater invertebrates (mussels, crayfishes),* ***(d)*** *pollinators (butterflies, bumblebees). Reproduced from (Hamilton et al., 2022) with permission from John Wiley and Sons.*

EDGE zones and *Cactaceae* maps

Based on the ED and EDGE scores for assessing Evolutionarily Distinct and Globally Endangered species (see section 2.2.1.2), the highest priority ED and EDGE zones were identified using species IUCN ranges (Safi et al., 2013). Two approaches were tested: one based on species richness and another based on randomisation, the maps being derived at rather coarse resolution (1° or 2°, see their Figure 1) to avoid commission errors.
Goettsch et al. (2015) carried out a convincing study on the threats to the *Cactaceae* family. They showed that this group is one of the most threatened assessed to date, with 31% of the 1,478 species at risk. They also derived maps of the proportion and richness of threatened species based on ranges generated in ArcGIS and validated by experts, see Figure 2.6.

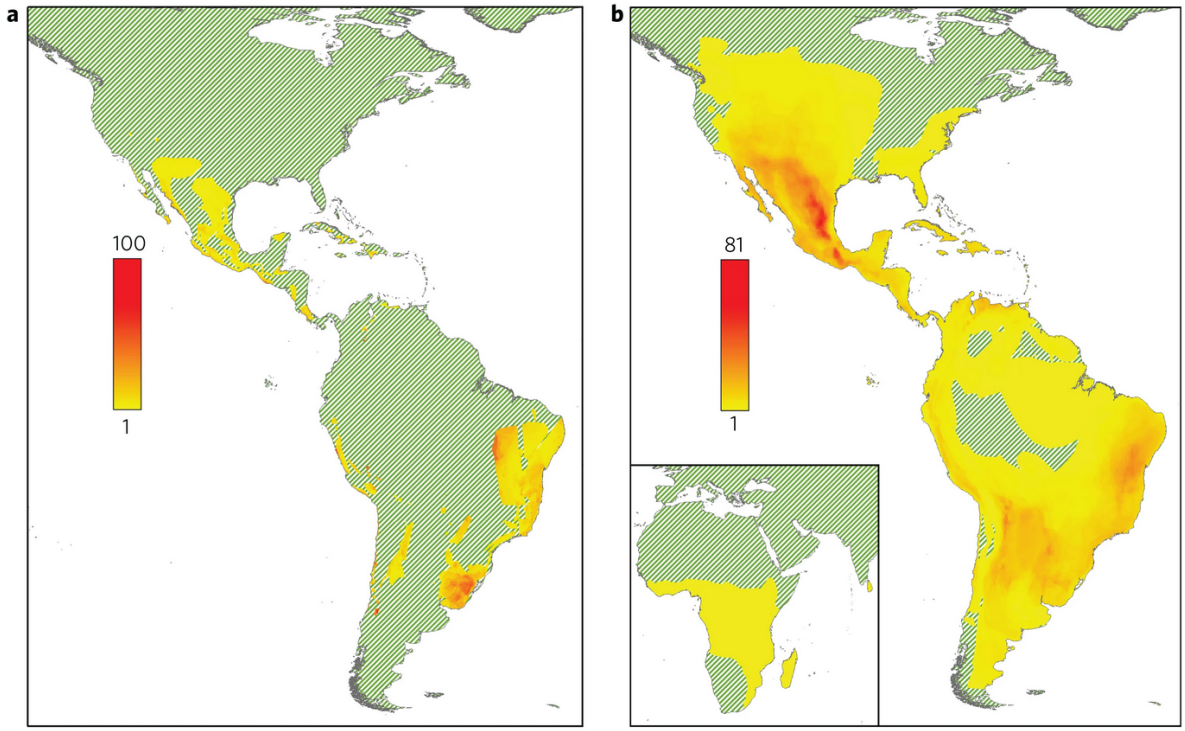


Figure 2.6: *Patterns of Cactaceae biodiversity.* ***a*** *Proportion of species that are threatened (Vulnerable, Endangered and Critically Endangered).* ***b*** *Total species richness. Reproduced from (Goettsch et al., 2015) with permission from Springer Nature.*

2.2.3.2 Mapping sources and mechanisms of threat

It is the ever-increasing level of threats and our inability as a society to compensate for them that has led to the current biodiversity crisis. Understanding and addressing threats to nature is therefore essential to ensure that action is targeted where it is most needed, although the ultimate goal of conservation is to maintain biodiversity. Attempting to address the sources and mechanisms of threats requires clear mapping. Threat maps are defined as "spatial representations of the distribution, intensity or frequency of threats to biodiversity across a landscape or seascape" (Tulloch et al., 2015). Using IUCN species or modelled ranges to model threats carries an implicit assumption: that species are threatened uniformly across their spatial range. Although this is often incorrect in practice, it is a convenient approximation that is found throughout the literature.
In Australia, Evans et al. (2011) led a compelling study of 1,700 species of vascular and non-vascular plants, vertebrates and invertebrates considered nationally threatened. They used SDMs to map the likely distribution of threatened species across the country (see Section 2.3). They highlighted the three main threats: habitat loss, inappropriate fire regimes and introduced species. Their result can be appreciated in the map Figure 2.7 depicting the predominant threats to biodiversity across Australia. Harfoot et al. (2021) used the IUCN Red List to map threats at a global scale based on 23,271 species representing terrestrial amphibians, birds and mammals. Species range maps were derived from BirdLife International, NatureServe and the IUCN. In addition, a probabilistic framework accounts for the fact that species are not uniformly threatened across their known ranges. This results in convincing global maps at a resolution of

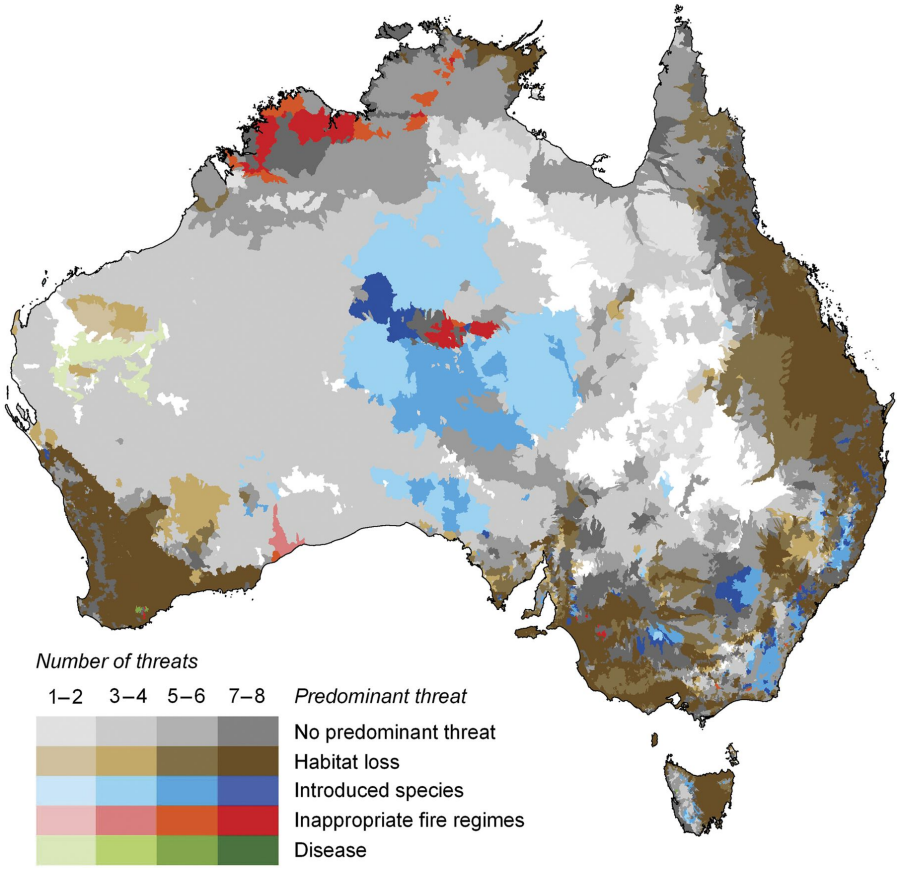


Figure 2.7: *The distribution of the predominant threats to biodiversity across Australia. The "predominant threat" is the threat affecting the greatest number of species in each subcatchment. White indicates areas where no threatened species occur. Reproduced from (Evans et al., 2011) with permission from Oxford University Press.*

50 km × 50 km representing threats to terrestrial vertebrates. The maps can also be refined by threat (deforestation, pollution, agriculture, invasive species, hunting, climate change). In addition, Han et al. (2015) used IUCN species ranges to produce impressive maps of current and future rodent disease reservoirs (see their Figure 2). Finally, much of Europe's native wildlife has disappeared. Therefore, rather than threat indicators based on remaining species, maps of cumulative drivers such as the terrestrial Human Footprint (Venter et al., 2016, see Section 2.3.1.4) may better reflect the true extent of human impact in certain regions.

2.2.3.3 Threat integration in conservation planning

Trying to integrate threats with other aspects of conservation planning is a delicate task. Knowing where and why threats are occurring is valuable information for making conservation decisions, but it is not in itself sufficient. Ideally, threat maps should be developed with an understanding of how species respond to threat mitigation measures in order to properly inform conservation decisions (Tulloch et al., 2015). In addition, social, political, economic and biodiversity outcomes need to be considered in order to allocate

resources and avoid unintended consequences (Tulloch et al., 2015). Figure 2.8 shows an example of how different aspects (here three threat maps and a conservation feature) can be combined to produce a synthetic decision support. Furthermore, in order to arrive at efficient decisions, the resource allocation process should be clearly structured, e.g. using a graph, and uncertainty accounting should be made explicit (Tulloch et al., 2015).

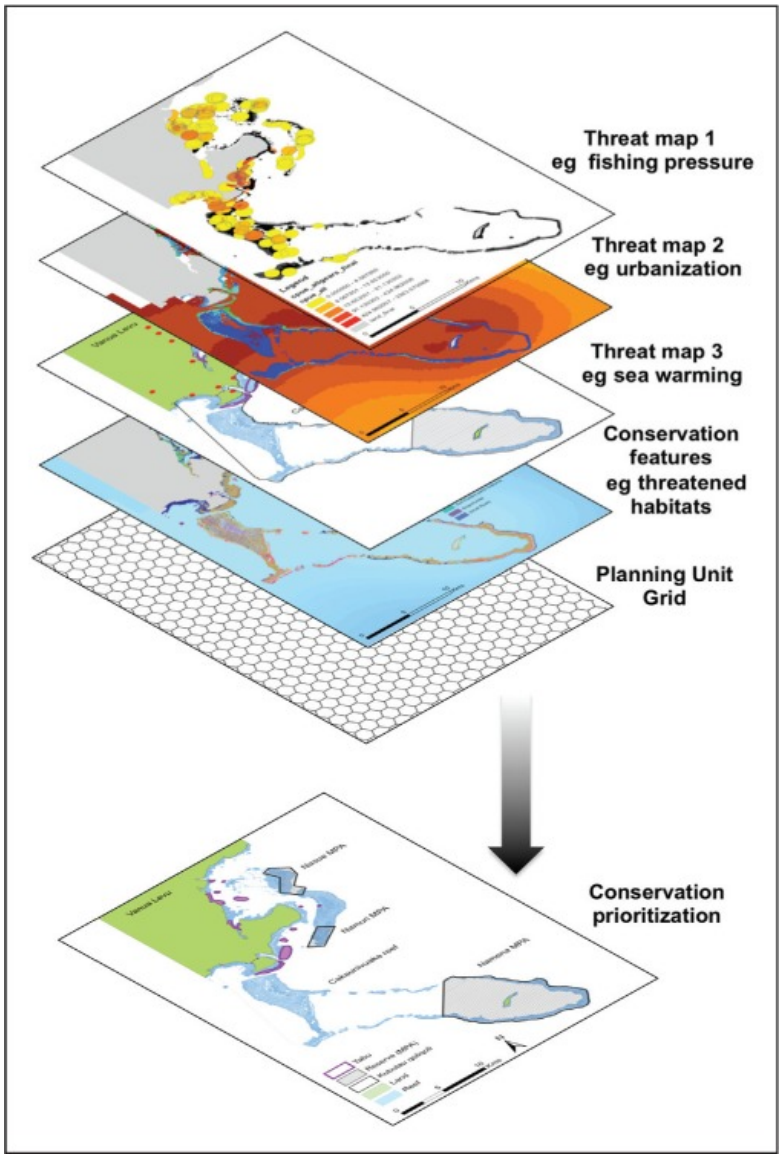


Figure 2.8: *Illustration of how different maps can be combined in conservation planning, based on a case study on marine protected areas in Fiji (Tulloch et al., 2013). Reproduced from (Tulloch et al., 2015) with permission from John Wiley and Sons.*

As an ultimate goal, biodiversity outcomes should be central to conservation planning. Mapping and mitigating threats is a lever to conserve biodiversity. Jetz et al. (2022) call for the inclusion of indicators reporting on biodiversity outcomes when assessing the effectiveness of area-based conservation targets. Indeed, international targets that are purely area-based can lead to the implementation of under-efficient Protected Areas (PA), Maxwell et al. (2020). Misleading observations could be made if a high proportion

of areas were protected but the network actually covered only a limited proportion of biodiversity. By linking PA implementation to biodiversity outcomes, this potential pitfall can be avoided. Similar to the 2°C climate target, a global biodiversity target based on species extinction is needed to galvanise biodiversity policy (Rounsevell et al., 2020). The inclusion of biodiversity indicators in area-based conservation targets would also encourage countries to protect their highly unequal share of biodiversity (Jetz et al., 2022). This process will require increased international cooperation. Indeed, countries have uneven capacities to finance conservation, with some developing countries holding some of the greatest biodiversity (Jetz et al., 2022).

Spatial analysis of threatened species also requires adapted tools. The Spatial Portal developed by the Atlas of Living Australia is an example of a coherent response with advanced filtering tools available online (Belbin et al., 2021). Such an initiative, with a clear emphasis on the spatial analysis involved in conservation planning, is essential to improving biodiversity protection.

Throughout this section we have seen how important the spatial distribution of species is. Spatial ranges are needed to assess species extinction risk, but also to map habitats defined by species composition, or simply to map species and threats. We have also seen how scarce spatial knowledge currently is. Attempts to map biodiversity usually result either in the study of taxonomic groups that are extensively assessed as vertebrates, or in the use of surrogates and models to also study underassessed groups. In this second part of the state of the art, we will be interested in the modelling of species distributions and, in particular, in the class of models that employ deep learning to deal with big data on biodiversity.

"Scientific illustrations should speak to the senses without fatiguing the mind"

Alexander von Humboldt



Figure 2.9: *Alexander von Humboldt's Tableau Physique (Humboldt & Bonpland, 1807). In 1802, physical geographer Alexander von Humboldt and botanist Aimé Bonpland climbed Chimborazo, an equatorial volcano that was then thought to be the world's highest mountain. They documented the mountain's flora, from the tropical rainforest at its base to the highest lichens. Humboldt's Tableau Physique organises these observations in an innovative diagram showing Chimborazo in cross-section, with text detailing which species live at which elevation. This impressive study is a testament to man's desire to map the natural world. It also is a unique data source for assessing vegetation shifts in response to climate change, and is the oldest existing dataset on altitudinal ranges of tropical mountain vegetation. Today, interdisciplinary work between historians and ecologists allows the validity of such pioneering work to be adjusted (Moret et al., 2019). CC0 license, public domain.*

2.3 Modelling species distribution

The study of species distribution is the fundamental goal of biogeography (Wallace, 1860). It improves our understanding of the natural world. If one argues that it is not enough intrinsically, it also allows a better management of natural resources at the theoretical benefit of both nature conservation and society. Here we will review the core components of species distribution models before introducing a new class of models: SDMs based on deep learning architectures (deep-SDMs), which have occupied a key position throughout my research project.

2.3.1 Species distributions

In this section, a brief introduction on how species distributions are represented precedes a review of standard species observation data, models and covariates relevance. Finally, we will summarise the use of SDMs to inform conservation.

2.3.1.1 Mapping species geographic range

Representing the geographic distribution of species is an incredibly difficult task for a number of reasons: i) it requires extensive geolocated data on species presence and observations, ii) the spatial range of species is dynamic, iii) range generalisation is based on ecological hypotheses that necessarily simplify the myriad interdependencies of nature.

How to define species geographic range?
As common as the notion of a species geographic range may be, it is difficult to find a consensus definition in the literature (Brown et al., 1996; Gaston, 1991). Broadly speaking, it corresponds to the region occupied by a particular species and is influenced by both biotic and abiotic factors. The IUCN has developed two surrogate measures of the geographic range of species: the extent of occurrence, which recognises the inclusion of unsuitable regions, and the area of occupancy, which is intended to include only regions actually occupied by a taxon within its EOO. The difficulty in properly defining the species geographic ranges stems from their inherently dynamic nature and the inability to determine how generalised or strictly matching observations it should be. We will now further discuss the geographic ranges by introducing three high-level types commonly employed (Beery et al., 2021).

Three common types of range maps
A simple representation of where species have been observed to occur is a first approximation of species geographic range. This raw display is hardly a geographic range, as observations are often punctual, but it can serve as a first surrogate to give a sense of species distribution. It is always possible to represent species in this way, as long as the observations have been georeferenced, e.g. in GBIF occurrence maps. Species observation data will be discussed further in the next Section 2.3.1.2.
A second possibility is to use statistical models to generalise the locations of sampled

species with environmental data. SDMs obviously fall into this category and will be defined in Section 2.3.1.3.
Finally, a third option is to rely on expert range maps. These are maps drawn by taxonomic experts with knowledge of terrain surveys, habitat preferences and species history. Importantly, this third type of representation is often the most trusted source of distribution information. However, expert maps are often scale dependent and tend to overestimate species with small ranges (Hurlbert & Jetz, 2007). They come in a wide variety of forms, from freehand drawing to maps resulting from data-intensive models and slightly refined by experts. Range maps from the RL and maps of iconic species are usually expert-based.

2.3.1.2 Species observation data

Definitions and examples
Any attempt to map species distributions relies primarily on observation data. A variety of sampling protocols and resulting data types coexist. *Presence-only* occurrences report that a particular species has been observed at a particular place and time. No information is given about where the species was not observed. In contrast, *presence-absence* data provide both the presence and absence of the species observed in space. This massive difference leads to different types of models to make the most of each type of observation. Other data types such as species abundance and species count exist and can feed species distribution modelling, but are not discussed further here (see Miller et al., 2019 for a review). Another distinction for species observation data is the sampling protocol. *standardised* data is defined as "data collected using a standardised sampling design and a fixed protocol at known sampling locations" and is opposed to non-standardised data that is "not collected under a standardised protocol, where sampling locations and sampling effort are often unknown and the sampling protocol varies" (Miller et al., 2019). The fixed sampling protocol in standardised data allows to partially mitigate the sampling biases introduced below. Examples of observation data and collection methods are listed in Table 2.1.

When it comes to mapping species distributions on a large scale, standardised data collection methods are too expensive to use. As a result, most large-scale distribution models heavily rely on presence-only and presence-absence data collected using non-standardised approaches. However, the integration of observation data opens up the opportunity to merge rich spatial data sources with more focused and standardised datasets (Miller et al., 2019). The GeoLifeCLEF 2023 dataset is an example dataset that aims to develop such integrated methods (Botella et al., 2023, see section 2.3.2.3).

Citizen science and new technologies
The recent surge in Citizen Science (CS) has generated massive amounts of opportunistic observations, which are ultimately collected on platforms such as the Global Biodiversity Information Facility (GBIF) portal[10]. Figure 2.10 illustrates this data explosion on GBIF for the *Plantae* kingdom. This trend results from the development of attractive and

[10] https://www.gbif.org/

Collection method	Samp. Pro.	Obs. type	Example	
CS observations	NS	P	Pl@ntNet	(Affouard et al., 2017)
Historic records	NS	P	Herbarium sheets	(Lutio et al., 2022)
CS checklists	S	P-A	eBird	(Sullivan et al., 2009)
Static sensors	S	P-A	Camera traps	(Trolliet et al., 2014)
Sample collection	S	P-A	Insect trapping	(Child & Pinniger, 1994)
Expert field surveys	S	P-A	Line transects	(Buckland et al., 2007)
	S	P-A	Vegetation plot	(Chytrý et al., 2016)
	NS	P	Distance sampling	(Buckland et al., 2005)
	NS	P-A	Site-occupancy	(MacKenzie & Nichols, 2004)

Table 2.1: *Sources of species observation data. Each of these examples represents a method of collecting or accessing observations of different species. Citizen science is denoted CS. Only presence-absence (P-A) and presence-only (P) observations observation types are listed in this table (no abundance or species count listed here for instance). The sampling protocol are either Standardised (S) or Nonstandardised (NS). Adapted from (Beery et al., 2021), CC-BY license.*

convenient ways to collect data, combined with increasing public engagement, including raising awareness from an early age (Hobbs & White, 2012). Citizens can indeed be highly motivated to contribute to biodiversity knowledge and conservation. National taxon-specific networks of passionate citizens, NGOs and online platforms[11] have largely contributed to the data explosion. Impressive advances in mobile phone cameras (Graham et al., 2011) and automatic species identification (Affouard et al., 2017; Unger et al., 2021) also play a big role. It is the use of deep learning that has enabled impressive species identification performance to be achieved.

Vegetation plots and specific datasets

Vegetation science has produced a variety of curated plant datasets with different focuses: taxonomic, geographic (e.g. national inventories), thematic, etc. For example, vegetation plots are an important source of standardised and presence-absence species observations. They are defined as "records of plant species composition, in plots of 1 m^2 to a few hundred m^2, collected by phytosociologists" (Zhongming et al., 2015). Global databases (Bruelheide et al., 2019; Dengler et al., 2011) and integrative programmes at continental (Chytrý et al., 2016) and regional (Schmidt et al., 2012) scales centralise these records and offer exciting opportunities for vegetation analysis and distribution modelling. A few examples among many other initiatives are: the PREDICTS database (Hudson et al., 2014), a global dataset collecting samples specifically exposed to human pressures, RAINBIO (Dauby et al., 2016), a project collecting observations of vascular plants in continental tropical Africa, and EU-Forest (Mauri et al., 2017), a dataset integrating forest plot surveys of European tree species.

[11] e.g. https://observation.org/

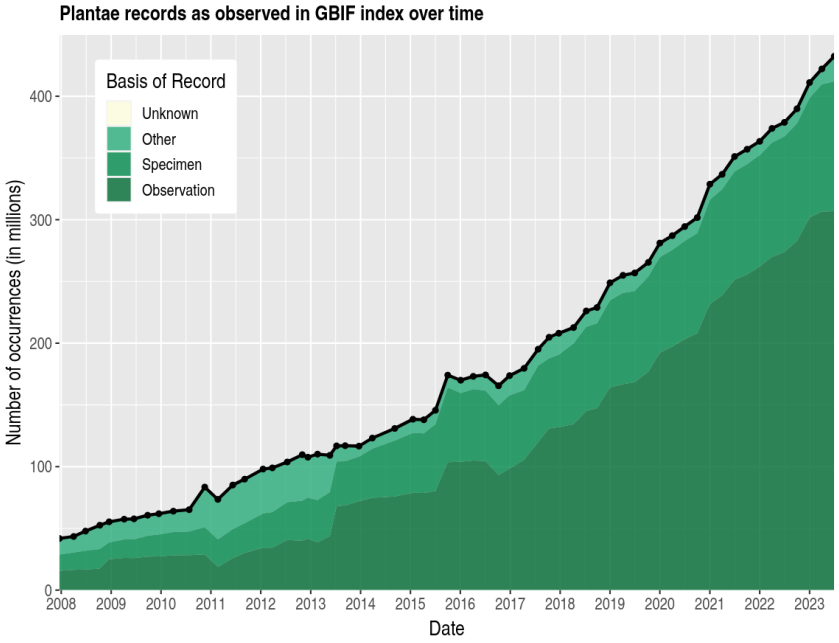


Figure 2.10: *GBIF cumulative number of plant records over time categorized by the basis of record. Reproduced from https://www.gbif.org/.*

Gaps, uncertainty and biases in observation data

Species observation data is conditioned by numerous gaps resulting from our limited biodiversity knowledge and non-standardised sampling protocols. First, biodiversity knowledge is conditioned by the taxonomic gap (Hortal et al., 2015). Distribution, abundance, evolution, abiotic tolerance, species traits and biotic interactions are the other dimensions of the biodiversity shortfall. Together they limit our understanding of biodiversity. Large taxa such as insects or invertebrates, and geographical areas such as the tropics, are still largely undersampled (Cayuela et al., 2009; Feldman et al., 2021).

Species observations are also subject to taxonomic, geographic and temporal uncertainties in their reporting (Meyer et al., 2016). Uncertainty is therefore propagated in all modelling attempts based on uncertain observations. However, certain modelling techniques (boosted regression trees and maximum entropy approach) have been shown to be particularly robust to spatial error (Graham et al., 2008).

Species observations are also conditioned by taxonomic, spatial and temporal sampling biases, as well as detection bias (Boakes et al., 2010; Troudet et al., 2017). Sampling biases result from unequal observation effort across species, space and time. Some places and species are sampled far more often than others. Biodiversity hotspots like the tropics are largely undersampled. Citizen science observations are classically concentrated at points and paths of interest, in spring or summer. Depending on the type of model, sampling biases can affect predictions to varying degrees. Standardised sampling protocols allow some compensation for sampling bias, e.g. through uniform or random sampling. Finally, detection bias also affects species observations (Botella, 2019). In a given place and time, species may be looked for but missed for various reasons: undetected presence, lack of expertise, fluctuating nature. Imperfect detection leads to omission errors (false negatives), which can damage the modelling process.

2.3.1.3 Models

Modelling species distributions involves learning a relationship between geolocated species observations and descriptors of the environment. Under precise ecological hypotheses, models can subsequently project the learned environmental preferences of species to spatially generalise their spatial distribution. In other words, SDMs are statistical tools that correlate species presence (and possibly absence) with environmental covariates. They can also be seen as a supervised learning problem, where the goal is to learn the function between species (labels) and their environment (qualified by some features).

Species distribution modelling is a very active area of research, boosted by new observational data, descriptive covariates and modelling techniques. We begin here with a formal definition of these models. The development of SDMs is underpinned by ecological concepts, methods and assumptions that are well summarised in (Soberón & Nakamura, 2009) and (Botella, 2019). Far from being able to do such a synthetic work in this section, we will next very briefly introduce the niche theory and the main assumptions of SDMs. We then describe the main properties, model classes and covariates of SDMs. Finally, we present their use in conservation planning. For a fluent overview on species distribution modelling, we refer the reader to the reference work of Elith and Leathwick (2009).

Formal definition

To simply define an SDM, we adapt the streamlined proposal from (Beery et al., 2021) to a presence-only dataset of C different species and n observations. In practice, other nested steps can occur but most models can be summarised by these three core elements: 1) species observation data, 2) a method to encode observation locations, and 3) a function mapping location encodings to species predictions:

1) *A dataset of species observations.* $\{(k_i, y_i)\}_{i=1}^{n} \subset \mathcal{K} \times \mathcal{Y}$
$\mathcal{K} = [0, 180) \times [0, 360) \times [0, 1)$ is the set of spatio-temporal locations including -but not restricted to- the dated locations of observations. $\mathcal{Y} = [\![1, C]\!]$ is the set of species labels.

2) *A location encoding function.* $h : \mathcal{K} \mapsto \mathcal{X}$
It is a simple function matching the geolocation of species observations with the chosen environmental features describing species habitat. $\mathcal{X}$ is commonly named the features space.

3) *A model mapping environmental encodings to species.* $f_\theta : \mathcal{X} \mapsto \mathcal{Y}$
θ is a parameter vector. The objective is then to optimise θ thanks to the environmental contexts of the observations and the species observed, to exploit this relation and predict likely species in previously unobserved environmental contexts. This can be framed as a supervised learning problem on the dataset $\{(h(k_i), y_i)\}_{i=1}^{n}$. The aim becomes to estimate θ to minimze the cost function over the $n' \leq n$ training set samples:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n'} J\big(h(k_i), y_i\big) \tag{2.2}$$

where J is a loss function (e.g. binary cross-entropy, mean square error, etc.) proportional to the prediction error.

In practice, model outputs are often conditional probability distributions of observing each possible species given an environmental context, and a thresholding step is required to select the most likely species (see Liu et al., 2005, 2013). Model outputs are often interpreted as habitat suitability or relative probability of presence conditioned on one observation in the case of presence-only data.

Niche theory and assumptions

The modelling of species distributions has its theoretical origins in Hutchinson's fundamental and realised niche concepts (Hutchinson, 1957). The *fundamental niche* is defined as "a hypervolume where a species can persist indefinitely in the absence of competition" (Hutchinson, 1957). In turn, the *realised niche* of a species is defined as "that part of the fundamental niche in which the species is not absent because of competition" (Hutchinson, 1957). The realised niche definition was subsequently refined to take into account the environments made available by colonisation-extinction phenomena and the species' dispersal capacities (active resource exploration for animals and dispersal mechanisms for plants). A third and more recent niche proposal is the *potential niche*, defined as "the set of environments where a species could survive for a reasonable lifetime if there were no dispersal constraints" (Jackson & Overpeck, 2000). There has been debate about what is modelled in SDMs, but no real consensus has been reached, partly due to the original ambiguities in niche concepts (Elith & Leathwick, 2009). The latter authors recommend "to retain a healthy scepticism about which components of the niche are represented by the predictions of an SDM". Araújo and Guisan (2006) argue for a simpler biological definition, where SDMs target "the environmental conditions that allow a species to satisfy its minimum requirements so that the birth rate of a local population is equal to or greater than its death rate". Finally, the binary nature of niches has been superseded by a more continuous *response function* between the likely presence of species and their environment. The environment is indeed highly spatially correlated, and we can easily see that in practice there are no fixed niche boundaries. It is ultimately this response function that SDMs aim to capture by correlating species observation data with a representation of the environment.

In modelling species distributions, we rely on two key hypotheses:

1. Species are in equilibrium with their environment (Araújo & Pearson, 2005). In other words, species are assumed to be observed in environments that are suitable for them (climate, resources, anthropogenic disturbances, etc.). The models actually learn the environmental preferences of species from their observed geolocations.

2. Relevant environmental gradients have been provided (Elith & Leathwick, 2009). In effect, SDMs correlate species sightings with the provided environmental descriptors, assuming that they play a role in species preferences. Models rely on any descriptor provided, as long as it can help to capture species dispersal. With a *reductio ad absurdum*, it was shown that using paintings (with patterns completely unrelated to species observations) as covariates could lead to correct SDM performance measured with the usual metrics (Fourcade et al., 2018). This is mainly due to the spatial autocorrelation of paintings - as with all environmental covariates - and highlights the need to provide relevant environmental covariates for SDMs.

SDM properties

Inspired by the review from Beery et al. (2021), here are some of the main SDM properties.

- **Presence-only & presence-absence models.** As mentioned above, dealing with one or the other type of observation data classically leads to the adoption of different models. However, research aims to bridge this gap and benefit from both the massive presence-only data and the highly informative presence-absence observations (Botella et al., 2023). When absences are not available, some methods use "pseudo-absence" or "background" points. These are artificial absence points that are often sampled at random in the study area or taking into account sampling bias (Botella et al., 2020; Phillips et al., 2009).

- **Single vs. multi-species models.** This is a key distinction from the SDM literature. Early models focused on one species at a time, while the trend has now moved towards modelling subsets of taxa. Increasing computational resources have logically played a key role in this shift.

- **Multi-species models: stacked vs. joint.** Again, this distinction is key to understanding the field. Stacked SDMs (SSDM) use single-species models before stacking their outputs (Schmitt et al., 2017). In contrast, Joint SDMs (JSDM) model multiple species in a single common step. Co-occurrence information and model parameters can be shared across species. Extensive data on more common species helps to shape the species environmental space and can ultimately benefit rare species modelling (Botella, 2019; Pollock et al., 2014; Zhang et al., 2020).

- **Spatially explicit models.** These models voluntarily include observations of geolocation within the model's covariates (Domisch et al., 2019). Unlike niche models, species are no longer modelled in a purely environmental space, but within an environment where location is also explicit. Araújo and Guisan (2006) specifies that niche models predict species *potential habitats* whereas spatially explicit models represent species *potential geographical distributions*. In practice, niche models *implicitly* incorporate spatial information through the spatial autocorrelation of environmental covariates.

- **Bias mitigation methods.** In response to the biases in observation data (see section 2.3.1.2), the community has developed a variety of methods to compensate for them. For example, sampling bias can be accounted for with Poisson point processes (Botella et al., 2021) and imperfect detection is modelled in site occupancy models (Bailey et al., 2014).

- **Uncertainty.** The use of SDMs in conservation decisions requires the modelling of error and uncertainty (Elith & Leathwick, 2009). Classically, we distinguish between uncertainty in the data (observations and covariates) propagating through the models, and uncertainty in the modelling process itself (Barry & Elith, 2006). (Beale & Lennon, 2012) Suggest a review on this still understudied topic. Zurell et al. (2020) propose a standard protocol for reporting SDM performance including uncertainty quantification. A convenient method to produce uncertainty estimates is to ensemble different models and examine the uncertainty around, for example, the median predictions (Marmion et al., 2009).

Model classes

Here is a brief description of the most commonly used models. For a more detailed overview, we refer the interested reader to (Botella, 2019). Classical methods, the widely used Maximum Entropy (MaxEnt) approach and ML methods are successively introduced. The aim is to capture potentially complex and non-linear relationships in a multi-dimensional environment. A recent study by Valavi et al. (2022) benchmarks the performance of numerous presence-only models of all types.

- **Classical statistical models.** Traditional species distribution models are i) the Generalised Linear Model (GLM) from Nelder and Wedderburn (1972), ii) its special case, logistic regression (Pearce & Ferrier, 2000), and iii) the non-parametric alternative called the Generalised Additive Model (GAM), Yee and Mitchell (1991). These methods use presence-absence data. GLMs are fitted using maximum likelihood and have the advantage of being both easy to implement and transparent. GAMs need fewer parameters to be adjusted and can express more complex functions without the risk of overfitting the training data. JSDM is an approach that considers species co-occurrence data in a multi-species SDM (Wilkinson et al., 2019). It was democratised by Pollock et al., 2014) and is based on a hierarchical multivariate GLM similar to logistic regression (probit regression).

- **The maximum entropy approach (MaxEnt).** This is the most commonly used model to infer species distributions from presence-only data (Phillips et al., 2006). Described by its authors as "a general-purpose machine learning method", it is based on Shannon's fundamental concept of entropy (Shannon, 1948). We present MaxEnt apart from ML models only because of its importance in the field of SDM. Its rationale is to maximise the entropy of the unknown species probability distribution while satisfying a set of constraints derived from the observed presence and background points (i.e. the maximum entropy principle, Jaynes, 1957). It is a single-species model. Elith et al. (2011) provided an alternative explanation of MaxEnt: the model minimises the relative entropy between two probability densities defined in environmental space: one estimated from the presence data and the other from the background -or landscape- points). The popularity of this method has been boosted by its ease of use and its ability to perform robust inference from limited occurrences (Phillips & Dudik, 2008).

- **Machine learning models.** The complex interactions between species and environment defy conventional statistical assumptions, such as linear covariate dependence or independent and identically distributed (i.i.d.) sampling. For this reason, machine learning methods have been adopted to model species distributions for decades. Decision tree algorithms have been appreciated for their impressive predictive performance, with first random forests (Breiman, 2001a; Cutler et al., 2007) and soon after Boosted Regression Trees (BRT, Elith et al., 2008). Support Vector Machines (SVM) have been used extensively since the early 2000s (Drake et al., 2006; Schölkopf & Smola, 2002) and, to a lesser extent, the k-Nearest Neighbours algorithm. Neural networks have also been used for a long time, well before the deep learning revolution (Özesmi & Özesmi, 1999). Deep learning has now brought new possibilities to improve multi-species distribution modelling, and this would be the specific topic of the Section 2.3.2. For more information, see the

impressive review by (Pichler & Hartig, 2023) and especially Table 1, which gives an overview of common supervised ML algorithms used in ecology.

The high predictive power of machine learning based models comes at the cost of model complexity and interpretability (Deneu, 2022). Model inputs cannot be easily linked to and explain prediction patterns. However, ecologists value transparent models that they can understand, as this trust and interpretability is crucial for policy decisions (see section 2.3.1.5). In response, extensive research is underway on interpretable machine learning or explainable AI (xAI) and will be further discussed in Section 2.3.2.1 (Murdoch et al., 2019). Some xAI studies are specifically focused on the field of species distribution modelling (Ryo et al., 2021).

The challenge of evaluating SDMs is not addressed in this state of the art. However, it is partially covered in the chapters and in the final discussion. For a clear and concise review of SDM metrics, we again refer to the article by Beery et al. (2021) and Allouche et al. (2006) for presence-absence models. Before diving into deep-SDMs, we will now review common covariates and how SDMs can inform conservation, with a particular focus on IUCN assessments.

2.3.1.4 Covariates relevance

Providing relevant environmental covariates to enable models to estimate species' sound environmental preferences and sensitivities is a fundamental assumption underlying distributional modelling. Mod et al. (2016) conducted a review of covariates from published plant SDM studies (2010-2015; n = 200). They concluded that i) ecophysiologically relevant environmental variables were neglected in the majority of studies, resulting in incomplete niche quantification and limited predictive power, ii) some of the missing predictors are already available across scales (e.g. soil moisture), while others are not (e.g. soil pH and nutrients), and therefore more attention should be paid to their development. This reveals a first characteristic of environmental predictors: They can have direct or indirect effects on species, i.e. they can be arranged on a gradient from proximal to distal predictors (Austin, 2002; Guisan & Thuiller, 2005). In addition, the latter authors classify environmental covariates as having three main types of influence on species: i) *limiting factors*, defined as factors that control the ecological physiology of species (e.g. temperature, water, soil composition), ii) *disturbances*, defined as any perturbation of environmental systems (natural or anthropogenic), and iii) *resources*, defined as all compounds that organisms can assimilate (e.g. energy and water). Another distinction concerns the format of covariates: they can be either *continuous*, e.g. temperature, or *categorical*, e.g. land cover. Most models are only adapted to handle continuous variables, so an adaptation step (or embedding) is needed to take advantage of categorical data.

Let's briefly review key environmental covariates. This is not intended to be an exhaustive list, but rather a synthetic entry point into SDM covariates.

- **Climatic variables.** They are the most widely used environmental covariates for two reasons: temperature and precipitation play a key role in the nature of ecosystems, and they are readily available at global and kilometre scale resolution (Fick & Hijmans, 2017). They are derived from a dense network of weather stations

(between 9,000 and 60,000 for the 1970-2000 averages) whose observations are interpolated with covariates such as altitude, distance from the coast and satellite information. In addition, global climate models are used to generate future climate variables and project impacts on biodiversity (see Section 2.4.3).

- **Pedological variables.** Plant growth is naturally dependent on soil characteristics. For example, the presence of mycorrhizal fungi in the soil affects the distribution of orchids (McCormick et al., 2018). *SoilGrids* produces global soil information up to 250 m spatial resolution, including soil organic carbon content, total nitrogen and pH among others (Poggio et al., 2021). Again, this scale is achieved through machine learning based interpolation from 240,000 sites and over 400 global environmental covariates. In a study by Descombes et al. (2020), combining CS data with edaphic and climatic conditions proved to be a cost-effective and relevant approach to model the distribution of Swiss plants.

- **Remote sensing imagery.** Remote sensing offers new opportunities to model species distributions. We have already motivated the use of remote sensing in the opening context of the manuscript section 1.1. We can add that, unlike climatic and pedological variables, RS imagery provides raw (non-interpolated) and extremely rich information on a global scale. Indeed, RS-derived covariates are used to interpolate station-based measurements around the world to reach the global scale. In addition, the increasing spatial, temporal and spectral resolution of satellite imagery places RS in a central position for large-scale distribution modelling. Thus, the contribution of satellite imagery time-series to SDM performance is analysed in chapter 3. Finally, remote sensing is crucial for SDM covariates in at least two other ways. First, the spatial context around species habitats has proven valuable for mapping species distributions (Deneu et al., 2021b). As described in section 2.3.2, convolutional neural networks can capture valuable spatial patterns of the environment around species sightings. Second, remote sensing is the source of information for many derived covariates. These include vegetation indices, land cover and human footprints. However, there is an abundance of RS-derived environmental covariates, see for example the European catalogue *EcoDataCube*[12].

  - **Vegetation indices.** From a computer vision perspective, they are essentially hand-crafted features of remotely sensed data that are used to qualify plant properties (Beery et al., 2021). The most famous example is the Normalised Difference Vegetation Index (NDVI). It is calculated from the red and near-infrared spectral bands and quantifies the health and density of vegetation. One motivation behind such an index is to make rich but heavy satellite data more manageable.

  - **Land use and land cover.** These are two close concepts, with the difference that land cover qualifies the physical nature of a terrain, whereas land use is intended to represent its function. They are typically categorical data. The MODIS global kilometre-scale land cover map is a prime product example (Friedl et al., 2002). While segmentation and classification have traditionally been performed with supervised algorithms, the addition of a self-supervised

[12] https://stac.ecodatacube.eu/

pre-training step holds promise to 1) achieve better performance and 2) reduce the need for labelled data (Scheibenreif et al., 2022).

– **Human impact.** Human settlements and activities exert strong pressure on biodiversity. For example, in a study of 4,867 terrestrial mammals, Di Marco and Santini (2015) showed that climatic variables and human pressures were the most influential predictors of geographic range size, ahead of biological traits. However, quantifying human impacts is a complex task that requires the integration of different sources of information. The Human Footprint Index is the reference product at the kilometre scale (Venter et al., 2016). It is consistent for 1993 and 2009 to allow for change detection and uses both satellite imagery and systematic ground surveys. Variables include 1) built environment, 2) population density, 3) electrical infrastructure, 4) cropland, 5) pastureland, 6) roads, 7) railways and 8) navigable waterways. Cumulative scores combine these rasters with a biome normalisation (Olson et al., 2001). In addition, recent work by Marconcini et al. (2020) combines open and free optical with radar satellite imagery to map human settlements globally at an unprecedented resolution of 10 m. Finally, human impact is not confined to land, with 41% of the world's marine ecosystems heavily impacted by multiple anthropogenic drivers (Halpern et al., 2008).

To conclude on SDM covariates, here are two important insights for us.

1. First, the influence of covariates is *scale and context dependent* in addition to being taxon dependent. At broad scales, climate was consistently found to be the main determinant of species distribution (Randin et al., 2020). However, at finer scales, non-climate predictors such as land cover (Luoto et al., 2007) or remotely sensed imagery (Deneu et al., 2021a) provide increasingly valuable information. They are also context dependent. We present an example where the relative importance of covariates varies with altitude. In a study of 2,616 vascular plant species in the European Alps, Chauvier et al. (2021) indeed showed that:

- Climate, although overall the most influential driver of spatial patterns, had decreasing importance from low to high elevation due to increasing species endemism and climate homogeneity with elevation.
- In compensation, the relative importance of soil and land cover increased with increasing altitude.
- Land cover shows strong and local effects in the lowlands due to human influence.

These authors concluded that while disentangling covariate effects remains a challenge, including soil and land cover covariates in species distribution models can *markedly benefit* predictions. This brings us to our second point.

2. The need for covariate selection depends on the model and the objective. Indeed, while traditional statistical models can be affected by the dimensionality of the covariates (curse of dimensionality, Giraud, 2021), the performance of machine learning based methods is not affected by inputs of very high dimensionality if

properly parameterised. Therefore, the selection of uncorrelated inputs is usually required in traditional statistical models. However, thanks to modern technologies, the number of potentially relevant SDM covariates is very large. Including them all allows learning from their innovative properties as well as from their interactions. This should ultimately lead to a better capture of species' environmental preferences (Botella, 2019). This is a key motivation for the use of deep-SDMs, which will be further discussed in Section 2.3.2.1. Nonetheless, this flexibility often comes at the expense of interpretability. Consequently, the objective of the application should be clearly identified before selecting models and covariates.

2.3.1.5 SDMs to inform conservation

In the first Section 2.2, we have seen how beneficial species range can be to construct indicators for biodiversity conservation at all levels (extinction risk assessment, habitat classification, mapping threat sources, etc.). Here, we will first look at how SDMs are used to support the red listing of species according to IUCN criteria. We then discuss other ways in which SDMs can be used to inform conservation decisions.

Assisting IUCN Criteria Assessment with SDMs

As introduced in section 2.2.1.1, the IUCN extinction risk assessment is based on five criteria, which include population size and dynamics, geographic range, and direct estimates of the probability of species extinction. SDMs have been used to estimate variables from criterion B on geographical range: EOO (Syfert et al., 2014), AOO (Jiménez-Alfaro et al., 2012) or both (Breiner et al., 2017; Moat et al., 2019). However, there are drawbacks to this approach. SDMs may tend to overestimate true range size if suitable habitat remains unoccupied due to ecological and/or biogeographical constraints (landscape barriers, biological interactions, dispersal limitations, see Guisan et al., 2013). Similarly, species ranges may be underestimated if known occurrences only partially cover the realised niche of a species. The IUCN actually provides official guidelines to be followed when estimating criterion variables from SDMs (Bland et al., 2017).

Approaches such as the GeoCAT and ConR packages compute EOO and AOO directly from species observations, without generalising species niches (Bachman et al., 2011, 2020; Camacho & Peyre, 2022; Dauby et al., 2017; Levin et al., 2022; Stévart et al., 2019). Santini et al. (2019) go a step further by estimating all five IUCN criteria from data on land cover change, species habitat preferences, population, and dispersal capacity. All these approaches respecting the official criteria thresholds for entering risk categories were called *index-based* methods by (Zizka et al., 2020). They are in contrast to the *prediction-based* methods that directly learn a correspondence between species features and status (see Section 2.4.2 for an overview).

Overview of conservation opportunities

SDMs are used for a wide range of conservation decisions (Sofaer et al., 2019). For a comprehensive overview, we refer the interested reader to two references: Guisan and Thuiller (2005) and Guisan et al. (2013). We will only touch on some of the main applications. The prediction of climate change impacts through the modelling process

can potentially benefit all of these applications (see Section 2.4.3).
A first application is the management of biological invasions. In Mexico, for example, SDMs have been used to predict the potential impact of the invasive moth cactus on native cacti (Soberón et al., 2001). In addition, a recent study employing SDMs highlights the role of human-mediated long-distance dispersal in plant invasions (Botella et al., 2022). SDM projections can also be used to identify critical habitats for threatened species. This was the case in Canada for the Ord's kangaroo rat (Heinrichs et al., 2010). Optimising PA implementation is a complex process with multiple inputs, as highlighted in Section 2.2.3.3. SDMs allow multiple aspects of biodiversity (species richness, threatened species, ecological or evolutionary originality, etc.) to be mapped before these outputs are fed into the optimisation process (see Section 2.4.4 for an overview). A recent example is the identification of priority areas for endangered plant species in Asian headwaters (Han et al., 2019). Other important applications include the translocation of species to survive pressures (Chauvenet et al., 2013) or natural disasters, and the orientation of prospective sampling based on model predictions of rare species (Le Lay et al., 2010). More indirect, but still instrumental, uses of SDMs for conservation planning include testing evolutionary hypotheses (Graham et al., 2004) or identifying disease reservoirs (Peterson et al., 2002).

We have now reviewed SDM principles and how these models can benefit conservation. As stated throughout the manuscript, deep learning brings new possibilities to the field by making the most of unprecedented amounts of data on biodiversity. This is the subject of the next section, which focuses on deep-SDMs.

2.3.2 Deep-SDMs

2.3.2.1 Motivations

Before introducing the concepts behind the efficiency of deep learning, we will first go back in time and present how machine learning and deep learning emerged. Finally, the trade-off between model performance and interpretability will conclude this section.

A brief history of machine learning
Machine learning relies on modern statistical principles. Maximum likelihood estimation, for example, dates back to the early 1920s, as depicted in Figure 2.11. The core principle of classic statistics is the assumption of an ideal data generating model, e.g. a Gaussian distribution, which allows parameters and probabilities to be estimated. Then the advent of computers and robust numerical algorithms such as Markov Chain Monte Carlo allowed more complex statistical models to be developed. However, the inference of traditional statistical methods is conditioned on simple model assumptions, making it difficult to approach the complexity of the natural world (Breiman, 2001b).

In the 1980s, the increasing availability of computing resources allowed to refine the numerical solutions of classic statistical methods, but also to develop a new modelling paradigm where the formulation of the data generation process is abandoned (Pichler & Hartig, 2023). Instead, machine learning is trained on data to perform a supervised

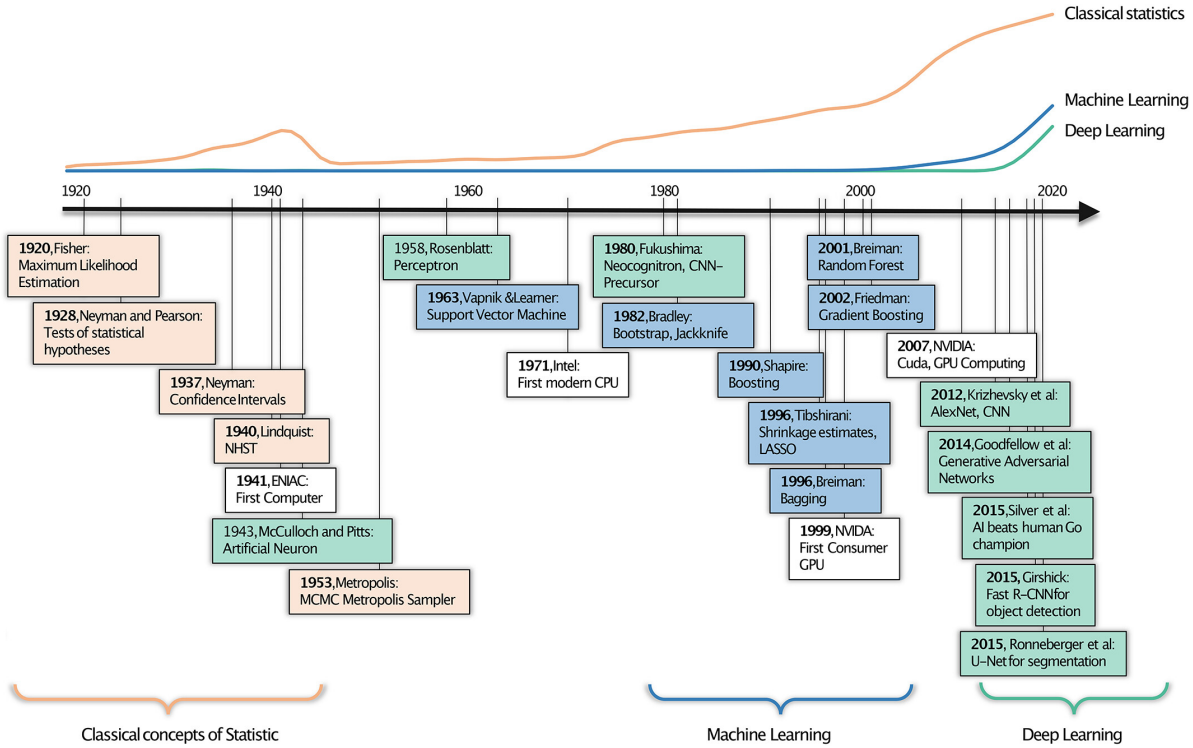


Figure 2.11: *The three eras of statistical learning. The classical concept of statistics was developed between the 1920s and the 1940s. Common machine learning algorithms or techniques were then discovered between 1980 and the early 2000s. While the theoretical foundations of deep learning were postulated in the 1960s, it has only gained popularity in recent years. The trend lines above the timeline correspond to the frequency of occurrence of each term in the scientific literature. Reproduced from (Pichler & Hartig, 2023), CC-BY license.*

(mainly classification or regression) or unsupervised task (dimension reduction, clustering, etc.) with the goal of minimising a given loss function. The objective of supervised machine learning is to use a training data set to learn a response function that generalises well to unseen data. In fact, ML models are excellent approximators of any measurable function, given inputs and target outputs. Hornik et al. (1989) showed that a neural network with enough neurons is a class of universal approximators. However, this comes at the cost of increasingly complex and over-parameterised methods that fit the training data but have become too specific and unrealistic to generalise to unseen data. This classic but undesirable phenomenon, called *overfitting*, has given rise to a number of compensatory *regularisation* techniques. Their aim is to avoid learning a response function that is only specific to the training data by limiting the complexity of the model. In addition, models are evaluated on unseen data (validation or test set, depending on the model), since the aim is to obtain a general, *flexible* function. The trade-off between model complexity and flexibility actually corresponds to the classical bias-variance trade-off discussed below, see Figure 2.12.

Deep learning is the branch of machine learning that deals with large amounts of data. It is distinct from ML because 1) it emerged later due to technical challenges, and 2) new principles governing model convergence have been identified. In practice, deep learning models are very large Neural Networks (NN), with hundreds to billions of

parameters. These architectures were described in the 1980s and 1990s (see Figure 2.11), but computational resources and optimisation techniques were too limited to train such a model. Finally, deep learning took off in the last decade thanks to three major technical breakthroughs (Botella, 2019):

i) The creation of huge learning datasets such as ImageNet (Deng et al., 2009)

ii) The democratisation of Graphical Processing Unit (GPU) computing since (Krizhevsky et al., 2012)

iii) Many advances in NN optimisation techniques such as the use of the ReLU activation function to speed up training and combat vanishing gradients (Nair & Hinton, 2010) or batch normalisation (Ioffe & Szegedy, 2015).

On the reasons behind deep learning performances

Considering the classical bias-variance trade-off that rules statistical and machine learning models, deep learning appears to be a puzzle. Indeed, according to this principle, the optimal trade-off between model error (or bias, computed on training data) and model variance (computed on unseen validation data) lies at intermediate model complexity, see Figure 2.12 **(a)**. This trade-off predicts that excessively large or *expressive* models (Raghu et al., 2017) should overfit the training data and thus generalise poorly. However, the distinctive behaviour of deep learning models is precisely to overcome this trade-off and continue to generalise well as model complexity increases, see Figure 2.12 **(b)**. Precise explanations of this phenomenon, often referred to as deep learning *double descent*, are still being debated by the research community (Sejnowski, 2020). An important distinction between *in-distribution* and *out-of-distribution* predictions is also acknowledged, with the latter logically experiencing higher generalisation error and uncertainty, see Figure 2.12. **(c)** and **(d)**. The field of domain adaptation focuses on improving these out-of-domain predictions, see (Farahani et al., 2021) for a review.

The commonly accepted reason is the combination of i) the efficient optimisation process based on Stochastic Gradient Descent (SGD), ii) DL expressivity resulting from the innumerable function compositions, and iii) the set of regularisation techniques applied. Regularisation prevent the model from learning an over-parameterised response function that reflects the training data. As a result, the number of parameters in a model is acknowledged to be a poor measure of its effective complexity, and various alternatives have been developed (Birdal et al., 2021). Regularisation techniques are numerous and intervene at different stages. For example, many ML algorithms produce ensemble predictions (Pichler & Hartig, 2023). The normalisation of layer inputs throughout the model, a widely used technique called batch normalisation (Ioffe & Szegedy, 2015), not only helps convergence but also has a strong regularisation effect. Furthermore, as DL models rely on large data to avoid overfitting, a number of data augmentation techniques have been developed to increase the size and quality of training datasets (see Shorten and Khoshgoftaar, 2019 for a survey on image data). In addition, dropout is a technique that consists of successively freezing random parts of the network during training (Srivastava et al., 2014). It can be interpreted as the indirect generation of many subnetworks that work together to regularise the output. Another well-known regularisation mechanism is the use of shrinkage penalties. These are additional terms in the loss function that bias

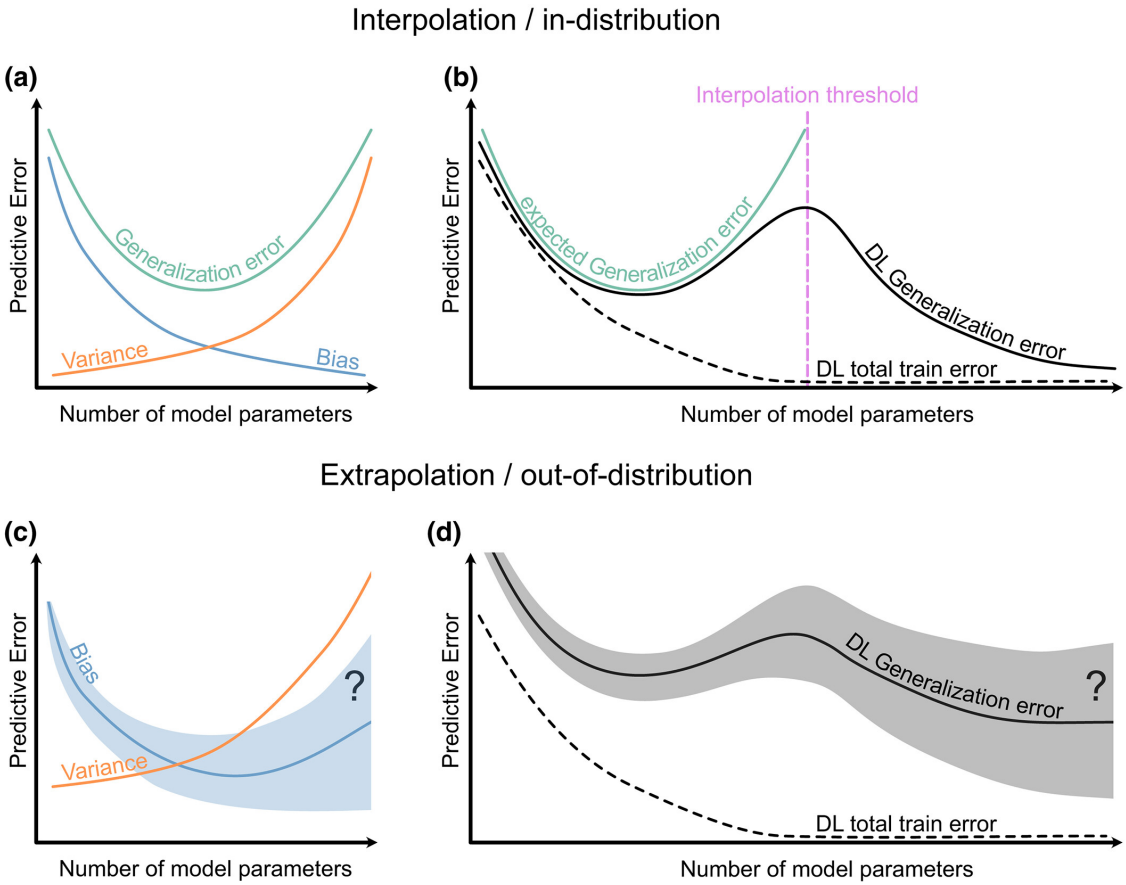


Figure 2.12: *Typical bias-variance trade-offs in classical machine learning* ***(left)*** *and deep learning* ***(right)*** *models for interpolation and extrapolation tasks. In contrast to the classical bias-variance trade-off in panel* ***(a)****, the bias-variance trade-off for DL in panel* ***(b)*** *shows that after the interpolation threshold (pink dotted line), the training loss is constant (i.e. bias is not improved by increasing model complexity), but the test loss (and thus variance) can still be reduced by increasing model size. For extrapolation tasks (**c** & **d**), the total generalisation error is usually higher and the optimal model complexity lower. Reproduced from (Pichler & Hartig, 2023), CC-BY license.*

the parameters updating to a certain value. L1, L2 and their combination elatic-net (Zou & Hastie, 2005) are typical examples, pushing the parameter norm to zero and limiting weight updates from being too specific.

Still, the regularisation techniques do not fully explain why the DL generalisation performance continues to improve after the interpolation threshold (see Figure 2.12 **b**), i.e. after the model has reached its minimum bias. In the case of multi-class tasks, Poggio et al. (2017) conjecture that i) SGD has a strong regularisation effect (see Ruder, 2016 for a review on gradient descent optimisation) and ii) the compositional nature of multi-class response functions plays a key role in DL success. In other words, deep neural networks are well suited to estimating complex response functions composed of many local and simpler functions that integrate lower-dimensional subsets of the input. In addition, deep learning seems to have an additional ability to share information between the classes involved, allowing well-represented classes to benefit classes with few samples.

A trade-off between performance and interpretability

Predictions in ecology and species distribution modelling can inform decision making

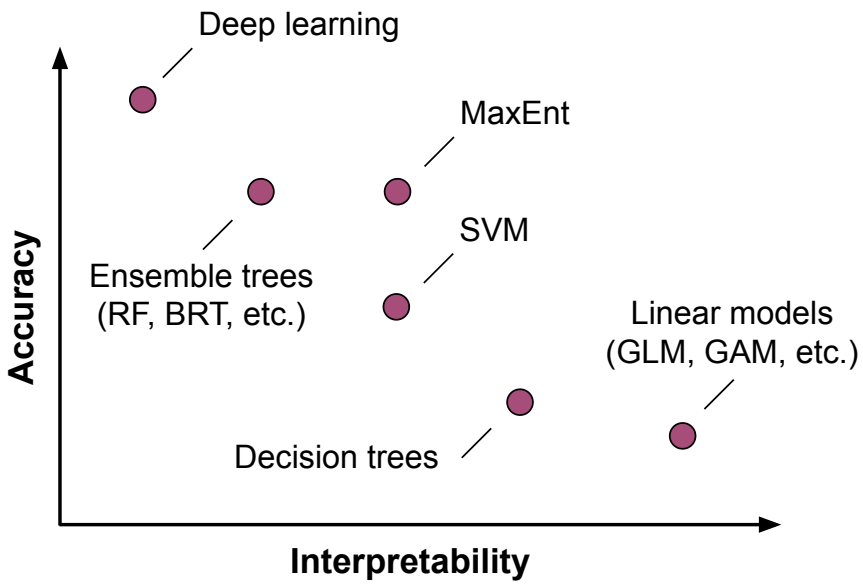


Figure 2.13: *Schematic trade-off between accuracy and interpretability of models. Overall, interpretability is inversely proportional to achievable accuracy. In this figure, the relative positions of the models are partly derived from (Valavi et al., 2022). However, they should not be considered as ground truth as they also reflect subjective views. Performance is application and scale dependent.*

(Sofaer et al., 2019), see section 2.3.1.5. In this sense, model outputs are expected to be transparent, understandable or at least interpretable. However, while deep learning has higher predictive power than other ML methods, it suffers from low interpretability, as shown in Figure 2.13. We have seen in the previous paragraph that DL's impressive generalisation power is still not fully understood. DL's poor interpretability is related to this point, and progress in one direction should benefit the other. While certain ML algorithms provide metrics of feature importance, e.g. RF from Breiman (2001a), they do not provide simple effect estimates, nor do they provide measures of confidence such as confidence intervals or p values (Pichler & Hartig, 2023). This has led to the development of xAI, the subfield of machine learning that is interested in explaining how models rely on inputs to make their predictions (Arrieta et al., 2020), the ultimate goal being to understand how the model works. A popular option for interpreting predictions is to use the xAI unified framework called SHAP (Lundberg & Lee, 2017). The majority of xAI tools are post-hoc, i.e. they operate once the model has been trained. They can be either model agnostic (valid for any model) or model specific. To get a sense of the contribution of features, another option is to include/exclude or randomise groups of features and run independent trainings to see how performance is affected (e.g. Han et al., 2015). Lucas (2020) provides an overview of interpretable machine learning in ecology and Ryo et al. (2021) focuses on xAI to interpret species distribution modelling. Finally, a promising research direction is the application of causal inference concepts to help interpret ML models and obtain proper effect estimates (Gonzalez et al., 2023; Zhao & Hastie, 2021).

Deep learning for species distribution modelling
Traditional statistical SDMs typically assume a relatively simple relationship between the outcome (e.g. presence or abundance of a species) and environmental covariates. Given a site represented by different values of environmental covariates, the response function is classically modelled as the sum of univariate covariate functions and simple bivariate functions (typically products) representing pairwise covariate interaction effects (Botella et al., 2018a). The strength of these traditional models, such as GAMs or MaxEnt, is the ability to isolate and understand the influence of each covariate on the model prediction. However, the complexity of the biotic and abiotic interactions behind the presence of plants certainly cannot be reduced to such simple functions. In this context, ML models and especially deep learning offer the possibility to learn complex and non-linear dependencies between environmental patterns and species presence. Indeed, as developed in the previous paragraph, DL models show exceptional generalisation power for multi-class classification problems such as multi-species distribution modelling.

A common principle in the multi-species SDM literature is to share most of the parameters between all species studied (Leathwick et al., 2006). This idea allows both to limit the number of model parameters - which was a bottleneck before the advent of advanced computing resources and optimisation techniques - and to share a common representation space of reduced dimension between species. In the case of NN architectures (see next section 2.3.2.2), it consists of shared layers throughout the model. The final fully conected NN layer constrains the different species responses with parameters common to all species. This results in species responses that are 1) organised along common environmental concepts and 2) hopefully crystallise the environmental information responsible for their likely presence/absence.

Another limitation of traditional models is their inability to capture the spatial patterns described by environmental variables. The rationale behind using Convolutional Neural Network (CNN) models to map species distributions is that spatial patterns (a water network, a forest edge, etc.) can have a significant impact on species occurrence. Such patterns cannot be inferred from point values of environmental covariates. Therefore, our choice is to build SDMs taking as input maps (2D arrays) for each environmental covariate. Captured with CNNs, such spatial and multidimensional patterns in environmental covariates have been shown to be informative about species distributions (Deneu et al., 2021b). CNNs are specific NN architectures originally developed for image analysis, see next section 2.3.2.2. They are theoretically able to capture patterns when applied to multi-dimensional spatial rasters, making them perfect candidates for modelling species distributions. CNNs have in fact found a variety of relevant applications in ecology, as reviewed in (Brodrick et al., 2019).

In conclusion, we believe that CNNs are especially appropriate for modelling species distributions because of their ability to 1) model complex and non-linear relationships between covariates, which are more likely to approximate the response of species to their environment, and 2) identify multidimensional and spatial patterns in environmental descriptor images (Botella et al., 2018a). Furthermore, the architecture of deep learning models can be easily adapted to combine different types of inputs. As a result, they are particularly suited to integrating and harnessing heterogeneous information, as illustrated in section 2.4.1 and finally discussed in the concluding chapter 6 of this manuscript.

2.3.2.2 Convolutional neural networks

Introduction

Convolutional neural networks are a class of deep learning models designed specifically for image processing. They can perform tasks such as image classification, segmentation and object detection thanks to their ability to capture shapes, edges, gradients and textures. Thanks to these capabilities, their use in ecology has exploded in recent years (see Figure 2 from Christin et al., 2019). Their development responded to two technical bottlenecks that prevented Fully Connected (FC) networks from processing very high-dimensional inputs such as images. First, FC networks assign weights to each input dimension and connections to all neurons from one layer to another (see definitions in next paragraph), quickly resulting in millions of parameters to be updated via backpropagation. This leads to intractable optimisation problems, especially with the computational resources available at the time. For example, a linear layer with a 256×256 RGB image as input and a similar output format would require $(256 \times 256 \times 3)^2 \simeq 3.87e + 10$ ($\simeq$150Gb!) parameters. Second, FC models are sensitive to translation. This problem alone makes it impossible to use FC models to analyse 2D arrays. Indeed, images have an unstructured format where the relevant information (e.g. a bicycle to be detected) can have very different positions, orientations and scales. However, if FC models have learnt to recognise bikes of a certain size and in the centre of the image through the training set, they would not be able to recognise bikes in any other configuration (i.e. spanning other pixels). In 1989, LeCun solved both technical locks with the development of CNNs (LeCun et al., 1989). Combined with the increasing availability of training datasets, computing resources and modern optimisation techniques, these models have since become state of the art for image processing tasks (Krizhevsky et al., 2012). More recently, models based on attention mechanisms such as the vision transformers are overtaking purely convolutional models in image classification tasks (Dosovitskiy et al., 2020a). Opportunities in ecology and species distribution modelling opened up by new DL architectures will be discussed in the final chapter 6. Here, we will only review some of the basic principles of deep learning and CNN architecture. For a comprehensive introduction to the field, please refer to:

i) the https://fleuret.org/dlc/ free and online deep learning course from François Fleuret for a complete introduction (CC BY-NC-SA license)

ii) *The Little Book of Deep Learning* (Fleuret, 2023) for a more gentle and concise introduction by the same author (a book that you can even easily read on your phone!)

iii) Borowiec et al. (2022) and Pichler and Hartig (2023) reviews for an introduction of DL as a tool for ecology. Moreover, Desprez et al. (2023) provide nine useful tips to help ecologists in implementing machine learning models.

CNN architecture

Here we present only some of the DL principles that we believe are essential for understanding and appreciating CNNs. First, the scalars of an input vector pass through **layers** of **neurons**. A neuron is a composition function made up of a linear combination of its inputs followed by a non-linear **activation** function. The weights assigned to each

input are parameters of the model, which are optimised by the **backpropagation** step. Classically, layers can be fully connected, i.e. composed of neurons with receiving fields covering all the outputs of the previous layer, or **convolutional**. Together with other operations such as pooling (aggregation) layers, which reduce the input dimension, they transform the input information through successive simple and parametric functions.

A **convolutional layer** is composed of a set of parametric kernels (or filters) spanning only local parts of their inputs with, like a neuron, a linear combination of its receptive field followed by an activation function, see the example Figure 2.14. They are applied

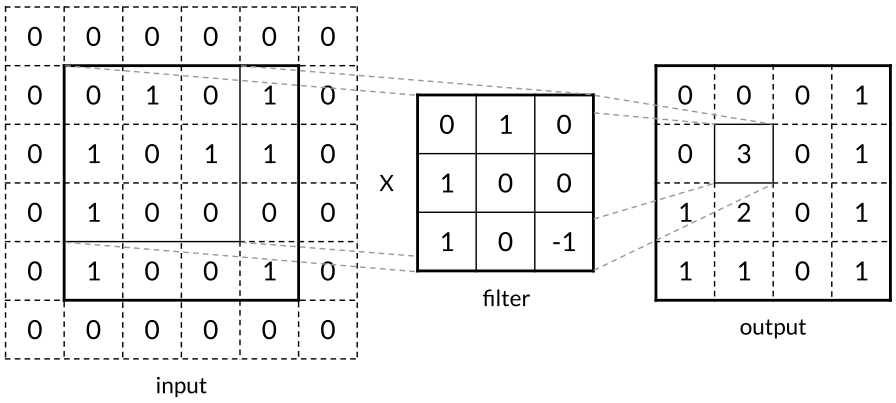


Figure 2.14: *Convolution example using a size 3 filter (padding of 1, step of 1). In this simple scheme there is no bias and the activation function is the identity function. Reproduced from (Deneu, 2022) with permission of the author.*

successively to all input pixels like a sliding window. Each kernel has a user-specified size (typically ranging from 3 x 3 to 11 x 11 pixels) and is applied to the inputs with common parameters. The fact that the parameters of each kernel are common to all input values when successively applied with a sliding window is the innovation that allows to massively reduce the number of parameters and thus the number of images to be processed.

In a supervised learning task, the model output obtained after all model transformations is compared to the target reference. This is done by applying a **loss function** that reflects how different the two entities are. Next, a mechanism is needed to find the set of model parameters that minimises the loss function. In other words, once the loss value has been computed for a set of training samples, the influence of each parameter on the error needs to be traced back. This is done via chain rule calculus and consists of finding the gradient of the loss function with respect to the trainable network parameters (LeCun et al., 1998). This mechanism is called **backpropagation**. Finally, the parameter values are updated as a function of (at least) the gradient and a **learning rate** setting the update step size. This process is highly iterative and is called **stochastic gradient descent** (SGD) when it is performed for each training batch (Amari, 1993). One **epoch** corresponds to one iteration of this process over all the **training set** samples. Tens to hundreds of epochs are run to optimise the weights of the model.

Between epochs, the model is frequently used in prediction-only mode (no backpropagation, only the forward pass) on an unseen **validation set** to evaluate its generalisation power. The best model is selected depending on the model validation performance computed along different **metrics**. The reported model performance is computed on a third independent set called **test set**. In practice, for a large dataset, the training, validation and test sets are split randomly or with a custom strategy, with fixed proportions such as 90/5/5%. The partitioning can be designed to avoid performance bias that occurs when the three sets are not completely independent. For geospatial data, it is common to perform a spatial split to reduce the spatial autocorrelation between the covariates sampled in each set.

Limits

CNNs also have some limitations that need to be acknowledged. First, training a CNN requires the manual setting of many hyperparameters adjusting the learning process. They can be set via a **grid search**, i.e. an expensive performance test that evaluates the model's ability to generalise after it has been trained with different hyperparameter values each time. Still, setting hyperparameter values can be challenging because of the number of possible combinations. The influence of the hyperparameters on the model is crucial, from the split of the dataset to the regularisation techniques and the optimisation process. Second, even though convolutions have made it possible to drastically reduce the number of parameters to be trained compared to an FC model, CNNs remain deep learning methods that can quickly become computationally expensive to optimise. In particular, they require the use of GPUs to reduce training times to reasonable orders of magnitude. Although this resource is increasingly available, as mentioned above, it can still be a source of complication. In addition, CNNs suffer from still limited interpretability, see *trade-off between performance and interpretability* in section 2.3.2.1 and the discussion 6.2 for more details. Finally, for deep-SDMs, translation invariance has the added advantage of compensating for the geolocation uncertainty of the observations. However, it may not be advantageous if the observation is central and the covariates have a large spatial extent. Indeed, in this case, the central environmental information is likely to be more informative than the peripheral information. The use of a modern CNN architecture that exploits attentional mechanisms, i.e. an adaptable receptive field, could help to overcome this concern.

2.3.2.3 Resources and inference

Deep-SDMs demonstrate their capabilities through the use of large datasets and specialised tools. Analysis-ready datasets are rare, with the exception of the GeoLifeCLEF challenge. Here we first present available resources and then illustrate how models can be used for spatial inference.

Datasets

The GeoLifeCLEF challenge (Botella et al., 2023) is the only analysis-ready dataset for the development of deep-SDMs. The 2023 edition includes five million presence-only plant observations distributed across Europe, associated with high resolution rasters:

remote sensing imagery, land cover, elevation, and coarse resolution data: climate, soil, and human footprint variables. The models are evaluated on 22,000 plots based on standardised surveys. Moreover, Gillespie et al. (2022) constituted an ambitious dataset of over half a million CS plant observations across California, paired with 1-metre resolution satellite imagery. The scripts to regenerate the dataset are available on the author's github. Another option is to manually associate environmental data with large datasets of species observations (see Section 2.3.1.2). In fact, this is how the datasets mentioned above were created. GBIF, for example, collects millions of geolocated observations. Finally, we can imagine using traditional analysis-ready datasets for multi-species SDMs such as (Norberg et al., 2019) or (Elith et al., 2020) to fine-tune deep learning models. Fine-tuning is a transfer learning approach where the weights of a model trained on a large dataset are reused and partially retrained on a smaller dataset for a specific task.

Modelling frameworks

Deep learning models are typically developed using PyTorch or TensorFlow Python libraries. GPUs can be accessed locally, on a private remote server, or online with initiatives such as google colab (Bisong, 2019). More recently, a PyTorch library called TorchGeo (Stewart et al., 2021) provides new tools to ease the manipulation of geospatial data in deep learning. In the field of deep-SDMs, a new framework called *Malpolon* facilitates model training and is available on GitHub at https://github.com/plantnet/malpolon. Finally, online resources on deep learning for ecology, such as the https://ecostat.gitlab.io/ website, can be a great help in getting started with deep-SDMs.

Model examples and spatial inference

Under appropriate modelling hypotheses, deep-SDMs can be used to infer the spatial distribution of species. Here, we briefly review the studies that successfully measure and illustrate their spatial generalisation power. In particular, we refer to the various distribution maps drawn in the articles, with links to the online versions. Their spatial resolution varies from kilometres to metres.

- Harris (2015) illustrated early on that realistic species assemblages could be predicted using neural networks. Their model can take advantage of complex, non-linear relationships between species occurrences, biotic interactions (via co-occurrences) and the environment.
- Chen et al. (2017) concatenated NN-based embeddings of environmental covariates with bird species co-occurrence data to feed a competitive multi-species SDM. The idea that justifies the separate modelling is that 1) species may share similar environmental preferences while exhibiting distinct inter-species associations, and 2) these two pieces of information have different spatial resolution.
- Botella et al. (2018a) showed that predictions of observed plant counts across France were closer to the ground truth using a CNN model than either MaxEnt or a FC model (see examples in their Figures 4-5).
- Deneu et al. (2021b) showed that deep-SDMs are particularly efficient for modelling species that are poorly represented in the training set (likely to be rare). With

only one training occurrence in France, their CNN was able to produce a response function representative of the species distribution (model output was compared with independent observations, see their Figs. 6-7).

- Deneu et al. (2022) illustrated with 1-metre resolution satellite imagery that deep-SDMs can capture landscape and habitat information at very fine scales. In particular, their Figure 7 shows coherent predictions of coastal species at 50 m resolution.
- Gillespie et al. (2022) showcase the interest of deep-SDMs with different high resolution (256 m) applications: prediction of keystone species (Fig. 2), but also detection of ecosystem changes from wildfires and urban biodiversity hotspots (Fig. 3).
- Brun et al. (2023) used a deep NN to jointly model the distributions of 2,477 plant species represented by 6.7 million Swiss observations. At a spatial resolution of 25 m, they estimated: species distributions, community composition, but also the seasonal variation of observation probability, i.e. a proxy for flowering phenology, see their Fig. 2.
- Finally, Bourhis et al. (2023) propose a novel NN architecture that combines species trait information with environmental covariates. Applied to UK butterfly and moth datasets, they predict species occurrence probabilities (see their Fig. 3) and use a popular interpretation method (Lundberg & Lee, 2017) to investigate the trait-mediated and species-specific model outputs.

The final section of this state of the art focuses on data-driven insights for applications in ecology and conservation. Building on the concepts introduced above (conservation indicators, SDMs and machine learning), we cover a selection of topics where the generalising power of Artificial Intelligence (AI) can make a difference to biodiversity conservation.

2.4 Insights for conservation planning

Our closing objective is to outline a number of areas where AI can benefit biodiversity conservation. Unlike previous efforts, the focus is on applications rather than concepts. Although some of these involve modelling species distributions, we do not limit this literature review to SDM-related studies and explore broader issues.

We will commence by identifying review articles interested in AI for ecology and conservation, and a selection of outstanding applications. Next, we will present data-intensive approaches for predicting the IUCN extinction risk status of unassessed species. The third focus will be on the future trajectories of conservation indicators influenced by climate change. Finally, optimising strategies for spatial conservation will be briefly introduced.

2.4.1 AI for ecology and conservation

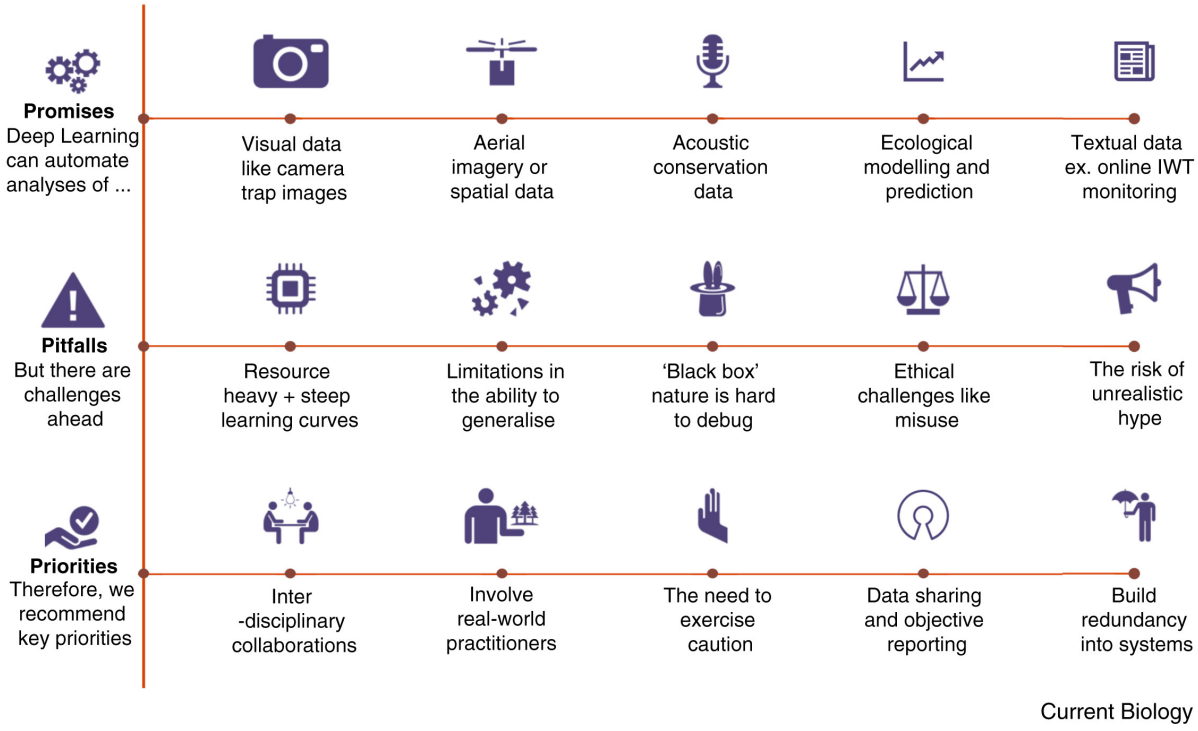


Figure 2.15: *The promises, pitfalls and priorities for deep learning in conservation. Reproduced from (Lamba et al., 2019), with permission from Elsevier.*

Review articles

First, we should start this section by mentioning again the two excellent reviews that introduce deep learning for ecology (Borowiec et al., 2022; Pichler & Hartig, 2023). They are an excellent starting point for ecologists who want to break into AI concepts. Next, Lamba et al., 2019 produced a short essay on the promises, pitfalls and priorities for deep learning in conservation, with a convincing summary Figure 2.15 reproduced here. While we have only focused on the subfield of "ecological modelling and prediction", we believe that most of the pitfalls and priorities illustrated resonate strongly with our work in this

manuscript. Similarly, we believe that the concerns of conservation scientists expressed in (Walker et al., 2020) are legitimate and important to clearly identify (e.g. the need for better metrics) in order to catalyse mitigation efforts. Furthermore, Tuia et al. (2022) provide insightful perspectives on machine learning approaches to wildlife conservation. They examine examples where ML has successfully impacted the field, as well as the specific resources available, sensors already in use, and those that hold promise. Finally, Mouquet et al. (2015) take a step back to review the history and interpret what *ecological predictions* mean and Stupariu et al. (2022) lead a litterature reivew on ML methods applied to landscape ecology.

While prediction has always been part of ecology (see Fig. 2.16 and Darwin's prediction on the Madagascan comet orchid), they recognise that new challenges have arisen. One of these is to clarify the distinction between *explanatory* and *anticipatory* predictions. The latter are not necessarily based on a mechanistic understanding of ecological drivers. They should be seen as guides to current action rather than actual descriptions of the future (Harfoot et al., 2014) - which is already a massive contribution. Moreover, an

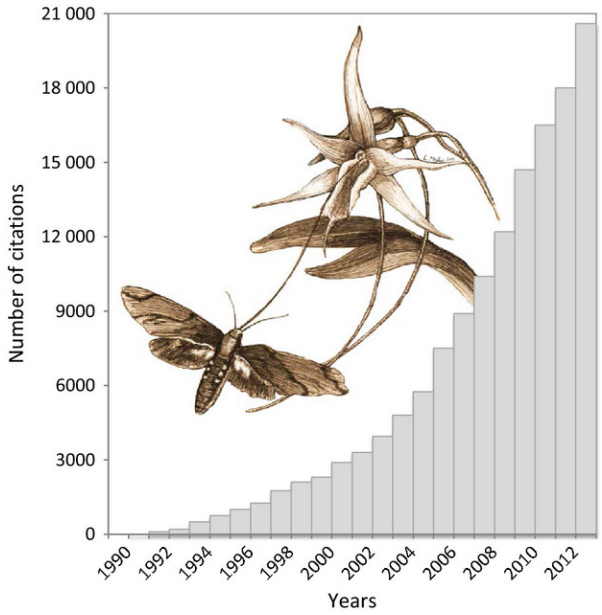


Figure 2.16: *Number of citations per year for articles with 'prediction' and 'ecology' as keywords (Source: Web of Science, search criteria used: Topic = prediction and ecology; Period = All Years). Predictive ecology has become increasingly important in recent years (especially in predicting species distributions). However, prediction is not new to ecology. In 1862, Charles Darwin received orchids from Madagascar. One species in particular, the Madagascan comet orchid Angraecum sesquipedale, with a surprisingly long nectar spur (20-35 cm), caught his attention. No insect with a proboscis of this length had ever been described, but Darwin was convinced of its existence because the plant could not reproduce without a suitable pollinator. The pollinator, a hawk moth, was indeed discovered in 1903, 41 years after Darwin's prediction. Illustration Laurence Meslin. Reproduced from (Mouquet et al., 2015), with permission from John Wiley and Sons.*

under-exploited avenue at the interface between IA, ecology and conservation is transfer learning. Indeed, ecological studies are often highly context-specific, resulting in relatively small (or focused) datasets providing valuable insights. Data science approaches such

as transfer learning, use of synthetic data and causal inference increase the value of such specific datasets beyond their original purpose (Todman et al., 2023). Therefore, not only should their collection continue, but their reusability and collection should be incentivised as well. To conclude this review paragraph and link to the following one, Christin et al. (2019) provides the perfect review article on deep learning applications in ecology. They cover the main applications of deep learning, but also suggest useful guidelines, recommendations and resources.

A selection of outstanding applications

- **Phenological research.** Katal et al. (2022) provided a systematic overview of the DL methods used in the field, identifying research trends and foreseeing promising directions. Lorieul et al. (2019) and Reeb et al. (2022), for example, trained a CNN to successfully detect pheno-phases in CS images and herbarium specimens. Brun et al. (2023) estimated national flowering phenology in Switzerland from seasonal observation probabilities.

- **Species identification.** Deep learning has become inevitable in the field of species identification, especially for citizen science (Affouard et al., 2017; Unger et al., 2021). However, difficult problems remain, such as high class imbalance and high ambiguity (both inter- and intra-class), as in the Pl@ntNet-300K dataset (Garcin et al., 2021). Other forms of automated species identification remain challenging. Herbarium sheets hold unique value for the world's botanical history, evolution and biodiversity, and are increasingly being digitised. However, their identification with DL is hampered by taxonomic mismatches, small datasets from different sampling protocols that are difficult to analyse together (Carranza-Rojas et al., 2017; Lutio et al., 2022).

- **Entomology.** While most animals on Earth are insects, and reports suggest they may be in drastic decline, this group remains highly understudied. Deep learning could help the field in a range of applications: automatic identification, but also estimation of abundance, biomass, diversity, quantification of phenotypic traits, behaviour and interactions, population monitoring, etc. (Høye et al., 2021).

- **Wildbooks.** Conservation of keystone species requires close monitoring at population and then individual level. Animal images collected on the Internet are rapidly becoming one of the richest sources of wildlife data. Wildbooks have been developed to harness this information and allow automatic identification of individuals of target species (Berger-Wolf et al., 2017). An example is the whale wildbook called *Flukebook*, see https://www.wildme.org/.

- **Earth science.** As a collection of all natural sciences related to the planet, Earth science includes ecology, but also atmospheric science, geology, geography, oceanography, etc. Deep learning is a logical tool to process the amount of data generated by Earth observation. First, Camps-Valls et al. (2021) provide a comprehensive book on this topic. Reichstein et al. (2019) argues that hybrid models coupling physical processes with ML are the next step in understanding the Earth system (see also physics-informed neural networks, Raissi et al., 2018). While much attention is

focused on optical data, Synthetic Aperture Radar (SAR) also has great potential (Zhu et al., 2021). Finally, Earth observation and DL have the potential to support the UN Sustainable Development Goals (Persello et al., 2022). For instance, the work of Metzger et al. (2022) allows the estimation of 100 m population maps. This information in turn benefits urban planning, environmental monitoring, public health and humanitarian operations.

2.4.2 Predicting the missing conservation status of species

Unlike the *index-based* methods already introduced in section 2.3.1.5, *prediction-based* methods bypass the official IUCN extinction risk criteria. Instead, such a method directly learns a mapping between features qualifying species (geographic range, bioclimatic preferences, exposure to anthropogenic pressures, species traits, phylogeny, etc.) and the IUCN extinction risk status. Exploring the covariates of species extinction risk (in particular the biological traits of animals) has been the subject of studies since the emergence of the Red List in the early 2000s (using regression analyses, Purvis et al., 2000). Boosted by machine learning and the increasing availability of data, this line of research is still very active. Novel methods and applications to new taxa are published regularly. Before presenting recommendations and challenges, we present a brief literature review organised by model class. In chapter 5, we propose a *prediction-based* method that uses deep-SDMs. The species features exploited are flexible enough to project the impact of future climate scenarios on species extinction risk. For a more comprehensive review, we refer to the work of Cazalis et al. (2022), which covers 73 prediction-based approaches and 25 index-based methods.

Statistical models and phylogenetic imputation

A classic approach is to 1) exploit the phylogeny of taxa to impute species missing biological or ecological traits, and 2) use this completed information (possibly alongside other spatial and widely available features) as input to a traditional statistical model. For example, Jetz and Freckleton (2015) assessed the global extinction risk of mammals using GLMs, and González-del-Pliego et al. (2019) highlighted that a thousand amphibians are threatened with extinction using a generalised least squares approach.

Random forest and decision trees

As a performing and easy-to-use ML model, the random forest algorithm and all decision trees are perfect candidates for predicting species extinction risk. Leão et al. (2014) tested how vegetation type, growth form and geographic range size relate to species extinction risk for Brazilian angiosperms using decision trees, but also standard and phylogenetic regression. RF has been applied to, among others, terrestrial mammals (Di Marco & Santini, 2015), bulbous monocot species described by coarse distribution data (Darrah et al., 2017) and 150,000 plant species using open-source geographic, ecological and morphological trait data (Pelletier et al., 2018). More recently, Caetano et al. (2022) used the XGBoost algorithm (Chen & Guestrin, 2016) to assess 4,369 reptiles, with impressive performance demonstrated on the validation data.

Neural networks and the *IUCNN* approach

Zizka et al. (2020) developed a method based on neural network classification and demonstrated its excellent performance on the orchid family. Species are associated with features representing their geographic, bioclimatic, human footprint and biome preferences. For raster covariates, the rationale is to take the species average across observations. A user-friendly R package called *IUCNN* has made this method a reference in the field (Zizka et al., 2021). In addition, further developments have opened up new possibilities for users, including quantifying uncertainty using a Bayesian approach and limit predictions with a confidence threshold. The method has now been applied to many taxa, including 21,000 globally distributed tree species (Silva et al., 2022) and 1,162 freshwater fish species in China (Chen et al., 2023).

Ensembling and representation learning

Ensembling is an effective way to improve the generalisation power of already good classifiers. Prediction of species extinction risk is no exception, as shown in (Borgelt et al., 2022a). The authors carried out a convincing work and trained an ensemble model based on 220 ML models (including GLMs, RFs, gradient boosted classification trees and deep NNs). The final model is based on 20 boosted trees and 3 deep neural networks. An impressive number of covariates (more than 400, geographic, bioclimatic, environmental, threats) are included, see their supplementary Table 2. Here, raster covariates are summarised by species with statistics (mean, median, min, max) taken from range maps and occurrence cells or native countries.

Finally, Mukadam et al. (2020) published an original study associating species embeddings generated by applying representation learning to Wikipedia text and animal taxonomy data. This work, involving natural language processing techniques for conservation biology, is an example of interdisciplinary research. Online text mining can provide valuable information for describing species.

Model comparison

Few independent studies have attempted to compare the predictive performance of models. Nic Lughadha et al. (2019) tested five different approaches (four index-based methods and RF) using herbarium-derived data for trees, shrubs and herbs. Random forest performed best, but the less data-intensive approaches also achieved good results. Besides, (Bland et al., 2015) compared seven ML models on terrestrial mammals. They concluded that classification trees and k-nearest neighbours (simpler and less computationally intensive) achieved lower classification performance than random forests, boosted trees, support vector machines and neural networks. Finally, new comparative studies, including recent developments, are needed to provide clear guidance on which modelling method is best (Cazalis et al., 2022). The field is indeed developing rapidly. Methods that have been evaluated on different taxa and with different species covariates cannot be compared, as the difficulty of the classification task can vary greatly.

Recommendations

Given the importance of assessing species extinction risk, Walker et al. (2020) recommends caution when using predictive methods. In particular, modellers should ensure that per-

formance is clearly reported, that the implications of modelling choices are explored, that threats are analysed and that limitations of biodiversity data (gaps, biases, uncertainties) are addressed. There is also a trade-off between occurrence cleaning and species coverage. Indeed, many species are represented by a very limited number of observations (long-tail distribution), and occurrence cleaning (especially with respect to geolocation errors) results in the removal of all observations of certain species. For a traditional threshold-based method, rigorous cleaning is necessary to achieve the best performance. However, the performance of ML methods (random forest) has been shown to be robust to minimal data cleaning (Walker et al., 2021), allowing a large species coverage to be maintained.

Challenges and opportunities

There are several challenges to automated extinction risk assessment. Taken together, they can explain why a research-implementation gap is observed (Cazalis et al., 2022). Indeed, while many prediction-based methods have been developed, few have been integrated into assessment practice. In response, the last authors urge academic researchers and Red List practitioners to collaborate (e.g. by involving Red List stakeholders early in the development process) and identify the need to develop and maintain user-friendly platforms. In addition, hybrid methods, which better incorporate Red List criteria and benefit from the generalisation power of ML, have the potential to facilitate uptake. Comparative work on environmental covariates and drivers of extinction risk is also needed. While threat exposure is highly taxon and context dependent, classic covariates are commonly used and can be summarised from the observation or range level to the species level in different ways. More generally, we believe that a standard protocol for covariate selection and integration for extinction risk prediction would benefit the community. Finally, uncertainty quantification, as proposed by IUCNN, is necessary because the ultimate goal of such an assessment is to aid conservation planning and resource guidance.

2.4.3 Future trajectories of conservation indicators

Climate change can affect species on at least three non-exclusive aspects (Bellard et al., 2012): i) physiology (Antala et al., 2022), i.e. their biological traits and behaviours, ii) phenology (Collins et al., 2021), e.g. with shifts in flowering timing (climate change indicator), and iii) geographic distribution (Lenoir et al., 2020). These three dimensions are sources of potential mismatch in biotic interactions (Renner & Zohner, 2018), or the required adaptation may be too demanding (e.g. large spatial shifts or rapid evolutionary response) for certain species (Corlett & Westcott, 2013; Parmesan, 2006). However, the vulnerability of species to climate change is diverse and the assessment methods are numerous (Pacifici et al., 2015). Ultimately, climate change has consistently been shown to be a driver of biodiversity loss (Urban, 2015). Furthermore, the redistribution of biodiversity under climate change has been shown to have impacts on ecosystem functioning, human well-being (e.g. shifts in the distribution of disease vectors such as mosquitoes) and the dynamics of climate change itself, creating amplifying feedbacks (Pecl et al., 2017). These effects may accumulate in certain highly exposed regions such as the Arctic.

Modelling trajectories of biodiversity requires scenarios to project future environmental conditions (Peterson et al., 2003). Such projections in turn depend heavily on socio-economic scenarios (see Pereira et al., 2010 Figure 1 for a schematic overview of how biodiversity scenarios are constructed). Changes in species distribution are then the most studied effect of climate change. SDMs are used to project future species distribution and range loss rates are compared to IUCN thresholds (Thuiller et al., 2005) or the species-area relationship (Thomas et al., 2004) is exploited to estimate likely species extinction rates. Biodiversity indicators, e.g. at species and population level, such as the IUCN extinction risk (Moat et al., 2019) and the Living Planet Index (see section 2.2.1) can therefore be projected into the future (Visconti et al., 2016). This is what we do in the chapter 5 of this thesis. The final objective is to communicate on the projected rates to try to influence policy, bend scenarios and mitigate biodiversity loss through conservation (Schwartz, 2012). Machine learning can help tackle climate change at many levels (Rolnick et al., 2022), including predicting biodiversity indicators through species distribution modelling (see section 2.3.2).

Climate change ecology is a young field full of challenges (Bellard et al., 2012). To effectively guide conservation and policy with the projected trajectories of biodiversity indicators, methodological developments are needed in a variety of facets. First, due to uneven data availability, climate impact studies are still highly biased in terms of taxonomic and geographic coverage (Feeley et al., 2017). As a result, tropical and marine ecosystems, as well as entire taxa such as plants, are underrepresented. Moreover, species are often the level of study, leaving the impact of climate change on functional, phylogenetic and genetic diversity rarely assessed (Thuiller et al., 2011, 2006b). Second, projecting biodiversity's response to climate change requires working with a variety of scenarios (socio-economic, emissions, climate) that strongly influence the results. Comparative and interdisciplinary studies are needed to identify the most likely directions. Ensembling and uncertainty quantification methods help to increase confidence in predictions. In addition, species response to climate change is highly taxon-specific and can have counterintuitive effects, adding to the complexity. For example, milder temperatures may have local positive effects on certain plants. The dispersal capacity of species also has a strong influence on projected distributions. Without this data, two extreme scenarios are often considered: no dispersal or universal dispersal capacity (Thomas et al., 2004). Overall, more focused studies of species characteristics and responses to climate change are critically needed (Todman et al., 2023). SDMs are good candidates for exploring species redistribution under climate change, but temporal and physiological responses are rarely studied. Finally, drivers of extinction risk are often studied independently, although they are likely to be interdependent and create amplifying patterns (e.g. land-use change through deforestation and climate change, see Sala et al., 2000 for a review). Given these sources of uncertainty and the specific responses of species, meta-analyses are valuable to provide synthetic messages on climate-induced biodiversity loss (Maclean & Wilson, 2011; Urban, 2015).

In conclusion, the impact of climate change on biodiversity is an active area of research that depends on critical data both at large scale (e.g. bioclimatic projections) and at

small scale (dispersal capacities, species responses and plasticity, etc.). With the evidence of spatial redistribution of biodiversity (both current and future), concerns have arisen about the effectiveness of protected areas (Araújo et al., 2011, 2004; Dobrowski et al., 2021; Hole et al., 2009). Indeed, there is a possibility that current PA implementation may be insufficient to protect species that are shifting their spatial range. Spatial conservation planning should ultimately anticipate this pitfall by accounting for likely biodiversity redistribution in the process of PA design.

2.4.4 Optimising spatial conservation

Spatial conservation prioritisation is a highly complex problem constrained by multiple trade-offs between implementation costs and benefits in terms of biodiversity, economic and social values. The basic objective is to identify areas that maximise complementary values at minimal or reasonable cost (Margules & Pressey, 2000). Surrogates and biodiversity indicators are used and spatial design constraints (e.g. connectivity) also come into play (Ferrier, 2002). There are many aspects to consider and assessing the effectiveness of PAs is remarkably challenging (Rodrigues & Cazalis, 2020). Citizen scientists can contribute to the monitoring of PA species thanks to automatic identification (Bonnet et al., 2020). As already mentioned, purely area-based international targets can lead to inefficient PA implementation (Maxwell et al., 2020) and leave areas of high conservation value under serious threat (Rodrigues et al., 2004). Success metrics including biodiversity protection and threat mitigation measures are therefore critical.
In our work we have not addressed the issue of spatial prioritisation. However, we believe that some of our findings could ultimately benefit such a task. Therefore, we briefly introduce the field as a final section of our state of the art. For the fundamentals of spatial conservation prioritisation we refer to (Wilson et al., 2009) and for a comprehensive review of the field to (Moilanen et al., 2009). Here we will present only a few existing frameworks, with a focus on two IA-based approaches.

Two of the reference frameworks currently in use are Marxan (Ball et al., 2009; Watts et al., 2009) and Zonation (Moilanen et al., 2005, 2022). Marxan[13] is the most widely used reserve planning software in the world. It is based on simulated annealing, a stochastic optimisation technique. It has been used to define many protected areas around the world (both marine and terrestrial), including the Great Barrier Reef, Australia and the Galapagos Islands.
Zonation[14] is another popular spatial prioritisation software. It is based on an iterative conditional sorting algorithm and has been continuously updated since 2005 to provide flexible and realistic decision support.

Artificial intelligence has embraced the problem of prioritising spatial conservation. Here we present two promising approaches: Conservation Area Prioritisation Through Artificial INtelligence (CAPTAIN) from Silvestro et al. (2022) and constraint programming.

[13] https://marxansolutions.org/
[14] https://zonationteam.github.io/

Based on reinforcement learning (a machine learning paradigm, see Sutton and Barto, 2018), CAPTAIN allows the exploration of multiple biodiversity metrics with a limited budget and accounts for dynamic changes in the system (e.g. in biodiversity monitoring). It has been shown to outperform state-of-the-art software, achieving conservation goals more reliably and producing more interpretable prioritisation maps.
Constraint programming is a hybrid AI technique that allows users to explicitly specify rules in an optimisation problem (Rossi et al., 2006). This expressive AI paradigm is promising in a field such as spatial conservation planning, where on-the-ground realities (social, economic, historical, etc.) can impose specific constraints on the implementation of PA. Incorporating socio-economic issues into the modelling process can also facilitate constructive dialogue with stakeholders (Justeau-Allaire et al., 2021). Reserve design (Justeau-Allaire et al., 2019) and restoration planning (with different underlying issues from conservation planning, Justeau-Allaire et al., 2023) have been successfully undertaken using this flexible technique.

CHAPTER 3

DEEP SPECIES DISTRIBUTION MODELING FROM SENTINEL-2 IMAGE TIME-SERIES: A GLOBAL SCALE ANALYSIS ON THE ORCHID FAMILY

Table of contents

3.1 Introduction 71
3.1.1 Context 71
3.1.2 Contributions 73
3.2 Materials and Methods 74
3.2.1 *DeepOrchidSeries* dataset 74
3.2.1.1 Raw input data description 74
3.2.1.2 Dataset construction 76
3.2.2 Species Distribution Models trained with satellite image time-series 80
3.2.2.1 Model definition and training procedure 80
3.2.2.2 Performance evaluation of the model 81
3.2.2.3 Interpretability experiments: quantifying the contribution of temporal information 84
3.2.2.4 Modality contribution on a global scale 86
3.3 Results 87
3.3.1 Model validation and performance 87
3.3.2 Results by number of species occurrences 88
3.3.3 Results by region and regional diversity index 89
3.3.4 Statistical tests 91
3.3.5 Model evaluation regarding time and spatial data mismatches . 92
3.3.6 Modality contribution on a global scale 93
3.4 Discussion 93

3.5 Conclusion 98

This chapter is an extended version of the article (Estopinan et al., 2022) published in *Frontiers in Plant Science*. It has been expanded to include a comparison of the performance of the global distribution model using Sentinel-2, WordlClim 2 and static variables.

Abstract

Species Distribution Models (SDMs) are widely used numerical tools that rely on correlations between geolocated presences (and possibly absences) and environmental predictors to model the ecological preferences of species. Recently, SDMs exploiting deep learning and remote sensing images have emerged and have demonstrated high predictive performance. In particular, it has been shown that one of the key advantages of these models (called deep-SDMs) is their ability to capture the spatial structure of the landscape, unlike prior models. In this paper, we ask whether the temporal dimension of remote sensing images can also be exploited by deep-SDMs. Indeed, satellites such as Sentinel-2 are now providing data with high temporal revisit and it is likely that the resulting time-series of images contain relevant information about the seasonal variations of the environment and vegetation. To confirm this hypothesis, we built a substantial and original dataset (called *DeepOrchidSeries*) aimed at modelling the distribution of orchids on a global scale based on Sentinel-2 Image Time-series. It includes around 1 million occurrences of orchids worldwide, each being paired with a twelve-month-long time-series of high resolution images (640 x 640 m RGB+IR patches centered on the geolocated observations). Thanks to this ambitious dataset, we trained several deep-SDMs based on Convolutional Neural Networks (CNNs) whose input was extended to include the temporal dimension. To quantify the contribution of the temporal dimension, we designed a novel interpretability methodology based on temporal permutation tests, temporal sampling and temporal averaging. We show that the predictive performance of the model is greatly increased by the seasonality information contained in the temporal series. In particular, occurrence-poor species and diversity-rich regions are the ones that benefit the most from this improvement, revealing the importance of habitats temporal dynamics to characterise species distribution.

Keywords

Species Distribution Modelling, Deep learning, Image time-series, Sentinel-2, Convolutional Neural Networks, Remote Sensing, Macroecology, Spaceborne Ecology, Biogeography, Data Science

3.1 Introduction

3.1.1 Context

Understanding and mapping species distributions is a major topic in conservation biology (Pecl et al., 2017). Species Distribution Models (SDMs) have recently become a key instrument: over the last 20 years, 6,000 peer-reviewed studies were found with this keyword according to (Araújo et al., 2019). These statistical algorithms learn the correlations between species presence (and possibly absence) records and some environmental predictors provided. Under certain modelling assumptions (Zurell et al., 2020), they can estimate species distribution by generalising learned habitat preferences over time and space (Phillips & Dudik, 2008; Thuiller et al., 2009). A major issue for the use of SDMs concerns the ecological relevance of the predictive variables used (Fourcade et al., 2018),

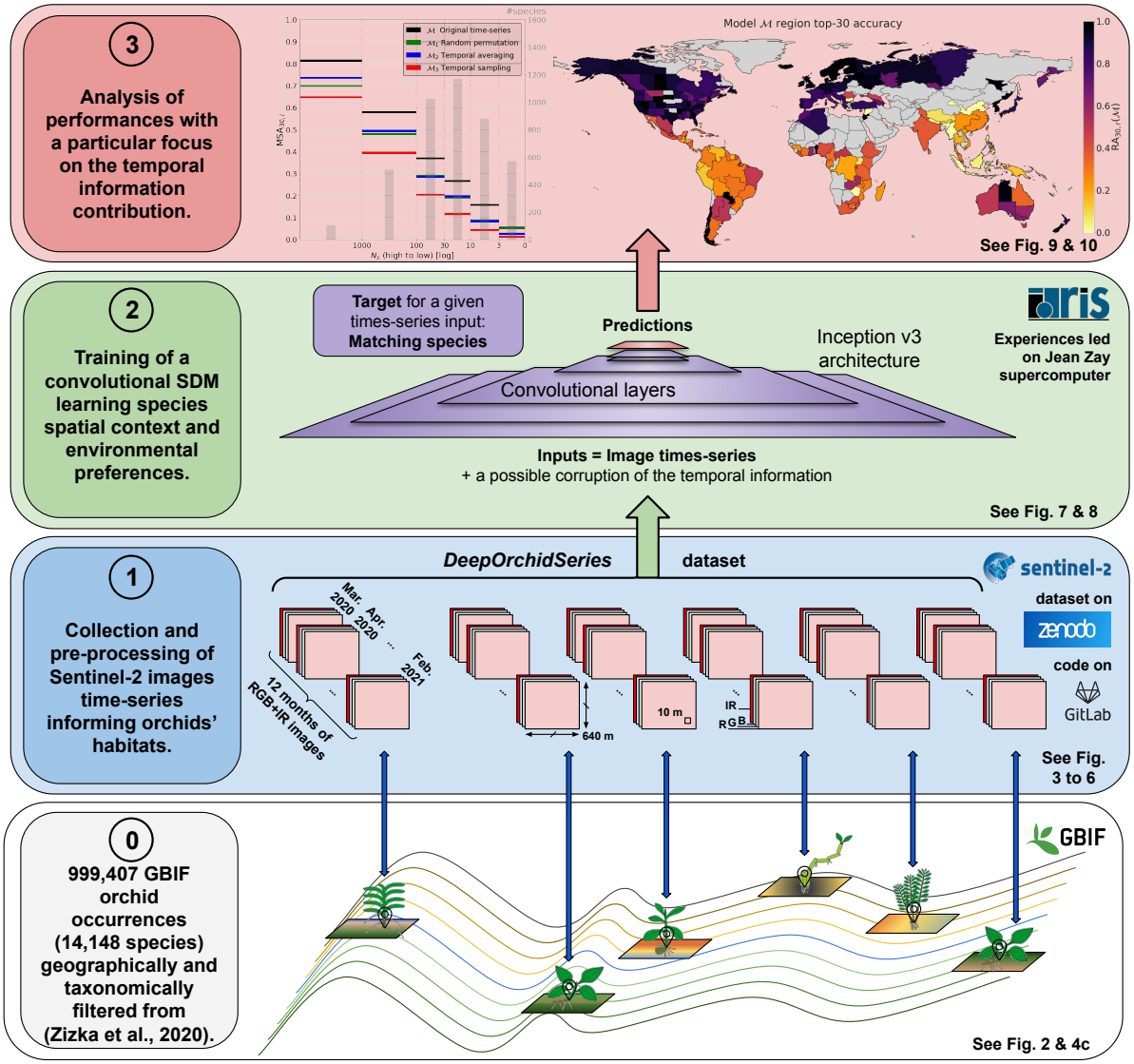


Figure 3.1: *Visual abstract of the method.* ***Layer 0 :*** *The dataset introduced in this paper (DeepOrchidSeries) is based on a filtered set of GBIF occurrences coming from the study of (Zizka et al., 2020).* ***Layer 1 :*** *Sentinel-2 image time-series were collected around each occurrence geolocation, keeping least cloudy data tiles every month between March 2020 and February 2021. Images are made of 640 x 640 m RGB+IR channels with 10 m spatial resolution. The dataset is available on Zenodo and the method to create it on the Gitlab.inria platform.* ***Layer 2 :*** *We then trained deep Species Distribution Models (deep-SDMs) based on convolutional neural network (Inception v3) to capture the spatio-temporal context and environmental preferences of species. Next, we conducted experiments where the input temporal dimension was modified (randomized, averaged or sampled) so as to measure its contribution on model performance.* ***Layer 3 :*** *the results are finally broken down into three main dimensions of analysis: species frequency in the dataset, bioregion, and species diversity in these bioregions. The analysis reveals that occurrence-poor species and diversity-rich regions are the ones that benefit the most from the improvement provided by the temporal information.*

see section 2.3.1.4. Furthermore, collecting appropriate data at large scale is usually very challenging. Global bio-climatic variables do not systematically provide enough information to draw conclusions on a species presence. Many other factors like species dispersal capacities (Monsimet et al., 2020) or shifts in land use actually come into play.

After having revolutionised computer vision, neural networks - and especially Convolutional Neural Networks (CNNs) - are also increasingly recognised in ecology (Botella et al., 2018a; Brodrick et al., 2019; Heikkinen et al., 2012; Williams et al., 2009). They allow to identify environmental patterns on images like tree crowns (Csillik et al., 2018) or forest type limitations (Wagner et al., 2019). Local environment spatial structure has already been proven to add relevant information to SDMs involving convolutional layers (Deneu et al., 2021b). In addition, remotely sensed data can grasp key features of vegetation functioning and thus convey relevant insights on species habitats (Adhikari et al., 2012; He et al., 2015; Remm & Remm, 2009). Unmanned aerial vehicles allow finer and finer-scale coverage at local, regional or even country scale (Kattenborn et al., 2020). Thanks to such imagery, the nature and spatial structure of ecosystems can be characterised and learned in SDM training. RGB and IR image patches around species occurrences (or digitized geolocated presence of species) are thus added to the environmental predictors, so as to include information on vegetation and land-use heterogeneity around the occurrences (Deneu et al., 2021a).

Satellite missions like Copernicus Sentinel-2 (S2) (Berger et al., 2012) now provide RGB and IR channels with fine spatial resolution and temporal revisit frequency worldwide (see subsection 3.2.1.1), which can feed high-resolution, CNN-based SDM models. However, there is still much potential ahead for bringing together remote sensing and deep learning (Camps-Valls et al., 2021). Remote sensing datasets that are *(i)* readily available for deep learning applications and *(ii)* exploiting the spatial, spectral and temporal dimensions of new satellite missions are still very few. For instance, among the twenty-three benchmark datasets implemented in *TorchGeo* (Stewart et al., 2021), only two encompass a temporal dimension. There is then an opportunity to build RGB+IR image time-series around occurrences spread worldwide. By sampling S2 data for a whole year, prominence is given to the seasonal evolutions of the plants habitats. These time-series are capturing the signature of ecosystems phenology and productivity. Our hypothesis is that this information can significantly help SDM predictions.

3.1.2 Contributions

This chapter contribution is threefold: First, we built a substantial and original dataset pairing nearly 1 million geolocated occurrences of the *Orchidaceae* family with satellite image time-series. This dataset and the associated method scripts, released as open data and code, should be useful for conservation biologists and SDM users in general. To our knowledge, no similar ready-to-use dataset is already available. Second, we designed interpretability tests of the deep-SDMs trained on this dataset in order to measure the importance of seasonal landscape variability in characterising species habitat and niche. Figure 3.1 provides the visual abstract of our method. Finally, we test which SDM modality allows our model to better capture species' environmental preferences on a global scale between satellite image time-series, bioclimatic and *static* variables.

Static should be considered in contrast to the other two modalities, which capture more dynamic information. It includes altitude, position (longitude and latitude, making this SDM spatial-explicit), human footprint and ecoregions. We hypothesise that in a global assessment, it is the bioclimatic variables that would be more valuable for drawing species distributions (Randin et al., 2020).

3.2 Materials and Methods

3.2.1 *DeepOrchidSeries* dataset

3.2.1.1 Raw input data description

Orchid occurrences dataset. The *Orchidaceae* family is of great interest because of its diversity (about 28 000 species estimated) and its aesthetic attractiveness (Chase et al., 2015). Orchids are of major concern for ecologists due to the numerous threats they are facing: habitat destruction, climate change, pollution, and illegal harvesting for horticulture and tourism industries (Wraith & Pickering, 2018). They are also considered as a relevant proxy of their ecosystem's health (Newman, 2009). Moreover, orchids are found on all continents in a wide range of habitats and they are blooming at very different altitudes. Such a range or environmental amplitude is difficult to achieve with other families, making the orchid family an excellent candidate for the purpose of our study (i.e., to measure the importance of seasonal variability in characterising species habitat and niche).

Rather than collecting a new set of orchid occurrences to build our image time-series dataset, we decided instead to re-use the one introduced in (Zizka et al., 2020). Their objective was different from ours (i.e. estimating the conservation status of orchids) but the set of occurrences they collected from GBIF meets two main criteria of interest for our study: (i) global scale, and (ii) suitable data quality thanks to several data filtering and cleaning processes (including the use of the R package *CoordinateCleaner* v. 2.0-9, Zizka et al., 2019). The complete process they use is summarised in the supplementary information Table 1 of their paper (Zizka et al., 2020). Another benefit of reusing (Zizka et al., 2020)'s occurrence data is to support the potential reuse of our deep-SDM for the automated assessment of orchid's IUCN status. In the long term, this will improve the reproducibility and comparability of newly developed methods in this regard.

In total, the dataset contains 999,407 occurrences of 14,148 species with: 70 records per species in average, 4 in median and 3,537 species (25%) with more than 13 observations. The (heavily-tailed) distribution of the number of occurrences per species is shown on Figure 3.2A (through a Lorenz curve). Figure 3.2B represents the temporal distribution of the occurrences in the dataset. Half of the observations dated from 1997, one quarter from 2010. 14.6% of the set (145,641 occurrences) came with no timestamp at all. The oldest occurrence was from 1901 as a result of the filtering process that got rid of data records older than 1900. Only observations with a position uncertainty higher than 100 km were discarded. Perspectives and limits related to the use of such large and imbalanced occurrence dataset will be discussed in the final Section 3.4.

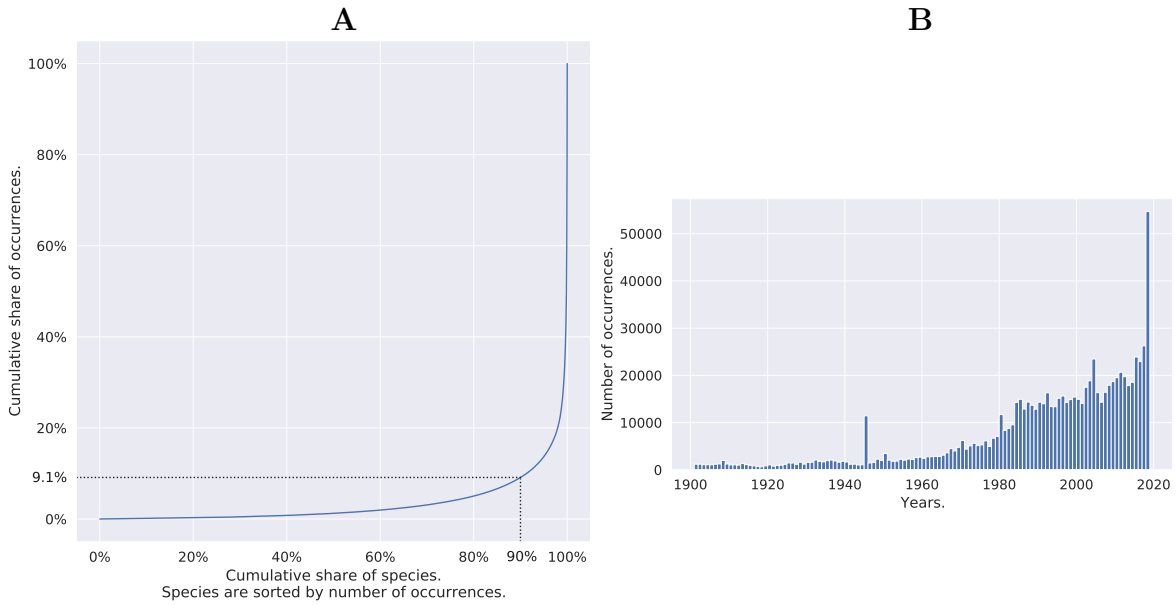


Figure 3.2: ***(A)*** *Occurrences' distribution. Species are ordered by frequency. The dotted lines are flagging that 90% of the species are only gathering 9.1% of the occurrences.* ***(B)*** *Occurrences' temporal distribution. The two graphs are based on all dataset's occurrences.*

Sentinel-2 multispectral images. Sentinel-2 multispectral data comes from two identical satellites in the same orbit but diametrically opposite to one another. Sentinel-2A was launched on 23 June 2015 and its counterpart Sentinel-2B on 7 March 2017. This satellite mission is part of the European Earth observation project Copernicus[1] (Drusch et al., 2012). Thirteen channels from the visible to short-wave infrared are monitoring the planet, with 10, 20 or 60 m spatial resolution and a 5-day temporal revisit above any point on Earth. Additional satellites 2C and 2D are planned to ensure the continuity in the coming years and the next generation of Sentinel-2 satellites are being prepared. We only kept four out of the thirteen channels, i.e. the three RGB channels and the Infrared (IR) channel (842 nm). These wavelengths are expected to convey the most relevant information about the environment (He et al., 2015) and are also the finer in terms of spatial resolution (10 m). The smallest geographic units downloadable via the *sentinelsat*[2] API are 109.8 x 109.8 km square data tiles in WGS84/UTM projection. They were defined following a military grid splitting Earth planisphere. The field square from a given satellite orbit at a given sensing time interval does not always cover a whole tile, so that several products must be merged and cropped to get an image of the whole tile.

Data products are made available to the user at two distinct levels: Top-of-Atmosphere (TOA) or and Bottom-of-Atmosphere (BOA). The important difference is the application of an atmospheric correction algorithm such as Sen2Cor (Ientilucci & Adler-Golden, 2019; Louis et al., 2016). Water vapour and other atmospheric components alter the

[1] https://www.copernicus.eu/en
[2] https://sentinelsat.readthedocs.io/en/stable/

satellite image caption with complex non-linear deformations. When and how atmospheric correction should be performed prior to exploit remote sensing data depends on the desired information and thus the targeted application. About classification and change detection tasks, a recognised work from (Song et al., 2001) advises to perform simple corrections only when multi-temporal data is used. Otherwise, having both training and test sets from the same relative scale proved to be sufficient: no significant performance gain would result from the addition of an atmospheric correction step. A more recent article estimating the relation between sea surface salinity and Sentinel-2 Imagery with a neural network and 2,700 points obtained better results with TOA than BOA imagery (Medina-Lopez, 2020). On their specific application, they found that the atmospheric correction entailed information loss due to alteration of actual multispectral relationships. They also observed that the time and computational resources spared by using the TOA products was an important element to consider. Using TOA products time-series, Rußwurm and Körner (2018) obtain state-of-the-art land cover classification performances. BOA products are not readily available at the global scale and, when needed, atmospheric corrections have in this case to be applied by users. Considering the conclusions of previous surveys and the large size of the targeted data, we decided to work with TOA products. Moreover, the atmosphere information could be valuable for our application and we believe deep-SDMs are capable of correctly learning without this additional filter.

3.2.1.2 Dataset construction

Figure 3.3 summarises the workflow followed to obtain image time-series from a set of geolocated occurrences. The first step is to define the set of Sentinel-2 tiles containing all targeted occurrences, for which more details are provided in the *Global scale processing* paragraph. Second and third steps are to define the *patch size* and the *time sampling strategy*. Our choices are presented in the two dedicated paragraphs hereafter. Finally a last paragraph introduces our method to select least cloudy S2 data.
We have furthermore considered only the four spectral bands available at 10 m resolution, but our workflow could be applied as well to bands at 20 m and 60 m resolution after a down-sampling step. Sentinel-2 queries and downloads were made with the Scihub Copernicus API[3]. We then extracted the patches by parallelizing the processing by UTM zone to gain speed. Code and details are available at *https://gitlab.inria.fr/jestopin/sen2patch*.

Global scale processing. First step consists then in defining the minimal set of Sentinel-2 tiles containing all our orchid observations. The *Sentinelsat* python API provides the option to query data by various geographical means, mainly: coordinates, polygons, tiles or satellite orbits. However, querying the API on an occurrence-by-occurrence basis for a dataset containing nearly one million occurrences is counterproductive. It is much more efficient to first download the tiles containing occurrences and then extract them locally (see Fig. 3.4A for the histogram of the number of occurrences per tile). To do so, we implemented the following two steps:

[3] https://scihub.copernicus.eu/, queries and downloads require an activated Scihub Copernicus account

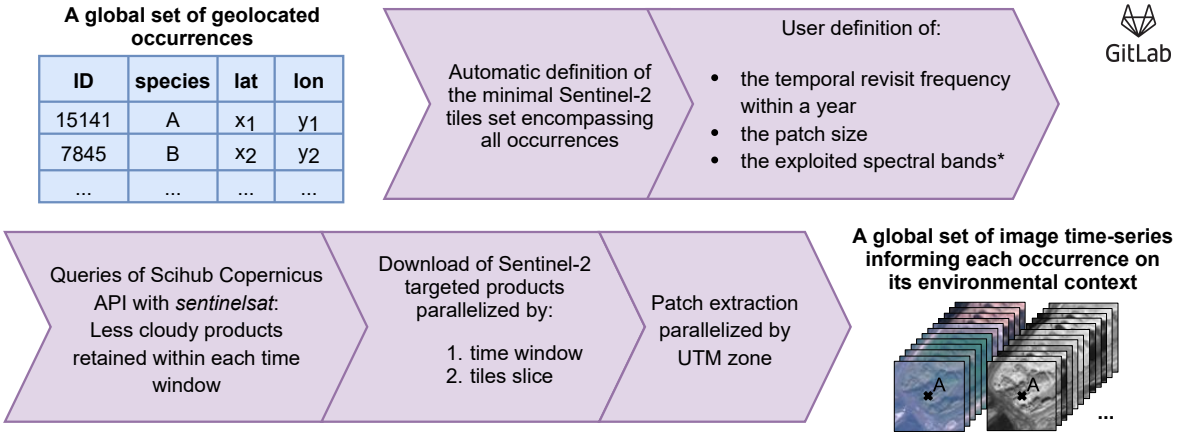


Figure 3.3: *Creation workflow of the DeepOrchidSeries dataset. Input is a set of geolocated occurrences, output gathers image time-series informing on species habitat preferences. Code and details available at https://gitlab.inria.fr/jestopin/sen2patch.*

- First, we created a dictionary linking each tile with its WGS84 geometry thanks to the *Sentinel-2 Level-1C tiling grid* provided by the ESA Sentinel-2 official portal[4].
- Then an iterative process on all occurrences was implemented, testing each time if the new observation is included in the union of the already retained tiles set. If not, a tile containing the occurrence location is downloaded and added to the set.

The final tiles set map is given on Figure 3.4C. It illustrates the full geographical scope of the dataset with 7,563 targeted tiles. 50% of all land areas (Antarctica excluded) were included in the collected data. The color scale proportional to the number of observations per tile (with a log10-scale) further shows a geographic (or observation) bias in the occurrences set: Europe, south Australia and New Zealand are gathering huge numbers of records.

Patch size. The size of the patches associated with each occurrence is an important hyper-parameter to set. Patches should be large enough to contain the most relevant spatial information, but not too large to avoid introducing patterns that are too distant from the occurrence. They should also be large enough to compensate for the geographic imprecision of the occurrences (see geolocation uncertainty distribution in Supplementary Information (SI) Fig. S1 and Wüest et al., 2020), but not too large to avoid computational issues. Considering all that constraints, our final choice was patches of size 640 x 640 m (only powers of two were considered to optimise memory usage). Figure 3.4B illustrates three different patch sizes around an observation on an island of the South Australian coast. It shows that the 640 x 640 m patch (40.96 ha) captures important landscape patterns around the record as well as potential threats due to surrounding land-use.

Time-series extent and temporal resolution. On of the main contributions of our study is to consider time-series of satellite images rather than a single date image, with

[4] https://sentinel.esa.int/web/sentinel/missions/sentinel-2

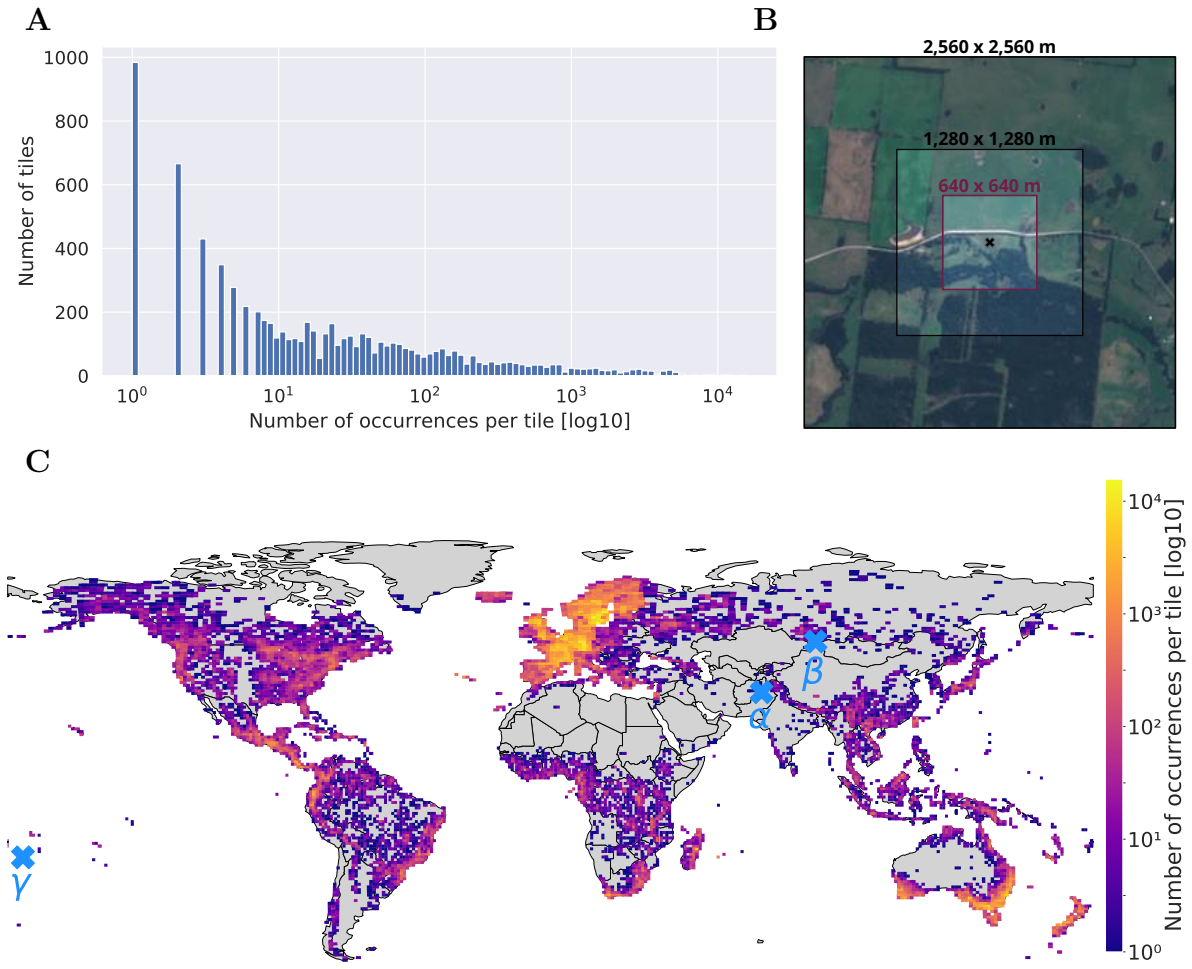
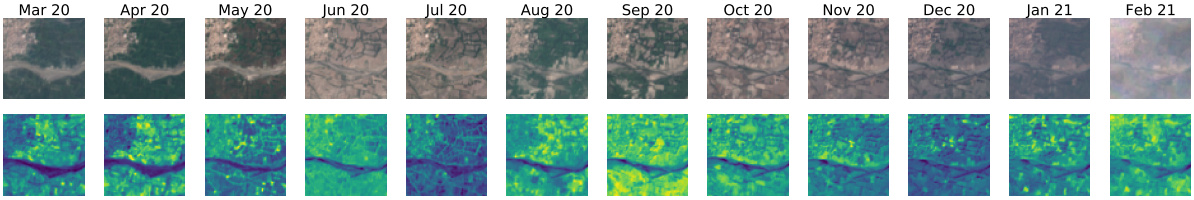


Figure 3.4: ***(A)*** *Histogram of the number of occurrences per tile,* ***(B)*** *different patch sizes comparison around an occurrence located at (-39.883306, 144.050000), decimal degree system,* ***(C)*** *map of the selected tiles coloured by number of records contained (log10 scale). Three occurrences are located by α, β and γ. Figure 3.5 provides the three associated image time-series.*

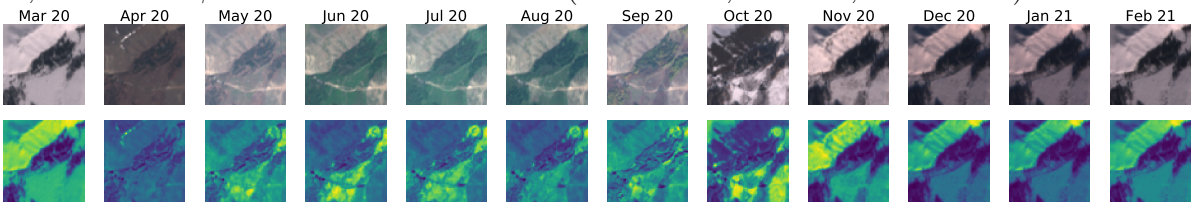
the objective of better characterising the habitat of species. Two important parameters in this regard are the temporal extent of the series and its resolution. Here too, there is a compromise to be made. The extent and resolution must be high enough to capture important (spatio-)temporal patterns, but cannot be too high due to computational constraints. We finally chose one-year time-series with a resolution of 1 month (i.e. twelve images, one per month).

Such twelve-month time-series allow to grasp the main seasonal variations of the environmental and ecological context including vegetation phenology, yearly weather variations as well as landscape annual variations linked to human activity (e.g. agriculture). Noticeably, such seasonal variations are often neglected in SDMs devised at global scale. Figures 3.5A and 3.5B show significant seasonal changes that can be meaningful to differentiate species habitats. In Figure 3.5A, the tree cover greatly vary depending on the season and in 3.5B snow covers the field half of the year. What if we only had one month of data? Environmental contexts would be characterised very partially and wrong inferences could be done on species ecological preferences (imagine having only one image

A, occurrence α located in Afghanistan near a river bed (lat:33.3667, lon:70.0167, alt:1092 m)



B, occurrence β located in South Russia (lat:50.1742, lon:87.9, alt:1524 m)



C, occurrence γ located in Tonga (lat:-21.3952, lon:-174.9271, alt:213 m)

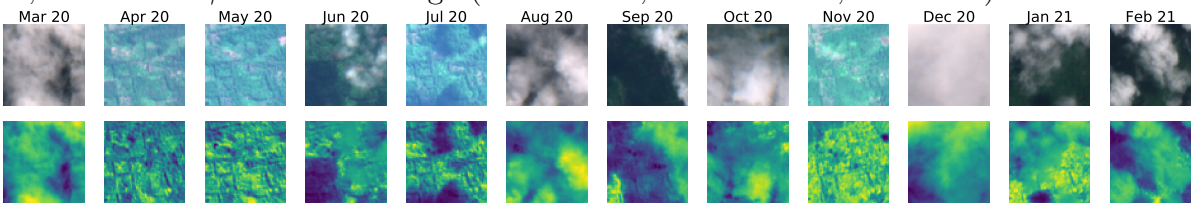


Figure 3.5: *Image time-series associated to the three occurrences located in Figure 3.4C map. RGB images are shown on the first line and IR patches on the second.* ***(A)*** *is almost cloud-free and globally normalised before visualisation (i.e. all months are divided by time-series maximum pixel),* ***(B)*** *is a cloudless time-series with a strong environmental gradient because of snow presence and is normalised by frame (i.e. each month data is divided by month maximum pixel, only for visualisation),* ***(C)*** *is an especially cloudy time-series also normalised by frame.*

covered by snow for Figure 3.5B). These examples illustrate the gain of ecologically relevant information when considering a twelve-month image-series.

Another parameter to be set, is the starting date of the time-series. Ideally, it should be chosen so that the date of the occurrences is included in the one-year period covered for the time-series. There are in practice various reasons impeding a perfect match between the occurrences dates and the associated predictive data. To begin with, Sentinel-2 satellite was launched only in 2015 so that older occurrences cannot be matched. Secondly all occurrences do not come with a precise date, some having no date information at all. Thirdly some S2 tiles from the defined minimal set would have to be downloaded a huge number of times to inform all observations at different dates. Lastly, there is no simple and open access to data older than a rolling year on Copernicus Open Access Hub. Because of all that constraints, we finally chose a fixed period for all twelve-month time-series, with a staring date on the 1st of March 2020 and an ending date on the 29th of February 2021 (the choice of the recent period being linked to the temporal distribution of the number of occurrences, see figure 3.2B).

Data selection based on cloud cover. Remote sensing data at RGB/IR channels are directly dependent on potential clouds covering the satellite's field of view. For-

tunately S2 products are including in their metadata a cloud cover percentage (cc%). Thereby when querying the *Sentinelsat* API over a given area and time window, one can ask to only keep the less cloudy products. The wider the chosen time window is, the more likely an almost cloud-free product will be available within. Based on this metadata, we selected the least cloudy S2 products within each month in the targeted time window. With this selection process, we expect the large majority of time-series to be cloud-free like Figure 3.5A or 3.5B. Figure 3.6 provides an overview of the cloud coverage distribution in selected products compared to all available products in the queried time window. When, despite our efforts to select least cloudy products, the obtained satellite data around an occurrence present many cloudy frames, it could nonetheless be interpreted as an information contributing to the species ecological niche. Furthermore, in this case the environment structure can still be captured from clear scenes at other dates of the time-series (see for instance April, May and November 2020 on Figure 3.5C).

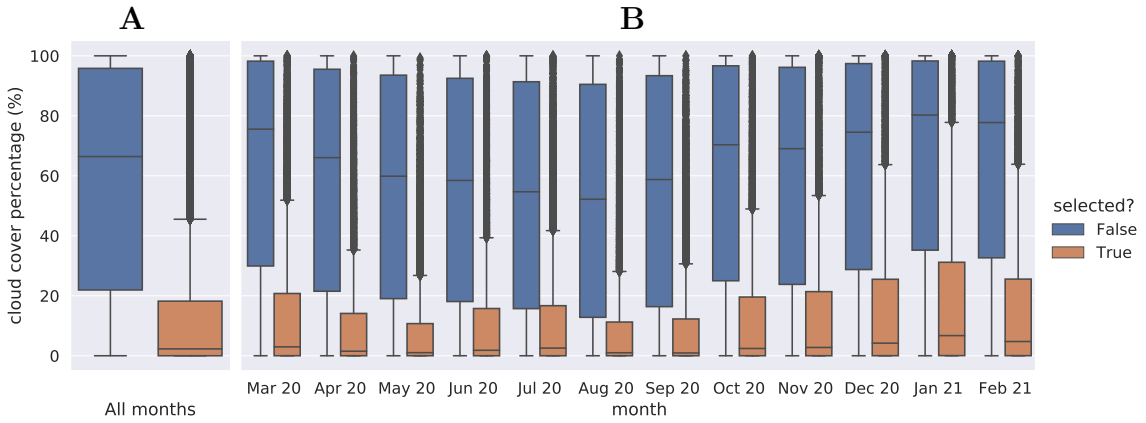


Figure 3.6: *Cloud cover percentages of the 1,067,989 tested products, 180,747 (16.9%) selected against 887,242 (83.1%) dismissed.* ***(A)*** *all months taken together,* ***(B)*** *detailed by month.*

3.2.2 Species Distribution Models trained with satellite image time-series

In this section, we describe the architecture and learning procedure of the deep-SDMs that we trained based on the *DeepOrchidSeries* dataset described above. Given an image time-series as input, the model estimates orchids *relative* probabilities of presence.

3.2.2.1 Model definition and training procedure

Model architecture. The model used is an extended version of the Inception v3 (Szegedy et al., 2016b) Convolutional Neural Network (CNN). Inception networks are appreciated because of their capacity to grasp patterns -here environmental patterns- at multiple scales. It has been shown in (Deneu et al., 2021b) that this architecture provides better species prediction performance than point neural networks, boosted trees or random forests. We use this work to justify our choice of model. Nevertheless, testing other recent neural architectures specifically designed to deal with spatio-temporal data

is an avenue to be exploited in the future, see the second perspective of the discussion. In particular, the performance gain was shown to be the most significant for rare species. In our context, the Inception v3 architecture was modified so as to accept not only RGB images but the full RGB+IR image time-series. Our inputs are of size (N_f, N_x, N_y) with N_f the number of features equal to $12 * 4 = 48$ (12 months x 4 RGB+IR channels), and $N_x = N_y = 64$ (corresponding to 640 x 640 m quadrats at 10 m resolution). To speed up the training and regularize the model, batch normalisation (Ioffe & Szegedy, 2015) was applied on the convolutional layer activations, just before the nonlinear ReLu function. Dropout (Srivastava et al., 2014) was finally used to prevent the network from overfitting (with a dropout probability of 0.5).

Model loss. The models were trained using the Label-Distribution-Aware Margin (LDAM) loss Cao et al. (2019) designed for strong class-imbalance multi-class classification problems. In our context, it allows pushing upwards rare species performance without deteriorating predictions on common species. The LDAM loss is a *label-distribution-aware* function that leads the model to an optimized trade-off between per-class margins. When considering two species only, say one rare and one common, the decision boundary drawn by this loss will be slightly shifted towards the common species in order to let the benefit of the doubt to the rare species (see Cao et al., 2019 Figure 1 for a meaningful scheme). The LDAM loss has been shown to perform very well in many deep learning benchmarks involving both a strong imbalance between classes and a high inter-class ambiguity.

Training procedure. The models were fitted using a stochastic gradient descent on multi-GPU nodes from the IDRIS supercalculator Jean Zay[5]. They were trained during 70 epochs with a batch size equal to 64. The training process took around 100h per model (with 8 gpus working in parallel). Convolutional and linear layers weights were initialised from a truncated normal continuous random variable. The Deferred Re-Weighting (DRW) training schedule associated to the LDAM loss was used. DRW is a vanilla empirical risk minimization until a given epoch, here 65. Then, the training ends with a re-weighted loss and SGD steps with a re-normalized learning rate, both by batch species frequency. The learning rate was initialised to 0.1 and later decayed by a ten factor at epochs 50 and 65. A trained model is approximately 600 MB.

3.2.2.2 Performance evaluation of the model

Data split. The *DeepOrchidSeries* dataset was split in three parts: (i) Training set (90%), (ii) Validation set (5%) and (iii), Test set (5%). Following the recommendations of (Roberts et al., 2017), the split was done using a spatial blocking strategy that enables a more robust estimation of the model's performance. The spatial blocks were defined in the spherical coordinate system according to a 0.025° grid, i.e. square blocks of 2.775 km at the equator. Splitting by block is important to impede the model from being validated or tested at locations very close to the training occurrences. In addition to the spatial blocking, we also used a stratified sampling strategy to ensure that any region

[5] http://www.idris.fr/annonces/annonce-jean-zay-eng.html

of the world has a minimal number of blocks in the training set. We therefore used the World Geographical Scheme for Recording Plant Distributions (WGSRPD) level 2 regions (Brummitt et al., 2001). Within each region, we randomly sampled 90% of the blocks present and assign them to the training set. The remaining blocks were assigned to either the validation set or the test set (at random). Validation and test occurrences from species not in the training set were removed. Table 3.1 provides the number of occurrences and species in each set.

Table 3.1: *Summary table of the number of occurrences and species in the training, validation and test sets.*

Set	Training	Validation	Test
#occurrences	897,296	51,116	50,375
#species	13,700	4,290	4,261

Evaluation metrics. Our model being trained with a multi-class classification loss on presence-only data, its output is a categorical probability distribution of the form $\eta_s(x) = \mathbb{P}(Y = s|X = x)$ where x is the input tensor (i.e. an RGB+IR image time-series), Y the observed species and $\eta_s(x)$ is the estimated probability that the observed species is s conditionally to x. Because the output is a categorical probability distribution, we have that the sum of probabilities over all species is equal to one ($\sum_{s=1}^{m} \eta_s(x) = 1$). To evaluate the model, we chose not to use pseudo-absences because of the bias induced by such methods (Botella et al., 2020; Phillips et al., 2009). Instead, we used a set-valued metric (Chzhen et al., 2021) to assess the quality of the species assemblage predicted by the model for a given input. Specifically, we chose the commonly used *top-k accuracy* as suggested in (Botella et al., 2019). It measures the success rate of the model when it returns the top-k most probable species for any input x. More formally:

$$\mathrm{A}_k = \frac{\sum_{i=1}^{n} \mathrm{A}_k(i)}{n} \tag{3.1}$$

where n is the number of occurrences in the test set (or validation set) and:

$$\mathrm{A}_k(i) = \begin{cases} 1 & \text{if } \eta_{y_i}(x_i) \geq \tilde{\eta}_k(x_i) \\ 0 & \text{otherwise} \end{cases}$$

with y_i the true species label of occurrence x_i and $\tilde{\eta}_k(x_i)$ the outputs of the model re-ordered in decreasing order of probabilities.
Because of the high class imbalance of our dataset, a shortcoming of this metric applied on all test occurrences taken together (or *micro-average*, Sokolova and Lapalme, 2009) is that it gives far too much importance to the most frequent species over the less frequent ones. To compensate for this imbalance, it is preferable to use the *macro-average* version of this metric (Sokolova & Lapalme, 2009) consisting in first calculating the score of each species and then averaging the scores over all species. More formally, the macro-average top-k accuracy (MSA) can be defined as:

$$\mathrm{MSA}_k = \frac{\sum_{s=1}^{l} \mathrm{SA}_{k,s}}{l} \tag{3.2}$$

where l is the number of species in the test and $\mathrm{SA}_{k,s}$ is the top-k accuracy for species s defined as:

$$\mathrm{SA}_{k,s} = \frac{\sum_{y_i=s} \mathrm{A}_k(i)}{n_s} \tag{3.3}$$

with n_s the number of occurrences of species s in the test set. During the training phase of the model, the macro-average top-k accuracy is computed on the validation set every two epochs for $k = 30$. The model selected in the end is the one with the highest value.

To analyse the performance of the model according to the number of occurrences available in the training set, we also measured the macro-average accuracy on subsets of species categorised by range of their number of occurrences. If we denote as N_s the number of occurrences of a species s in the training set, we can define as $\mathrm{S}_I = \{s \mid N_s \in I, n_s > 0\}$ the set of species in the test set having a number of training occurrences in a given interval I. The macro-average accuracy for a given interval I is then defined as:

$$\mathrm{MSA}_{k,I} = \frac{\sum_{s \in \mathrm{S}_I} \mathrm{SA}_{k,s}}{|\mathrm{S}_I|} \tag{3.4}$$

Another batch of experiences will focus on performances per geographic region. Spatial units are taken from the WGSRPD. The level 3 division defines the *botanical countries* that we exploit. Performance per region r is denoted as $\mathrm{RA}_{k,r}$ and is defined as the micro-average top-k accuracy computed only on the occurrences encompassed in r:

$$\mathrm{RA}_{k,r} = \frac{\sum_{x_i \in r} \mathrm{A}_k(i)}{n_r} \tag{3.5}$$

where n_r is the number of test occurrences in r. Regions with n_r fewer than 50 occurrences were excluded as statistically insignificant. Further, performance per region is compared with regions species diversity. Therefore, we computed the *diversity index* ${}^q\mathrm{D}_r$ of each region r according to the definition of Hill (1973) and Jost (2006). It is a quantitative measure of biodiversity combining, in a given region, species richness with species relative prevalence. The term prevalence is used instead of abundance to account for the observation bias in our data. Species richness corresponds to the number of distinct species observed (denoted L_r). Species relative prevalence is the share of species occurrences compared to all region's observations: $p_{s,r}$ equals $\frac{n_{s,r}}{n_r}$, with $n_{s,r}$ the number of test occurrences from species s in r. The general expression of region's diversity index is:

$${}^q\mathrm{D}_r = \begin{cases} \left(\sum_{s=1}^{\mathrm{L}_r} p_{s,r}^q\right)^{\frac{1}{1-q}} & \text{if } q \neq 1 \\ \exp\left(-\sum_{s=1}^{\mathrm{L}_r} p_{s,r} \ln(p_{s,r})\right) & \text{if } q = 1 \end{cases} \tag{3.6}$$

where q is a parameter weighting the trade-off between the importance granted to species richness (small value) VS relative prevalence (big value). ${}^0\mathrm{D}_r$ results in region species richness and ${}^1\mathrm{D}_r$ is the exponential of the Shannon entropy (Shannon, 1948). Performance per region is then averaged per categories I on the diversity index and written as $\mathrm{MRA}_{k,I}$. In the literature, the majority of studies involving species diversity use it as *response*

variable. They are focusing on its potential drivers like bio-climatic variables, topographic heterogeneity or forest structure (Hakkenberg et al., 2016; Thuiller et al., 2006a). Here, we exploit species diversity as an *explanatory* variable possibly explaining our model performances. In a similar manner, (Emerson & Kolm, 2005) defended that species diversity is a driver of speciation and (Dawud et al., 2016) examined its influence on soil carbon stocks among others.

3.2.2.3 Interpretability experiments: quantifying the contribution of temporal information

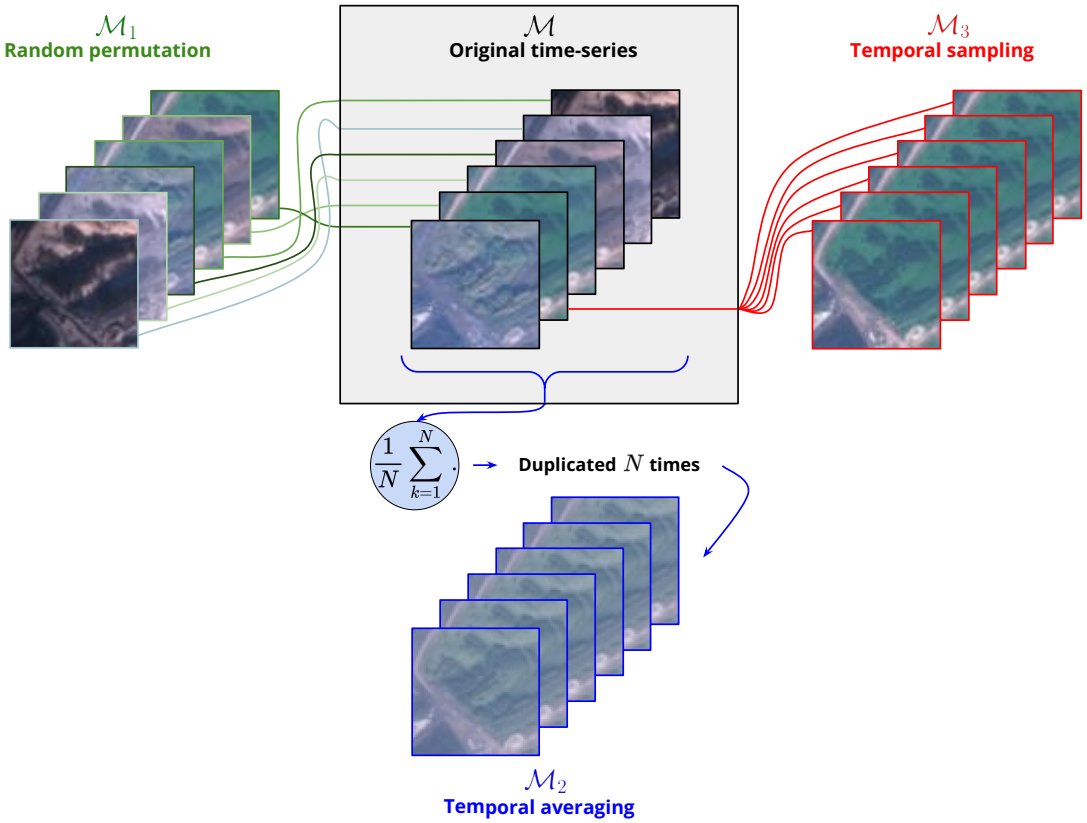


Figure 3.7: *Scheme illustrating the three transformations applied to the input image time-series towards interpreting the contribution of the temporal information. Only 6 RGB images are depicted but these procedures are applied on the whole twelve-month-long time-series, IR channel included (here $N = 6$ but would normally equal 12). The central image time-series $\mathcal{M}$ in black corresponds to the original data, i.e. to the images stacked in chronological order. The image series $\mathcal{M}_1$ is obtained by randomly permuting the original time-series. The image series $\mathcal{M}_2$ is constructed by averaging the 12 images of the original time-series and replicating the resulting mean image N times. The image series $\mathcal{M}_3$ is made of one month picked at random and replicated N times. Please note that the same legend's colours will be used in the figures of the paper presenting the results of these experiments.*

We designed several tests to analyse to what extent the trained model uses the temporal information contained in the image time-series. The general principle is to transform

the input data in order to suppress some information and to retrain a new model based on this transformed data. The comparison of the model deprived of information with the original model then allows to quantify the importance of the suppressed information. Figure 3.7 gives a comprehensive overview of the procedure detailed hereafter:

$\mathcal{M}$ **Original time-series.** This is the default original model where the input image time-series are kept unchanged (stacked in chronological order). Here the model can learn from the temporal dynamics present in the series. The filters learned by the Inception v3 model are themselves ordered feature maps time-series of 12 months and are likely to capture spatio-temporal redundancies in the input data (e.g. seasonal variations of the environment or phenological patterns).

$\mathcal{M}_1$ **Random permutation.** In this model, the 12 images of the original time-series are randomly shuffled so that the model can no longer base its predictions on the actual temporal sequencing (Garnot et al., 2019). All input variance and spatial information remain nonetheless in the input. The filters learned by the Inception v3 model can no longer be specialised by month. Nor can the model differentiate relations between months input. It actually learns from the block of twelve months considering them all equally. This procedure is comparable to the variable importance technique where a given input variable is randomised across samples to test how the model perform without its contribution. However, here we do not randomise a given feature across samples, but features order independently for each sample.

$\mathcal{M}_2$ **Temporal averaging.** In this model, the input image series are reduced to the mean over the twelve months replicated twelve times. Only the first moment of the distribution over the time dimension is kept and the model only "sees" a mean landscape averaged along the year. The objective here is to test to what extent a simple temporal averaging is sufficient to sum up most of the temporal variation. Each month contributes equally to the mean and the result is blurry. Variance between months has been totally removed. Ecological gradients of the different patch elements are reduced to their sum divided by twelve.

$\mathcal{M}_3$ **Temporal sampling.** In this model, the input image series are reduced to only one month picked at random and replicated twelve times. The neural network is being provided with only a twelfth of the predictive data and is deprived of any temporal information.

Please notice that for each of the cases ($\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$), the data transformation is applied once on the whole dataset (including training, validation and test set) before the model is trained and evaluated.

Model $\mathcal{M}_1$ being deprived only of the months order information, its comparison with Model $\mathcal{M}$ can be interpreted as a statistical test of the hypothesis that the composition of species depends on the existence of months specific features, in particular the ones resulting from yearly seasonality cycles. The comparison between $\mathcal{M}$ and $\mathcal{M}_3$ can be interpreted as a test of the hypothesis that the species composition does or does not depend on any temporal variability. Model $\mathcal{M}_2$ can be seen as an intermediate scheme where the temporal variability is summarised only by the mean of the distribution.

Accordingly, the comparison between $\mathcal{M}_1$ and $\mathcal{M}_2$ allows to assess how useful statistical moments of higher order than the mean are for characterising the temporal variability.

To compare the performances of two different models, say $\mathcal{M}$ and $\mathcal{M}_i$ with $i \in \{1, 2, 3\}$, for a given species s in the test set, we set a metric down called *relative performance change* of $\mathcal{M}_i$ compared to $\mathcal{M}$, defined as:

$$\mathrm{S}\Delta_{k,s}(\mathcal{M}, \mathcal{M}_i) = \frac{\mathrm{SA}_{k,s}(\mathcal{M}) - \mathrm{SA}_{k,s}(\mathcal{M}_i)}{\mathrm{SA}_{k,s}(\mathcal{M})} \tag{3.7}$$

where $\mathrm{SA}_{k,s}$ is the top-k accuracy of species s (see Eq. 3.3).

In the same manner that we defined the macro-average accuracy per category I on species training set number of occurrences, we can now consider the mean relative performance change per category between two models:

$$\mathrm{MS}\Delta_{k,I}(\mathcal{M}, \mathcal{M}_i) = \frac{\sum_{s \in \mathrm{S}_I} \mathrm{S}\Delta_{k,s}(\mathcal{M}, \mathcal{M}_i)}{|\mathrm{S}_I|} \tag{3.8}$$

Relative region performance change $\mathrm{R}\Delta_{k,r}(\mathcal{M}, \mathcal{M}_i)$ is also calculated as $\frac{\mathrm{RA}_{k,r}(\mathcal{M}) - \mathrm{RA}_{k,r}(\mathcal{M}_i)}{\mathrm{RA}_{k,r}(\mathcal{M})}$. This measure is averaged per categories I on the diversity index as well and is represented by $\mathrm{MR}\Delta_{k,I}(\mathcal{M}, \mathcal{M}_i)$.

When computing $\mathrm{S}\Delta_{k,s}(\mathcal{M}, \mathcal{M}_i)$ (resp. $\mathrm{R}\Delta_{k,r}(\mathcal{M}, \mathcal{M}_i)$) between $\mathcal{M}$ and $\mathcal{M}_i$ models for a given species s (resp. a given region r), it is beforehand necessary to make sure that the denominator, $\mathrm{SA}_{k,s}(\mathcal{M})$ (resp. $\mathrm{RA}_{k,r}(\mathcal{M})$), is not null. It can sometimes be when model $\mathcal{M}$ fails to predict the correct label for all s occurrences (resp. all occurrences in r). In this case, no performance change can be calculated since it is already null. Species s (resp. region r) is then removed from the calculation of the mean performance change by categories on species training set number of occurrences (resp. on regions diversity index). This is why there is a drop of support between Figure 3.9 (resp. Fig. 3.10) left and right graphs, i.e. there is fewer species (resp. regions) encompassed in the categories, as indicated on the horizontal axis. This effect is a lot more important on the support of the species mean performance change than on the region's one. To sum up, relative performance change can not be calculated for species or regions having already the lowest possible score with the whole temporal information. They are in that case discarded from the mean performance change calculation.

3.2.2.4 Modality contribution on a global scale

We train three models separately, one with each of the three modalities tested:

- The model trained with satellite imagery is the $\mathcal{M}$ model described above (original time-series).
- The bioclimatic model is trained with the 19 variables from WorldClim 2 (Fick & Hijmans, 2017). These are average monthly climate data based on summary statistics (minimum, mean, maximum) of temperature and precipitation for 1970-2000 at a spatial resolution of 30 arc seconds (1 km^2). As we keep the same

input data format as for satellite imagery, bioclimatic values are repeated to fill 640 x 640 m patches of 10 m resolution.

- The third model is trained with:

  i) Elevation from the ASTER global digital elevation model (Tachikawa et al., 2011) at 1 arc second (30 m^2)
  ii) Latitude and longitude of observation
  iii) Human Footprint individual rasters (Venter et al., 2016, built environment, population density, electrical infrastructure, cropland, pastureland, roads, railways and navigable waterways) for the years 1993 and 2009 at 1km^2 resolution.
  iv) Olson's ecoregions (Olson et al., 2001)

Again, all variables are upsampled to respect the input format of 640 x 640 m at 10 m resolution. The training procedure is the same as for the satellite image time-series experiments.

3.3 Results

3.3.1 Model validation and performance

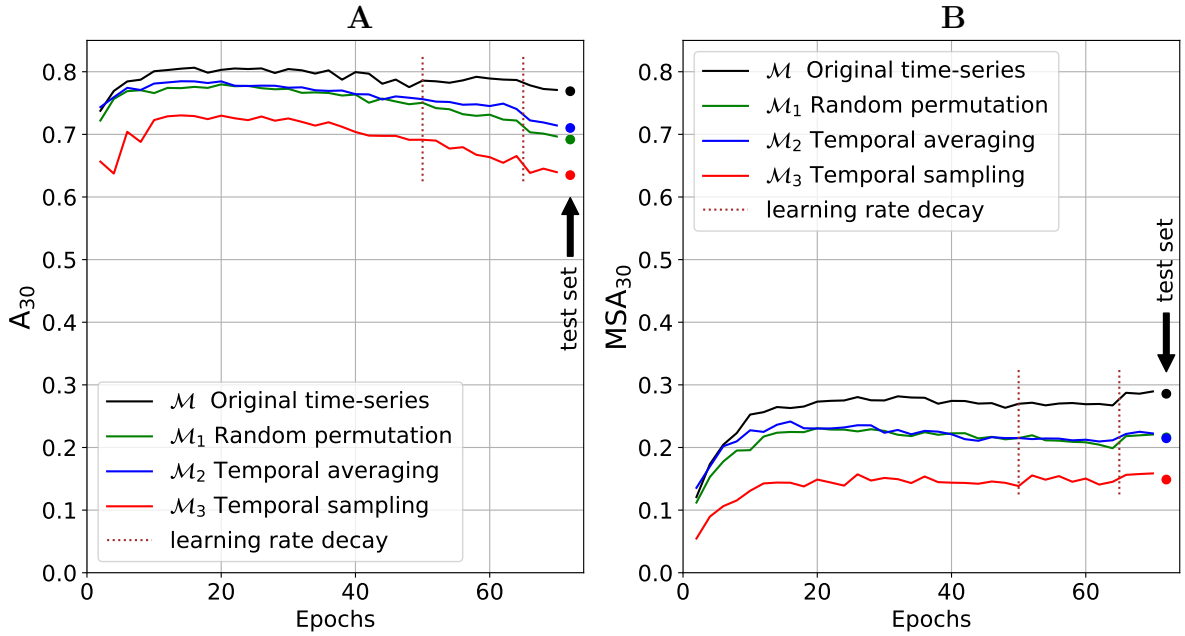


Figure 3.8: *Micro* ***(A)*** *and macro* ***(B)*** *average top-30 accuracy for model validation and test sets. Micro-average results tend to represent common species whereas macro-average performances are more representative of rare species.*

The top-30 and macro-average top-30 accuracy of the four models ($\mathcal{M}$, $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$) are presented on Figure 3.8 (at each epoch of the training phase for the validation set and on the test set for the final selected model). Due to the long-tail distribution

of species occurrences (Fig. 3.2A), the top-30 accuracy A_{30} is representative of the performance on the most common species whereas the macro-averaged top-30 accuracy MSA_{30} is more representative of the performance on the rare species. The final increase of the MSA_{30} score at epoch 65 is due to the DRW optimizer previously described: re-weighting the loss towards training's end enables a boost on rare species performances (Cao et al., 2019). The top-30 accuracy A_{30} tends to slightly decrease after the first quarter of the training phase. Our hypothesis is that this is mainly due to the use of the LDAM loss: as the training goes by, the models are reaching a better estimation of rare species ecological niche and tend to predict them more often to the detriment of common species that were chosen by default.

The model $\mathcal{M}$ trained and tested with the original time-series provides better results than the three other models deprived from temporal information. $\mathcal{M}$ is the only one where the temporal dynamics are undamaged and hence fully exploitable to statistically draw predictions. The macro-average top-30 accuracy is 0.286 for the unaltered model $\mathcal{M}$, against 0.216 for $\mathcal{M}_1$ trained on shuffled data, 0.215 for $\mathcal{M}_2$ trained on the yearly mean and 0.149 for $\mathcal{M}_3$ trained on a single random month. The following analyses can be made of these results:

1. The strong performance decrease between $\mathcal{M}$ and $\mathcal{M}_3$ shows that the temporal information contained in the time-series is a key factor of the predictive performance. For most species, it appears to be as important as the spatial information alone (cf. macro-average accuracy plot MSA_{30}).

2. The comparison between $\mathcal{M}$ and $\mathcal{M}_1$ shows that the decisive temporal information is largely related to the order of the images in the time-series, i.e. to the months specific features captured by the model (such as the ones resulting from yearly seasonality cycles).

3. The comparison between models $\mathcal{M}_1$ and $\mathcal{M}_2$ shows that their performances is almost identical (cf. MSA_{30} plot). This means that the decisive information related to the unordered temporal variability can be synthesised efficiently by the mean of the time-series. In other words, higher order statistical moments of the temporal dynamic independent from the time of the year are likely to be useless for predicting species composition (e.g. the standard deviation of acquisition noise).

4. The comparison between models $\mathcal{M}_1$ and $\mathcal{M}_3$ shows that the decisive temporal information is also largely explained by the unordered temporal variability of the images (typically due to some stochastic processes independent from the time of year).

3.3.2 Results by number of species occurrences

Figure 3.9A displays the performance of the four models as a function of the number N_s of species occurrences in the training set (cf. equation 3.4). Not surprisingly, we can observe that the accuracy of the model is positively correlated with the number of occurrences. The more occurrences in the training set and the better the top-30 accuracy. It should be noted, however, that the performance on the rarest species remains much

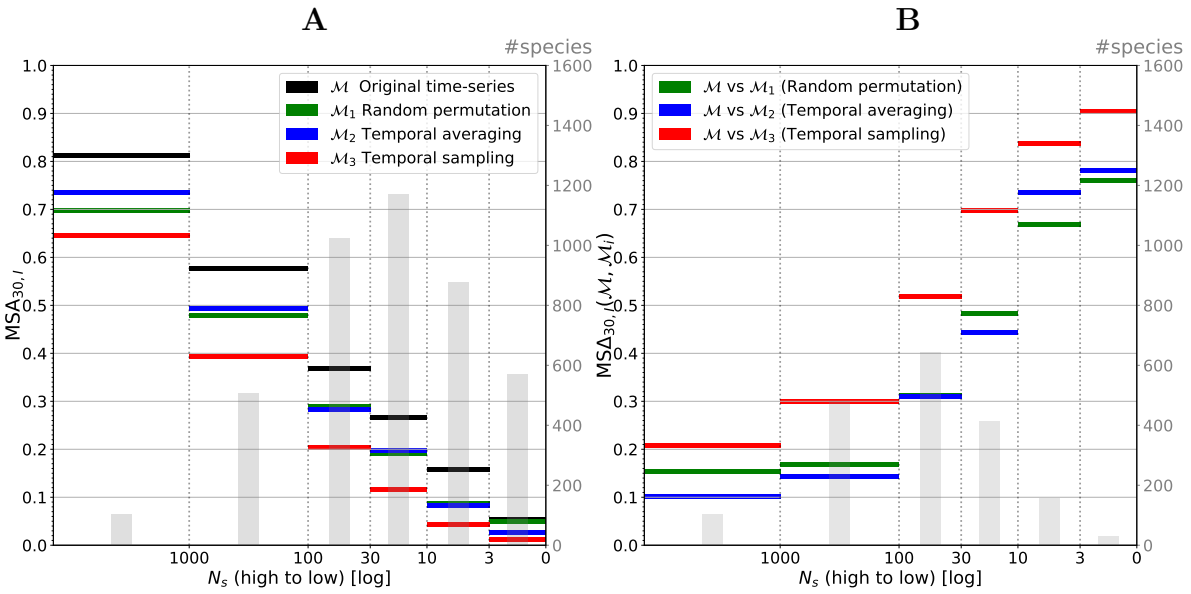


Figure 3.9: *Macro-average top-30 accuracy* ***(A)*** *and relative top-30 accuracy change* ***(B)*** *averaged per categories of number of species occurrences in the training set. All models performances are following the drop of N_s when relative performance changes are inversely proportional to it.*

better than that of a random predictor. Species having between 3 and 10 occurrences, for instance, are predicted in the set of the top-30 most probable species in 17% of the cases. A random predictor over the 13,700 species of the training set would have a top-30 accuracy below 0.22 %.
Figure 3.9B displays the mean relative performance change between the unaltered model $\mathcal{M}$ and the three models $\mathcal{M}_i$ ($i \in \{1, 2, 3\}$) as a function of the number of species occurrences (see Equation 3.8). It shows that the relative performance drop is inversely correlated with the species number of occurrences. In other words, the rarer the species (in the data), the higher the performance gain obtained thanks to the temporal information. This can be explained by the fact that this is precisely on rare species predictions that the room for improvement is the bigger, as depicted on graph 3.9A. The use of time-series thus makes it possible to compensate for the lack of occurrence data by an increased knowledge of the temporal dynamics of the environment.

3.3.3 Results by region and regional diversity index

Figure 3.10 displays all results related to the regional analysis of our models.

The first sub-graph 3.10A shows that the predictive performance of the four models is negatively correlated with the regional diversity index. Regions with small diversity indexes ${}^1D_r$ are the ones where the model predictions are the better. On the contrary, regions with high diversity see the models achieve poor performance. With $q = 1$, the diversity index equals the Shannon entropy exponential. This measure strongly depends on species richness. Hence areas with high diversity are where there is a lot of possible different orchids. This means many possible classes for the models and a high risk of confusion between species with similar environmental preferences. Moreover, these areas

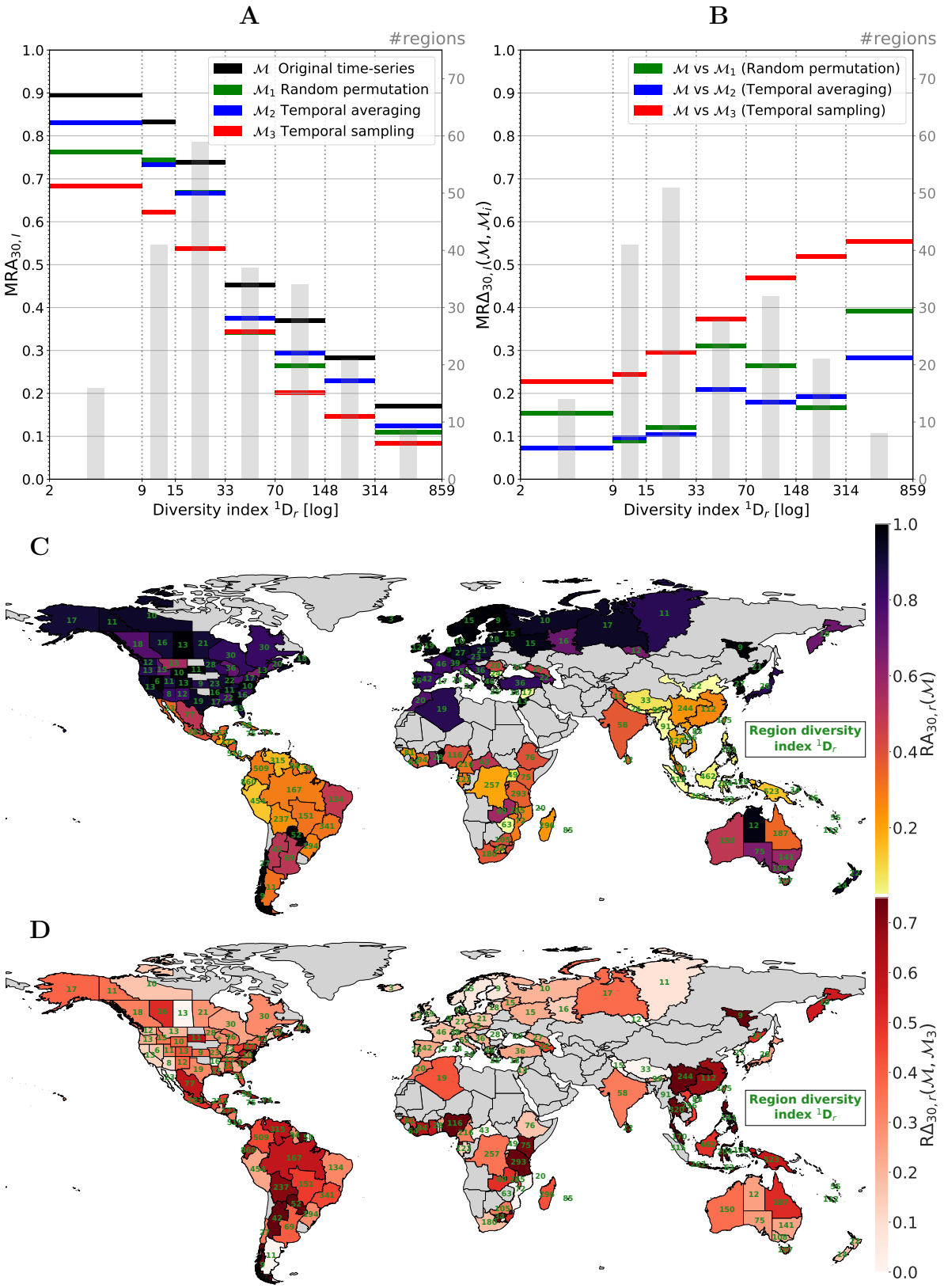


Figure 3.10: *Region top-30 accuracy* ***(A)*** *and relative top-30 accuracy change* ***(B)*** *averaged per cat. of* ^1D_r*. Map* ***(C)*** *presents region top-30 accuracy with* ^1D_r *indicated in green. Map* ***(D)*** *illustrates spatial decreases of performance when comparing* $\mathcal{M}_3$ *to* $\mathcal{M}$*, i.e. without/with the temporal information.*

are often including a lot of rare species and/or are still poorly observed. Regions with low ${}^1\mathrm{D}_r$ values are regions with relatively low species richness and tend to encompassed common species that the models are predicting well (see Fig. 3.9A).

The second sub-graph 3.10B displays the relative performance change when comparing model $\mathcal{M}$ to $\mathcal{M}_i$ models, as a function of the regional diversity index. The most obvious trend is the red curve: when totally deprived from the habitat temporal dynamics, predictions on most diverse regions are proportionally more impacted than on low diversity regions. The tendency is more irregular for $\mathcal{M}_1$ and $\mathcal{M}_2$ but is globally valid too. It implies that, similarly than for rare species on Figure 3.9B, the temporal information especially benefits highly diverse areas. An enlightenment of this tendency also is that this is where the room for improvement is the largest. Models especially take advantage of further temporal information to progress on hard tasks. Figure S2 presents the results of the same experience but with categories formed on regions' number of occurrences in the training set N_r, the total number of occurrences entailed in region r during training. Unlike Figure 3.10A and 3.10B, no tendency can be drawn. It reaffirms our idea that it is regions diversity that is driving results spatially and not only the observation bias.

The map displayed in 3.10C depicts the top-30 accuracy per region achieved by model $\mathcal{M}$ (i.e. the unaltered model with original time-series). A clear difference of performances can be observed between southern and northern hemispheres. Looking at regions' diversity index ${}^1\mathrm{D}_r$, written in green on the map, allows a better understanding of this gap. Northern regions (especially northern Europe) are presenting less species and are well sampled whereas regions around and below the equator (Australia excepted) are a lot more diverse and still insufficiently observed. Models average performances are actually quite consistent on the Earth parallels. This map is the direct illustration of the 3.10A black curve.

Finally, the map 3.10D shows where the loss of the temporal information impacts the more the performances. It corresponds to Figure 3.10B red curve, when the model trained with only one randomly picked and duplicated data month is compared to the reference model trained with full time-series. Relative performance decreases in very diverse regions like southern China or in Bolivia are really pronounced. On the contrary performances in countries with low orchid diversity and well observed like Norway of Finland are relatively spared by the input reduction.

3.3.4 Statistical tests

A t-test between $\mathcal{M}$ and $\mathcal{M}_1$ species micro-average accuracies $\mathrm{SA}_{30,s}(\mathcal{M})$ and $\mathrm{SA}_{30,s}(\mathcal{M}_1)$ does confirm that results are notably different (p-value of 5e$-$42). The same conclusion arises from the comparison of the average top-30 accuracy per region: $\mathrm{MRA}_{30}(\mathcal{M}) = 0.591$ with ordered data against $\mathrm{MRA}_{30}(\mathcal{M}_1) = 0.509$ without, p-value of 3.5e$-$9. This confirms that the order of the images in the time-series does matter and that providing the data stacked in chronological order leads to significantly better performances than when providing data in random order.

3.3.5 Model evaluation regarding time and spatial data mismatches

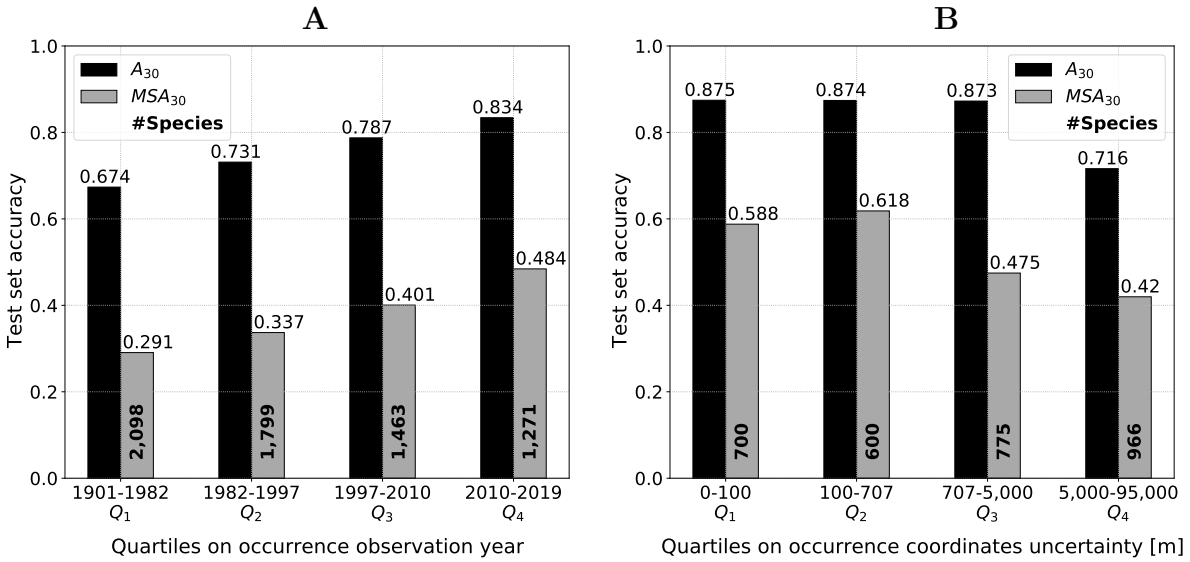


Figure 3.11: *Model performances on the test set divided by quartiles Q_i on* ***(A)*** *occurrence observation year and* ***(B)*** *occurrence coordinates uncertainty. Test accuracy is higher on more recent observations and on observations with reasonably low coordinates uncertainty.*

Figure 3.11A reveals a marked gradient of performance depending on test occurrence observation year. This analysis discarded 15% of the 50,375 test occurrences presenting no observation date information. Each quartile includes approximately 11,000 points. Both micro and macro top-30 accuracy seem to be linearly correlated to the occurrence observation year quartile. The linear behaviour is confirmed when choosing a division with a thinner percentile. Top-30 performances on the last quartile 2010-2019 are impressive: 0.834/0.484 of micro/macro average accuracy. When cutting the test set data at the median 1997, i.e. considering separately the oldest and the most recent half of test observations, performances are of 0.703/0.281 (oldest half) and 0.811/0.409 (most recent half). Moreover it should be noted that all macro-average performances calculated on test set's subsets are comparatively higher than overall performances because less distinct species are considered (see Fig. 3.11 species number in bold, against $4,261$ in the entire test set).

Figure 3.11B focuses on the influence of test occurrence coordinates uncertainty on model performance. Test set is divided by quartiles on the studied variable, likewise Fig. 3.11A. 31% of test observations do not include any information on coordinates uncertainty and are consequently put aside. Each quartile contains approximately 9,000 observations. Micro-average top-30 accuracy is identical on first three quartiles and only drops when uncertainty is higher or equal than 5,000 meters. Macro-average top-30 accuracy is similar when uncertainty is kept under 707 meters, i.e. for the first two quartiles only (it is even slightly higher for the second one). Then the macro-average performance goes a step down starting from the median 707 meters. Both micro and macro average performance are severely diminished when coordinates uncertainty is superior or equal to 5 km.

3.3.6 Modality contribution on a global scale

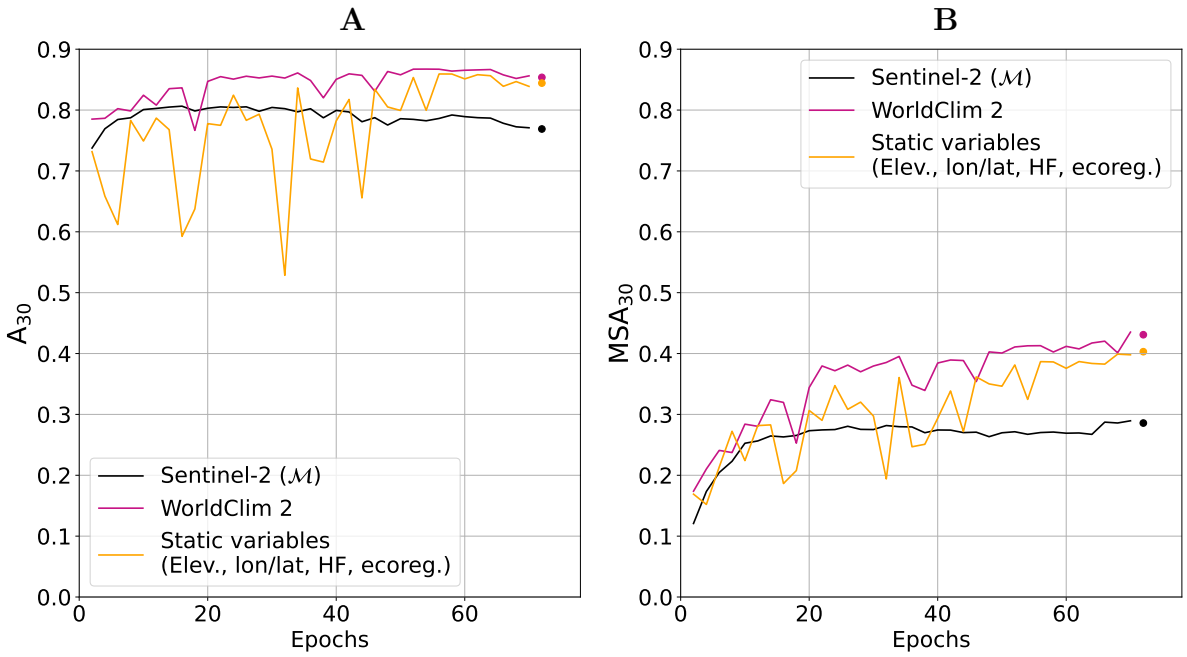


Figure 3.12: *Global comparison of micro* ***(A)*** *and macro* ***(B)*** *average top-30 accuracy on model validation and test sets (final dots) between three SDM modalities: Satellite imagery, bioclimatic variables and static variables.*

Bioclimatic variables outperform Sentinel-2 image time-series in determining global orchid distribution. Static variables also achieve better test performance than satellite data, with positional information likely to make a large contribution. This is true for both micro and macro average test performance (and therefore for both common and rare species). The training process of the model fed with static variables fluctuates strongly before stabilising after epoch 50.

3.4 Discussion

SDMs and satellite data.
Remote sensing is an invaluable source of predictive features for SDMs, and more widely for deep learning based earth observation applications (Borowiec et al., 2022; He et al., 2015; Zhu et al., 2017). Combined together, they offer a key opportunity in monitoring biodiversity facing climate change (Randin et al., 2020).

SDMs coupled with remote sensing data are often exploiting the widespread vegetation indexes EVI or NDVI (Bannari et al., 1995). These indices are computed from satellite channels and are intended to reflect vegetation properties. The NDVI is said to assess photosynthetic activity and productivity (Pettorelli et al., 2011). Texture measures derived from satellite EVI were proven adapted to map habitat heterogeneity and bird species richness patterns (Farwell et al., 2020).

The WorldClim variables (weather station data interpolated with satellite-derived covariates, Fick and Hijmans, 2017; Hijmans et al., 2005) certainly are the most widely used global SDM predictors (Nogués-Bravo, 2009; Svenning et al., 2011). This bio-climatic data approaches habitats annual trends (e.g. annual precipitation) and seasonality (e.g. temperature annual range and standard deviation). Contrary to our one-year *DeepOrchidSeries* dataset, here the variables are averaged across several decades.

SDMs and remote sensing data can also help rare species detection by capturing the biophysical conditions driving their distributions (Cerrejón et al., 2021). Recent studies have successfully leverage the spatial structure of satellite images as input to CNN-based SDMs (Deneu et al., 2021b). Trained on fine scale tensors, these models were proven able to learn and cluster species ecological preferences like annual mean temperature (Deneu et al., 2021a).

Regarding the use of the temporal dimension of satellite data in SDMs, few studies actually take advantage of it as underlined in (Randin et al., 2020). We can cite (Cord & Rödder, 2011) who tried in 2011 to include EVI seasonality information in their SDMs inputs. Their study was however on a totally different range than us since they focused on eight Mexican anurans and used one-dimensional predictors.

Benefits of Deep-SDMs trained on remote sensing image time-series.
The main outcome of our study is that using time-series of satellite images significantly improve Deep-SDM performance, in particular for rare species and in most diverse regions, supporting the interest of the approach for conservation science. Rare species are almost always threatened because of their small numbers and lack of conservation measures. Moreover, the World most diverse regions include nearly all undiscovered species (Joppa et al., 2011). Better knowledge of the ecological niche of rare or little-prospected species should foster more appropriate and effective conservation measures to ensure their survival.

We collected time-series of remote-sensing images to grasp the temporal variation in habitat properties. Our results confirm that this information is of high value to capture species ecological niches and potential distributions. Our time-series are also providing SDMs with the spatial structure of species habitats, a key information to enhance predictive performances (Deneu et al., 2021b). Recent satellite missions offer both high temporal revisit frequency and high spatial resolution at the global scale, supporting the use of such data for niche modelling. Exploiting even more intensive remote sensing data, e.g. all products without any selection by month or on a wider time window, would probably allow even better ecological niche estimation. That said, the Sentinel-2 data curation we devised here represents a good trade-off to acknowledge the phenology of orchid habitats at a broad spatial scale. Trying to avoid as much as possible clouds on selected images was also a sensitive point in our dataset creation workflow. A thinner temporal resolution would have resulted in richer time-series, but also a higher number of cloud frames. The question of whether the presence of clouds is in itself a relevant information for characterising the environment was not addressed in our study and remains nonetheless an open question.

Comparison with other open remote sensing datasets for deep learning.
Remote sensing datasets for deep learning applications are currently gaining much interest and are more and more accessible. The very recent launch of *TorchGeo* (Stewart et al., 2021), a Python library to easily handle geospatial datasets in the PyTorch environment, illustrates the recent and still ongoing progresses. However, the available datasets remain currently few and the temporal information provided by satellite revisits is almost never used (Sumbul et al., 2019). The available datasets are mostly used for land-cover classification (Helber et al., 2017) or semantic segmentation (Schmitt et al., 2019), as described in the benchmark datasets provided in TorchGeo (see Stewart et al., 2021 Table 1). *Sen12MS* is for instance a global dataset including 180,662 patches of Sentinel-1/2 256 x 256 m images and MODIS-derived land cover maps (Schmitt et al., 2019). Another dataset, similar to ours in terms of spatial coverage, is named Seasonal Contrast (*SeCo*) (Mañas et al., 2021) and was released in 2021. It gathers 2.65 x 2.65 km Sentinel-2 image time-series around about 200K locations worldwide. Time-series include 5 images separated by approximately 3 months. The objective was to learn an encoder that can be used for a variety of tasks, from land-cover classification to change detection. *SeCo* includes images from all over the world to represent a wide variety of landscapes. Among the currently available and open datasets, our dataset is, to the best of our knowledge, the only one providing monthly image data at so many points worldwide. In order to allow its reuse and the reproducibility of our experiments, the entire dataset is made publicly available with the Zenodo DOI *10.5281/zenodo.4972593*. We also share the scripts that allowed to create it at https://gitlab.inria.fr/jestopin/sen2patch. In particular, these can be used to collect new image time-series at locations other than those covered by our dataset.

Interpretability: in which cases is the modelling of the temporal dynamics the most beneficial ?
One of the major conclusions of our study is that the regions benefiting the most from a performance gain due to the modelling of the temporal dynamics of satellite images are those with the highest species diversity. This conclusion may seem counterintuitive at first. Indeed, the regions with the highest diversity are often located towards the tropics and are not those with the most pronounced seasonal patterns. Consequently, the image time-series in these regions are not expected to be the ones with the strongest temporal signal. However, it is important to understand that the model operates on a global scale with thousands of habitats to discriminate from each other. Whatever the temporal signature of a given habitat, it is a useful information for distinguishing it from other habitats. At the extreme, the temporal signature of a constant habitat throughout the year is a strong marker of that habitat. A study lead in Mediterranean natural habitats analyzed habitat discrimination from a variety of multispectral sensors answers simulated from field measurement, including Sentinel-2 (Féret et al., 2015). They showed that multi-temporal acquisitions outperform single data acquisition to discriminate habitats.

The reason for the higher performance gains in high diversity regions is actually more related to the higher model uncertainty in that regions. Species from these regions are indeed those for which there is the least amount of occurrence data available and our study clearly demonstrates that the performance gain is strongly correlated with this variable. In other words, our study shows that the addition of the temporal information

allows to reduce the model uncertainty related to the lack of occurrence data in high diversity region. This result appears particularly interesting since habitats with the highest diversity and the rarest species are also the most threatened ones and modelling them is essential to put in place adapted conservation measures.

Key considerations for building new models with our method or using existing ones.
Our method could be readily applied to other taxonomic groups than the orchids family. The ease and cost of implementation will mainly depend on the geographical distribution of the occurrences of the target taxon. With a family as large and widespread as the orchids, our method requires significant computing resources. Downloading Sentinel-2 tiles at a very large extent demands a lot of storage available (about 100Tb). To keep model training time reasonable, GPUs have to be used too. A computing cluster is more than welcome and the technical requirements can be a limitation for some researchers. However, once the dataset is built and the model is trained, predictions can perfectly be run on standard local machines. To this end, the model built for our study is shared publicly in the same Zenodo repository as the dataset (DOI 10.5281/zenodo.4972593). Providing new S2 image time-series as input, it can be used to predict species orchids composition anywhere on earth or to build high-resolution maps of specific orchid species at global scale. It may also be used for other ecological tasks via transfer learning approaches (i.e. keeping unchanged all the weights of the model except those of the last layer dedicated to species classification, Torrey and Shavlik, 2010).

On temporal and spatial biases.
In the context of species and habitats distribution modelling in general, a recurrent challenge is the possible mismatch, both in time and space, between the occurrences and the environmental variables (Phillips et al., 2006). As shown in Figure 3.2B, in particular, a fraction of the occurrences in our dataset date from several decades ago, while the satellite data is from March 2020 to February 2021. If the environment changed since the observation, e.g. because of a housing project or deforestation, the model may learn incorrect relationships. Figure 3.11A focuses on this particular issue and acknowledges the influence of occurrence observation date on model performances. Top-30 test accuracy is gradually higher on more recent occurrences than older ones. Interestingly, common and rare species predictions seem to respond in the same manner to temporal shifts between predictive habitat data and species observation dates.

Spatial mismatch can also happen because of the occurrences position uncertainty (See Fig. S1). However, our model being based on convolutional filters, it is highly robust to such spatial shifts until the true occurrence position does not exceed the extent of the input image (here, 640 x 640 m). Ideally, only occurrences with a position uncertainty of less than 320 m (half of the patch size) should be considered with our method. Figure 3.11B traduces the impact of test occurrence coordinates uncertainty on model performance. As expected, top-30 accuracy drops when uncertainty is substantial and there is actually very little chance that the predictive data is anywhere near the actual observation place (see performances on Q_4 quartile). Besides performance on both

common and rare species remains almost constant when uncertainty is inferior to the median equal to 707 meters. Thereby, when the maximum uncertainty is of the order of the patch size, the model performs as well as on very precise occurrences. Finally, the Q_3 marked difference of evolution between micro/macro top-30 accuracy could be explained by the following hypothesis: rare species predictions are more affected by a growing coordinates uncertainty than common species because of more locally specific habitat preferences.

In machine learning, such mismatch between labels and predictive data is called *label noise* (Frenay & Verleysen, 2014) and is actively studied (Ghosh et al., 2017; Lee et al., 2018). The strength of our dataset in counteracting this noise is its very large size, as demonstrated in (Rolnick et al., 2017). Their work showed that deep learning models can learn correct generalizations even with massively noisy datasets.

At last, the strong spatial bias present in the *DeepOrchidSeries* dataset influences SDMs predictions (Beck et al., 2014). Such bias results from a very uneven sampling effort (See Fig. 3.4C map) and not from orchids distribution. Use of methods to mitigate spatial bias at the cost of occurrence number is a promising direction to exploit *DeepOrchidSeries* (see abovementioned publication). Nonetheless true understanding of orchids distribution and health will only be reached with significant and uniform observation effort. Having access to constructive and global predictive data is remarkably valuable but not sufficient. Biodiversity hotspots (Myers et al., 2000) urgently need to be sampled with high standards of care to limit human disturbance. Citizen science initiatives are also contributing to enhance biodiversity monitoring worldwide (Affouard et al., 2017; Kobori et al., 2016).

Modality contribution on a global scale

Bioclimatic variables have been consistently identified to be the main driver of plant species distributions at large scales (Randin et al., 2020; Woodward, 1990). At fine scales, satellite imagery provides critical information on the habitat (environmental layout, land cover and patterns, Deneu et al., 2022). Our results confirm what is expected from the literature: bioclimatic conditions are the major determinant of vegetation distribution at large scales, ahead of non-climatic predictors such as habitat information provided by satellite imagery time-series. In turn, we assume that such information is most useful at finer scales, as highlighted by Luoto et al. (2007) and Deneu et al. (2022). The static variables provided in the third model result in a global species distribution that is, on average, almost as good as that produced by the bioclimatic model. We hypothesise that the combination of elevation, position and ecoregion can be considered a good proxy for bioclimatic state. Bioclimatic information is needed to map plant species on a global scale. By capturing habitat information, satellite imagery provides complementary rather than alternative information. However, collecting satellite imagery on a global scale is an expensive task, which is affordable when it is tailored to a specific dataset, but becomes extremely challenging when the goal is to cover global terrestrial data at the kilometre scale. Ultimately, we decided to conduct the following global-scale analyses without satellite imagery, as the cost-benefit ratio seemed unfavourable for the tasks at hand.

Perspective 1: enriching the input with other predictors informing orchids habitat.
An exciting future development is to add other relevant predictors to our models. Other image time-series like the frequently used bio-climatic variables from WorldClim[6] or ecosystem functional attributes (Arenas-Castro et al., 2018, although not independent since they also are computed from satellite data) would bring complementary information on species ecological niche. Altitude[7], available global human footprint rasters[8], soil properties variables[9] (Batjes et al., 2020) or ecoregions (Olson et al., 2001) would help to crystallise species preferences and vulnerabilities as well.

Perspective 2: Using NN architectures designed to extract long-term temporal dependencies
An active research avenue concerns adapting neural networks architectures to best analyze satellite image time-series with broad temporal and spatial coverages. Recurrent convolutional neural networks (Lai et al., 2015) achieve significant performance gain in land-cover classification tasks (Garnot et al., 2019; Rußwurm & Körner, 2018), and we anticipate it should also be relevant for the analysis of species distributions and spatio-temporal dynamics. In our case, we can suggest an hybrid architecture relying on an Inception v3 model to first extract the spatial features at each week or month and then a Recurrent Neural Network (RNN) to encode the temporal dimension over a long period of time. 3D CNNs are another promising candidate architecture but, as pointed out by (Garnot et al., 2019), convolutions in the temporal dimension are not well adapted to grasp long-term dependencies and assume a regular sampling of occurrences in time, which we do not have. Lastly, spatio-temporal encoders with temporal attention could be worth investigating time when seeing their success on other tasks like satellite time-series segmentation (Garnot & Landrieu, 2021). For now, our CNN architecture is considering the stacked time-series of size twelve as a global temporal context. It was proven suited to grasp the local landscape yearly dynamics and globally improve species relative probability of presence prediction. But with larger time-series, attributing more modelling weight to the temporal dimension will be a must. This seems all the more relevant given that predictions of rare species and in very diverse regions benefit in particular from the temporal information.

3.5 Conclusion

In this paper, we studied for the first time a worldwide species distribution model based on high resolution remote sensing image time-series. Therefore, we built and shared a substantial dataset (called *DeepOrchidSeries*) aimed at modelling the distribution of orchids on a global scale from Sentinel-2 data. The spatial structure and phenology of species habitat are captured over a whole year for 999,258 occurrences. We then

[6] http://www.worldclim.org
[7] https://lpdaac.usgs.gov/products/srtmgl1v003/
[8] https://sedac.ciesin.columbia.edu/data/set/wildareas-v3-2009-human-footprint and 1993 version
[9] https://soilgrids.org/

trained deep-SDMs resting on an Inception v3 architecture whose input was modified to deal with twelve months time-series of RGB+IR images. The analysis of the resulting model reveals that the temporal information contained in the time-series enables a strong improvement of the predictive performance compared to a purely spatial model. Thanks to interpretability experiments, we did show that seasonal patterns, in particular, are well captured, resulting in a better discrimination of habitats all over the world. We also demonstrated that occurrence-poor species and diversity-rich regions are the ones that benefit the most from this improvement, revealing the importance of habitat temporal dynamics to characterise biodiversity. We hope that this work will pave the way for even more elaborate spatio-temporal models allowing to predict future trajectories of ecosystems.

Acknowledgments

We warmly thank Alexander Zizka for providing us with the filtered set of orchid occurrences. Our dataset contains modified Copernicus Sentinel data and Copernicus Service information (2020, 2021). Sentinel-2 MSI data used were available at no cost from ESA Sentinels Scientific Data Hub. This work was granted access to the HPC resources of IDRIS under the allocation 20XX-AD011011389R1 made by GENCI. Finally, we would like to thank the reviewers for their insightful comments that helped us improve our manuscript.

Data Availability Statement

- Code is available on the *sen2patch* gitlab: https://gitlab.inria.fr/jestopin/sen2patch
- Occurrences initial GBIF query is https://doi.org/10.15468/dl.4bijtu (accessed August 2019)
- Dataset and models are available on Zenodo with the DOI: 10.5281/zenodo.4972593.

CHAPTER 4

AI-BASED MAPPING OF THE CONSERVATION STATUS OF ORCHID ASSEMBLAGES AT GLOBAL SCALE

Table of contents

4.1 Introduction 103
4.2 Materials and Methods 105
4.2.1 Taxonomic focus: the *Orchidaceae* family 105
4.2.2 Species assemblage prediction model 106
4.2.2.1 Definition 106
4.2.3 Validation 107
4.2.4 Conservation indices for species assemblages 108
4.2.4.1 Indices definition 108
4.2.4.2 Missing status completion 109
4.2.5 High-resolution maps construction 110
4.2.5.1 Global grid design 110
4.2.5.2 Maps definition and construction 110
4.2.6 Zonal statistics 110
4.2.6.1 Region spatial coverage of the most critical IUCN status 110
4.2.6.2 Region average proportions 111
4.2.7 Data 111
4.2.7.1 Orchid occurrences 111
4.2.7.2 Predictive features 111
4.3 Results 113
4.3.1 $\mathcal{I}_{\mathcal{O}}$ indicator: most critical status of the species in the assemblage 113
4.3.1.1 Global patterns 113
4.3.1.2 Country-level analysis 113
4.3.2 $\mathcal{I}_c$ indicator: proportion of species in the assemblage with a given status 114

4.3.2.1 Global patterns . . . 114
4.3.2.2 Country-level analysis . . . 116
4.4 Discussion . . . 119
4.4.1 Modelling choices . . . 119
4.4.2 Considerations on covariates . . . 120
4.4.3 Our indicators originality . . . 120
4.4.4 Orchids conservation . . . 122
4.4.5 Conclusions . . . 122

This chapter is currently **under review** in the journal *Ecological Informatics*.

Abstract

Although increasing threats on biodiversity are now widely recognised, there are no accurate global maps showing whether and where species assemblages are at risk. We hereby assess and map at kilometre resolution the conservation status of the iconic orchid family, and discuss the insights conveyed at multiple scales. We introduce a new deep species distribution model trained on 1M occurrences of 14K orchid species to predict their assemblages at global scale and at kilometre resolution. We propose two main indicators of the conservation status of the assemblages: (i) the proportion of threatened species, and (ii) the status of the most threatened species in the assemblage. We show and analyze the variation of these indicators at World scale and in relation to currently protected areas in Sumatra island. Global and interactive maps available online at https://mapviewer.plantnet.org/ show the indicators of conservation status of orchid assemblages, with sharp spatial variations at all scales. The highest level of threat is found at Madagascar and the neighbouring islands. In Sumatra, we found good correspondence of protected areas with our indicators, but supplementing current IUCN assessments with status predictions results in alarming levels of species threat across the island. Recent advances in deep learning enable reliable mapping of the conservation status of species assemblages on a global scale. As an umbrella taxon, orchid family provides a reference for identifying vulnerable ecosystems worldwide, and prioritising conservation actions both at international and local levels.

Keywords

spatial indicator, species assemblage, deep learning, species distribution modelling, IUCN status, orchids

4.1 Introduction

Nearly a million species will face extinction in the coming decades (Díaz et al., 2019), many of which having high value for medicine, food, materials, etc (Pollock et al., 2020). The Post-2020 Global Biodiversity Framework requires assessing current biodiversity state and quantifying conservation measures impacts (Nicholson et al., 2021). However, the distribution of many species is little known (Wallacean shortfall), and there is lack of comprehensive enough information on species conservation status (Schatz, 2009). Land managers still need accurate indicators of species extinction risk that should be available both at a large scale (to allow comparisons between regions) and at a sufficiently fine spatial resolution. Recent automatic assessment of conservation status (Borgelt et al., 2022a; Zizka et al., 2020) have proved promising to complement the assessment based on informing IUCN criteria, which should help tackle the major objective of intensive prediction at broad taxonomic and spatial coverage.

Species distribution and richness patterns are complex, habitat and scale dependent, which entails that species conservation status must be assessed and acknowledged at multiple spatial scales and depending on habitat variation. According to Whittaker et al. (2005), protected areas design based on species distribution and richness may be sensitive

to spatial scale, and the conservation challenges must be addressed at both global scale and fine-resolution (Puglielli & Pärtel, 2023). Here we perform (i) multiscale assessment of conservation status, based on (ii) high-resolution characterization of habitat properties, in the case of the emblematic orchid family.

Deep learning offers an unprecedented opportunity to characterize complex, scale-dependent relationships between species and their environment (Deneu et al., 2021b). In addition, the ever-increasing volume of data stemming from citizen science observations on one hand, and from remote sensing characterization of environmental heterogeneity on the other hand, requires adapted DL workflows (Borowiec et al., 2022). DL models can learn from complex effects and interactions between environmental predictors (Puglielli & Pärtel, 2023), and Cai et al. (2023) have shown that DL can help to isolate relationships between biodiversity and ecological drivers.

Understanding how threatened species are distributed is a task that ecologists have been working on since the nineteenth century (Gaston & Blackburn, 1997; Moret et al., 2019). Yet there are few quantitative studies of the distribution of threatened species (Orme et al., 2005). Successful attempts to design anthropogenic threat index at the regional scale (Paukert et al., 2011) or even worldwide with the Human Footprint (Venter et al., 2016) have lead the community to adopt this information as model predictor. However, several major questions remain unsatisfactorily answered: how do anthropogenic and bioclimatic pressures relate to species environmental niches, at what scale and to what degree? New studies in that regard consist in combining species IUCN status with known or predicted range of species and produce conservation priority maps (Han et al., 2019; Mair et al., 2021; Verones et al., 2022). For an overview of recent successful attempts, see the state of the art section 2.2.3.1. Species included in these indices must have been previously assessed and their extinction risk status officially recognised by the IUCN. However, as of 2022, only 7% of the world's described species have an IUCN status (15% for the world's known plants) (IUCN, 2022). Ultimately, there is a strong case to be made for including unassessed species in the design of spatial threat indicators.

In order to widen the currently narrow IUCN coverage, automatic classification methods have made a breakthrough. A major research avenue has emerged from this urgent task (Walker et al., 2020). Two families of methods coexist: approaches that estimate IUCN criteria variables in advance to compare with official thresholds (Dauby et al., 2017; Stévart et al., 2019), and models that directly predict IUCN status after being trained with predictors and already assessed species (Borgelt et al., 2022a; González-del-Pliego et al., 2019; Nic Lughadha et al., 2019; Zizka et al., 2021). Methods in the first category are easier to interpret by construction. However, newer predictive models achieve impressive performance. Research is also exploring the use of SDMs to inform conservation status thanks to their niche modelling capabilities (Breiner et al., 2017; Syfert et al., 2014).

SDMs are correlative models learning from the association of species observations with environmental predictors (Elith & Leathwick, 2009). These statistical tools are now widely used and ongoing methodological work continue to improve their convergence and predictive power (Lembrechts et al., 2019; Pollock et al., 2014; Powell-Romero et al., 2023). Applications at all scales contribute to grasp diversity patterns and help to hold invasive

species back (Botella et al., 2021), highlight biodiversity hotspots (Hamilton et al., 2022) or orient Protected Areas (PAs) design (Guisan et al., 2013). Deep-SDMs embrace deep learning vision architectures to leverage rare and critical environment spatial patterns (Deneu et al., 2021b; Leblanc et al., 2022). Indeed, spatial and temporal (Estopinan et al., 2022) contexts were proven significant to model rare species niches and species-rich regions diversity. These models capture the shared environmental preferences between multiple species and let information flow from the most common to the rarest species without corrupting their specific features (Botella et al., 2018b). Spatially Explicit Models (SEM) integrate the location of observations as a predictor variable. While ecologists discourage its use when modelling species' environmental preferences, it has been shown to significantly improve prediction performance and influence conservation planning (Domisch et al., 2019). SEMs can incorporate local heterogeneities, creating positive feedbacks and allowing patterns to emerge at larger scales (DeAngelis & Yurek, 2017).

Our main contribution is to produce kilometre-scale extinction risk maps of species assemblages on a global scale. A species assemblage is simply defined as *members of a community that are phylogenetically related*, where a community is *a collection of species that occur in the same place at the same time* (Fauth et al., 1996). To do this, we trained a deep-SDM model on 1M observations of 14K species distributed worldwide. We then developed a novel method to estimate species assemblages. Coupled with the species' IUCN status, the assemblages are then characterised by extinction risk indicators. Interactive maps are available online at https://mapviewer.plantnet.org/?config=apps/store/orchid-status.xml. To our knowledge, this is the first realisation of SDM-derived spatial indicators at such resolution, taxonomic and geographic coverage. Four levels of analysis are also discussed: i) How is the extinction risk of orchid assemblages distributed at different scales? ii) Which zones appear to contain the most threatened assemblages? iii) Is there a correlation between the diversity of orchids in a country and the proportion of threatened species? and finally iv) In Sumatra, how do our indicators relate to current PA implementation?

4.2 Materials and Methods

4.2.1 Taxonomic focus: the *Orchidaceae* family

The *Orchidaceae* family is a perfect taxon to guide our research, both because of its inherent nature and because of its large data coverage (Cribb et al., 2003). This uniquely diverse taxon comprises around 31,000 species, making it one of the largest flowering plant families (KEW, 2023). Diversity and aesthetic appeal of orchids have made them the focus of attention for botanists and enthusiasts for decades. This has resulted in both a rich scientific literature (Cozzolino & Widmer, 2005; Givnish et al., 2016) and a wealth of observations: 8M raw GBIF observations, including 6.8M with coordinates (GBIF, 2023). Orchids are present on all continents and are flowering in a very wide range of altitudes and habitats. This is a crucial aspect as our modelling approach aims to capture and project species preferences worldwide. The threats they face - habitat destruction, climate change, pollution and intensive harvesting - make them singularly vulnerable. Moreover orchids are a relevant indicator of the health of their environ-

ment (Newman, 2009). This well-known and change-sensitive family can be used as a proxy to identify ecosystem conservation priorities (Yousefi et al., 2020). Understanding threats, monitoring populations and distributions, and raising awareness are other key conservation objectives for the group (Wraith et al., 2020). Orchids are widely used by international institutions as flagship species to lead and give visibility to the conservation debate (Cribb et al., 2003). The challenge of orchid conservation cannot be tackled at the species level alone. Large-scale and broad approaches should necessarily complement studies carried out on emblematic species with a high risk of extinction (Fay, 2018).

4.2.2 Species assemblage prediction model

4.2.2.1 Definition

Our model for predicting species assemblages is derived from what is called *set-valued prediction* (or *set-valued classification*) in the machine learning community (Chzhen et al., 2021; Mortier et al., 2021). The model is trained on presence-only (single-label) data, but is then used to predict a set of labels by thresholding the output categorical probabilities. In more details, let us consider the following species assemblage prediction problem with C distinct species. The input set made of the predictive features associated to each occurrence location is denoted $\mathcal{X} = \{x_1, ..., x_n\}$. The matching species label set is $\mathcal{Y} = \{1, ..., C\}$. The objective is to learn a species assemblage predictor on a training dataset composed exclusively of presence-only occurrences $(x_1, y_1), ..., (x_{n_t}, y_{n_t}) \in \mathcal{X} \times \mathcal{Y}$. The pairs (x_i, y_i) are supposed to be independently sampled from a unknown probability measure $\mathbb{P}_{X,Y}$. This joint measure can be decomposed into the marginal distribution measure over $\mathcal{X}$, $\mathbb{P}_X$, and the conditional distribution of y given an input x denoted $\eta(x) = (\eta_1(x), ..., \eta_C(x))$ and equal to

$$\eta_k(x) = \mathbb{P}_{X,Y}(Y = k|X = x)$$

Then, the assemblage of species likely to be present conditionally to x can be defined as:

$$S^*_\lambda(x) := \{k \in \mathcal{Y} : \eta_k(x) \geq \lambda\}$$

where λ is a threshold on the conditional probability of species optimised to return precautionary assemblages (see next section on model validation).
In practice, the true conditional probability $\eta(x)$ is unknown and we assume we are given an estimator $\hat{\eta}(x)$ from which we can derive the following *plug-in* estimator of the species assemblage:

$$S_\lambda(x) := \{k \in \mathcal{Y} : \hat{\eta}_k(x) > \lambda\} \tag{4.1}$$

One approach to get a good estimator $\hat{\eta}_k(x)$ of the conditional probability is to fit a model using the negative log-likelihood which is known to be a strictly proper loss (Gneiting & Raftery, 2007), i.e. it is minimized only when the model predicts η. The negative log-likelihood loss is defined as:

$$l_{log}(k, \hat{\eta}) = -log\ \hat{\eta}_k(x) \tag{4.2}$$

In the context of deep learning, $\hat{\eta}(x)$ is typically chosen as a softmax function on top of a deep neural network $f_\theta(x) : \mathcal{X} \to \mathbb{R}^C$ so that:

$$\hat{\eta}_k(x) = \frac{exp(f_\theta^k(x))}{\sum_j exp(f_\theta^j(x))}$$

where θ is the set of parameters of the neural network to be optimized by minimizing the loss function of equation 4.2.
Using this very common deep learning framework, it is possible to show that the species assemblage predictor $S_\lambda(x)$ of Equation 4.1 is consistent (Lorieul, 2020), i.e. it tends towards the optimal set $S_\lambda^*(x)$ when the number of training samples increases. In other words, our species assemblage predictor is as simple as training a deep neural network with a *cross-entropy* loss function on the presence-only samples and thresholding the output softmax probabilities to get the assemblage of predicted species.

Our backbone model is an adaptation of the Inception v3 (Szegedy et al., 2016b). This convolutional neural network learn spatial patterns from two-dimensional predictors (Botella et al., 2018a; Deneu et al., 2021b). A spatial block hold-out strategy is used to limit the effect of spatial autocorrelation in the data when evaluating the model (Roberts et al., 2017). Blocks are defined in the spherical coordinate system according to a 0.025° grid (2.8 km square blocks at the equator). The split of the training/validation/test spatial blocks (90%/5%/5%) is stratified by region to ensure that all regions are represented within each set. We use the regions defined by the WGSRPD level 2 (Brummitt et al., 2001). Training is done on Jean Zay, an IDRIS supercomputer. A full description of the model architecture, dataset spatial split and training procedure can be found in supplementary information (SI) Box A. Finally, the species assemblages are post-processed. i) Predictions outside the continents where species are known to occur (according to our observation dataset) are removed, and ii) conditional probabilities associated with orchids are normalised, see Box B.

4.2.3 Validation

The species assemblage model is calibrated and assessed on the unseen occurrences from the validation spatial blocks (see dataset split in Box A). The objective is to guarantee that the true species is included within the kept species assemblage. This optimises recall rather than model precision. It results in species assemblages that are potentially larger than in reality, and consequently in aggregated indicators at species level that are potentially overestimated but precautionary (see next section).
Our dataset is highly unbalanced in terms of the number of occurrences per species (see Box F). It is therefore difficult to calibrate a specific threshold for many species. However, this would have been appropriate if we wanted to guarantee an error per species rather than per observation point. The aim is indeed to reduce the marginal error of classification per observation (i.e. we want assemblages with little error on the species observed). The optimal solution is given by a common threshold per species (Fontana et al., 2023).

The threshold value λ is then an important hyper-parameter of the method. Theoretically, we could consider that any species with a non-null conditional probability $\eta_k(x)$ is potentially present in the assemblage (i.e. by chosing $\lambda = 0$). However, in practice, the estimator $\hat{\eta}_k(x)$ is never null even for the most unlikely species. Thus, it is required to adjust the value of λ so that only the relevant species are returned in the assemblage. Therefore, we use a subset of the training dataset that is used only for this calibration step. It allows estimating the average error rate for a given value λ:

$$\mathcal{E}(S_\lambda) = \mathbb{P}_{X,Y}[Y \notin S_\lambda(X)]$$

by computing the percentage of samples x_i in the calibration set for which the true observed species y_i is not in $S_\lambda(x_i)$.

Finally, we can chose λ so as to minimize the average species assemblage size $\mathbb{E}[|S_\lambda(X)|]$ - which is equivalent to maximize λ - while guarantying that the average error rate is lower than an ϵ objective:

$$\begin{aligned} &\arg\min_{\lambda \in [0,1]} \mathbb{E}[|S_\lambda(X)|] \\ &\quad s.t.\ \mathcal{E}(S_\lambda) \leq \epsilon \end{aligned} \quad \Leftrightarrow \quad \begin{aligned} &\max_{\lambda \in [0,1]} (\lambda) \\ &\quad s.t.\ \mathcal{E}(S_\lambda) \leq \epsilon \end{aligned} \tag{4.3}$$

This is equivalent to what is called conformal prediction in machine learning (Fontana et al., 2023) and guarantees that the actual species is contained within the set with probability at least $1 - \epsilon$.

In practice, we choose $\epsilon = 0.03$ as explained in more details in the Box C. We predict assemblages that have been validated to contain the initial species in 97% (at the point level) and 80% (at the species level) of cases. The performance at the species level shows the robustness of our assemblages and the performance at the point level its validity in space.

4.2.4 Conservation indices for species assemblages

4.2.4.1 Indices definition

We define two indices characterising the extinction risk of a predicted species assemblage, $\mathcal{I}_c$ and $\mathcal{I}_\mathcal{O}$. They respectfully render the proportion of threatened species in the assemblage and the most critical IUCN status in the assemblage. Let's break down their construction.

IUCN status notations Our indices partly rely on the extinction risk classification scheme from the IUCN Red List of threatened species, https://www.iucnredlist.org/ (Mace et al., 2008). IUCN categories are limited to *Least Concerned* (LC), *Near Threatened* (NT), *Vulnerable* (VU), *Endangered* (EN) and *Critically Endangered* (CR). We set the ensemble $E_{\text{status}} = \{\text{LC}, \text{NT}, \text{VU}, \text{EN}, \text{CR}\}$ with the relation order LC<NT<VU<EN <CR. Additionally, we introduce a general THREAT category corresponding to the union of VU, EN and CR categories. We denote as $\varphi(y)$ the function that provides the extinction risk status of a species y.

Indicator $\mathcal{I}_{\mathcal{O}}(S)$: most critical status of the species in the assemblage For a given species assemblage S, our first indicator consists in taking on the most critical species extinction risk status. This is a concise and precautionary index. It aims at providing an information easy to understand and represent. Here is its formal definition:

$$\begin{array}{rccc} \mathcal{I}_{\mathcal{O}}: & \mathcal{P}(\mathcal{Y}) & \mapsto & E_{\text{status}} \\ & S & \to & \max\limits_{s_j \in S} \{\varphi(s_j)\} \end{array} \tag{4.4}$$

Indicator $\mathcal{I}_c(S)$: proportion of species in the assemblage with a given status Our second indicator $\mathcal{I}_c(S)$ measures the proportion of species from a given category c in an assemblage S. Let us consider a species assemblage with its associated probability distribution (S, η). $\mathcal{I}_c$ is defined as the proportion of species with status c in S, with the species being weighted by their relative probability of presence η (see Equation 4.5). The proportion of critically endangered species is for instance denoted $\mathcal{I}_{\text{CR}}(S)$. And so on for the four other IUCN status in E_{status} and the overall THREAT category.

$$\begin{array}{rccc} \mathcal{I}_c: & \mathcal{P}(\mathcal{Y}) \times \mathbb{R}^C & \mapsto & \mathbb{R}^{[0,1]} \\ & (S, \eta) & \to & \sum\limits_{j \in \varphi^{-1}(c)} \eta_j \end{array} \tag{4.5}$$

The Shannon index $\mathcal{I}_{\mathcal{H}}(S)$ The Shannon index is one of the most popular measures of biodiversity. It originates from the famous communication theory (Shannon, 1948), but was adopted in ecology as early as 1955 (Ricotta, 2005). Denoted $\mathcal{I}_{\mathcal{H}}$, this metric evaluates the quantity of information of a set. Both the set richness (number of distinct classes) and evenness (classes ratio) influence the index (Marcon, 2015). Let (S, η) be a species assemblage, with η its associated conditional probability distribution:

$$\mathcal{I}_{\mathcal{H}}(S) = -\sum_{l \in S} \eta_l \cdot \log(\eta_l) \tag{4.6}$$

4.2.4.2 Missing status completion

Only 889 of our 14,129 orchid species have an official IUCN status in 2021, i.e. 6.3%. It therefore seems unreasonable to ignore all unassessed species in our indicator calculation. We decide to supplement the status information with an automatic preliminary assessment method from the literature called IUCNN (Zizka et al., 2021). The distributions of the IUCN-assessed and predicted IUCN status are shown in Figure S6. Both indicators can then be computed considering only IUCN-assessed species or the entire species assemblage. By default, the indicators are on all the orchid species from our assemblage, i.e. considering both known IUCN status and predicted IUCN status. When they are restrained on the IUCN-assessed species only, the indicators are denoted with an *IUCN* superscript: $\mathcal{I}^{\text{IUCN}}$.

4.2.5 High-resolution maps construction

4.2.5.1 Global grid design

The aim now is to create a global grid to support our spatial indicators. This is done in two steps. First, we create a regular grid covering all longitudes and latitudes. We sample the longitude range [-180°,180°] and the latitude range [-90°,90°] at 30-second intervals. One second equals $1/3600$ degrees, hence $r = 30/3600$ degrees. Let $\mathcal{M} = \{-180, -180 + r, ..., 180 - r, 180\}$ and $\mathcal{N}$ be its latitudinal counterpart. The grid support is then obtained by crossing the two sampled axes $\mathcal{M} \times \mathcal{N}$. Secondly, we spatially intersect the grid with the land areas of the world. We are indeed only interested in terrestrial regions. The geometry used is the *Esri* grid of world country boundaries (Esri, 2023). The intersection contains 221M points. Finally, predictive features are assigned to each land grid position. This results in $\mathcal{G} = \{x_{m,n} \,|\, m, n \in \mathcal{M} \times \mathcal{N}\}$.

4.2.5.2 Maps definition and construction

Maps are constructed in two steps: First, the species assemblages associated to each $\mathcal{G}$ grid point are predicted by batch with our model: $\hat{\mathcal{S}}_\lambda(\mathcal{G})$. Second, the spatial indices defined in section 4.2.4.1 are computed on the predicted assemblages: $\mathcal{I}(\mathcal{G}) = \{\mathcal{I}(S) \,|\, S \in \hat{\mathcal{S}}_\lambda(\mathcal{G})\}$. This set of indicators $\{\mathcal{I}(\mathcal{G})\}$ constitute our global and kilometre-scale maps (*reminder:* by default all orchid species are considered and predicted IUCN status thus employed). Within worldwide predicted species assemblages:

- $\mathcal{I}_\mathcal{O}(\mathcal{G})$ highlights the most critical IUCN status
- $\mathcal{I}_c(\mathcal{G})$ represents the proportion of species with IUCN status c (five maps)
- $\mathcal{I}_{\text{THREAT}}(\mathcal{G})$ maps the proportion of threatened species
- $\mathcal{I}_\mathcal{H}(\mathcal{G})$ draws the global patterns of predicted orchid diversity.

Details on predictions batch processing and on the website solution are available in Box D.

4.2.6 Zonal statistics

Spatial analysis can necesit aggregated regional indicators. With a kilometre scale resolution, $\mathcal{I}_\mathcal{O}$ and $\mathcal{I}_c$ can be dissolved at different organization levels. Municipalities, protected areas, states or biodiversity units: the choice depends on the application. To illustrate this method at the global scale, we aggregate our indicators at the WGSRPD level 3. It corresponds to *botanical countries* which can ignore political borders (Brummitt et al., 2001). We selected countries of at least 2,000 km² to highlight large area priorities (65 countries out of 369 removed).

4.2.6.1 Region spatial coverage of the most critical IUCN status

This measure is based on $\mathcal{I}_\mathcal{O}$, the spatial indicator of the most critical IUCN status in the species assemblage. In a given region r, areas with distinct worst IUCN status

coexist. Focusing on a given status c, its spatial coverage proportion in r is denoted $\text{Area}_{\%}[\mathcal{I}_{\mathcal{O}}](r, c)$. By default, this variable is computed on the entire species assemblage. Nonetheless, it can also be expressed considering only IUCN-assessed species.

4.2.6.2 Region average proportions

Second zonal statistic consists in taking $\mathcal{I}_c$ average for a given region r and status c. It represents region's average proportion of species with c as IUCN status and is written down $\mu[\mathcal{I}_c](r, c)$. The entire species assemblage is taken into account. Such statistic allows direct comparison between arbitrary zones. For the sake of simplicity, square brackets precising the spatial indicator can be dropped in both zonal statistics.

4.2.7 Data

4.2.7.1 Orchid occurrences

The orchid occurrence dataset comes from Zizka et al. (2020), whose authors queried GBIF in August 2019. This dataset has the advantage of being both global and already geographically/taxonomically curated. Nearly 1 million occurrences of 14,129 different species were used to build our model (999,258 observations after duplicate checking). The average number of observations per species is 70, while the median is 4. 25% of species have more than 13 species. Date distribution summary statistics are $\min = 1901, \text{Q1} = 1982, \text{med} = 1997, \text{Q3} = 2010$ and $\max = 2019$. The cumulative number of occurrences per species, the distribution of observation dates, the distribution of georeferencing uncertainty, the observation map and the species richness maps are all available in Box F and Fig. S1.

4.2.7.2 Predictive features

A large environmental context around each observation is collected and provided to the model: 64 x 64 2D tensors sampled at the kilometre-scale resolution and centred on the observation. Predictors include WorldClim2 bioclimatic variables, Soilgrids pedological variables, human footprint rasters, terrestrial ecoregions of the world and the observation location, see Box G for details. Examples of input are shown in Figure S9 and the full list of predictors is given in Table S2.

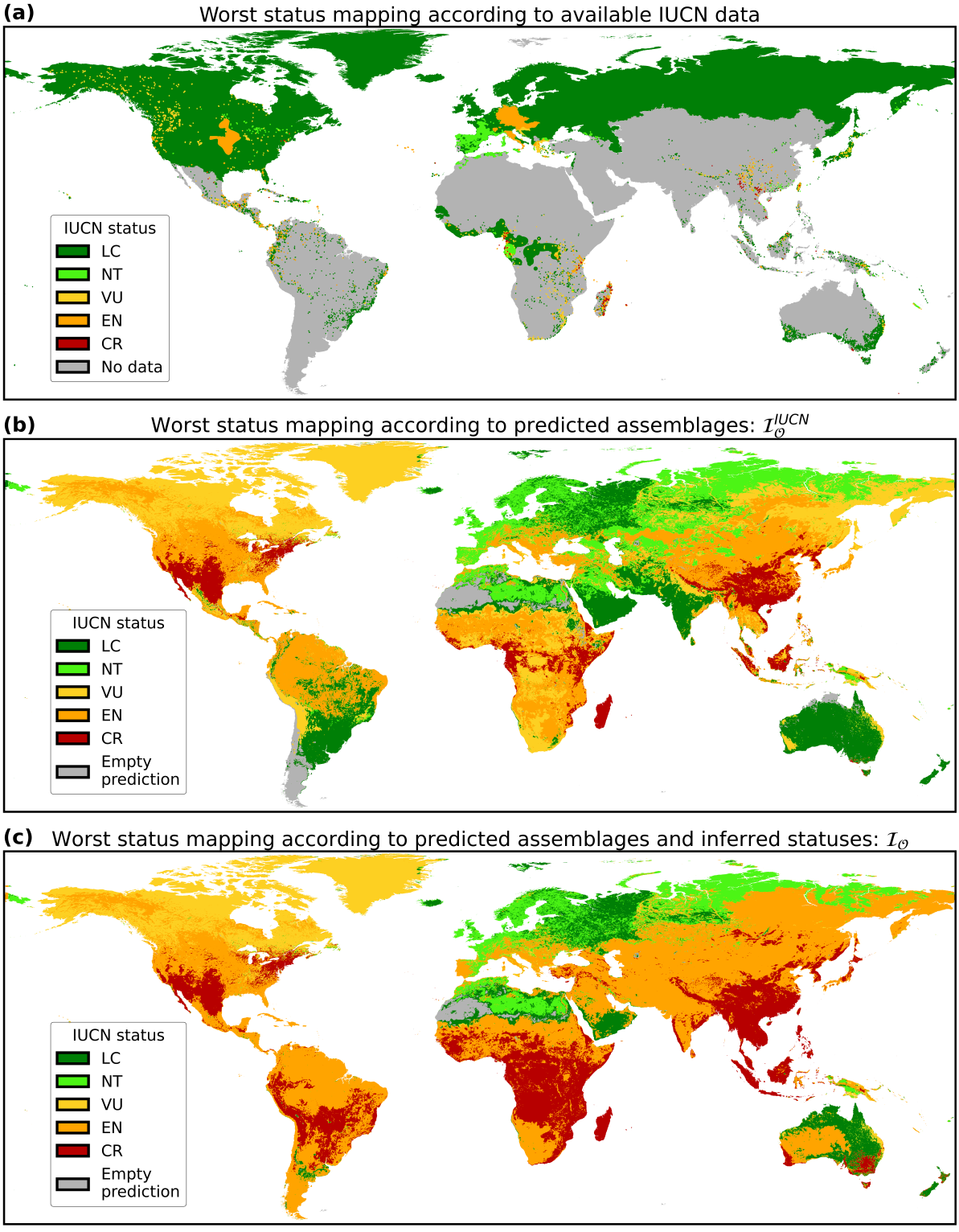


Figure 4.1: *Global comparison of the most critical IUCN status indicator according to three methods. (a) represents the IUCN information on our dataset: observations and available spatial data (polygons and points from https://www.iucnredlist.org/resources/spatial-data-download) taken together. Spatial data is available for only 167 IUCN-assessed orchids from our dataset, i.e. 1.2% of all species. (b) is the result of our species assemblage prediction model coloured by the most critical known IUCN status whereas (c) includes predicted IUCN status too in the indicator calculation. [Figure maps are under-sampled, see the website for full-resolution]*

4.3 Results

4.3.1 $\mathcal{I}_{\mathcal{O}}$ indicator: most critical status of the species in the assemblage

4.3.1.1 Global patterns

Considering the worst status of a species assemblage, Figure 4.1 compares (a) currently available IUCN information with (b,c) our model results $\mathcal{I}_{\mathcal{O}}^{\text{IUCN}}$ and $\mathcal{I}_{\mathcal{O}}$. IUCN species range data are still very scarce (only 1.2% of species in our dataset have IUCN ranges) and of variable quality: some species have raw model outputs as official IUCN range maps whereas others will have tailored expert-designed maps. Our species assemblage model combined with known IUCN status results in a consistent and contrasted map Fig. 4.1b.

Predictions in tropical Africa, East and South-East Asia and North America include CR species assessed by the IUCN. The presence of CR species in North America may be surprising at first, but given that i) this continent is comparatively well assessed and ii) this indicator is both sensitive and precautionary (only one species is sufficient to reach the CR category), it is reasonable. No CR species are predicted in South America if only known IUCN status are considered. However, when predicted IUCN status are included on Fig. 4.1c, the value of $\mathcal{I}_{\mathcal{O}}$ across South America is drastically different. Indeed, EN and CR species predictions lead the indicator to change to higher categories of risk. According to our model taking into account predicted IUCN status, Brazil and the Andes are for instance hosts to CR-estimated species on a large part of the territory. On Figure 4.1c, new global patterns are highlighted. These include India and temperate Asia presenting EN species, the Western Ghats and Southeast Asia hosting CR species, and Portugal, western Spain and the French Landes turning orange due to the prediction of EN species. Overall, the differences are more pronounced in the southern hemisphere than in the northern hemisphere. This illustrates the fact that IUCN assessments are biased towards northern countries and that large assessment gaps remain.

4.3.1.2 Country-level analysis

Table 4.1 shows the botanical countries with the largest $\mathcal{I}_{\mathcal{O}}$ coverage as CR or as EN. There are many islands in this ranking. All top fifteen countries are almost completely covered by only one status. See supplementary file T3 for the full table. High on the $\text{Area}_{\%}$(CR) ranking are Equatorial Guinea, Réunion, Mauritius, Madagascar, Comoros and Laos. CR species are present throughout these countries. By construction, countries with a high CR coverage status cannot also have a high EN coverage. Therefore, countries with high $\text{Area}_{\%}$(EN) are different from the first column. European territories such as Corse or Portugal appear in the ranking and Caribbean islands are well represented.

Table 4.1: *Top-15 countries with the largest share of their area covered by CR (left) or EN (right) as most critical IUCN status.*

	CR		EN	
	B. country	**Area$_{\%}$**	**B. country**	**Area$_{\%}$**
1	Eq. Guinea	100.00	Jamaica	100.00
2	Réunion	100.00	Dominican R.	100.00
3	Mauritius	100.00	Haiti	99.95
4	Madagascar	99.76	Cuba	99.86
5	Comoros	99.60	Afghanistan	99.74
6	Laos	99.38	French Guiana	99.65
7	Connecticut	98.71	Guyana	99.45
8	Vietnam	98.59	Surinam	99.29
9	Rhode I.	98.49	Costa Rica	99.15
10	Cambodia	98.26	Portugal	99.02
11	Jawa	97.93	Corse	98.98
12	Massachus.	97.25	Tadzhikistan	98.79
13	E Himalaya	97.07	Puerto Rico	98.71
14	Thailand	96.99	Windward Is.	98.64
15	Sumatra	96.93	Galápagos	98.50

B. country, botanical country (WGSRPD level 3).

4.3.2 $\mathcal{I}_c$ indicator: proportion of species in the assemblage with a given status

4.3.2.1 Global patterns

Figure 4.2a shows the Shannon index calculated on our species assemblage predictions (full resolution on the website). As expected, the tropics appear to contain the richest areas. This map can be read in parallel with the Box F second map: the species richness map of our occurrence dataset stratified by botanical country (WGSRPD level 3). The resolution gain is clear. Moreover, some biases in the initial observations set explain $\mathcal{I}_{\mathcal{H}}$ patterns. Colombia orchid richness, estimated for instance at 4,327 species according to *World Plants* (Hassler, 2023), is for instance under-represented within our occurrence set with only 1,375 species. Global orchid diversity patterns can also be appreciated in relation to the three following maps, which reflect the extinction risk of the predicted species assemblages.

High proportions of threatened species appear in East Africa, South and Southeast Asia on Figure 4.2b $\mathcal{I}_{\text{THREAT}}$. The Sahel also has a particularly high proportion of threatened species. Orchids in central North America also appear to have relatively high rates of threatened species, given the low observed and predicted diversity in this region. The threat levels in the Amazon Basin are high. However, compared to East Africa or tropical

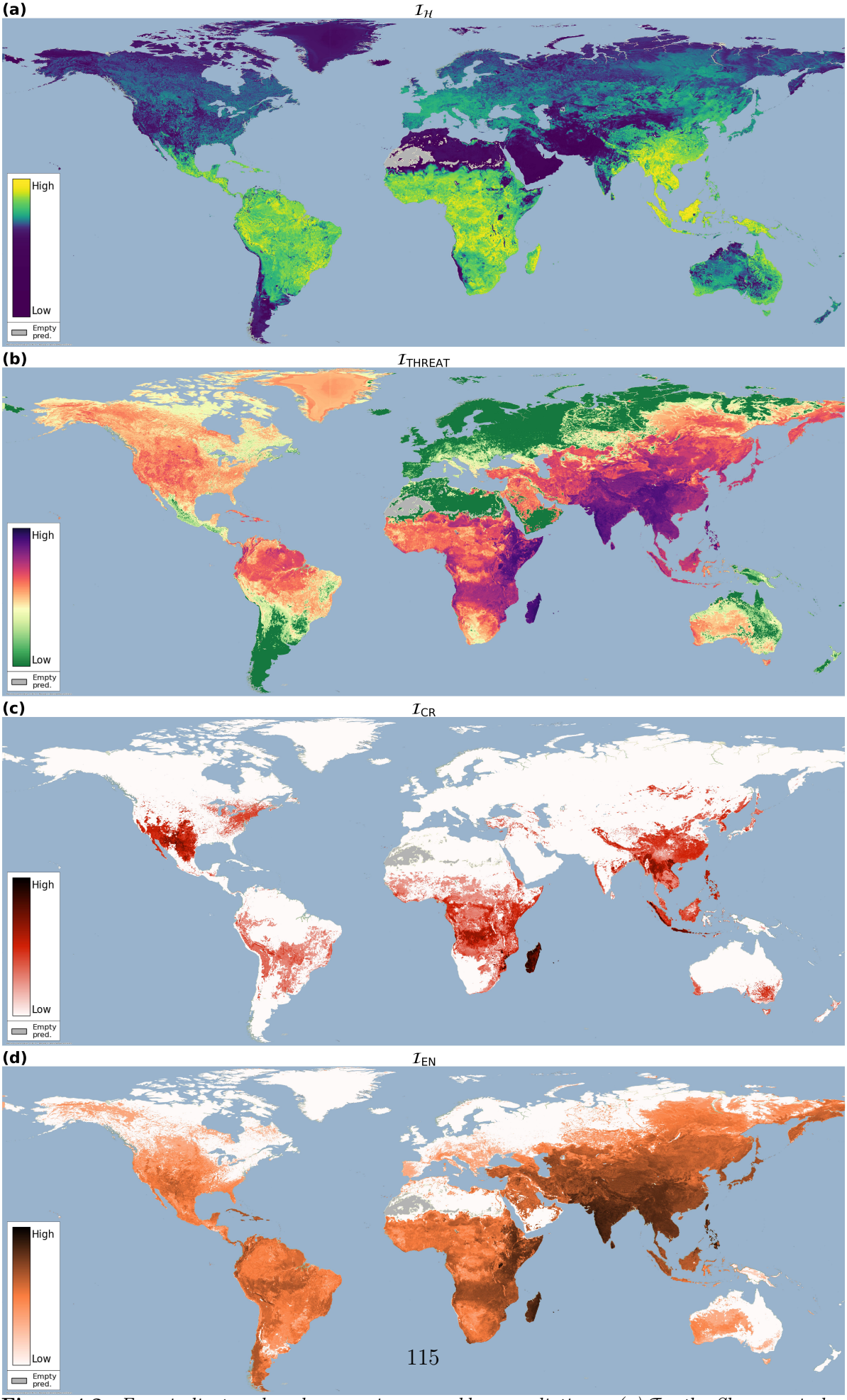


Figure 4.2: *Four indicators based on species assemblage predictions. (a) $\mathcal{I}_{\mathcal{H}}$ the Shannon index, (b) $\mathcal{I}_{THREAT}$ the weighted proportion of threatened species, (c) and (d) the weighted proportions of respectively CR species $\mathcal{I}_{CR}$ and EN species $\mathcal{I}_{EN}$. [see website for full-resolution]*

Asia, they are not as high as the region's impressive orchid richness would suggest. This result is quantified on the scatter plot Figure 4.3. High diversity does not necessarily imply high threat levels.

On Figure 4.1c map (proportion of CR species), the first striking element is certainly the strong emphasis on Madagascar. The patterns in the Himalayan belt, Indonesia and Southeast Asia are both more contrasted and appear more localised than on the $\mathcal{I}_{\text{THREAT}}$ (b) map. In northern Mexico and the southwestern United States of America, high levels of CR species are appealing and contrasting with the Shannon index. In South America, our model predicts relatively high levels of CR species along the Andes, in Bolivia, Paraguay and southern Brazil. If we compare $\mathcal{I}_{\text{CR}}$ with $\mathcal{I}_{\text{CR}}^{\text{IUCN}}$ (see website), we can see that the presence of CR species in South America is almost entirely due to predictions whose IUCN status has been automatically classified.

Finally, $\mathcal{I}_{\text{EN}}$ levels (Fig. 4.2d) are important throughout sub-Saharan Africa, Central and South America, South and Southeast Asia. The patterns observed here are closer to $\mathcal{I}_{\text{THREAT}}$ than $\mathcal{I}_{\text{CR}}$. With these maps we can better understand how the patterns of $\mathcal{I}_{CR}$, $\mathcal{I}_{EN}$ and $\mathcal{I}_{VU}$ indicators combine to produce the $\mathcal{I}_{\text{THREAT}}$ map.

4.3.2.2 Country-level analysis

Table 4.2: *Top-15 average status proportions per botanical country. From left to to right: threatened species all taken together (THREAT), Critically endangered species (CR) and Endangered species (EN). In average, 60% of the predicted species in Madagascar are threatened by extinction (63% in Réunion island).*

	THREAT		**CR**		**EN**	
	B. country	$\mu[\mathcal{I}_c]$	**B. country**	$\mu[\mathcal{I}_c]$	**B. country**	$\mu[\mathcal{I}_c]$
1	Réunion	0.63	Réunion	0.15	Réunion	0.44
2	Madagascar	0.60	Madagascar	0.12	Madagascar	0.39
3	Mauritius	0.55	Mauritius	0.10	Mauritius	0.38
4	Comoros	0.48	Comoros	0.10	India	0.36
5	Kenya	0.46	Jawa	0.07	Philippines	0.35
6	Myanmar	0.45	Sumatra	0.04	Taiwan	0.34
7	Nepal	0.45	Azores	0.03	Myanmar	0.33
8	E Himalaya	0.44	Philippines	0.03	Sri Lanka	0.33
9	Somalia	0.44	Vietnam	0.03	E Himalaya	0.33
10	India	0.44	Laos	0.03	Nepal	0.33
11	Laos	0.43	Arizona	0.03	Laos	0.32
12	Assam	0.43	New Mexico	0.03	Assam	0.32
13	China SC	0.42	Myanmar	0.03	Comoros	0.30
14	W Himalaya	0.40	Mozambique	0.03	Thailand	0.30
15	Taiwan	0.40	Lesser Sunda Is.	0.03	Cambodia	0.29

In Table 4.2, the top three botanical countries with the highest average proportion of threatened species, species classified as CR and species classified as EN are common: Réunion Island, Madagascar and Mauritius Island. Overall, 60% of the species predicted for Madagascar are threatened with extinction. All $\mu[\mathcal{I}_{\text{THREAT}}]$ top fifteen countries have an overall predicted proportion of threatened species greater than or equal to 40%. Again, the three columns are dominated by East African and tropical Asian countries. See supplementary file T3 for the full table.

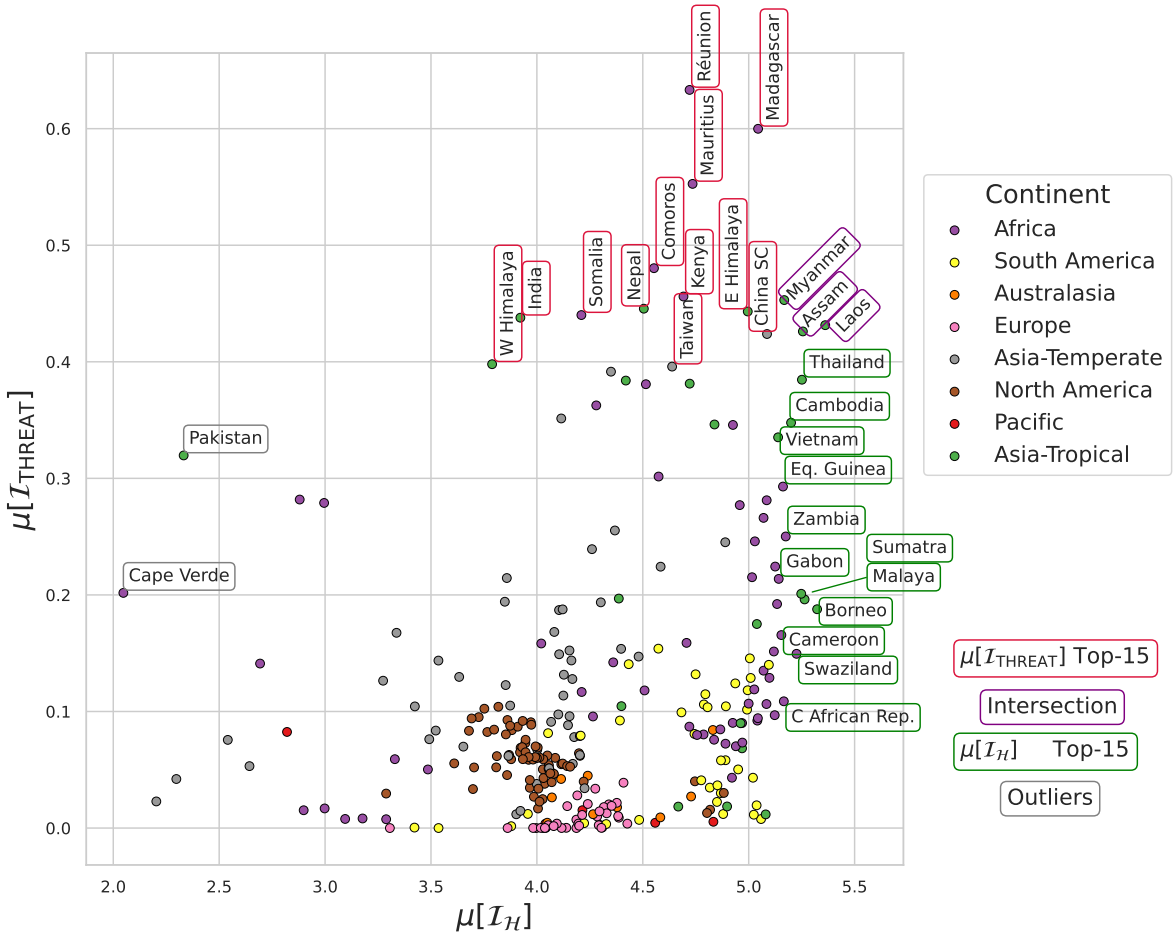


Figure 4.3: *Average proportion of species predicted as threatened by botanical country (WGSRPD level 3) versus average Shannon index. Countries are coloured in function of their continent (WGSRPD level 1) and top-15 countries of both variables are highlighted. Myanmar, Assam and Laos are the only three regions in the top-15 intersection whereas Pakistan and Cape Verde show especially high threatened species proportions with low diversity indices.*

The scatterplot Figure 4.3 tests the relation between the average rate of threatened species and the Shannon index at the level of botanical countries. The Spearman ρ value is 0.29 ($p = 2.5\text{e}{-7}$), indicating a positive but relatively law global correlation. The colour code, indexed by continent, reveals different patterns per continent. North American (brown) and European (pink) countries are clearly clustered on the graph, with a medium diversity index and low threat levels on average. The top fifteen $\mu[\mathcal{I}_{\text{THREAT}}]$ countries (Table 4.2 first column) are this time marked with red borders. The top fifteen

$\mu[\mathcal{I}_{\mathcal{H}}]$ are framed in green and the intersection includes Myanmar, Assam and Laos. African (purple), Asian temperate (grey) and Asian tropical (green) countries present more variation in this graph and represent the extremes. The South American countries (yellow) at the bottom right of the graph confirm the observation made with Figure 4.2: this continent is highly diverse with relatively low levels of threat to its species assemblages. A Venn diagram crossing $\mu[\mathcal{I}_{\mathcal{H}}]$ and $\mu[\mathcal{I}_{\text{THREAT}}]$ top-30 countries plus the Spearman correlations per continent are available at Figure S7.

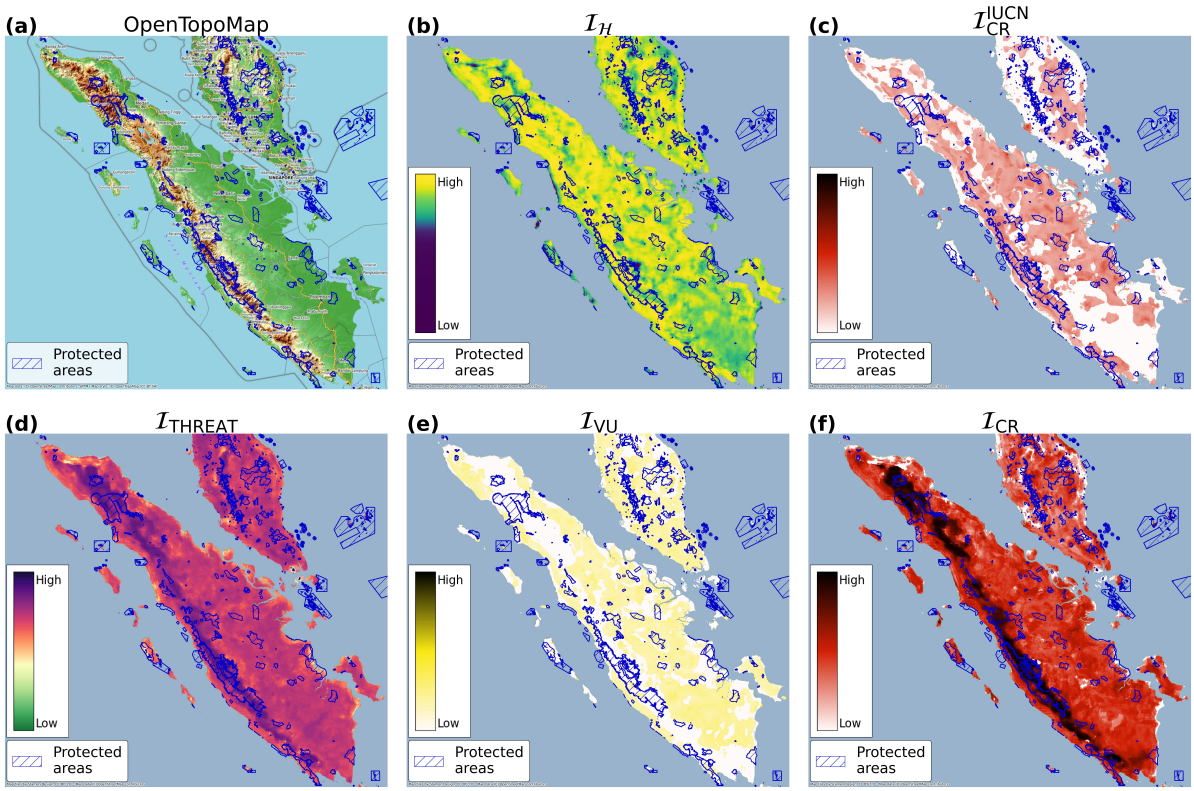


Figure 4.4: *Five indicators of species assemblage extinction risk applied on Sumatra island. Elevation is also provided and protected areas are hashed in blue (downloaded from https://www.protectedplanet.net/). (a) elevation map, (b) Shannon index, (c) proportion of IUCN-assessed CR species in the predicted species assemblages. On the second line, species proportion of: (d) threatened species, (e) VU species only, and (f) CR species only (all statuses combined). [Maps in figures are under-sampled, see the website for full-resolution]*

Sumatra case study

On the western side of Sumatra, the Barisan Mountains form a sharp relief (see Figure 4.4a). The *elevational diversity gradient* theory would suggest that species richness is particularly high along the mountainous area. However, according to the $\mathcal{I}_{\mathcal{H}}$ indicator on (b), the predicted orchid diversity appears to be fairly constant across the island. Considering only the known IUCN assessments, the presence of CR species (c) is not clearly correlated with the mountain range. In addition, there are areas where no CR species are predicted, for example in the northern and southern regions of the island. When the predicted IUCN status are included in the indicator calculation with $\mathcal{I}_{\text{CR}}$ on (f) map, high proportions of CR species are predicted across the island. There is a sharp

pattern following the Barisan Mountains. By construction, a similar trend is drawn on the (d) map representing $\mathcal{I}_{\text{THREAT}}$. Such a difference between $\mathcal{I}_{\text{CR}}$ and $\mathcal{I}_{\text{CR}}^{\text{IUCN}}$ at the regional scale confirms the need to include automatic IUCN assessments when designing extinction risk indicators. Finally, $\mathcal{I}_{\text{VU}}$ on Fig. 4.2e map indicates the likely presence of VU species inhabiting the lower elevations of the islands.

Protected areas cover 12.7% of the island of Sumatra. Three national parks on the spine of the Barisan Mountains were inscribed on UNESCO's World Heritage List in 2004, forming the Tropical Rainforest Heritage of Sumatra. They are the three largest protected areas on the island. From north to south: Gunung Leuser National Park, Kerinci Seblat National Park and Bukit Barisan Selatan National Park. Since 2011, these parks have been placed on a Danger List to help combat numerous threats, including poaching, illegal logging and agricultural encroachment.

Let's look at the zonal statistics for PAs. We calculate the ratio of two indicators, both averaged across PAs: i) the proportion of *all* CR species (known IUCN status + predicted status combined) and ii) the proportion of *IUCN-assessed* CR species: $\frac{\mu[\mathcal{I}_{\text{CR}}]}{\mu[\mathcal{I}_{\text{CR}}^{\text{IUCN}}]}(\text{PAs}) = 3.1$. This ratio is even greater when all threatened species are considered together: $\frac{\mu[\mathcal{I}_{\text{THREAT}}]}{\mu[\mathcal{I}_{\text{THREAT}}^{\text{IUCN}}]}(\text{PAs}) = 7.1$ The level of threat in Sumatra's PAs is then significantly higher than the IUCN information alone would suggest. Now let's compare the average CR proportion inside *versus* outside PAs: $\mu[\mathcal{I}_{\text{CR}}](\text{PAs}) = 0.108$ and $\mu[\mathcal{I}_{\text{CR}}](\overline{\text{PAs}}) = 0.036$. Thus the average proportion of CR species is 3 times higher in PAs than outside PAs. The current design of PAs therefore seems to well match habitats hosting particularly threatened orchids. However, looking closely at the map reveals that many areas with a specially high proportion of CR species are still outside PAs, so that the ratio could be consistently improved. With IUCN-assessed species only, the average proportion of CR species in PAs is 3.4%. It is similar to the proportion of CR species *outside* PAs with the completed Red List. Again, enriching the current IUCN information within our method changes the narrative on PA efficiency.

4.4 Discussion

4.4.1 Modelling choices

Our species assemblage predictor has theoretical guarantees that we have validated on a previously unseen observation set (see Box C). However, some bias in the input data could prejudice its predictions. Unlike some methods, it has the advantage of not being biased by the heterogeneous sampling effort. Indeed, it depends only on the conditional probability $\mathbb{P}_{X,Y}(Y = k|X = x)$ and not on the marginal distribution $\mathbb{P}_X$. Nonetheless, it is impacted by species detection bias, i.e. by the fact that some species might be observed more than others conditionally to a given x. Largely under-observed species, in particular, may be excluded from the predicted assemblage. Conversely, some over-observed species could be predicted at locations where they are not present. In future work, it would be interesting to study the impact of this type of bias on the assemblage-level indicators introduced in this paper.

When calibrating the species assemblage model, the choice of average error rate translates into a trade-off between model generalisation and over-prediction. Indeed, imposing a lower error rate results in a lower probability threshold and larger predicted assemblages. But are the newly retained species likely to be present, or are they unreasonably predicted? This is a difficult question of model calibration, which we believe deserves more attention in future studies.

Over the validation set, an error is defined as the absence of the true label within the returned species assemblage. What is effectively measured is the recall of the model, i.e. the proportion of relevant species that are successfully retained. In contrast, precision, i.e. the proportion of retained species that are relevant to the test point, is not directly measured (and could not be without absence data). However, precision is a positive function of the conditional probability threshold. Therefore, by maximising the threshold for a given target error rate, we also maximise precision. Finally, a recent study confirms that as long as models are flexible enough and well fine-tuned to avoid overfitting, they make coherent predictions on spatially separated test data (Valavi et al., 2023).

4.4.2 Considerations on covariates

Nature's myriad of elements are interfaced to produce heterogeneous patterns of diversity, unpredictable at a given point, but statistically structured. Measuring some of these factors and feeding them into our model will hopefully allow us to capture biodiversity shapes. However, it is essential to remember that no single mechanism fully explains a given pattern, that inter-scale dependencies and local historical events strongly influence biodiversity, and that no pattern is exempt from variation and exceptions (Gaston, 2000). Other ecological variables contain valuable information influencing the distribution of orchids. They have not been included because of the currently limited spatial and taxonomic coverage or for practical reasons. Remote sensing is a natural perspective for improvement (Gillespie et al., 2022; He et al., 2015). The inclusion of biological and functional traits of orchids is another exciting perspective (Bourhis et al., 2023; Puglielli & Pärtel, 2023; Weigelt et al., 2020), as well as mycorrhizal fungi or pollinator distribution (McCormick et al., 2018).

We believe that predictors of large spatial patterns may play a significant role in the regional diversity of orchids, and that the computer vision model can learn such information. The model's strength is to rely on the best possible input set and exploit complex interactions in order to be as predictive as possible. The trade-off is interpretability, but the AI community is investing heavily in this area and our understanding is getting finer (Linardatos et al., 2021). For example, deep-SDMs have been shown to construct a feature space with structured functional traits and bioclimatic preferences, even though only remote sensing data were provided (Deneu et al., 2022).

4.4.3 Our indicators originality

One of the main strengths and originality of our indicators is their scalability. An analysis can start at the country level with zonal statistics before delving deep into regional patterns thanks to the interactive maps online, see the Figure 4.5 for an overview of

available tools. For example, India ranks fourth in terms of its average proportion of CR species (Table 4.2 last column). Looking at the $\mathcal{I}_{\text{CR}}$ indicator, the Western Ghats and eastern India appear to be the main hosts of CR species. Finally, the interactive map allows you to zoom in on patterns, explore and look for terrain correspondence with the base maps. The case study of Sumatra also shows that mountainous regions can host particularly high proportions of CR species. One of the main shortcomings of our

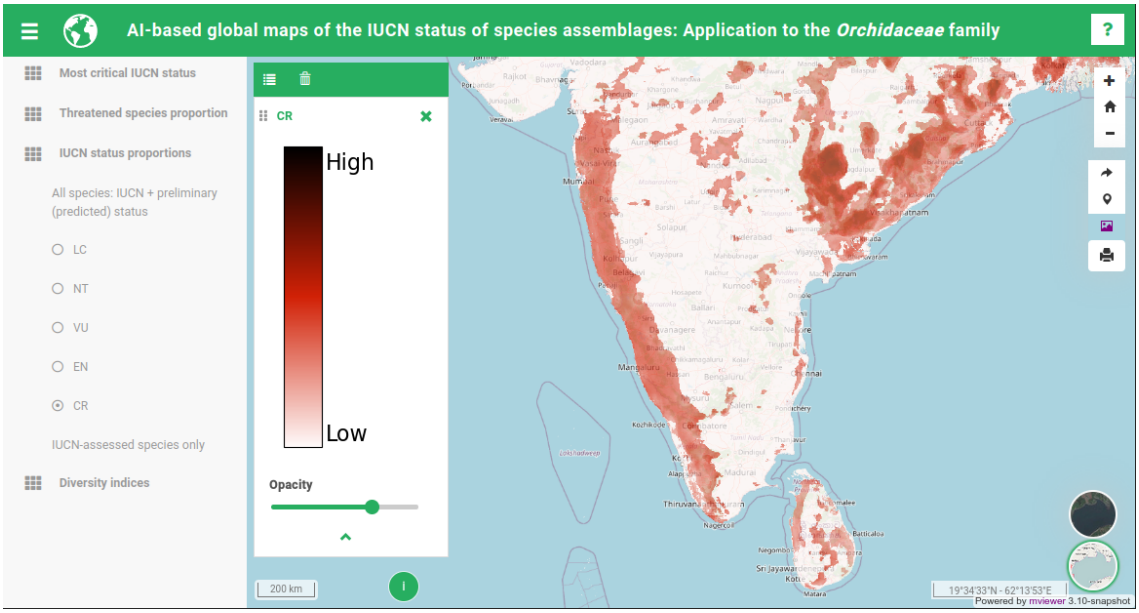


Figure 4.5: *Screenshot example of the website tools. To start, use the left panel to select the desired indicator. You can then zoom in and out directly on the regions of interest. To explore relationships between terrain and indicators, you can 1) adjust the layer opacity to see the base maps, 2) change the base map (bottom right) between OpenStreetMap or Esri imagery. The imagery is particularly useful for zooming in on local patterns. Finally, you can export the map as a .png using the commands at the top right.*

indicators is their lack of transparency. A first direct perspective for improvement is to return, for a given point, the names and IUCN status of the species assemblage. However, this is a technical challenge given the global support size of 221M points. Another drawback is the interpretability of deep-SDMs. Feature importance experiments would provide a sense of which features the model relies on most. Again, this is a very active area of research and future work will complement this point (Ryo et al., 2021).

Orchids have specific characteristics that make them valuable indicators of ecosystem health (Newman, 2009). They are sensitive to climate change and environmental disturbances (Kull & Hutchings, 2006), and their interactions with pollinators and mycorrhizal associations contribute to ecosystem functioning (Swarts & Dixon, 2009). In addition, orchids are easy to monitor in the sense that once a population has been established, it is easy to find it every year. Therefore, as defined by (Jørgensen et al., 2016), orchids can be considered as suitable ecological indicators of ecosystem health. The family is i) easy to monitor, ii) sensitive to small-scale environmental changes, whose response can be quantified and predicted, and iii) globally dispersed. They also are umbrella species and their local disappearance may be an early warning of environmental disturbance

(Gale et al., 2018). However, they don't encompass all aspects of ecosystem biodiversity. While orchids can be used as surrogate species for biodiversity planning, they can't fully represent overall ecosystem health. Taking these elements into account, orchid-based indicators such as $\mathcal{I}_{\mathcal{O}}$ and $\mathcal{I}_c$ can be considered to have a wider scope than just qualifying their family, but also a degree of habitat quality. Nonetheless, we do not pretend to be able to fully capture ecosystem health through a single family of indicators. In practice, achieving this goal would require a large number of indicators and measurements. Comparisons with established indicators are provided in Box E.

4.4.4 Orchids conservation

Spatial indicators can be used to identify priority areas and support the design of PAs (Almpanidou et al., 2021). An intuitive method is to select the k-highest percentiles of the indicator as hotspots. In Sumatra, the creation of corridors extending PAs along the Barisan Mountains seems a natural improvement to conserve CR species. While this approach is easy to understand, there is a risk that some aspects of biodiversity will be missed by the indicator and left unprotected (Orme et al., 2005). It is fair to ask: if the current PAs preserve key aspects of biodiversity and are representative of the other areas identified as most at risk, where is the next priority? The combination of complementary indicators is the key to designing effective PAs with a limited budget (Silvestro et al., 2022).

Manual extinction risk assessments should be carried out extensively in the tropics and on islands. Indeed, it is well known that the tropics are poorly assessed, although they host most of the world's biodiversity (Collen et al., 2008). The orchid family follows the same trend. Automated assessment methods will continue to improve, hand in hand with the quality of IUCN assessments in terms of taxonomic coverage, geographical extent and consistency. Finally, special attention must be paid to the assessment and protection of islands: all our indicators point to them as hosts of particularly threatened species assemblages.

4.4.5 Conclusions

Based on deep-SDMs architectures, we have developed global indicators that qualify the extinction risk of species assemblages at an unprecedented kilometre resolution. This allows multiscale analysis from global patterns down to country statistics or landscape discrepancies. The indicators are available as interactive maps online. Although our results show how our novel indicators can be successfully employed, working closely with decision-makers would ultimately allow for more effective guidance of conservation actions (Guisan et al., 2013). To enable efficient technology transfer, interdisciplinary studies between computer science and conservation science need dialogue with conservation practitioners (Gale et al., 2018).

ACKNOWLEDGEMENTS

The research described in this paper was funded by the European Commission via the GUARDEN and MAMBO projects, which have received funding from the European Union's Horizon Europe research and innovation programme under grant agreements 101060693 and 101060639. The opinions expressed in this work are those of the authors and are not necessarily those of the GUARDEN or MAMBO partners or the European Commission. The INRIA exploratory action CACTUS fund also supported this work. This work was granted access to the HPC resources of IDRIS under the allocation 20XX-AD011013648 made by GENCI. Finally, we warmly thank Alexander Zizka for providing us with the filtered set of orchid occurrences.

DATA AVAILABILITY STATEMENT

The data and code that support the findings of this study are openly available in figshare at http://doi.org/10.6084/m9.figshare.22803431 [reserved DOI, active if article published]. Private link in the meantime: https://figshare.com/s/15404886eb3b62363a5f

CHAPTER 5

EXPLOITING DEEP-SDMS TO PREDICT PLANT EXTINCTION RISK AND TEST CLIMATE CHANGE INFLUENCE

Table of contents

5.1 Introduction . . . 126
5.2 Materials and methods . . . 129
5.2.1 Method motivation . . . 129
5.2.2 Data on the *Orchidaceae* family . . . 130
5.2.2.1 Species observations . . . 130
5.2.2.2 IUCN Red List status . . . 130
5.2.2.3 Predictive features . . . 130
5.2.3 Method definition . . . 131
5.2.3.1 Deep species distribution modelling . . . 131
5.2.3.2 Dispersal scenarios . . . 132
5.2.3.3 Species niche features for IUCN classifcation . . . 132
5.2.3.4 Classifying extinction risk status from species niche features . . . 134
5.2.3.5 Projections within the CMIP6 SSP5-8.5 scenario . . . 135
5.2.4 Model validation . . . 136
5.3 Results . . . 136
5.3.1 Continents . . . 137
5.3.2 Latitude . . . 138
5.3.3 Altitude . . . 138
5.4 Discussion . . . 139
5.4.1 Interpretations . . . 139
5.4.2 Limitations . . . 140
5.4.3 Perspectives . . . 141
5.4.4 Conclusion . . . 142

This chapter is under review for the *Predictive Biogeography* special issue of *Ecography*.

Abstract

The post-2020 global biodiversity framework needs ambitious, research-informed targets. Estimating the accelerated extinction risk due to climate change is critical. Species distribution modelling based on deep learning (deep-SDMs) offers exciting opportunities to harness species-level information from rich biodiversity data. Field observations, which are massively collected on online platforms, are highly biased. However, by applying compensatory techniques and formulating adapted hypotheses, deep learning models have become robust enough to learn species features representative of their environmental niche. Our curated observation dataset comprises 1M occurrences of 14K orchid species distributed worldwide. Predictive features include bioclimatic and pedological variables, human footprint rasters, ecoregions and location. We evaluate a novel method for classifying the International Union for Conservation of Nature (IUCN) extinction risk status of species benefiting from the generalisation power of deep-SDMs. Cross-validation shows that our method matches state-of-the-art classification performance while relying on flexible SDM-based features that capture species environmental preferences. 889 orchids were assessed by the IUCN. Tenfold cross-validation yields average accuracies of 0.61 for status classification and 0.78 for binary classification (threatened or not). Climate change will reshape future species distributions. Under the species-environment equilibrium hypothesis, SDM projections approximate plausible future outcomes. Two extremes of species dispersal capacity are considered: unlimited or null, with the most likely species trajectories assumed to lie in between. The projected species distributions are translated in the features feeding our IUCN classification method. Finally, trends of threatened species are analysed in time and i) by continent and as a function of average ii) latitude or iii) altitude. The proportion of threatened species is increasing globally, with critical rates in Africa, Asia and South America. Furthermore, the proportion of threatened species is predicted peak around both Tropics, at the Equator, in the lowlands and in the 800-1,500 m altitudinal range. Taking only climate change into account, we predict that the total number of threatened orchid species will increase by a third by 2100, with a marked acceleration in the second half of the century, important geographical disparities and an irregular increase along elevation.

Keywords

deep learning, species distribution modelling, IUCN status, extinction risk, orchids

5.1 Introduction

Failure to meet any of the 2020 Aichi Biodiversity Targets is a clear signal that transformative change is urgently needed (IPBES, 2019). The post-2020 global biodiversity framework must set ambitious targets with quantified, measurable objectives underpinned by research (Mace et al., 2018). Biodiversity targets based on species extinction rates and measures of ecosystem services, for example, should be included (Reyers et al., 2013; Rounsevell et al., 2020). Indeed, this could help to galvanise policy in a similar way to the 2°C maximum climate change target. Finally, such goals are inherently interlinked

and should be set together to allow parallel progress and avoid contradictions (Díaz et al., 2020). Quantifying the acceleration of extinction risk due to climate change appears to be a top priority in this context (Mace et al., 2018).

Extinction risk and climate change. The increase in extinction risk due to climate change is an active area of research (Carpenter et al., 2008; Maclean & Wilson, 2011; Malcolm et al., 2006; Thomas et al., 2004), see the state of the art section 2.4.3. A common practice is to use Species Distribution Models (SDMs) to first learn species' environmental preferences and then project the learned relationship into future climates (Guisan & Thuiller, 2005). Species are then predicted to become extinct if their potential habitat is reduced below minimal thresholds. Habitat loss rates can also be compared to the official extinction risk criteria of the IUCN (Mace et al., 2008; Moat et al., 2019). While this approach is largely dominant, other ways of estimating extinction risk include using process-based models of physiology or demography, or relying on species-area relationships and expert opinion, as presented in this meta-study (Urban, 2015). Their literature review concludes that under Representative Concentration Pathway 8.5 (RCP 8.5), one in six species will be at risk of extinction due to climate change.

The IUCN Red list of Threatened Species. As a reminder, the IUCN Red List of Threatened Species (RL) is the reference classification scheme for the extinction risk of species (see Section 2.2.1.1). The risk assessment is uniform for all living organisms (except micro-organisms, Bland et al., 2017). It is a strength in terms of coherence and visibility. The Red List is the basis for biodiversity indicators used in international agreements. It allows monitoring of Parties' commitment to conservation. Starting with *least concern* (LC) and *near threatened* (NT), the threat categories are ordered by increasing risk of extinction. Threatened species are either *vulnerable* (VU), *endangered* (EN) or *critically endangered* (CR) before possibly becoming *extinct* (EXT).

IUCN status classification. Manual assessment of extinction risk cannot keep up with the current levels of threats, so species are disappearing before they have been assessed or even discovered (Pimm & Joppa, 2015). As of 2022, only 15% of the world's known plant species have an IUCN status[1]. Research has responded to this massive concern by developing compensatory automated assessment methods (Stévart et al., 2019). Such methods are designed to provide preliminary extinction risk categories and focus manual effort where it is urgently needed (Bachman et al., 2020). Methods can either estimate IUCN variables to be compared with the official criteria thresholds (index-based methods, see Section 2.3.1.5 for a review) or directly learn a correspondence between species features and IUCN status (*prediction-based* methods, see Section 2.4.2), Zizka et al. (2020). Classifiers such as the Random Forest (RF) algorithm and ensemble methods show high performance in differentiating threatened species within this active research area (Borgelt et al., 2022a; Pelletier et al., 2018). Species distribution modelling can inform IUCN assessments by estimating IUCN range variables (as recommended in the Guidelines, *index-based* methods). Alternatively, SDMs and habitat modelling can be used to test new predictive methods of IUCN categories (Breiner et al., 2017; Brooks

[1]https://www.iucnredlist.org/about/barometer-of-life

et al., 2019). The latter authors concluded that SDM-based niche size estimates provide valuable complementary information to range size in assessing species extinction risk, but are not a good proxy for extent of occurrence (EOO) or area of occupancy (AOO). Other approaches therefore need to be explored.

Contributions. Our main contribution is a novel method for extracting species traits predictive of IUCN extinction risk status, which allows modelling and testing the impact of bioclimatic projections. Based on a deep species distribution model, it achieves state-of-the-art classification performance while allowing to explore climate change scenarios and test how status distributions might evolve. Thanks to this approach, we proceed by analysing how threatened species would be distributed across continents, latitudes and altitudes under RCP 8.5 and two extreme dispersal scenarios. While the number of threatened species is projected to increase globally, some areas will be particularly affected: Africa, Asia, South America, latitudes around both Tropics and the Equator, and finally the lowlands and intermediate altitudes between 800 and 1500m.

Background information

CMIP6 SSP5-8.5 scenario and climate models. Scenarios that project socio-economic behaviours called Shared Socioeconomic Pathways (SSP) and corresponding emissions rates Representative Concentration Pathways (RCP) into the future are necessary to model future climate alternatives. Developed in 2011, RCPs are a set of four main scenarios that are intended to span the range of plausible outcomes until 2100 (Van Vuuren et al., 2011). RCP 8.5 is the scenario with the highest assumed fossil fuel use, but also the best match to our current emissions levels and stated policies (Schwalm et al., 2020). A wide range of climate models coexist within the Coupled Model Intercomparison Project Phase 6 (CMIP6). Fortunately, tools are available to assist modellers in selecting the most appropriate model for their application in terms of region and season (Parding et al., 2020). Scenario planning provides a framework for developing more resilient biodiversity models, and hence conservation policies, in the face of inherent future uncertainty (Pereira et al., 2010; Peterson et al., 2003).

Species distribution models. SDMs are statistical tools interlinking terrain observations with predictive variables (Elith & Leathwick, 2009). Observations can include opportunistic and/or field surveys presence and absence data, abundances, herbarium collections. Predictive features can be considered on a panel from those directly involved in species presence to those thought to have a loose influence on species preferences, or in other words from proximal to distal predictors (Austin, 2002). They are classified according to their function: *limiting factors* (e.g. elevation), *disturbances* (e.g. competition) and *resources* (e.g. solar radiation) (Guisan & Thuiller, 2005).

Modelling species distributions involves strong assumptions and limitations. First, species are assumed to be in equilibrium with their environment when learning their potential niche and to remain so in the future (Araújo & Pearson, 2005). However, training data can be misleading if species have been observed outside their natural niche, e.g. due to extreme climatic events or mismatches in the predictive data alignment. In

addition, species' resilience to climate change, genetic adaptation, competition or pollinator dependence, among other factors, could cause species to thrive under unexpected conditions or, conversely, to disappear from their current environmental niche. Second, data on dispersal capacity and proximal features are often lacking. Other considerations may militate against the use of SDMs to estimate extinction risk, such as approximations to IUCN criteria, inappropriate spatial modelling scale or variable selection (Akçakaya et al., 2006; Fourcade et al., 2018). Nevertheless, it should not be forgotten that i) the design of SDM is rapidly evolving to mitigate acknowledged biases and limitations (Rocchini et al., 2023; Valavi et al., 2022), ii) the generalisation of biodiversity knowledge through species distribution modelling is already at the origin of notable successes (Elith et al., 2011; Guisan et al., 2013), and iii) the urgency of raising general awareness and convincing decision-makers to embrace transformative change let us short of other options than working diligently with imperfect but demonstrative tools inherited from decades of research (IPBES, 2019).

Deep-SDM introduction. Building on its many successes in computer vision over the last decade, deep learning is now being applied to ecology to tackle complex tasks (Lamba et al., 2019). The explosion of biodiversity data, resulting from both new techniques (citizen data science, remote sensing) and efforts to pool freely available data such as the GBIF or the World Checklist of Vascular Plants (WCSP), also requires adapted frameworks that the deep learning (DL) community can provide (Borowiec et al., 2022). SDM based on DL techniques, in particular Convolutional Neural Networks (CNNs) for spatial patterns, allow to learn complex relationships between species and their environment (Botella et al., 2018a; Deneu et al., 2021b). Providing deep-SDMs with access to both spatial and temporal contexts of species observations proved to be particularly valuable for shaping rare species distributions (Deneu et al., 2021b; Estopinan et al., 2022). Finally, research to extend the predictive power of DL and to overcome challenging problems such as class imbalance, model interpretability, multimodality fusion or label noise directly benefits applications in ecology (Benedetti et al., 2018; Cao et al., 2019; Dosovitskiy et al., 2020b; Rolnick et al., 2017; Ryo et al., 2021).

5.2 Materials and methods

5.2.1 Method motivation

Current extinction risk assessments of vascular plants mostly rely on their geographic distribution with IUCN's B criterion and the EOO/AOO measures. As a result, the geolocation of observations takes up a central position in species subsequent IUCN category. This affects automated methods that logically obtain the geographic information to be the best predictor of species extinction risk. However, over-relying on species observation locations prevents exploring future scenarios to project species distribution shifts and forecast likely trends in species extinction risks. Using an alternative IUCN criterion is often impossible because of data scarcity. For instance, the application of criterion A measuring reductions of population size over time is hampered by the lack of repeated assessments and knowledge of species generation length. A sensitive response is

then to model future species distribution, retain most likely species presence according to a validated selection strategy and apply EOO/AOO thresholds. Yet, changes in SDM-predicted range size were found to be a poor surrogate for species EOO and AOO (Breiner et al., 2017). Further research is therefore needed to determine if and how spatial predictions of SDMs can be used to adequately complement current IUCN assessments.

In response of this limitations, our goal is to extract species classification features with high generalisation capabilities in space and time. To this end, we use a deep-SDM to reduce the dimension and capture critical environmental information that correlates with species observations. The successive convolutional filters and pooling layers greatly reduced the input dimension, with the aim of capturing and preserving characteristic patterns. The resulting N-dimensional space is hereafter called the *feature space*. A fundamental assumption in our method is that these features are not only informative about the species likely to be present conditionally on an observation, but also informative about the species' environmental and spatial niche.

5.2.2 Data on the *Orchidaceae* family

5.2.2.1 Species observations

Orchid observations have been filtered from GBIF in (Zizka et al., 2020) thanks to the R package *CoordinateCleaner*. (Zizka et al., 2019). The global dataset contains 999,258 occurrences of 14,129 species. It is highly unbalanced: a few common species have thousands of observations, while most orchids are represented by a few opportunistic samples. The average number of occurrences per species is 70, while the median is 4. Only a quarter of the species have more than 13 occurrences. The oldest occurrences date back to 1901, but three quarters were observed after 1982 (and half after 1997). Metadata on the orchid observation dataset are provided in the Box F.

5.2.2.2 IUCN Red List status

In July 2023, there are 1,970 orchid species assessed on the Red List, or 6.3% of the estimated 31,000 species of this uniquely diverse family (KEW, 2023). When we checked our orchid observations against the Red List as of December 2021, we found 889 species that had already been assessed. Figure 5.2A shows the distribution of status. These species will be our reference for training our IUCN extinction risk classifier and evaluating its performance. While the binary status distribution (threatened or not) is balanced, LC and EN species are largely dominant at the status level. Together, they represent more than two-thirds of the species assessed by IUCN.

5.2.2.3 Predictive features

Our predictive features include only global rasters at kilometre resolution. This makes the method easily transferable to other taxa. Large spatial contexts (64 x 64 km tensors) centred on each observation are provided to the model. The available predictors are

bioclimatic variables from WorldClim2 variables, Soilgrids pedological variables, human footprint rasters, terrestrial ecoregions of the world and the observation location (see Box G for details, input examples Fig. S9 and full list of predictors Tab. S2). No variable selection step was performed as deep convolutional neural networks do not overfit their predictive features under the right settings (Poggio et al., 2017). Within the CMIP6 SSP5-8.5 future scenario, the 19 WorldClim2 bioclimatic variables (averages from 1970-2000) are replaced by the corresponding projections from the *EC-Earth3-Veg* climate model (Döscher et al., 2022). Four time periods are considered: 2021-2040, 2041-2060, 2061-2080 and 2081-2100. More details can be found in the subsection 5.2.3.5.

5.2.3 Method definition

5.2.3.1 Deep species distribution modelling

The first step consists in using a trained deep-SDM to encode the high-dimensional predictive data around each species occurrence into a reduced feature space (Figure 5.1 **Step 1**). The model used is an Inception v3 convolutional neural network (Szegedy et al., 2016a). The data set was divided into training/validation/test sets with spatial blocks of 0.025 degrees in the spherical coordinate system. The 90/5/5% block allocation was further stratified by region to optimise the diversity of the sets. At the occurrence level this results in a set distribution of 902,174 / 46,290 / 50,794. At the species level this leads to a distribution of 14,129 / 4,037 / 4,166. Training was performed on two V100 GPUs from the Jean Zay supercomputer. The model is trained with the LDAM loss, a modified cross-entropy function that gives more weight to rare species during training (Cao et al., 2019). Training on 70 epochs took 42 hours with a batch size of 128 and an initial learning rate of 0.01. Model performance is evaluated every two epochs. The final test set performance is reported for the best validation epoch. Finally, the deep-SDM is retrained on the entire dataset for the best validation epoch prior to feature extraction.

Inception v3's multiple convolution and pooling layers allow the input information to be reduced to 2048 dimensions in the original version of the network. However, 2048 dimensions is still too many to perform classification on a few hundred samples. To reduce the feature dimensionality, the Fully Connected (FC) layer with dimensions $(2048, \#labels)$ before the final softmax layer of the model was replaced by two layers $FC_1 = (2048, N), FC_2 = (N, \#labels)$, creating a dimensional bottleneck. ReLU activations are also appended after each FC. Finally, the feature associated with an observation o is the N-dimensional test activation extracted after FC_1 and noted as f(o).

Deep-SDM validation. For a given observation, top-k accuracy assesses whether the model returns the true label among the k most likely species. Success rates can then be calculated for all classes together (micro-average) or first by class and then averaged together (macro-average). This means that the micro-average is more representative of performance on common species, whereas the macro-average is a better representation of performance on rare species. Validation performance has plateaued since epoch 66, i.e. after the LDAM loss reweighting scheme in epoch 65. The micro-average top-30 accuracy stabilises around 0.82 and the macro-average top-30 accuracy stabilises around

0.42, with the same performance on the test set. The final deep-SDM is then retrained on the full dataset for 70 epochs.

5.2.3.2 Dispersal scenarios

As a reminder, a fundamental assumption when inferring species distributions with an SDM is that species maintain the same environmental niche (Bakkenes et al., 2002). The assumption that species have unlimited dispersal capacity leads to species niches shifting with climate change. In practice, however, species have specific and limited dispersal capacities that prevent them from following climate change (Schloss et al., 2012). As data on plant dispersal capacity is extremely scarce, we worked with two extreme scenarios and assumed that the truth lies in between (Thomas et al., 2004; Urban, 2015):

- *No dispersal.* Species can only be re-predicted by the deep-SDM at locations of true observations. By construction, this results in species potential presences (*support points*) prior to SDM inference that can only be fewer in number than in the present.
- *Unlimited dispersal.* Species can now be predicted at every location in the dataset (999,258 occurrences). In fact, the dataset is considered large enough to be used as an approximation of every possible location an orchid species could occupy.

In both dispersal scenarios, the relative probabilities of presence $\mathbb{P}(Y|X)$ returned by the SDM are thresholded to retain only the most likely species predictions (this is traduced by the indicator function in Equation 5.2). The threshold λ was optimised in section 4.2.3 to return precautionary species assemblages. More specifically, the value was set on a calibration set to allow only a 3% error in assessing whether the true label was retained in the assemblage - while keeping the threshold as high as possible. Then the model recall is optimised, which corresponds to the conformal prediction setting (Fontana et al., 2023). As the threshold has been optimised to return likely overestimated but precautionary species assemblages, our estimates can be considered as lower bounds on species extinction risk. Finally, species are classified as extinct if they are not predicted to occur at any point. Only the no dispersal scenario leads to this case. A species that is predicted to become extinct in a given time period cannot return to any other status thereafter.

5.2.3.3 Species niche features for IUCN classifcation

We need an information at the species level to feed an extinction risk classifier (Figure 5.1 **Step 3**). However, SDMs successively process environmental data at the level of an observation (Figure 5.1 **Step 1**) before returning the most likely species in these conditions. This is why we need to aggregate the information initially at observation level to species level. To do so, we use summary statistics on vectors from the model's feature space (Figure 5.1 **Step 2**) corresponding to a set of *support points*: known points in the present or future potential presences. In a future scenario, the choice of support points to aggregate features at the species level depends on species assumed dispersal capacities.

We now describe in more details the two first steps of our method for extracting species features. The item numbers are matching the steps of Figure 5.1.

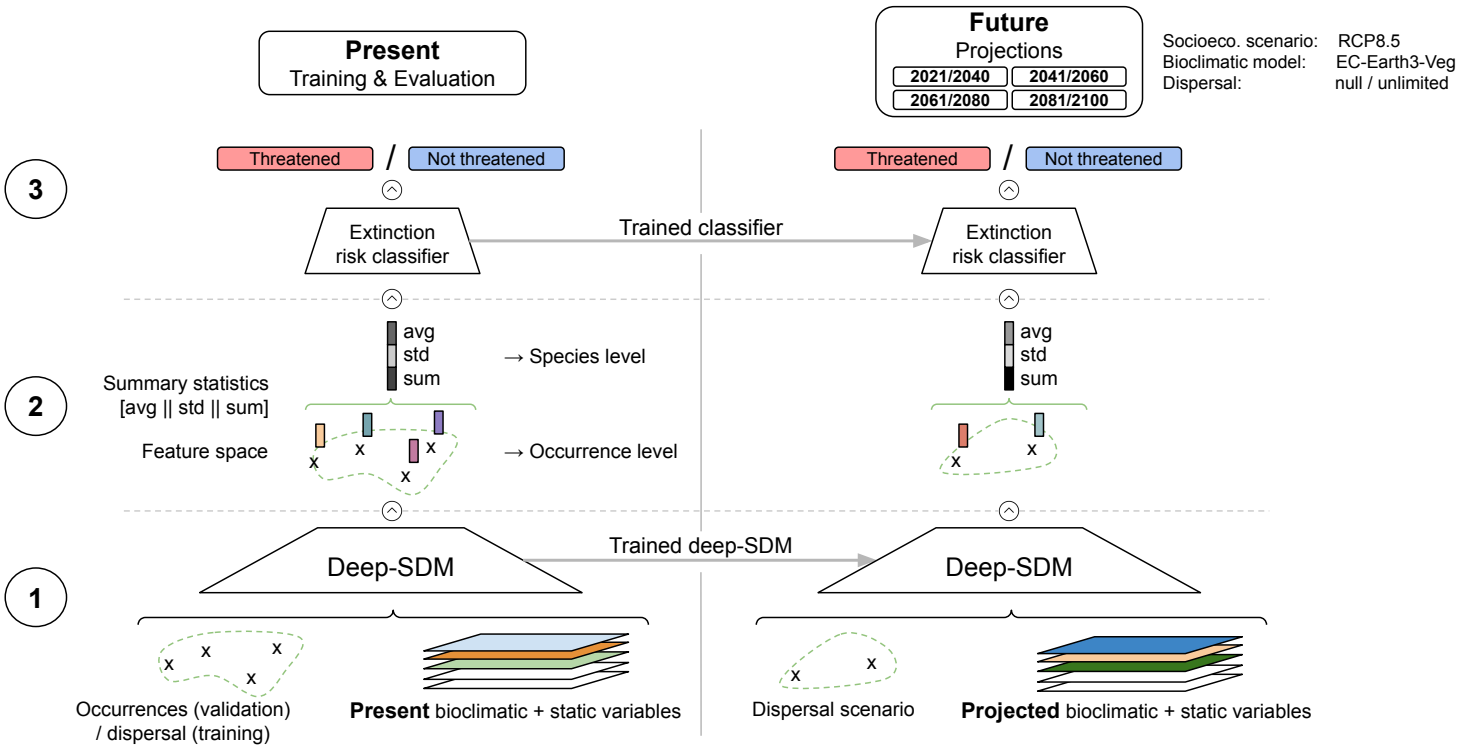


Figure 5.1: *Method scheme. The first column represents the current time frame (training and evaluation), while the right column represents a future scenario (projection).* ***Step 1*** *consists in associating a given set of support points (true observations or dispersal scenario) with environmental covariates using deep-SDM inference. Sites where the species has been observed and is predicted to remain are indicated by 'x', abandoned sites by 'o' and new predicted sites by 'X'. Species niches are indicated by dashed circles. In* ***Step 2****, the predicted features are summarised by taking their mean, standard deviation and sum, and the result is concatenated. This operation allows the information that was at the point level to be condensed to the species level. Finally, the* ***Step 3*** *is the mapping between the species summary feature and its conservation status. After training and validation in the present with real observations (see Results), the random forest classifier is trained within dispersal scenarios to ensure coherence with future projections. Classification can be either binary, as shown, or at the IUCN status level.*

(1) **Identification of the support points of the species and deep-SDM inference.** At present, it is the species known sightings for model evaluation and status inference. In the future, the set of support points depends on the dispersal hypothesis made as detailed in section 5.2.3.2. Assuming no dispersal, the set of support points of a given species only includes the geolocations where the species has already been observed. Assuming conversely unlimited dispersal, the set of support points for a species is approximated by the 1M geolocations encompassed in our orchid observation dataset. Another possibility would have been to exploit a global regular grid as support points. However, we preferred to use the proxy of all the locations covered in our orchid dataset to limit computation resources. The deep-SDM was trained beforehand and evaluated with true occurrences and present covariates as described in section 5.2.3.1.

(2) **Calculation of summary statistics on the resulting features.** Weighted mean,

standard deviation and sum are computed over the features, along the N dimensions. The final species feature is the concatenation of the three statistics above (dimension $3N$). Different summary statistics were evaluated. This concatenation led to the best results and will be further discussed in the final section. In addition, different weighting schemes for support points were considered when using a dispersal scenario. First, only points where the given species is predicted to be present with a minimum relative probability are retained, as explained in section 5.2.3.2. Second, among the retained points, their contribution is weighted according to the prediction rank of the target species. Indeed, a first implementation was to weight the contribution of the retained support points by the relative probability of presence of the target species. However, this strategy was not efficient to differentiate the support point contribution according to the most likely species. This is due to the deep-SDM calibration. The model was trained with presence-only observations and, as a result, the distribution of species relative probabilities has low dynamic. Therefore, for a given species, we preferred to weight the contribution of its support points with the inverse of its rank when ordering the most likely species conditional on the observation (see Equation 5.2).

Let $K_{s,t}$ be the set of support points of a species s for the time period t that depends on the dispersal scenario considered (see Section 5.2.3.2). Based on the outputs of the deep-SDM at all points in $K_{s,t}$, we construct an aggregated feature vector $\mathrm{h}_{s,t}$ to be used as input of the final species status classifier:

$$\mathrm{h}_{s,t} = \underset{k \in K_{s,t}}{\mathrm{Stats}}[w_k.\mathrm{f}(k)] \tag{5.1}$$

with:

- Stats = [avg ‖ std ‖ sum], ‖ being the concatenation operator
- f(k) the N-dimensional feature vector associated to the environmental covariates $x(k)$ through the deep-SDM.
- w_k the weight attributed to support point k. The higher the weight, the more likely the species is present at this support point. w_k expression depends on the species support points. We denote rank$[k_{s,t}]$ the position of species s within the ordered list of the most likely species returned by the deep-SDM at the point k and time period t.

$$w_k = \begin{cases} 1 & \text{if true observations are used} \\ \mathbb{1}_{P[Y=s|X=x(k_{s,t})] \geq \lambda} \cdot \frac{1}{\mathrm{rank}[k_{s,t}]} & \text{within a dispersal scenario} \end{cases} \tag{5.2}$$

5.2.3.4 Classifying extinction risk status from species niche features

The final step of our method is to classify the extinction risk status of species from their aggregated feature vector $\mathrm{h}_{s,t}$:

③ **Extinction risk classification.** Therefore, a random forest classifier is trained using official IUCN status and the present occurrences of the species as support points. In the present, it is then used in inference to determine the preliminary extinction risk status of unassessed species using their observations as support

points. To be coherent with the future dispersal scenario and for the calibration of the extinction risk classifier, the reference classifier used to predict future IUCN status is also retrained in the present with the same dispersal scenario (either null or unlimited). Otherwise, a classifier that is i) trained with species features computed using true observations as support points, and ii) used to predict the future IUCN categories associated with species features aggregated from unlimited dispersal support points, will result in iii) severely underestimated extinction risk levels. Finally, two levels of classification are considered: one at the binary level (threatened or not, as shown in the scheme) and another at the IUCN status level.

Other classifiers were also considered before selecting the random forest: a shallow multi-layer perceptron, a multinomial log-linear regression and linear classifiers with Stochastic Gradient Descent (SGD) training. Performance was evaluated using a 10-fold cross-validation strategy on the 889 species assessed by IUCN. The best classifier, the random forest, was then compared with the state-of-the-art *IUCNN* extinction risk classifier (Zizka et al., 2021) at both binary and status levels (see model valiation Section 5.2.4). Once the relationship between species features and extinction risk status has been learned for the 889 IUCN-assessed species, risk status can be predicted for the remaining 13,240 species in the dataset.

5.2.3.5 Projections within the CMIP6 SSP5-8.5 scenario

Climate model choice. The choice of climate model was made using the GCMeval tool (Parding et al., 2020). The focus region was set to global and the skill assessment weights were left at their default values (equal importance given to temperature and precipitation, idem for all seasons and skill scores). Finally, the emission scenario was set to SSP5-8.5 and only climate models whose projections were available for download at i) 30 seconds spatial resolution, ii) for all time periods and iii) for the 19 bioclimatic variables on https://www.worldclim.org/ were considered. This led to the adoption of the EC-Earth3-Veg system model projections (Döscher et al., 2022).

Levels of analysis over time: Continents, latitude and altitude. In addition to the temporal dimension, the predicted status distributions are crossed with three other variables. One is categorical in the inhabited continents and two are continuous in the species mean latitude and mean altitude. Inhabited continents were obtained by spatially intersecting occurrences with WGSRPD level 3 zones. A given species may span several continents. Mean latitude was calculated directly from the observation coordinates. Finally, elevation values were obtained from a global raster with 15 arc-second spatial resolution downloaded by tile at http://www.viewfinderpanoramas.org/ and processed using GDAL command lines. The predictions of extinction risk are identical to the three facets of the analysis. The only difference is in the presentation of the results. In practice, it is the variable used to group species that varies from one representation to another.

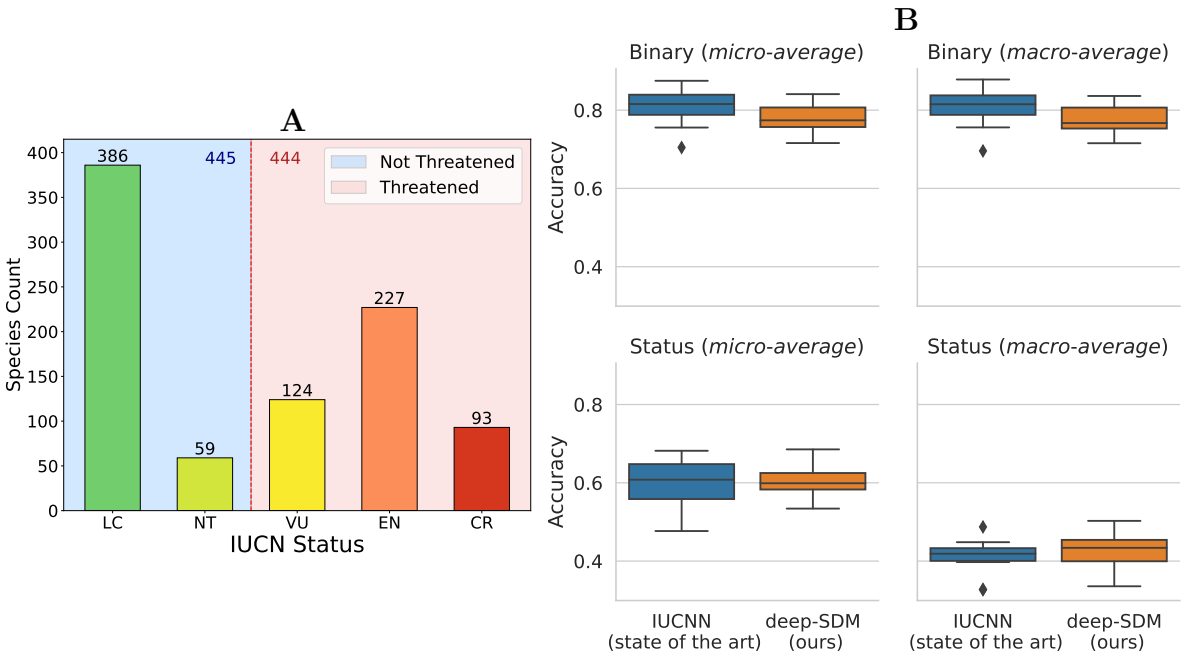


Figure 5.2: ***A.*** *Distribution of the 889 IUCN extinction risk status from our dataset.* ***B.*** *Classification performance comparison between the state-of-the-art IUCNN method and ours (Zizka et al., 2021). The 10-fold cross-validation results in an accuracy distribution represented by boxplots. The first row shows the binary classification and the second row the status classification. Micro- and macro-average accuracies are identical, as the binary status distribution is balanced. The IUCNN method achieves an average accuracy of 0.81 for binary classification and our deep SDM method achieves 0.78. However, for status classification, our method gives a micro-average accuracy of 0.61 and a macro-average accuracy of 0.43 and the IUCNN method 0.60 and 0.41 respectively.*

5.2.4 Model validation

Classification performance and comparison with the state-of-the-art IUCNN method are shown in Figure 5.2B (Zizka et al., 2021). While each method has a slight advantage in either binary or status classification, their performances are close enough to consider the deep SDM-based method as competitive with the state-of-the-art method. In addition, deep-SDM confusion matrices for the two classification levels are provided Fig. S10.

5.3 Results

We analyse the global dynamics of orchid IUCN status distribution over time and in function of three aspects: continents, latitude and altitude. In the figures below:

i) the results from the two dispersal scenarios are averaged to provide synthetic trends (Figure 5.3 represents their difference with error bars)
ii) all species are considered to be aiming at broad conclusions at the family level (i.e. both those already assessed by the IUCN and those that are not)
iii) binary statuses are reported as their prediction is more robust.

The same analysis, restricted to species assessed by the IUCN, is presented in Figures S12-S14. Comparing these gives a sense of the generalising power of our approach.

5.3.1 Continents

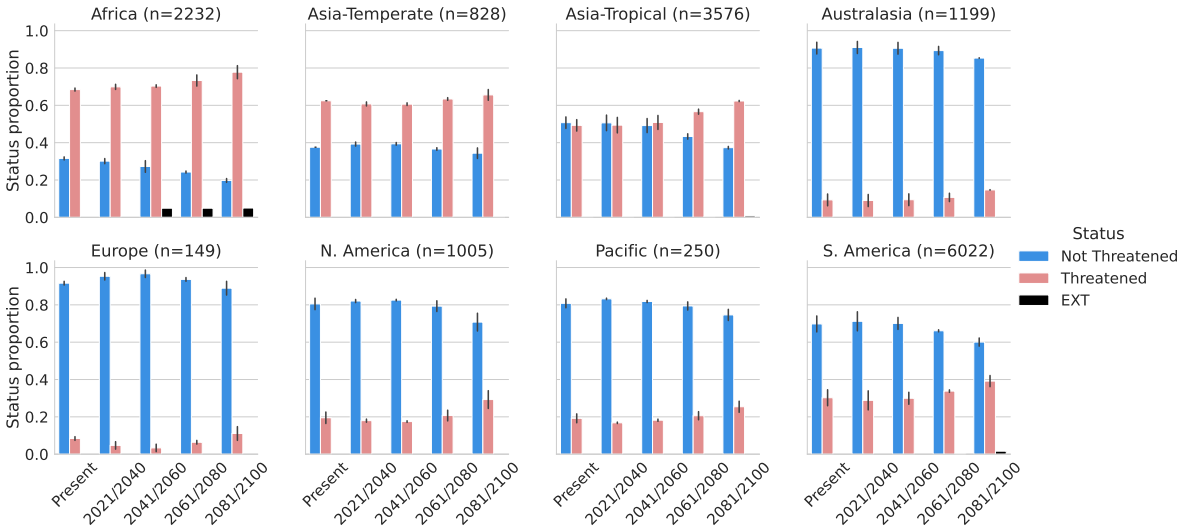


Figure 5.3: *Binary status proportions per continent and time period. All species are included and their number per continent is given in the subtitles. Error bars account for differences between the two dispersal scenarios.*

The global trend in Figure 5.3 shows an increasing proportion of threatened species across all continents. Individual patterns also appear:

- Africa and Asia-Temperate are the only two continents with a current majority of threatened orchid species.
- Five percent of African species are predicted to become extinct between 2041 and 2060 (11% of IUCN assessed species).
- In tropical Asia, threatened species become the majority by mid-century, reaching 60% by the end of the century.
- Several continents - notably Europe and North America - see their proportion of threatened species decline in the first half of the century, before recovering to overtake current levels.
- Although South America may appear relatively spared, its increase in the proportion of threatened species is significant and covers six thousand species.

On a global scale, the number of threatened species is expected to increase by an average of one third by the end of the century (14-40% rise depending on dispersal scenario). All status predictions are provided in the supplementary file *ALL_species_status.csv*. The .csv file is described in Box H. A smaller increase in threatened species is predicted for IUCN-assessed species, see Fig. S11. Unassessed species could therefore be expected to be at even greater risk of extinction than those already assessed.

We also predict that 234 species will become extinct, of which 42 are IUCN-listed: 111 in Africa, 96 in South America, 26 in tropical Asia and one in temperate Asia. The isolated list is also given as supplementary file *EXT_species.csv* and is described in Box H. A small number of species have been identified in GBIF as having very few and old occurrences.

5.3.2 Latitude

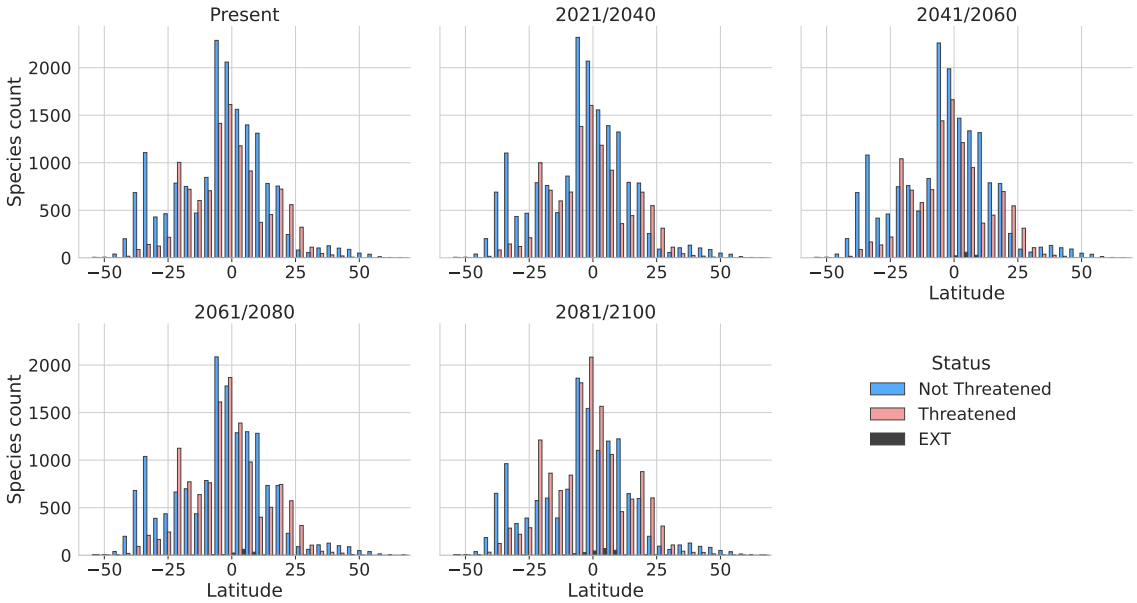


Figure 5.4: *Species count histograms as a function of average latitude and time period. All species are included and colours indicate the binary extinction risk status. Bins cover four degrees of latitude.*

In Figure 5.4, the number of threatened species clearly peaks around the equator and in the tropics, i.e. at latitudes with high species diversity. Species going extinct are also found around the equator. Both high and low latitudes are dominated by non-threatened species.

5.3.3 Altitude

Figure 5.5 crosses the number of classified species over time with their average altitude. The profile of threatened species along the altitudinal gradient appears to be stable over time. It shows a peak at low elevations and a concentration of threatened species in the 800-1500 m range. However, the number of threatened species increases at all altitudes. They outnumber non-threatened species in their peaks and also at very high elevations, about 2500 m.

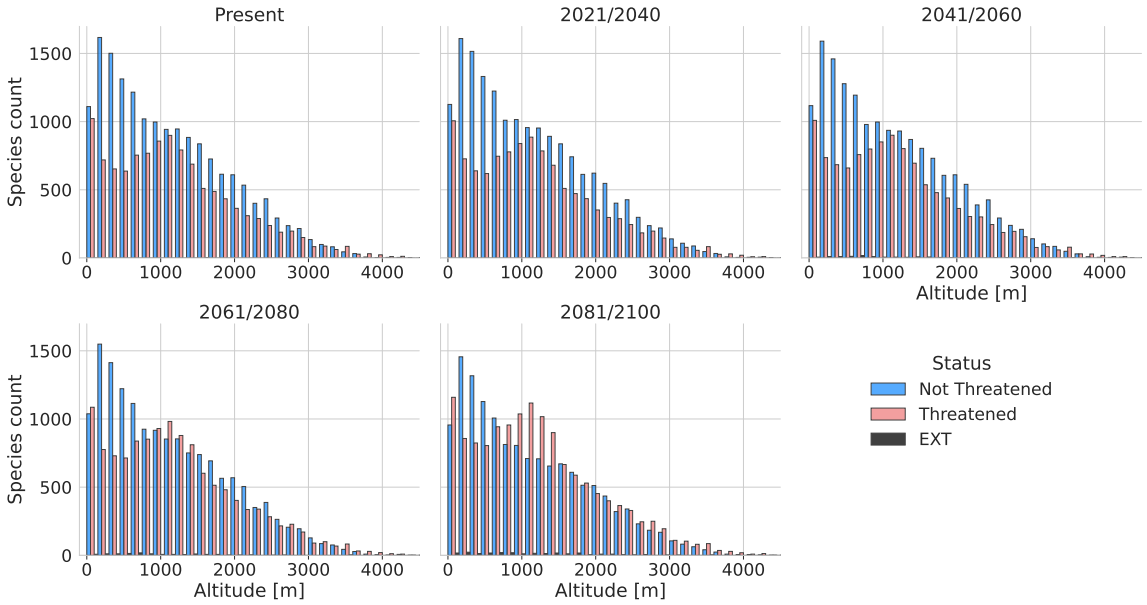


Figure 5.5: *Species count histograms as a function of their average altitude and over time. All species are included and colours indicate the binary extinction risk status. Bins cover 150 metre elevation ranges.*

5.4 Discussion

5.4.1 Interpretations

We now propose ecological interpretations that account for the different patterns observed. The relative decline of threatened species by mid-century - particularly pronounced in Europe and North America - is rooted in the unlimited dispersal scenario. Indeed, in these projections, the support for species training and then potential niches can be largely overestimated, leading to lower rates of threatened species. However, this phenomenon is mitigated in the second half of the century, where we assume that bioclimatic change significantly limits species' potential niches. The large proportion of threatened species currently predicted for Africa and Asia-Temperate could be partly explained by a different accounting of LC species. Further analysis at this level is needed to test this possibility. This seems all the more relevant when considering the proportion of threatened species restricted to IUCN-assessed species in Fig. S12.

On average, considering both dispersal scenarios together, the number of threatened species is expected to increase by a third with our method (14-40% rise interval, see Figure S11). This trend may seem small compared to other studies on the extinction risk due to climate change. For example, Urban (2015) concluded that under current emissions trajectories, up to one in six species could be threatened with extinction by climate change. However, the global trends we have expressed are relative, i.e. the increase in threatened species relative to current levels. In absolute terms, we predict an 11% increase in the number of all threatened species worldwide (6% if only IUCN-assessed species are considered). This is closer to the 16% absolute increase from (Urban, 2015). Furthermore, our method construction leads to lower bounds on extinction risk levels as already formulated. Finally, as observed in the analysis of trends by continent, the

impact of climate change on species loss is predicted to be greater in the second half of the century than in the first. Similarly, Vuuren et al. (2006) predicts that the impact of climate change will become increasingly important after 2050.

Species that are already IUCN-assessed are predicted to experience a smaller increase in their threatened share than species that are not. We formulate two hypotheses for this pattern. First, plant assessment targets include a majority of species expected to be threatened (Bachman et al., 2019), resulting in an overall proportion of threatened plant species of 48%. In comparison, the global estimate is only 21% (Brummitt et al., 2015). This assessment bias therefore contributes to explaining the more stable proportion of threatened species among IUCN-assessed species. Second, since the classifiers were trained on the IUCN-assessed species with current bioclimatic values, we can expect some overfitting on these species. Even if bioclimatic variables change in the future, the classifiers are also provided with static variables, which could indeed explain a higher tendency to re-predict status quo for these species.

The overall latitudinal species distribution follows, as expected, the latitudinal gradient of biodiversity (Willig et al., 2003). However, we found no satisfactory hypothesis to explain the trident shape of the distribution of threatened species along the latitudinal gradient. This shape is also present when only considering IUCN-assessed species, but with the highest number of threatened species peaking around the -20° parallel.

In the case of lowland species at risk, a direct assumption is that land-use change and high exposure to anthropogenic threats are at play. The hump-shaped pattern of threatened species corresponds to the diversity peak known to occur at intermediate altitudes (Whittaker, 1960). Indeed, these altitudes are known to be rich transition zones between different habitats and with specific interactions between temperature and water gradients (Zhao et al., 2005). However, this does not explain why only threatened species follow this pattern. There may be several reasons. First, speciation rates and endemicity also peak at intermediate altitudes, thanks to an optimal combination of area and isolation for the persistence and divergence of native species according to Lomolino (2001). This would result in species with few individuals and/or restricted geographical ranges that are likely to be threatened. Secondly, there is almost no more primary forest at low elevations, in contrast to intermediate elevations. These forests may have been orchid refuges in the past, but may now be increasingly threatened by extensive land change and climate change.

Preliminary results show that terrestrial Human Footprint appears to correlate with the predicted proportion of threatened species, see Fig. S17. This is consistent with previous work showing a strong correlation between human footprint and species extinction risk (Di Marco et al., 2018. This study even found anthropogenic pressure to be more predictive of extinction risk than environmental or life-history variables.

5.4.2 Limitations

Our method has biases and limitations, which we will acknowledge below. As mentioned previously, the threshold for species re-prediction may be too permissive. In both dispersal scenarios, this may lead to an overestimation of geographic support for learning species

features. A more restrictive choice could have the effect of increasing the number of species predictions at risk - further work is indeed necessary. Similarly, the weighting scheme used to compute species features from activations needs more attention. Classification performance was indeed found to be sensitive to changes at this level and validation would allow the most appropriate scheme to be set.

Furthermore, our species features are based on points projected in time and space for the dispersal scenario. In the worst case, future bioclimatic conditions combined with static variables create previously unseen contexts leading to out-of-domain model inference. In a conservative scenario, static variables lead to automatic re-prediction of true labels plus likely species, regardless of bioclimatic conditions. This results in constant or underestimated extinction risk depending on the dispersal scenario. Finally, in the targeted scenario, bioclimatic projections combined with static variables shift species contexts in a large enough and structured representation domain, leading to coherent inferences. Thanks to the size of our dataset and previous interpretability study on the generalisation power of deep-SDMs, we believe that the behaviour of the model does indeed tend towards this third scenario.

At the level of status classification, performance drops for NT and VU statuses, i.e. for transition categories (see confusion matrices Fig. S10). While these statuses have relatively low training support (59 and 124 species respectively), we believe that this is largely due to the rather ambiguous definition of the NT category:

> *"A taxon is Near Threatened when it has been evaluated against the criteria but does not qualify for Critically Endangered, Endangered or Vulnerable now, but is close to qualifying for or is likely to qualify for a threatened category in the near future."*
>
> (Bland et al., 2017)

As the reference delimitation between NT and VU species is potentially confused, their respective learning is assumed to be worse than other classes and performance optimisation leads to their abandonment in favour of LC/EN/CR predictions.

5.4.3 Perspectives

The perspectives opened up by this study are many. The SDM-based species features extracted before the final softmax layer of the model were shown to be predictive of extinction risk status. This feature space close to the SDM output is meant to be linearly separable by the different classes. In contrast, activations closer to the SDM input might be less informative about the classes, but more representative of the predictive features. Extracting species features earlier may be advantageous given that the activations flow on a predictors-class information gradient across the model layers.

One of the strengths of our model is its scalability to thousands of species. However, it is not adapted to follow the IUCN species-specific guidelines when using an SDM to estimate the different criteria (Bland et al., 2017). A direct exchange with the conservation community on their specific needs would certainly help to guide future development efforts. Another natural area for improvement is to focus on a few well-documented

species and subject our method to the IUCN guidelines. Furthermore, it is appealing to test our method on another suitable taxon to evaluate its taxonomic generalisation power.

Finally, including additional information that predicts the likely presence and extinction risk of species are other promising directions: species traits (Bourhis et al., 2023), ecosystem functional attributes (Arenas-Castro et al., 2018) and other threat/habitat rasters such as urban expansion forecasts, forest cover predictions or protected areas (Borgelt et al., 2022a; Vieilledent et al., 2022). Indeed, the inclusion of climate change alone inevitably leads to an underestimation of future threats to biodiversity (Brook et al., 2009).

5.4.4 Conclusion

The prediction of species extinction risk from SDM-based features achieves state-of-the-art performance while being flexible enough to allow testing climate change scenarios. This means that the valuable information provided by the predictors has been successfully encoded by the SDM. Future projections of orchid extinction risk averaged over two dispersal scenarios provide biodiversity trends to support global conservation targets (Nicholson et al., 2019). Indeed, this classification framework allows to investigate the impact of climate change on the distribution of species extinction risk. While the proportion of threatened species is increasing globally, analysis by continent, latitude or altitude reveals specific and escalating patterns.

ACKNOWLEDGEMENTS

The research described in this paper was funded by the European Commission via the GUARDEN and MAMBO projects, which have received funding from the European Union's Horizon Europe research and innovation programme under grant agreements 101060693 and 101060639. The opinions expressed in this work are those of the authors and are not necessarily those of the GUARDEN or MAMBO partners or the European Commission. The INRIA exploratory action CACTUS fund also supported this work. This work was granted access to the HPC resources of IDRIS under the allocation 20XX-AD011013648 made by GENCI. Finally, we warmly thank Alexander Zizka for providing us with the filtered set of orchid occurrences.

CHAPTER 6

CONCLUSION AND PERSPECTIVES

Table of contents

6.1 Results synthesis 144

6.2 Limits and perspectives 145

6.2.1 Contribution of our work to conservation 145

6.2.2 Uncertainty and confidence in predictions for conservation . . . 147

6.2.3 Research directions 148

6.2.3.1 Species distribution modelling with deep learning . . . 148

6.2.3.2 IUCN status prediction 149

6.2.4 IA, Ecology & Conservation 150

6.3 Conclusion 151

6.1 Results synthesis

In this doctoral project we explored the application of deep learning based species distribution models to conservation science. While the fields of biodiversity modelling and conservation are still largely independent (Pollock et al., 2020), we believe that our work contributes to identifying and opening up research directions that could ultimately benefit conservation.

First, in **chapter 3** we showed that the potential of satellite imagery for modelling species distributions remains largely untapped. To this end, we collected a novel dataset *DeepOrchidSeries* linking 1M orchid occurrences with image time-series of their habitat (twelve-month RGB+IR series of Sentinel-2 imagery, 640 x 640 m extent and 10 m resolution). We then demonstrated how the habitat phenology captured by these time-series helps models to shape species distributions. Interestingly, the performance gain is greatest for hard predictions of rare species and in species-rich regions. Furthermore, we have shown that even when the model is trained on partially noisy data (geolocation and temporal mismatches), it generalises well to unseen data and performs best when mismatches are closed. This highlights the value of the increasing amount of accurate observations collected by experts and citizens. In addition, this first chapter provides a comparison between three SDM covariates: bioclimatic conditions, habitat imagery and static variables (elevation, position, human footprint and ecoregions). On a global scale, bioclimatic covariates are identified as the best driver of orchid distribution. Compared to bioclimatic variables, fine-scale habitat imagery provides complementary, rather than alternative, information at very high spatial resolution. Furthermore, static variables cause the spatial-explicit model to perform close to the bioclimatic reference.

Next, in **chapter 4**, we investigated how spatial inference from deep-SDMs can fill the knowledge gap on the spatial range of understudied taxa (Wallacean shortfall). Applied to a global dataset of orchid observations, our method first predicts likely species assemblages at the kilometre scale from globally available information: bioclimatic, pedological, human footprint variables, ecoregions and position. On a validation set, the true species is guaranteed to be within the predicted assemblage with 97% confidence (conformal prediction). This in turn allows the construction of global, scalable indicators of extinction risk of species assemblages and of the Shannon diversity index. More specifically, our two indicators span (i) the proportion of threatened species and (ii) the status of the most threatened species in the predicted assemblages. In addition, we have shown how the addition of status predictions to current IUCN assessments leads to substantially increased overall threat levels and reveals strong spatial patterns in understudied regions. Analyses can then be carried out at regional, national or international level, thanks to summary statistics and an interactive website displaying our results at https://mapviewer.plantnet.org/?config=apps/store/orchid-status.xml. Using these tools, we have identified regions and countries with highly threatened species assemblages. Southeast Asian countries accumulate very high levels of orchid diversity and extinction risk, while the islands of Madagascar, Reunion and Mauritius host the most threatened orchid assemblages (60% of them are predicted to be threatened). We have also illustrated how the indicators developed relate to the current implementation of protected areas in Sumatra.

Finally, in **chapter 5** we used a trained deep-SDM to predict the IUCN extinction risk status of species and forecast future status dynamics. In this context, the model can be seen as a dimension reduction algorithm. It benefits from the generalisation power of deep-SDMs and results in a rich embedding of the species niche. Our classification scheme aims to mitigate the over-reliance on geographic information in flora assessments, thus allowing prediction of future extinction risk patterns. The critical information found to correlate with species observations is propagated through the model layers and is considered representative of the species' environmental preferences. The resulting model activations are structured in a *feature space* of reduced dimension. We aggregate the features per species and use the summary statistics in a supervised classification task where the objective is to predict the IUCN status of species. This method leads to a competitive classification performance, illustrating the interest of deep-SDMs for downstream tasks. Furthermore, our framework allows us to project the impact of climate change on IUCN status predictions. Under a business-as-usual scenario of pollution rates, bioclimatic conditions are projected into the future. Species are assumed to have either zero or unlimited dispersal capacity. We examined the consequences of climate change under these conditions, analysing the results by continent, latitude and altitude. More than 80% of African orchids could be threatened with extinction by the end of the century. Tropical Asia and South America - the two most species-rich continents considered here - would see a significant increase in the proportion of threatened species. In addition, the predicted number of threatened species peaks at the Equator and the two Tropics, but also in the lowlands and at middle altitudes (800-1,500m), where species used to be relatively unaffected.

6.2 Limits and perspectives

6.2.1 Contribution of our work to conservation

To start with, as pointed out in Pollock et al. (2020), modelling plays a crucial but under-appreciated role in conservation planning and practice, particularly in setting global conservation targets. The outputs of single-species SDMs are already being used for conservation (Guisan et al., 2013). Modern models such as deep SDMs address the same goal of mapping species distributions, but rely on more advanced techniques to harness large amounts of data and help overcome assessment biases along with knowledge shortfalls about biodiversity (Hortal et al., 2015). One can therefore ask what is slowing down the adoption of modern SDMs by the conservation community? As already mentioned, the reason certainly lies in the "black box" appearance of the model. However, there are solutions and the next section 6.2.2 will explore some of them.

Again, models are needed to help broaden our limited view of biodiversity. In this respect, Pollock et al. (2020) distinguishes two classes of models: *imputation models*, which aim to fill gaps in our knowledge of biodiversity properties (e.g. biological traits, functions), and *spatial biodiversity models*, which aim to spatially generalise our knowledge to unsampled sites (often for the benefit of conservation planning). In **chapter 3**, our work on the contribution of satellite image time-series aims to improve such spatial biodiversity models. The multiplication of remotely sensed views of species habitat

targets a better characterisation of species environmental preferences and ultimately a better mapping of species distributions. Our research on species assemblage prediction and indicator design (**chapter** 4) also aims to widen our knowledge of biodiversity in space. In fact, our method lies at the intersection of the two classes of models outlined above: we first use an imputation model to complete the IUCN Red List coverage of a given taxon, and then generalise the knowledge of extinction risk in space. While such a combined method may seem ambitious, we believe that the effort is worthwhile to further our understanding of biodiversity and to illustrate modern modelling capabilities. Finally, our IUCN status classification scheme developed in **chapter** 5 belongs to the class of imputation models.

Here we discuss additional benefits of our **chapter** 4 maps that motivate their uptake in conservation. First, our maps are based on a multi-species approach that combines information at the species and assemblage level. Pollock et al. (2020) argue that this is highly relevant for conservation applications, as it allows the simultaneous estimation of species-relevant ecological processes and larger-scale attributes. Second, the habitat-based metrics often used in conservation may overlook important facets of biodiversity, such as trophic-level species traits (Decker et al., 2017; Marshall et al., 2020), hence the need for complementary metrics. Furthermore, when multiple species are considered, this is usually done through species richness. This metric alone cannot represent the spatial heterogeneity in extinction risk patterns. Finally, conservation priorities at local, national or global scales can often differ because they take into account different extents, objectives and data (Pollock et al., 2017). Multi-scale indicators prevent such inconsistencies from hampering the conservation planning process.

Naturally, there is room for improvement to increase the relevance of our work for conservation. Regarding our *DeepOrchidSeries* dataset, adopting the formalism of Essential Biodiversity Variables (EBV, see section 2.2.2.3) would improve its reuse by the community. Next, comparing the orchid assemblages predicted by our method with other initiatives such as the SESAM framework (Spatially Explicit Species Assemblage Modelling, Guisan and Rahbek, 2011) could assist in the adoption of our indicators. In addition, further exploration of the implications of our work with orchid conservationists is a logical direction. Defining hotspots of extinction risk for species assemblages would be an easy first step to start the dialogue. Comparing our findings on orchids with the recently reviewed global conservation priorities of the family (Fay, 2018; Gale et al., 2018; Vitt et al., 2023; Wraith et al., 2020) is definitely an area for improvement in our work. Furthermore, an analysis of the discrepancies between the extinction risk predictions of different automated methods would allow a better understanding of the strengths and weaknesses of the classifiers. Finally, improving the interpretability and reliability of models is a general perspective for improvement. As we believe that this is a pivotal point in our work, but also in the field of biodiversity modelling, we dedicate the next section to this topic. In addition, the research directions on models and analysis of results presented in section 6.2.3 would ultimately help to inform conservation as well.

6.2.2 Uncertainty and confidence in predictions for conservation

In order to properly inform conservation, quantifying the uncertainty associated with our model predictions and, when appropriate, increasing confidence in the predictions is a priority perspective for improvement (Rapacciuolo, 2019). Few directions can be identified at the level of SDMs, IUCN status classification or, more conceptually, for all biodiversity models.

Applying existing xAI tools to explain our SDM predictions is a promising direction (Ryo et al., 2021). Options are numerous, be it with the popular shapeley values method (SHAP, Lundberg and Lee, 2017), with local interpretable model-agnostic explanation (LIME, Ribeiro et al., 2016), or with the recently unified *Captum* PyTorch library (Kokhlikyan et al., 2020) among others. With regard to our species assemblage predictions and the indicators derived from them, it would be of great interest to produce the list of retained species in order to increase transparency. This would be a challenging task when considering the global 221M point, but it is definitely worth exploring for a regional or random under-sample. In addition, using the distance of the relative probability of species presence to the selection threshold would give a sense of model confidence. For example, for our spatial indicator of the most critical IUCN status, model confidence could be represented by the probability/threshold distance of the most likely species of that status. As a result, confidence maps could be produced to aid the use of our indicators. Finally, advances and standards in SDM evaluation would ensure more reliable use of SDM predictions for conservation. Model evaluation affects not only final performance reports, but also model training through successive validation steps, hence the need for ecologically relevant metrics (Mouton et al., 2010). A first challenge concerns the ground truth to evaluate the models: while presence-absence observation data is the best reference, it is really scarce and therefore surrogates such as presence-only data from citizen science or expert range maps are used (Mainali et al., 2020). Other challenges include the influence of species prevalence on model assessment (long-tail distribution in the number of observations per species) and covariate spatial autocorrelation, which can lead to overestimation of performance if not controlled (Schratz et al., 2019).

With regard to our IUCN extinction risk classification method, the use of xAI tools is complicated by the fact that the inputs are not easily interpretable as they are encoded in a reduced dimensional space. However, this representation space has been shown to be structured by species' environmental preferences and bioclimatic information (Deneu et al., 2022). Model ensembling is an option to produce more robust predictions, as shown by Borgelt et al. (2022a). We can imagine ensembling models trained with different climate change scenarios to cover the range of possibilities, or different climate models for a given scenario. The use of a Bayesian approach to quantify uncertainty, as in the *IUCNN* approach, is also a perspective that holds promise. Furthermore, our IUCN prediction method includes a range generalisation step by relying on SDM predictions. We believe that testing the robustness of our classifiers to a degraded number of occurrences would provide useful insights. Similarly, Breiner et al. (2017) tested the impact of local extinction events on species EOO/AOO. In addition, further analysis of the predicted patterns of extinction risk would help to confirm or disprove the method. This is discussed in more detail in section 6.2.3.2.

To conclude, we present three general perspectives on the uptake of biodiversity modelling results. First, conservation actors need flexible tools and platforms to interact with. Rigid products and frameworks are unlikely to be appropriate for specific case studies. In this respect, we believe that the interactive website presenting our indicators is an asset. This leads to the second point raised by Pollock et al. (2020). To avoid inefficient method development and to optimise the results of biodiversity modelling, conservation stakeholders should be involved early in the model development process. This would allow appropriate conservation objectives, targets and scenarios to be considered. Such dialogue would ultimately increase confidence and acceptance of models. We fully acknowledge this as a specific shortcoming of our research products, as no conservation actor was involved in the early stages of method design. Third, incorporating causal modelling techniques into biodiversity models could greatly support effective conservation action (Gonzalez et al., 2023; Pichler & Hartig, 2023). Causal inference helps machine learning models become more accurate and interpretable by identifying and explaining cause-effect relationships between variables. Already widely used in medicine, economics and political science, its adoption in biodiversity modelling is both promising and necessary.

6.2.3 Research directions

In this section, we detail a few research directions that we believe could directly improve the performance of our models and clarify our results. In contrast, the next section 6.2.4 is broader and addresses general perspectives on the interplay between AI, ecology and conservation.

6.2.3.1 Species distribution modelling with deep learning

In terms of deep-SDM architecture and readily available covariates, the field is rapidly evolving and the directions of progress are many. First, environmental covariates are often heterogeneous, with different formats and scales, which can make their interaction complex. One possibility is to train independent first layers for each input scale/format and fuse the intermediate representation later (late fusion). Embedding categorical data before model input has several advantages: it speeds up model training, but more importantly it can help to generalise better with sparse data (Guo & Berkhahn, 2016). In response to the class imbalance problem, machine learning has produced different loss functions and strategies that help improve rare species predictions, so exploring alternatives and complements to LDAM loss would provide valuable insights (Cao et al., 2019; Wang et al., 2021). A natural perspective for improvement is to exploit novel deep learning architectures. There are many candidate models from computer vision, with vision transformers and their adaptive receptive fields and the next generation of CNNs showing great promise (Liu et al., 2021; Liu et al., 2022). Models specifically designed to capture spatio-temporal dynamics in satellite image time-series should also be considered to capture habitat dynamics and model species distributions (Hong et al., 2021; Ji et al., 2018; Rußwurm & Körner, 2018). Furthermore, habitat characterisation through remote sensing to feed multi-species SDMs is still in its infancy. We have shown in this first chapter how the temporal dimension can help to differentiate species habitat on a global scale, but many other aspects could be explored. Combining the spatial, temporal and

spectral dimensions of satellite imagery with radar data is an exciting opportunity (Zhu et al., 2021). Analysing the test performance as a function of the cloud cover percentage of the image time-series would be an informative validation experiment. Closing the temporal and spatial mismatches between each species observation and its remotely sensed covariates is challenging when dealing with large amounts of data, but promising given the results in section 3.3.5. Assuming the observations are in the centre of the patch, models could attribute particular importance to the central spatial information and gradually diffuse the weight towards the patch periphery. Finally, providing species traits, phylogeny and life forms as SDM inputs, when available, could help shape species distributions and ultimately benefit the IUCN status classification with SDM-encoded features (Bourhis et al., 2023).

6.2.3.2 IUCN status prediction

In addition to the required uncertainty quantification already detailed, our IUCN status classification method would benefit from efforts in species modelling and results analysis. Starting with modelling perspectives, providing current forest cover as SDM input and informing future scenarios with deforestation projections (Vieilledent et al., 2022) could help to capture threats to orchids and ultimately benefit the IUCN classification, especially for epiphyte species. Species traits, life form, phylogeny and any other species-level information can also be concatenated with the SDM-based features to aid status classification. In the case of epiphyte orchids (about 70% of species according to Atwood, 1986), taking advantage of observational data on their host trees in addition to providing life form information is likely to benefit models (Flores-Tolentino et al., 2020). As our approach focuses primarily on the quality and extent of species' habitats, it is also worth testing restricting the training support to species that are specifically threatened by habitat conversion. Another direction is to explore SDM's intermediate representation spaces to compute species features. Considering that the input environmental information is progressively encoded to translate species presence, intermediate layers could indeed represent relevant trade-off states. Finally, increasing the number and consistency of manual extinction risk assessments will also improve automatic classification. Indeed, they serve as reference for model training and thus have a large influence on automatic classification.

The understanding of IUCN status predictions can be enhanced by further analysis of our results. First, phylogenetic trees provide a great opportunity to examine whether the status distribution is phylogenetically structured (with or without phylogeny explicitly provided as classifier input, González-del-Pliego et al., 2019). The same observation applies to species' life form: does it make a difference to IUCN status predictions? IUCN predictions under climate change also merit further analysis. Investigating which species are predicted to change status and where is a first perspective. Moreover, we can imagine calculating the indicators built in chapter 4 with the predicted species distributions and statuses from chapter 5. Highlighting areas of high indicator change could then be revealing.

6.2.4 IA, Ecology & Conservation

In this final section, we address high-level perspectives on the interdisciplinary field of our studies. They apply to our work as well, but their scope is broader than species distribution modelling based on deep learning.

First, we have seen throughout this manuscript how taxonomic, spatial coverage, and extinction risk assessments, for example, are still hampered by knowledge gaps. In other words, reference labels to supervise learning algorithms are lacking. The machine learning community faces this problem in many fields, and compensatory strategies are grouped under the terms weak supervision, i.e. supervision with noisy or coarse-grained labels, and semi-supervision: only a subset of the training data has labels (Zhou, 2018). In ecology, where standardised data can be expensive to collect (Todman et al., 2023), and meta-databases such as the GBIF are still noisy and biased towards certain regions, taxa and facets of biodiversity, the application of these techniques seems appropriate and necessary. Self-supervised learning is another learning paradigm where the training goal is to learn, for example, to detect input modifications (e.g. degree of image rotation) or even to reconstruct masked parts of the input (e.g. masked image subset or words in a sentence). It allows learning the structure of huge amounts of unlabelled data in a first step, and then possibly fine-tuning models with few labels in downstream tasks. It has great potential for remote sensing where labels are scarce but data is abundant (Scheibenreif et al., 2022). These different learning paradigms are not exclusive and can be combined to make the most of the available data. Finally, unsupervised learning approaches are already used in ecology (Derkarabetian et al., 2019; Sonnewald et al., 2020) and still have a lot to offer for data exploration, representation, dimension reduction, clustering and anomaly detection (see Pichler and Hartig, 2023 Box 2).

Another promising avenue in ecological modelling is hybrid AI. Physics-informed neural networks (PINN, Raissi et al., 2018) have recently taken off by constraining model learning with proven physical laws. Similarly, ecology-informed models could benefit from field knowledge (Pichler & Hartig, 2023). This would also increase confidence in model predictions. Furthermore, we believe that the flexibility and transparency of constraint programming has potential for ecology in general and for implementing effective conservation measures in particular (Justeau-Allaire et al., 2021).

As a closing remark, there is a strong case for developing deep learning models directly with ecological data. Although it may seem naive, there is sufficient data volume, complexity and diversity to challenge DL research with ecological data. It would also lead to the development of specific architectures that avoid the need to adapt or retrain models for ecological tasks. For example, computer vision models could be benchmarked against a plant identification dataset such as Pl@ntNet-300K instead of ImageNet (Deng et al., 2009; Garcin et al., 2021).

Bringing together data science and conservation science is an ambitious task, not only from a practical point of view, but also on a social plan. The research communities are animated by challenges that are difficult to coordinate and link, and that span a wide spectrum from theory to concrete implementation. Moreover, it implies in fact bringing together not two but three fields: data science, biodiversity modelling and biodiversity conservation, as the latter two are still largely independent (Guisan et al., 2013; Pollock

et al., 2020). However, many additional benefits would result from such a union. As noted by Pollock et al. (2020), modern modelling techniques would benefit conservation, but precise prediction targets for conservation would also catalyse modelling research. The same observation applies between data science and biodiversity modelling: while AI capabilities are often praised, precise modelling targets, case studies and ecological theory provide a meaningful challenge to the data science community.
Finally, I would like to end this general perspective with a quote from Jetz et al. (2012) that resonates with my position. It is taken from a paragraph in which the authors outline the challenges and opportunities of their *Map of Life* project to integrate knowledge of species distributions:

> *"Although challenges remain in developing statistically robust models to integrate heterogeneous distribution data types, they are unlikely to represent a permanent obstacle. All computational tools to implement the envisioned cyberinfrastructure already exist. Thus, the greatest challenge for fully realizing our vision might be more sociological than technological."*
>
> (Jetz et al., 2012)

Considering how deep learning has taken off since then, with the breakthrough of deep convolutional neural networks in the same year (Krizhevsky et al., 2012, see timeline Figure 2.11), the plurality of technological advances, and today's global concerns, I find it a particularly perceptive thought.

6.3 Conclusion

Embracing deep learning and satellite imagery to model species distribution opens up new possibilities for informing conservation. Yet this interdisciplinary journey is still in its infancy. Data science offers the opportunity to extract critical information from massive, heterogeneous data describing the state of nature and provide flexible decision support. However, significant efforts are needed to improve model interpretability and quantify uncertainties if we are to properly guide conservation actions with deep learning results. Nevertheless, we believe that the results generated in this project have value to motivate further research, which is urgently needed. In addition, we believe that challenging and complementing the results presented here with alternative strategies would provide additional benefits.

Before closing this manuscript, I would like to mention a few challenging aspects that I encountered during my research. First, publishing research that straddles the boundaries of data science, biodiversity and conservation is likely to require increased attention. It demands effective communication of the significance of the findings to a diverse audience. Second, despite increasing resources, the processing and analysis of global remote sensing datasets still presents computational and logistical hurdles. Finding open, free solutions to facilitate the efficient handling of such promising and public data is imperative. Finally, the limited time frame of PhD programmes can sometimes limit the scope of research. Balancing the need for comprehensive investigation with the constraints of the doctoral

format can be difficult. I also believe this is exacerbated for exciting, exploratory projects at the growing intersection of fields such as data science and biodiversity conservation.

The interdisciplinary nature of our work provides a unique vantage point for contributing to pressing conservation issues. By collaborating across disciplines, the potential of species distribution models and data science can be harnessed to drive positive change in conservation.

Bibliography

Adhikari, Yagya Prasad, Anton Fischer, & Hagen Siegfried Fischer (2012). "Micro-site conditions of epiphytic orchids in a human impact gradient in Kathmandu valley, Nepal". In: *Journal of mountain Science* 9.3, pp. 331–342 (cit. on p. 73).

Affouard, Antoine, Hervé Goëau, Pierre Bonnet, Jean-Christophe Lombardo, & Alexis Joly (2017). "Pl@ntnet app in the era of deep learning". In: *ICLR: International Conference on Learning Representations* (cit. on pp. 38, 61, 97).

Akcakaya, H. Resit, Scott Ferson, Mark A. Burgman, David A. Keith, Georgina M. Mace, & Charles R. Todd (Aug. 15, 2000). "Making Consistent IUCN Classifications under Uncertainty". In: *Conservation Biology* 14.4, pp. 1001–1013. ISSN: 0888-8892, 1523-1739. DOI: 10.1046/j.1523-1739.2000.99125.x (cit. on p. 19).

Akçakaya, H. Resit, Stuart H. M. Butchart, Georgina M. Mace, Simon N. Stuart, & Craig Hilton-Taylor (Nov. 2006). "Use and misuse of the IUCN Red List Criteria in projecting climate change impacts on biodiversity: IUCN RED LIST AND CLIMATE CHANGE IMPACTS". In: *Global Change Biology* 12.11, pp. 2037–2043. ISSN: 13541013. DOI: 10.1111/j.1365-2486.2006.01253.x (cit. on p. 129).

Allan, Blake M., Dale G. Nimmo, Daniel Ierodiaconou, Jeremy VanDerWal, Lian Pin Koh, & Euan G. Ritchie (2018). "Futurecasting ecological research: the rise of technoecology". In: *Ecosphere* 9.5, e02163. ISSN: 2150-8925. DOI: 10.1002/ecs2.2163 (cit. on p. 3).

Allouche, Omri, Asaf Tsoar, & Ronen Kadmon (2006). "Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS)". In: *Journal of Applied Ecology* 43.6, pp. 1223–1232. ISSN: 1365-2664. DOI: https://doi.org/10.1111/j.1365-2664.2006.01214.x (cit. on p. 44).

Almpanidou, Vasiliki, Aggeliki Doxa, & Antonios D. Mazaris (Nov. 1, 2021). "Combining a cumulative risk index and species distribution data to identify priority areas for marine biodiversity conservation in the Black Sea". In: *Ocean & Coastal Management* 213, p. 105877. ISSN: 0964-5691. DOI: 10.1016/j.ocecoaman.2021.105877 (cit. on p. 122).

Amari, Shun-ichi (1993). "Backpropagation and stochastic gradient descent method". In: *Neurocomputing* 5.4-5, pp. 185–196 (cit. on p. 55).

Andermann, Tobias, Alexandre Antonelli, Russell L Barrett, & Daniele Silvestro (2022). "Estimating alpha, beta, and gamma diversity through deep learning". In: *Frontiers in plant science* 13, p. 839407 (cit. on p. 25).

Antala, Michal, Radoslaw Juszczak, Christiaan van der Tol, & Anshu Rastogi (June 25, 2022). "Impact of climate change-induced alterations in peatland vegetation phenology and composition on carbon balance". In: *Science of The Total Environment* 827, p. 154294. ISSN: 0048-9697. DOI: 10.1016/j.scitotenv.2022.154294 (cit. on p. 64).

Araújo, Miguel B., Diogo Alagador, Mar Cabeza, David Nogués-Bravo, & Wilfried Thuiller (2011). "Climate change threatens European conservation areas". In: *Ecology letters* 14.5, pp. 484–492 (cit. on p. 66).

Araújo, Miguel B., Mar Cabeza, Wilfried Thuiller, Lee Hannah, & Paul H. Williams (2004). "Would climate change drive species out of reserves? An assessment of existing reserve-

selection methods". In: *Global Change Biology* 10.9, pp. 1618–1626. ISSN: 1365-2486. DOI: 10.1111/j.1365-2486.2004.00828.x (cit. on p. 66).

Araújo, Miguel B. & Antoine Guisan (2006). "Five (or so) challenges for species distribution modelling". In: *Journal of biogeography* 33.10, pp. 1677–1688 (cit. on pp. 41, 42).

Araújo, Miguel B. & Richard G. Pearson (Oct. 2005). "Equilibrium of species' distributions with climate". In: *Ecography* 28.5, pp. 693–695. ISSN: 09067590. DOI: 10.1111/j.2005.0906-7590.04253.x (cit. on pp. 41, 128).

Araújo, Miguel B. et al. (Jan. 1, 2019). "Standards for distribution models in biodiversity assessments". In: *Science Advances* 5.1, eaat4858. ISSN: 2375-2548. DOI: 10.1126/sciadv.aat4858 (cit. on p. 71).

Arenas-Castro, Salvador, João Gonçalves, Paulo Alves, Domingo Alcaraz-Segura, & João P. Honrado (June 18, 2018). "Assessing the multi-scale predictive ability of ecosystem functional attributes for species distribution modelling". In: *PLOS ONE* 13.6, e0199292. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0199292 (cit. on pp. 98, 142).

Arrieta, Alejandro Barredo, Natalia Diaz-Rodriguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. (2020). "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information fusion* 58, pp. 82–115 (cit. on p. 52).

Atwood, John T (1986). "The size of the Orchidaceae and the systematic distribution of epiphytic orchids". In: *Selbyana*, pp. 171–186 (cit. on p. 149).

Austin, M. P (Nov. 30, 2002). "Spatial prediction of species distribution: an interface between ecological theory and statistical modelling". In: *Ecological Modelling* 157.2, pp. 101–118. ISSN: 0304-3800. DOI: 10.1016/S0304-3800(02)00205-3 (cit. on pp. 44, 128).

Bachman, Steven P., Richard Field, Tom Reader, Domitilla Raimondo, John Donaldson, George E. Schatz, & Eimear Nic Lughadha (June 2019). "Progress, challenges and opportunities for Red Listing". In: *Biological Conservation* 234, pp. 45–55. ISSN: 00063207. DOI: 10.1016/j.biocon.2019.03.002 (cit. on pp. 10, 11, 16–19, 140).

Bachman, Steven P., Justin Moat, Andrew Hill, Javier de la Torre, & Ben Scott (Nov. 28, 2011). "Supporting Red List threat assessments with GeoCAT: geospatial conservation assessment tool". In: *ZooKeys* 150, pp. 117–126. ISSN: 1313-2970. DOI: 10.3897/zookeys.150.2109 (cit. on p. 47).

Bachman, Steven P., Eimear M Nic Lughadha, & Malin C Rivers (2018). "Quantifying progress toward a conservation assessment for all plants". In: *Conservation Biology* 32.3, pp. 516–524 (cit. on p. 17).

Bachman, Steven P., Barnaby Eliot Walker, Sara Barrios, Alison Copeland, & Justin Moat (Jan. 23, 2020). "Rapid Least Concern: towards automating Red List assessments". In: *Biodiversity Data Journal* 8. ISSN: 1314-2828. DOI: 10.3897/BDJ.8.e47018 (cit. on pp. 47, 127).

Bailey, Larissa L, Darryl I MacKenzie, & James D Nichols (2014). "Advances and applications of occupancy models". In: *Methods in Ecology and Evolution* 5.12, pp. 1269–1279 (cit. on p. 42).

Bakkenes, M., J. R. M. Alkemade, F. Ihle, R. Leemans, & J. B. Latour (2002). "Assessing effects of forecasted climate change on the diversity and distribution of European higher plants for 2050". In: *Global Change Biology* 8.4, pp. 390–407. ISSN: 1365-2486. DOI: 10.1046/j.1354-1013.2001.00467.x (cit. on p. 132).

Ball, Ian R, Hugh P Possingham, & Matthew Watts (2009). "Marxan and relatives: software for spatial conservation prioritisation". In: *Spatial conservation prioritisation: Quantitative methods and computational tools* 14, pp. 185–196 (cit. on p. 66).

Balmford, Andrew, Pete Carey, Valerie Kapos, Andrea Manica, Ana S. L. Rodrigues, Jörn P. W. Scharlemann, & Rhys E. Green (2009). "Capturing the Many Dimensions of Threat: Comment on Salafsky et al." In: *Conservation Biology* 23.2, pp. 482–487. ISSN: 0888-8892 (cit. on pp. 8, 28).

Balmford, Andrew, Rhys E Green, & Martin Jenkins (2003). "Measuring the changing state of nature". In: *Trends in Ecology & Evolution* 18.7, pp. 326–330 (cit. on p. 20).

Bannari, Abdou, Daniel Morin, F Bonn, & AjRsr Huete (1995). "A review of vegetation indices". In: *Remote sensing reviews* 13.1-2, pp. 95–120 (cit. on p. 93).

Barbault, R & M Loreau (2005). *Actes de la conférence internationale Biodiversité, sciences et gouvernance* (cit. on p. 24).

Barnosky, Anthony D. et al. (Mar. 2011). "Has the Earth's sixth mass extinction already arrived?" In: *Nature* 471.7336, pp. 51–57. ISSN: 1476-4687. DOI: 10.1038/nature09678 (cit. on p. 2).

Barry, Simon & Jane Elith (2006). "Error and uncertainty in habitat models". In: *Journal of Applied Ecology* 43.3, pp. 413–423 (cit. on p. 42).

Batjes, Niels H, Eloi Ribeiro, & Ad Van Oostrum (2020). "Standardised soil profile data to support global mapping and modelling (WoSIS snapshot 2019)". In: *Earth System Science Data* 12.1, pp. 299–320 (cit. on p. 98).

Beale, Colin M & Jack J Lennon (2012). "Incorporating uncertainty in predictive species distribution modelling". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1586, pp. 247–258 (cit. on p. 42).

Beck, Jan, Marianne Böller, Andreas Erhardt, & Wolfgang Schwanghart (Jan. 2014). "Spatial bias in the GBIF database and its effect on modeling species' geographic distributions". In: *Ecological Informatics* 19, pp. 10–15. ISSN: 15749541. DOI: 10.1016/j.ecoinf.2013.11.002 (cit. on p. 97).

Beery, Sara, Elijah Cole, Joseph Parker, Pietro Perona, & Kevin Winner (June 28, 2021). "Species Distribution Modeling for Machine Learning Practitioners: A Review". In: *ACM SIGCAS Conference on Computing and Sustainable Societies*. COMPASS '21. New York, NY, USA: Association for Computing Machinery, pp. 329–348. ISBN: 978-1-4503-8453-7. DOI: 10.1145/3460112.3471966 (cit. on pp. 36, 38, 40, 42, 44, 45).

Belbin, Lee, Elycia Wallis, Donald Hobern, & Andre Zerger (2021). "The Atlas of Living Australia: History, current state and future directions". In: *Biodiversity Data Journal* 9 (cit. on p. 34).

Bellard, Céline, Cleo Bertelsmeier, Paul Leadley, Wilfried Thuiller, & Franck Courchamp (2012). "Impacts of climate change on the future of biodiversity". In: *Ecology Letters* 15.4, pp. 365–377. ISSN: 1461-0248. DOI: 10.1111/j.1461-0248.2011.01736.x (cit. on pp. 64, 65).

Benedetti, Paola, Dino Ienco, Raffaele Gaetano, Kenji Ose, Ruggero G. Pensa, & Stephane Dupuy (Dec. 2018). "$M^3\textFusion$: A Deep Learning Architecture for Multiscale Multimodal Multitemporal Satellite Data Fusion". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 11.12, pp. 4939–4949. ISSN: 2151-1535. DOI: 10.1109/JSTARS.2018.2876357 (cit. on p. 129).

Bennun, Leon et al. (2018). "The Value of the IUCN Red List for Business Decision-Making". In: *Conservation Letters* 11.1, e12353. ISSN: 1755-263X. DOI: 10.1111/conl.12353 (cit. on p. 16).

Benoit, Laelia, Isaiah Thomas, & Andrés Martin (2022). "Review: Ecological awareness, anxiety, and actions among youth and their parents – a qualitative study of newspaper narratives". In: *Child and Adolescent Mental Health* 27.1, pp. 47–58. ISSN: 1475-3588. DOI: 10.1111/camh.12514 (cit. on p. 2).

Berger, Michael, Jose Moreno, Johnny A. Johannessen, Pieternel F. Levelt, & Ramon F. Hanssen (May 2012). "ESA's sentinel missions in support of Earth system science". In: *Remote Sensing*

of Environment 120, pp. 84–90. ISSN: 00344257. DOI: 10.1016/j.rse.2011.07.023 (cit. on p. 73).

Berger-Wolf, Tanya Y., Daniel I. Rubenstein, Charles V. Stewart, Jason A. Holmberg, Jason Parham, Sreejith Menon, Jonathan Crall, Jon Van Oast, Emre Kiciman, & Lucas Joppa (Oct. 24, 2017). "Wildbook: Crowdsourcing, computer vision, and data science for conservation". In: *arXiv:1710.08880 [cs]* (cit. on p. 61).

Betts, Jessica, Richard P. Young, Craig Hilton-Taylor, Michael Hoffmann, Jon Paul Rodríguez, Simon N. Stuart, & E.J. Milner-Gulland (June 2020). "A framework for evaluating the impact of the IUCN Red List of threatened species". In: *Conservation Biology* 34.3, pp. 632–643. ISSN: 0888-8892, 1523-1739. DOI: 10.1111/cobi.13454 (cit. on p. 16).

Birdal, Tolga, Aaron Lou, Leonidas J Guibas, & Umut Simsekli (2021). "Intrinsic dimension, persistent homology and generalization in neural networks". In: *Advances in Neural Information Processing Systems* 34, pp. 6776–6789 (cit. on p. 50).

Bisong, Ekaba (2019). "Google colaboratory". In: *Building machine learning and deep learning models on google cloud platform: a comprehensive guide for beginners*, pp. 59–64 (cit. on p. 57).

Bland, Lucie M., Ben Collen, C. David L. Orme, & Jon Bielby (2015). «Predicting the conservation status of data-deficient species». In: *Conservation Biology* 29.1, pp. 250–259. ISSN: 1523-1739. DOI: https://doi.org/10.1111/cobi.12372 (cit. on p. 63).

Bland, Lucie M., DA Keith, RM Miller, NJ Murray, & JP Rodriguez (2017). "Guidelines for the application of IUCN Red List of Ecosystems Categories and Criteria, version 1.1". In: *International Union for the Conservation of Nature, Gland, Switzerland* (cit. on pp. 47, 127, 141).

Boakes, Elizabeth H, Philip JK McGowan, Richard A Fuller, Ding Chang-qing, Natalie E Clark, Kim O'Connor, & Georgina M Mace (2010). "Distorted views of biodiversity: spatial and temporal bias in species occurrence data". In: *PLoS biology* 8.6, e1000385 (cit. on p. 39).

BON, GEO (2015). "Global Biodiversity Change Indicators. Version 1.2". In: *Leipzig: Group on Earth Observations Biodiversity Observation Network Secretariat* (cit. on p. 28).

Bonnet, Pierre et al. (Nov. 2020). "How citizen scientists contribute to monitor protected areas thanks to automatic plant identification tools". In: *Ecological Solutions and Evidence* 1.2. ISSN: 2688-8319, 2688-8319. DOI: 10.1002/2688-8319.12023 (cit. on p. 66).

Borgelt, Jan, Martin Dorber, Marthe Alnes Høiberg, & Francesca Verones (Aug. 4, 2022a). "More than half of data deficient species predicted to be threatened by extinction". In: *Communications Biology* 5.1, pp. 1–9. ISSN: 2399-3642. DOI: 10.1038/s42003-022-03638-9 (cit. on pp. 63, 103, 104, 127, 142, 147).

Borgelt, Jan, Jorge Sicacha-Parada, Olav Skarpaas, & Francesca Verones (2022b). "Native range estimates for red-listed vascular plants". In: *Scientific Data* 9.1, p. 117 (cit. on p. 29).

Borowiec, Marek L, Rebecca B Dikow, Paul B Frandsen, Alexander McKeeken, Gabriele Valentini, & Alexander E White (2022). "Deep learning as a tool for ecology and evolution". In: *Methods in Ecology and Evolution* 13.8, pp. 1640–1660 (cit. on pp. 54, 59, 93, 104, 129).

Botella, Christophe (2019). "Méthodes statistiques pour la modélisation de la distribution spatiale des espèces végétales à partir de grandes masses d'observations incertaines issues de programmes de sciences citoyennes". PhD thesis (cit. on pp. 39, 40, 42, 43, 47, 50).

Botella, Christophe, Pierre Bonnet, Cang Hui, Alexis Joly, & David M Richardson (2022). "Dynamic species distribution modeling reveals the pivotal role of human-mediated long-distance dispersal in plant invasion". In: *Biology* 11.9, p. 1293 (cit. on p. 48).

Botella, Christophe, Benjamin Deneu, Diego Marcos, Maximilien Servajean, Joaquim Estopinan, Théo Larcher, César Leblanc, Pierre Bonnet, & Alexis Joly (2023). "The GeoLifeCLEF

2023 Dataset to evaluate plant species distribution models at high spatial resolution across Europe". In: *arXiv preprint arXiv:2308.05121* (cit. on pp. vii, 37, 42, 56).

Botella, Christophe, Alexis Joly, Pierre Bonnet, Pascal Monestiez, & François Munoz (2018a). "A deep learning approach to species distribution modelling". In: *Multimedia Tools and Applications for Environmental & Biodiversity Informatics*, pp. 169–199 (cit. on pp. 4, 53, 57, 73, 107, 129, 191).

— (Feb. 2018b). "Species distribution modeling based on the automated identification of citizen observations". In: *Applications in Plant Sciences* 6.2, e1029. ISSN: 21680450. DOI: 10.1002/aps3.1029 (cit. on p. 105).

Botella, Christophe, Alexis Joly, Pierre Bonnet, François Munoz, & Pascal Monestiez (2021). "Jointly estimating spatial sampling effort and habitat suitability for multiple species from opportunistic presence-only data". In: *Methods in Ecology and Evolution* 12.5, pp. 933–945 (cit. on pp. 42, 105).

Botella, Christophe, Alexis Joly, Pascal Monestiez, Pierre Bonnet, & François Munoz (2020). "Bias in presence-only niche models related to sampling effort and species niches: Lessons for background point selection". In: *PLoS One* 15.5, e0232078 (cit. on pp. 42, 82).

Botella, Christophe, Maximilien Servajean, Pierre Bonnet, & Alexis Joly (2019). "Overview of GeoLifeCLEF 2019: plant species prediction using environment and animal occurrences". In: *Working Notes of CLEF 2019-Conference and Labs of the Evaluation Forum*. 2380 (cit. on p. 82).

Bourhis, Yoann, James R. Bell, Chris R. Shortall, William E. Kunin, & Alice E. Milne (2023). "Explainable neural networks for trait-based multispecies distribution modelling—A case study with butterflies and moths". In: *Methods in Ecology and Evolution* n/a (n/a). ISSN: 2041-210X. DOI: 10.1111/2041-210X.14097 (cit. on pp. 58, 120, 142, 149).

Breiman, Leo (Oct. 1, 2001a). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324 (cit. on pp. 43, 52).

— (Aug. 2001b). "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)". In: *Statistical Science* 16.3, pp. 199–231. ISSN: 0883-4237, 2168-8745. DOI: 10.1214/ss/1009213726 (cit. on p. 48).

Breiner, Frank T. (2016). "Revision of the Red List of fern and vascular plants - evaluation and application of statistical methods". PhD thesis (cit. on p. 20).

Breiner, Frank T. & Ariel Bergamini (July 1, 2018). "Improving the estimation of area of occupancy for IUCN Red List assessments by using a circular buffer approach". In: *Biodiversity and Conservation* 27.9, pp. 2443–2448. ISSN: 1572-9710. DOI: 10.1007/s10531-018-1555-5 (cit. on p. 19).

Breiner, Frank T., Antoine Guisan, Michael P. Nobis, & Ariel Bergamini (2017). "Including environmental niche information to improve IUCN Red List assessments". In: *Diversity and Distributions* 23.5, pp. 484–495. ISSN: 1472-4642. DOI: 10.1111/ddi.12545 (cit. on pp. 19, 47, 104, 127, 130, 147).

Brodrick, Philip G., Andrew B. Davies, & Gregory P. Asner (Aug. 2019). "Uncovering Ecological Patterns with Convolutional Neural Networks". In: *Trends in Ecology & Evolution* 34.8, pp. 734–745. ISSN: 01695347. DOI: 10.1016/j.tree.2019.03.006 (cit. on pp. 53, 73).

Brook, Barry W., H. Resit Akçakaya, David A. Keith, Georgina M. Mace, Richard G. Pearson, & Miguel B. Araújo (July 22, 2009). "Integrating bioclimate with population models to improve forecasts of species extinctions under climate change". In: *Biology Letters* 5.6, pp. 723–725. DOI: 10.1098/rsbl.2009.0480 (cit. on p. 142).

Brooks, Thomas M. et al. (Nov. 2019). "Measuring Terrestrial Area of Habitat (AOH) and Its Utility for the IUCN Red List". In: *Trends in Ecology & Evolution* 34.11, pp. 977–986. ISSN: 01695347. DOI: 10.1016/j.tree.2019.06.009 (cit. on pp. 19, 127).

Brown, James, George C. Stevens, & Dawn M Kaufman (1996). "The geographic range: size, shape, boundaries, and internal structure". In: *Annual review of ecology and systematics* 27.1, pp. 597–623 (cit. on p. 36).

Bruelheide, Helge et al. (2019). "sPlot – A new tool for global vegetation analyses". In: *Journal of Vegetation Science* 30.2, pp. 161–186. ISSN: 1654-1103. DOI: 10.1111/jvs.12710 (cit. on p. 38).

Brummitt, Neil A. et al. (Aug. 7, 2015). "Green Plants in the Red: A Baseline Global Assessment for the IUCN Sampled Red List Index for Plants". In: *PLOS ONE* 10.8. Ed. by Clinton N Jenkins, e0135152. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0135152 (cit. on pp. 20, 140).

Brummitt, Richard Kenneth, Francisco Pando, S Hollis, & NA Brummitt (2001). *World geographical scheme for recording plant distributions*. International working group on taxonomic databases for plant sciences (TDWG) (cit. on pp. 82, 107, 110, 191).

Brun, Philipp, Dirk N. Karger, Damaris Zurell, Patrice Descombes, Lucienne C. de Witte, Riccardo de Lutio, Jan Dirk Wegner, & Niklaus E. Zimmermann (May 31, 2023). *Rank-based deep learning from citizen-science data to model plant communities*. DOI: 10.1101/2023.05.30.542843 (cit. on pp. 58, 61).

Buckland, Stephen T, David R Anderson, Kenneth P Burnham, & Jeffrey L Laake (2005). "Distance sampling". In: *Encyclopedia of biostatistics* 2 (cit. on p. 38).

Buckland, Stephen T, David L Borchers, A Johnston, Peter A Henrys, & Tiago A Marques (2007). "Line transect methods for plant surveys". In: *Biometrics* 63.4, pp. 989–998 (cit. on p. 38).

Burkhard, Benjamin & Joachim Maes (2017). "Mapping ecosystem services". In: *Advanced books* 1, e12837 (cit. on p. 15).

Butchart, Stuart H. M., H Resit Akçakaya, Janice Chanson, Jonathan EM Baillie, Ben Collen, Suhel Quader, Will R Turner, Rajan Amin, Simon N Stuart, & Craig Hilton-Taylor (2007). "Improvements to the red list index". In: *PloS one* 2.1, e140 (cit. on p. 20).

Butchart, Stuart H. M., Alison J Stattersfield, Leon A Bennun, Sue M Shutes, H Resit Akçakaya, Jonathan E M Baillie, Simon N Stuart, Craig Hilton-Taylor, & Georgina M Mace (2004). "Measuring global trends in the status of biodiversity: Red List Indices for birds". In: *PLoS biology* 2.12, e383 (cit. on p. 20).

Butchart, Stuart H. M. et al. (May 28, 2010). "Global Biodiversity: Indicators of Recent Declines". In: *Science* 328.5982, pp. 1164–1168. DOI: 10.1126/science.1187512 (cit. on p. 20).

Caetano, Gabriel Henrique de Oliveira, David G. Chapple, Richard Grenyer, Tal Raz, Jonathan Rosenblatt, Reid Tingley, Monika Böhm, Shai Meiri, & Uri Roll (May 26, 2022). "Automated assessment reveals that the extinction risk of reptiles is widely underestimated across space and phylogeny". In: *PLOS Biology* 20.5, e3001544. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.3001544 (cit. on p. 62).

Cai, Lirong, Holger Kreft, Amanda Taylor, Pierre Denelle, Julian Schrader, Franz Essl, Mark van Kleunen, Jan Pergl, Petr Pyšek, Anke Stein, et al. (2023). "Global models and predictions of plant diversity based on advanced machine learning techniques". In: *New Phytologist* 237.4, pp. 1432–1445 (cit. on pp. 104, 200).

Camacho, Francisco & Gwendolyn Peyre (Jan. 1, 2022). "Red List and Vulnerability Assessment of the Páramo Vascular Flora in the Nevados Natural National Park (Colombia)". In: *Tropical Conservation Science* 15, p. 19400829221086958. ISSN: 1940-0829. DOI: 10.1177/19400829221086958 (cit. on p. 47).

Camps-Valls, Gustau, Devis Tuia, Xiao Xiang Zhu, & Markus Reichstein (Aug. 16, 2021). *Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate*

Science and Geosciences. John Wiley & Sons. 436 pp. ISBN: 978-1-119-64614-3 (cit. on pp. 61, 73).

Cao, Kaidi, Colin Wei, Adrien Gaidon, Nikos Arechiga, & Tengyu Ma (Oct. 27, 2019). "Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss". In: *arXiv:1906.07413 [cs, stat]* (cit. on pp. 81, 88, 129, 131, 148, 191).

Carpenter, Kent E. et al. (July 25, 2008). "One-Third of Reef-Building Corals Face Elevated Extinction Risk from Climate Change and Local Impacts". In: *Science* 321.5888, pp. 560–563. DOI: 10.1126/science.1159196 (cit. on p. 127).

Carranza-Rojas, Jose, Herve Goeau, Pierre Bonnet, Erick Mata-Montero, & Alexis Joly (2017). "Going deeper in the automated identification of Herbarium specimens". In: *BMC evolutionary biology* 17.1, pp. 1–14 (cit. on p. 61).

Cavender-Bares, Jeannine et al. (Mar. 24, 2022). "Integrating remote sensing with ecology and evolution to advance biodiversity conservation". In: *Nature Ecology & Evolution*, pp. 1–14. ISSN: 2397-334X. DOI: 10.1038/s41559-022-01702-5 (cit. on p. 4).

Cayuela, L, DJ Golicher, AC Newton, M Kolb, FS De Alburquerque, EJMM Arets, JRM Alkemade, & AM Pérez (2009). "Species distribution modeling in the tropics: problems, potentialities, and the role of biological data for effective species conservation". In: *Tropical Conservation Science* 2.3, pp. 319–352 (cit. on p. 39).

Cazalis, Victor et al. (Jan. 19, 2022). "Bridging the research-implementation gap in IUCN Red List assessments". In: *Trends in Ecology & Evolution*. ISSN: 0169-5347. DOI: 10.1016/j.tree.2021.12.002 (cit. on pp. 62–64).

Cerrejón, Carlos, Osvaldo Valeria, Philippe Marchand, Richard T. Caners, & Nicole J. Fenton (2021). "No place to hide: Rare plant detection through remote sensing". In: *Diversity and Distributions* n/a (n/a). ISSN: 1472-4642. DOI: https://doi.org/10.1111/ddi.13244 (cit. on p. 94).

Challender, Daniel W. S. et al. (July 6, 2023). "Identifying species likely threatened by international trade on the IUCN Red List can inform CITES trade measures". In: *Nature Ecology & Evolution*. ISSN: 2397-334X. DOI: 10.1038/s41559-023-02115-8 (cit. on p. 16).

Chase, Mark W., Kenneth M. Cameron, John V. Freudenstein, Alec M. Pridgeon, Gerardo Salazar, Cássio van den Berg, & André Schuiteman (Feb. 1, 2015). "An updated classification of Orchidaceae". In: *Botanical Journal of the Linnean Society* 177.2, pp. 151–174. ISSN: 0024-4074. DOI: 10.1111/boj.12234 (cit. on p. 74).

Chauvenet, ALM, JG Ewen, DP Armstrong, TM Blackburn, & N Pettorelli (2013). "Maximizing the success of assisted colonizations". In: *Animal Conservation* 16.2, pp. 161–169 (cit. on p. 48).

Chauvier, Yohann, Wilfried Thuiller, Philipp Brun, Sébastien Lavergne, Patrice Descombes, Dirk N. Karger, Julien Renaud, & Niklaus E. Zimmermann (2021). "Influence of climate, soil, and land cover on plant species distribution in the European Alps". In: *Ecological Monographs* 91.2, e01433. ISSN: 1557-7015. DOI: 10.1002/ecm.1433 (cit. on p. 46).

Chen, Di, Yexiang Xue, Shuo Chen, Daniel Fink, & Carla Gomes (Feb. 21, 2017). *Deep Multi-Species Embedding*. DOI: 10.48550/arXiv.1609.09353 (cit. on p. 57).

Chen, Jinnan, Chengzhi Ding, Dekui He, Liuyong Ding, Songhao Ji, Tingqi Du, Jingrui Sun, Minrui Huang, & Juan Tao (July 8, 2023). "Assessing the conservation status of Chinese freshwater fish using deep learning". In: *Reviews in Fish Biology and Fisheries*. ISSN: 1573-5184. DOI: 10.1007/s11160-023-09792-5 (cit. on p. 63).

Chen, Tianqi & Carlos Guestrin (2016). "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794 (cit. on p. 62).

Child, RE & DB Pinniger (1994). "Insect trapping in museums and historic houses". In: *Studies in Conservation* 39.sup2, pp. 129–131 (cit. on p. 38).

Christin, Sylvain, Éric Hervet, & Nicolas Lecomte (2019). "Applications for deep learning in ecology". In: *Methods in Ecology and Evolution* 10.10, pp. 1632–1644. ISSN: 2041-210X. DOI: 10.1111/2041-210X.13256 (cit. on pp. 4, 54, 61).

Chytrý, Milan, Stephan M Hennekens, Borja Jiménez-Alfaro, Ilona Knollová, Jürgen Dengler, Florian Jansen, Flavia Landucci, Joop HJ Schaminée, Svetlana Aćić, Emiliano Agrillo, et al. (2016). "European Vegetation Archive (EVA): an integrated database of European vegetation plots". In: *Applied vegetation science* 19.1, pp. 173–180 (cit. on pp. 26, 38).

Chytrý, Milan et al. (2020). "EUNIS Habitat Classification: Expert system, characteristic species combinations and distribution maps of European habitats". In: *Applied Vegetation Science* 23.4, pp. 648–675. ISSN: 1654-109X. DOI: 10.1111/avsc.12519 (cit. on p. 26).

Chzhen, Evgenii, Christophe Denis, Mohamed Hebiri, & Titouan Lorieul (Feb. 24, 2021). *Set-valued classification – overview via a unified framework*. DOI: 10.48550/arXiv.2102.12318 (cit. on pp. 82, 106).

Collen, Ben, Jonathan Loh, Sarah Whitmee, LOUISE McRAE, Rajan Amin, & Jonathan EM Baillie (2009). "Monitoring change in vertebrate abundance: the Living Planet Index". In: *Conservation Biology* 23.2, pp. 317–327 (cit. on p. 22).

Collen, Ben, Mala Ram, Tara Zamin, & Louise McRae (June 2008). "The Tropical Biodiversity Data Gap: Addressing Disparity in Global Monitoring". In: *Tropical Conservation Science* 1.2, pp. 75–88. ISSN: 1940-0829, 1940-0829. DOI: 10.1177/194008290800100202 (cit. on pp. 3, 122).

Collen, Ben et al. (Apr. 30, 2016). "Clarifying misconceptions of extinction risk assessment with the IUCN Red List". In: *Biology Letters* 12.4, p. 20150843. DOI: 10.1098/rsbl.2015.0843 (cit. on p. 19).

Collins, Courtney G. et al. (June 11, 2021). "Experimental warming differentially affects vegetative and reproductive phenology of tundra plants". In: *Nature Communications* 12.1, p. 3442. ISSN: 2041-1723. DOI: 10.1038/s41467-021-23841-2 (cit. on p. 64).

Commission, European, Directorate-General for Environment, & K Sundseth (2015). *The EU birds and habitats directives : for nature and people in Europe*. Publications Office. DOI: doi/10.2779/49288 (cit. on pp. 21, 26).

— (2017). *An action plan for nature, people and the economy : the EU Habitats and Birds Directives*. Publications Office. DOI: doi/10.2779/242535 (cit. on pp. 21, 26).

Commission, IUCN Species Survival et al. (2001). "IUCN Red List Categories and Criteria: Version3. 1". In: *http://www. iucnredlist. org/documents/redlist_ cats_ crit_ en. pdf* (cit. on pp. 16, 17, 19).

Cord, Anna & Dennis Rödder (2011). "Inclusion of habitat availability in species distribution models through multi-temporal remote-sensing data?" In: *Ecological Applications* 21.8, pp. 3285–3298. ISSN: 1939-5582. DOI: 10.1890/11-0114.1 (cit. on p. 94).

Corlett, Richard T. & David A. Westcott (Aug. 1, 2013). "Will plant movements keep up with climate change?" In: *Trends in Ecology & Evolution* 28.8, pp. 482–488. ISSN: 0169-5347. DOI: 10.1016/j.tree.2013.04.003 (cit. on p. 64).

Costello, Mark J., Robert M. May, & Nigel E. Stork (Jan. 25, 2013). "Can We Name Earth's Species Before They Go Extinct?" In: *Science* 339.6118, pp. 413–416. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1230318 (cit. on p. 2).

Cozzolino, Salvatore & Alex Widmer (Sept. 1, 2005). "Orchid diversity: an evolutionary consequence of deception?" In: *Trends in Ecology & Evolution* 20.9, pp. 487–494. ISSN: 0169-5347. DOI: 10.1016/j.tree.2005.06.004 (cit. on pp. 11, 105).

Crain, Benjamin J. & Melania Fernández (2020). "Biogeographical analyses to facilitate targeted conservation of orchid diversity hotspots in Costa Rica". In: *Diversity and Distributions* 26.7, pp. 853–866. ISSN: 1472-4642. DOI: 10.1111/ddi.13062 (cit. on p. 11).

Cribb, Phillip J, Shelagh P Kell, Kingsley W Dixon, & Russell L Barrett (2003). "Orchid conservation: a global perspective". In: *Orchid conservation* 124 (cit. on pp. 11, 105, 106).

Cribb, Phillip J. & Johan Hermans (2007). "The conservation of Madagascar's orchids . A model for an integrated conservation project". In: *Lankesteriana: International Journal on Orchidology*. ISSN: 2215-2067. DOI: 10.15517/lank.v7i1-2.19514 (cit. on p. 16).

Csillik, Ovidiu, John Cherbini, Robert Johnson, Andy Lyons, & Maggi Kelly (2018). "Identification of citrus trees from unmanned aerial vehicle imagery using convolutional neural networks". In: *Drones* 2.4, p. 39 (cit. on p. 73).

Cutler, D. Richard, Thomas C. Edwards Jr., Karen H. Beard, Adele Cutler, Kyle T. Hess, Jacob Gibson, & Joshua J. Lawler (2007). "Random Forests for Classification in Ecology". In: *Ecology* 88.11, pp. 2783–2792. ISSN: 1939-9170. DOI: 10.1890/07-0539.1 (cit. on p. 43).

Darrah, Sarah E., Lucie M. Bland, Steven P. Bachman, Colin P. Clubbe, & Anna Trias-Blasi (Apr. 2017). "Using coarse-scale species distribution data to predict extinction risk in plants". In: *Diversity and Distributions* 23.4. Ed. by Kenneth Feeley, pp. 435–447. ISSN: 13669516. DOI: 10.1111/ddi.12532 (cit. on p. 62).

Dauby, Gilles et al. (Nov. 7, 2016). "RAINBIO: a mega-database of tropical African vascular plants distributions". In: *PhytoKeys* 74, pp. 1–18. ISSN: 1314-2011. DOI: 10.3897/phytokeys.74.9723 (cit. on p. 38).

Dauby, Gilles et al. (2017). "ConR: An R package to assist large-scale multispecies preliminary conservation assessments using distribution data". In: *Ecology and Evolution* 7.24, pp. 11292–11303. ISSN: 2045-7758. DOI: https://doi.org/10.1002/ece3.3704 (cit. on pp. 47, 104).

Davies, Cynthia E, Dorian Moss, & Mark O Hill (2004). "EUNIS habitat classification revised 2004". In: *Report to: European environment agency-European topic centre on nature protection and biodiversity*, pp. 127–143 (cit. on p. 26).

Dawud, Seid Muhie, Karsten Raulund-Rasmussen, Timo Domisch, Leena Finér, Bogdan Jaroszewicz, & Lars Vesterdal (June 1, 2016). "Is Tree Species Diversity or Species Identity the More Important Driver of Soil Carbon Stocks, C/N Ratio, and pH?" In: *Ecosystems* 19.4, pp. 645–660. ISSN: 1435-0629. DOI: 10.1007/s10021-016-9958-1 (cit. on p. 84).

DeAngelis, Donald L. & Simeon Yurek (Mar. 1, 2017). "Spatially Explicit Modeling in Ecology: A Review". In: *Ecosystems* 20.2, pp. 284–300. ISSN: 1435-0629. DOI: 10.1007/s10021-016-0066-z (cit. on p. 105).

Decker, Emilia, Simon Linke, Virgilio Hermoso, & Juergen Geist (2017). "Incorporating ecological functions in conservation decision making". In: *Ecology and evolution* 7.20, pp. 8273–8281 (cit. on p. 146).

Deneu, Benjamin (Nov. 2022). "Interprétabilité des mod'eles de distribution d'esp'eces basés sur des réseaux de neurones convolutifs". Theses. Université de Montpellier (UM), FRA (cit. on pp. 44, 55).

Deneu, Benjamin, Alexis Joly, Pierre Bonnet, Maximilien Servajean, & François Munoz (Jan. 10, 2021a). "How Do Deep Convolutional SDM Trained on Satellite Images Unravel Vegetation Ecology ?" In: ICPR 2020 - International Conference on Pattern Recognition. Vol. 12666. Springer, p. 148. DOI: 10.1007/978-3-030-68780-9_15 (cit. on pp. 46, 73, 94).

— (2022). "Very High Resolution Species Distribution Modeling Based on Remote Sensing Imagery: How to Capture Fine-Grained and Large-Scale Vegetation Ecology With Convolutional Neural Networks?" In: *Frontiers in Plant Science* 13. ISSN: 1664-462X (cit. on pp. 5, 58, 97, 120, 147, 191).

Deneu, Benjamin, Maximilien Servajean, Pierre Bonnet, Christophe Botella, François Munoz, & Alexis Joly (Apr. 19, 2021b). "Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment". In: *PLOS Computational Biology* 17.4, e1008856. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1008856 (cit. on pp. 6, 45, 53, 57, 73, 80, 94, 104, 105, 107, 129, 191).

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, & Li Fei-Fei (June 2009). "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848 (cit. on pp. 50, 150).

Dengler, Jürgen, Florian Jansen, Falko Glöckler, Robert K Peet, Miquel De Cáceres, Milan Chytrỳ, Jörg Ewald, Jens Oldeland, Gabriela Lopez-Gonzalez, Manfred Finckh, et al. (2011). "The Global Index of Vegetation-Plot Databases (GIVD): a new resource for vegetation science". In: *Journal of Vegetation Science* 22.4, pp. 582–597 (cit. on p. 38).

Derkarabetian, Shahan, Stephanie Castillo, Peter K Koo, Sergey Ovchinnikov, & Marshal Hedin (2019). "A demonstration of unsupervised machine learning in species delimitation". In: *Molecular phylogenetics and evolution* 139, p. 106562 (cit. on p. 150).

Descombes, Patrice, Lorenz Walthert, Andri Baltensweiler, Reto Giulio Meuli, Dirk N. Karger, Christian Ginzler, Damaris Zurell, & Niklaus E. Zimmermann (2020). "Spatial modelling of ecological indicator values improves predictions of plant distributions in complex landscapes". In: *Ecography* 43.10, pp. 1448–1463. ISSN: 1600-0587. DOI: 10.1111/ecog.05117 (cit. on p. 45).

Desprez, Marine, Vincent Miele, & Olivier Gimenez (2023). "Nine tips for ecologists using machine learning". In: *arXiv preprint arXiv:2305.10472* (cit. on p. 54).

Di Marco, Moreno & Luca Santini (2015). "Human pressures predict species' geographic range size better than biological traits". In: *Global Change Biology* 21.6, pp. 2169–2178. ISSN: 1365-2486. DOI: https://doi.org/10.1111/gcb.12834 (cit. on pp. 46, 62).

Di Marco, Moreno, Oscar Venter, Hugh P. Possingham, & James E. M. Watson (Nov. 5, 2018). "Changes in human footprint drive changes in species extinction risk". In: *Nature Communications* 9.1, p. 4621. ISSN: 2041-1723. DOI: 10.1038/s41467-018-07049-5 (cit. on p. 140).

Di Marco, Moreno et al. (Apr. 2017). "Changing trends and persisting biases in three decades of conservation science". In: *Global Ecology and Conservation* 10, pp. 32–42. ISSN: 23519894. DOI: 10.1016/j.gecco.2017.01.008 (cit. on p. 8).

Díaz, Sandra et al. (Dec. 13, 2019). "Pervasive human-driven decline of life on Earth points to the need for transformative change". In: *Science* 366.6471, eaax3100. DOI: 10.1126/science.aax3100 (cit. on pp. 3, 103).

Díaz, Sandra et al. (Oct. 23, 2020). "Set ambitious goals for biodiversity and sustainability". In: *Science* 370.6515, pp. 411–413. DOI: 10.1126/science.abe1530 (cit. on p. 127).

Dobrowski, Solomon Z, Caitlin E Littlefield, Drew S Lyons, Clark Hollenberg, Carlos Carroll, Sean A Parks, John T Abatzoglou, Katherine Hegewisch, & Josh Gage (2021). "Protected-area targets could be undermined by climate change-driven shifts in ecoregions and biomes". In: *Communications Earth & Environment* 2.1, p. 198 (cit. on p. 66).

Domisch, Sami, Martin Friedrichs, Thomas Hein, Florian Borgwardt, Annett Wetzig, Sonja C. Jähnig, & Simone D. Langhans (2019). "Spatially explicit species distribution models: A missed opportunity in conservation planning?" In: *Diversity and Distributions* 25.5, pp. 758–769. ISSN: 1472-4642. DOI: 10.1111/ddi.12891 (cit. on pp. 42, 105).

Döscher, Ralf et al. (Apr. 8, 2022). "The EC-Earth3 Earth system model for the Coupled Model Intercomparison Project 6". In: *Geoscientific Model Development* 15.7, pp. 2973–3020. ISSN: 1991-9603. DOI: 10.5194/gmd-15-2973-2022 (cit. on pp. 131, 135).

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. (2020a). "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (cit. on p. 54).

Dosovitskiy, Alexey et al. (Oct. 22, 2020b). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *arXiv:2010.11929 [cs]* (cit. on p. 129).

Drake, John M., Christophe Randin, & Antoine Guisan (2006). "Modelling ecological niches with support vector machines". In: *Journal of Applied Ecology* 43.3, pp. 424–432. ISSN: 1365-2664. DOI: 10.1111/j.1365-2664.2006.01141.x (cit. on p. 43).

Drusch, Matthias, Umberto Del Bello, Sébastien Carlier, Olivier Colin, Veronica Fernandez, Ferran Gascon, Bianca Hoersch, Claudia Isola, Paolo Laberinti, Philippe Martimort, et al. (2012). "Sentinel-2: ESA's optical high-resolution mission for GMES operational services". In: *Remote sensing of Environment* 120, pp. 25–36 (cit. on p. 75).

Elith, Jane, Catherine Graham, Roozbeh Valavi, Meinrad Abegg, Caroline Bruce, Simon Ferrier, Andrew Ford, Antoine Guisan, Robert J Hijmans, Falk Huettmann, et al. (2020). "Presence-only and presence-absence data for comparing species distribution modeling methods". In: *Biodiversity informatics* 15.2, pp. 69–80 (cit. on p. 57).

Elith, Jane & John R. Leathwick (Dec. 1, 2009). "Species Distribution Models: Ecological Explanation and Prediction Across Space and Time". In: *Annual Review of Ecology, Evolution, and Systematics* 40.1, pp. 677–697. ISSN: 1543-592X, 1545-2069. DOI: 10.1146/annurev.ecolsys.110308.120159 (cit. on pp. 40–42, 104, 128).

Elith, Jane, John R. Leathwick, & T. Hastie (2008). "A working guide to boosted regression trees". In: *Journal of Animal Ecology* 77.4, pp. 802–813. ISSN: 1365-2656. DOI: 10.1111/j.1365-2656.2008.01390.x (cit. on p. 43).

Elith, Jane, Steven J. Phillips, Trevor Hastie, Miroslav Dudík, Yung En Chee, & Colin J. Yates (Jan. 2011). "A statistical explanation of MaxEnt for ecologists: Statistical explanation of MaxEnt". In: *Diversity and Distributions* 17.1, pp. 43–57. ISSN: 13669516. DOI: 10.1111/j.1472-4642.2010.00725.x (cit. on pp. 43, 129).

Emerson, Brent C. & Niclas Kolm (Apr. 2005). "Species diversity can drive speciation". In: *Nature* 434.7036, pp. 1015–1017. ISSN: 1476-4687. DOI: 10.1038/nature03450 (cit. on p. 84).

Esri (2023). *World Countries*. https://hub.arcgis.com/datasets/esri::world-countries/about (cit. on p. 110).

Estopinan, Joaquim, Maximilien Servajean, Pierre Bonnet, François Munoz, & Alexis Joly (2022). "Deep Species Distribution Modeling From Sentinel-2 Image Time-Series: A Global Scale Analysis on the Orchid Family". In: *Frontiers in Plant Science* 13. ISSN: 1664-462X (cit. on pp. vii, 70, 105, 129).

Evans, Megan C., James E. M. Watson, Richard A. Fuller, Oscar Venter, Simon C. Bennett, Peter R. Marsack, & Hugh P. Possingham (Apr. 2011). "The Spatial Distribution of Threats to Species in Australia". In: *BioScience* 61.4, pp. 281–289. ISSN: 1525-3244, 0006-3568. DOI: 10.1525/bio.2011.61.4.8 (cit. on pp. 28, 31, 32).

Farahani, Abolfazl, Sahar Voghoei, Khaled Rasheed, & Hamid R Arabnia (2021). "A brief review of domain adaptation". In: *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, pp. 877–894 (cit. on p. 50).

Farley, Scott S, Andria Dawson, Simon J Goring, & John W Williams (Aug. 1, 2018). "Situating Ecology as a Big-Data Science: Current Advances, Challenges, and Solutions". In: *BioScience* 68.8, pp. 563–576. ISSN: 0006-3568. DOI: 10.1093/biosci/biy068 (cit. on pp. 3–5).

Farwell, Laura S., Paul R. Elsen, Elena Razenkova, Anna M. Pidgeon, & Volker C. Radeloff (2020). "Habitat heterogeneity captured by 30-m resolution satellite image texture predicts

bird richness across the United States". In: *Ecological Applications* 30.8, e02157. ISSN: 1939-5582. DOI: 10.1002/eap.2157 (cit. on p. 93).

Fauth, JE, J Bernardo, M Camara, WJ Resetarits Jr, Jr Van Buskirk, & SA McCollum (1996). "Simplifying the jargon of community ecology: a conceptual approach". In: *The American Naturalist* 147.2, pp. 282–286 (cit. on p. 105).

Fay, Michael F. (June 5, 2018). "Orchid conservation: how can we meet the challenges in the twenty-first century?" In: *Botanical Studies* 59.1, p. 16. ISSN: 1999-3110. DOI: 10.1186/s40529-018-0232-z (cit. on pp. 11, 12, 106, 146).

Feeley, Kenneth J., James T. Stroud, & Timothy M. Perez (2017). "Most 'global' reviews of species' responses to climate change are not truly global". In: *Diversity and Distributions* 23.3, pp. 231–234. ISSN: 1472-4642. DOI: 10.1111/ddi.12517 (cit. on p. 65).

Feldman, Mariano J., Louis Imbeau, Philippe Marchand, Marc J. Mazerolle, Marcel Darveau, & Nicole J. Fenton (Mar. 11, 2021). "Trends and gaps in the use of citizen science derived data as input for species distribution models: A quantitative review". In: *PLOS ONE* 16.3, e0234587. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0234587 (cit. on p. 39).

Féret, Jean-Baptiste, Christina Corbane, & Samuel Alleaume (May 2015). "Detecting the Phenology and Discriminating Mediterranean Natural Habitats With Multispectral Sensors—An Analysis Based on Multiseasonal Field Spectra". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8.5, pp. 2294–2305. ISSN: 2151-1535. DOI: 10.1109/JSTARS.2015.2431320 (cit. on p. 95).

Ferrier, Simon (Mar. 1, 2002). "Mapping Spatial Pattern in Biodiversity for Regional Conservation Planning: Where to from Here?" In: *Systematic Biology* 51.2, pp. 331–363. ISSN: 1063-5157. DOI: 10.1080/10635150252899806 (cit. on p. 66).

Fick, Stephen E. & Robert J. Hijmans (2017). "WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas". In: *International Journal of Climatology* 37.12, pp. 4302–4315. ISSN: 1097-0088. DOI: 10.1002/joc.5086 (cit. on pp. 44, 86, 94, 203).

Fleuret, François (2023). "The Little Book of Deep Learning". In: *A lovely concise introduction* (cit. on p. 54).

Flores-Tolentino, Mayra, Raúl García-Valdés, Cuauhtémoc Saénz-Romero, Irene Ávila-Díaz, Horacio Paz, & Leonel Lopez-Toledo (2020). "Distribution and conservation of species is misestimated if biotic interactions are ignored: the case of the orchid Laelia speciosa". In: *Scientific reports* 10.1, p. 9542 (cit. on p. 149).

Foley, Jonathan A. et al. (July 22, 2005). "Global Consequences of Land Use". In: *Science* 309.5734, pp. 570–574. DOI: 10.1126/science.1111772 (cit. on p. 2).

Fontana, Matteo, Gianluca Zeni, & Simone Vantini (Feb. 2023). "Conformal prediction: A unified review of theory and new challenges". In: *Bernoulli* 29.1, pp. 1–23. ISSN: 1350-7265. DOI: 10.3150/21-BEJ1447 (cit. on pp. 107, 108, 132).

Fourcade, Yoan, Aurélien G. Besnard, & Jean Secondi (2018). "Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics". In: *Global Ecology and Biogeography* 27.2, pp. 245–256. ISSN: 1466-8238. DOI: https://doi.org/10.1111/geb.12684 (cit. on pp. 41, 71, 129).

Frenay, Benoit & Michel Verleysen (May 2014). "Classification in the Presence of Label Noise: A Survey". In: *IEEE Transactions on Neural Networks and Learning Systems* 25.5, pp. 845–869. ISSN: 2162-2388. DOI: 10.1109/TNNLS.2013.2292894 (cit. on p. 97).

Friedl, Mark A, Douglas K McIver, John CF Hodges, Xiaoyang Y Zhang, D Muchoney, Alan H Strahler, Curtis E Woodcock, Sucharita Gopal, Annemarie Schneider, Amanda Cooper, et al. (2002). "Global land cover mapping from MODIS: algorithms and early results". In: *Remote sensing of Environment* 83.1-2, pp. 287–302 (cit. on p. 45).

Gale, Stephan W, Gunter A Fischer, Phillip J Cribb, & Michael F Fay (Mar. 27, 2018). "Orchid conservation: bridging the gap between science and practice". In: *Botanical Journal of the Linnean Society* 186.4, pp. 425–434. ISSN: 0024-4074. DOI: 10.1093/botlinnean/boy003 (cit. on pp. 12, 122, 146).

Garcia, Claude A. et al. (May 9, 2022). "Strategy games to improve environmental policymaking". In: *Nature Sustainability* 5.6, pp. 464–471. ISSN: 2398-9629. DOI: 10.1038/s41893-022-00881-0 (cit. on p. 2).

Garcin, Camille, Alexis Joly, Pierre Bonnet, Jean-Christophe Lombardo, Antoine Affouard, Mathias Chouet, Maximilien Servajean, Titouan Lorieul, & Joseph Salmon (2021). "Pl@ntNet-300K: a plant image dataset with high label ambiguity and a long-tailed distribution". In: *NeurIPS 2021-35th Conference on Neural Information Processing Systems* (cit. on pp. 61, 150).

Garnot, Vivien Sainte Fare, L. Landrieu, S. Giordano, & N. Chehata (July 2019). "Time-Space Tradeoff in Deep Learning Models for Crop Classification on Satellite Multi-Spectral Image Time Series". In: *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*. IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, pp. 6247–6250. DOI: 10.1109/IGARSS.2019.8900517 (cit. on pp. 85, 98).

Garnot, Vivien Sainte Fare & Loic Landrieu (July 26, 2021). "Panoptic Segmentation of Satellite Image Time Series with Convolutional Temporal Attention Networks". In: *arXiv:2107.07933 [cs]* (cit. on p. 98).

Gaston, Kevin J (1991). "How large is a species' geographic range?" In: *Oikos*, pp. 434–438 (cit. on p. 36).

— (May 2000). "Global patterns in biodiversity". In: *Nature* 405.6783, pp. 220–227. ISSN: 1476-4687. DOI: 10.1038/35012228 (cit. on p. 120).

Gaston, Kevin J. & Tim M. Blackburn (Jan. 1997). "The spatial distribution of threatened species: macro-scales and New World birds". In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 263.1367, pp. 235–240. DOI: 10.1098/rspb.1996.0037 (cit. on p. 104).

GBIF (2023). *Orchidaceae*. https://www.gbif.org/species/7689 (cit. on p. 105).

Ghosh, Aritra, Himanshu Kumar, & PS Sastry (2017). "Robust loss functions under label noise for deep neural networks". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 31. 1 (cit. on p. 97).

Gillespie, Lauren, Megan Ruffley, & Moises Exposito-Alonso (Aug. 16, 2022). *An image is worth a thousand species: combining neural networks, citizen science, and remote sensing to map biodiversity*. DOI: 10.1101/2022.08.16.504150 (cit. on pp. 57, 58, 120).

Giraud, Christophe (Aug. 25, 2021). *Introduction to High-Dimensional Statistics*. CRC Press. 410 pp. ISBN: 978-1-00-040835-5 (cit. on p. 46).

Givnish, Thomas J. et al. (2016). "Orchid historical biogeography, diversification, Antarctica and the paradox of orchid dispersal". In: *Journal of Biogeography* 43.10, pp. 1905–1916. ISSN: 1365-2699. DOI: 10.1111/jbi.12854 (cit. on p. 105).

Gneiting, Tilmann & Adrian E Raftery (2007). "Strictly proper scoring rules, prediction, and estimation". In: *Journal of the American statistical Association* 102.477, pp. 359–378 (cit. on p. 106).

Goettsch, B. et al. (Oct. 5, 2015). "High proportion of cactus species threatened with extinction". In: ISSN: 2055-026X. DOI: 10.1038/nplants.2015.142 (cit. on pp. 30, 31).

Gonzalez, Andrew, Jonathan M Chase, & Mary I O'Connor (2023). "A framework for the detection and attribution of biodiversity change". In: *Philosophical Transactions of the Royal Society B* 378.1881, p. 20220182 (cit. on pp. 52, 148).

González-del-Pliego, Pamela, Robert P. Freckleton, David P. Edwards, Michelle S. Koo, Brett R. Scheffers, R. Alexander Pyron, & Walter Jetz (May 6, 2019). "Phylogenetic and Trait-Based Prediction of Extinction Risk for Data-Deficient Amphibians". In: *Current Biology* 29.9, 1557–1563.e3. ISSN: 0960-9822. DOI: 10.1016/j.cub.2019.04.005 (cit. on pp. 62, 104, 149).

Goring, Simon J, Kathleen C Weathers, Walter K Dodds, Patricia A Soranno, Lynn C Sweet, Kendra S Cheruvelil, John S Kominoski, Janine Rüegg, Alexandra M Thorn, & Ryan M Utz (2014). "Improving the culture of interdisciplinary collaboration in ecology by expanding measures of success". In: *Frontiers in Ecology and the Environment* 12.1, pp. 39–47. ISSN: 1540-9309. DOI: 10.1890/120370 (cit. on p. 3).

Graham, Catherine H., Jane Elith, Robert J Hijmans, Antoine Guisan, A Townsend Peterson, Bette A Loiselle, & NCEAS Predicting Species Distributions Working Group (2008). "The influence of spatial errors in species occurrence data used in distribution models". In: *Journal of Applied Ecology* 45.1, pp. 239–247 (cit. on p. 39).

Graham, Catherine H., Santiago R Ron, Juan C Santos, Christopher J Schneider, & Craig Moritz (2004). "Integrating phylogenetics and environmental niche models to explore speciation mechanisms in dendrobatid frogs". In: *Evolution* 58.8, pp. 1781–1793 (cit. on p. 48).

Graham, Eric A., Sandra Henderson, & Annette Schloss (2011). "Using mobile phones to engage citizen scientists in research". In: *Eos, Transactions American Geophysical Union* 92.38, pp. 313–315 (cit. on p. 38).

Griffin, J. N., F. Leprieur, D. Silvestro, J. S. Lefcheck, C. Albouy, D. B. Rasher, M. Davis, J.-C. Svenning, & C. Pimiento (May 10, 2020). *Functionally unique, specialised, and endangered (FUSE) species: towards integrated metrics for the conservation prioritisation toolbox.* DOI: 10.1101/2020.05.09.084871 (cit. on p. 21).

Guisan, Antoine & Carsten Rahbek (2011). *SESAM–a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages* (cit. on p. 146).

Guisan, Antoine & Wilfried Thuiller (2005). "Predicting species distribution: offering more than simple habitat models". In: *Ecology Letters* 8.9, pp. 993–1009. ISSN: 1461-0248. DOI: 10.1111/j.1461-0248.2005.00792.x (cit. on pp. 44, 47, 127, 128).

Guisan, Antoine et al. (2013). "Predicting species distributions for conservation decisions". In: *Ecology Letters* 16.12, pp. 1424–1435. ISSN: 1461-0248. DOI: 10.1111/ele.12189 (cit. on pp. 4, 47, 105, 122, 129, 145, 150).

Gumbs, Rikki, Claudia L Gray, Monika Böhm, Ian J Burfield, Olivia R Couchman, Daniel P Faith, Félix Forest, Michael Hoffmann, Nick JB Isaac, Walter Jetz, et al. (2023). "The EDGE2 protocol: Advancing the prioritisation of Evolutionarily Distinct and Globally Endangered species for practical conservation action". In: *PLoS Biology* 21.2, e3001991 (cit. on p. 20).

Guo, Cheng & Felix Berkhahn (2016). "Entity embeddings of categorical variables". In: *arXiv preprint arXiv:1604.06737* (cit. on p. 148).

Gurevitch, J & D Padilla (Sept. 2004). "Are invasive species a major cause of extinctions?" In: *Trends in Ecology & Evolution* 19.9, pp. 470–474. ISSN: 01695347. DOI: 10.1016/j.tree.2004.07.005 (cit. on p. 2).

Haider, L. Jamila et al. (Jan. 1, 2018). "The undisciplinary journey: early-career perspectives in sustainability science". In: *Sustainability Science* 13.1, pp. 191–204. ISSN: 1862-4057. DOI: 10.1007/s11625-017-0445-1 (cit. on p. 3).

Hakkenberg, Christopher R., Conghe Song, Robert K. Peet, & Peter S. White (2016). "Forest structure as a predictor of tree species diversity in the North Carolina Piedmont". In: *Journal of Vegetation Science* 27.6, pp. 1151–1163. ISSN: 1654-1103. DOI: 10.1111/jvs.12451 (cit. on p. 84).

Halpern, Benjamin S, Shaun Walbridge, Kimberly A Selkoe, Carrie V Kappel, Fiorenza Micheli, Caterina d'Agrosa, John F Bruno, Kenneth S Casey, Colin Ebert, Helen E Fox, et al. (2008). "A global map of human impact on marine ecosystems". In: *science* 319.5865, pp. 948–952 (cit. on p. 46).

Hamilton, Healy et al. (2022). "Increasing taxonomic diversity and spatial resolution clarifies opportunities for protecting US imperiled species". In: *Ecological Applications* 32.3, e2534. ISSN: 1939-5582. DOI: 10.1002/eap.2534 (cit. on pp. 28–30, 105).

Hampton, Stephanie E, Carly A Strasser, Joshua J Tewksbury, Wendy K Gram, Amber E Budden, Archer L Batcheller, Clifford S Duke, & John H Porter (2013). "Big data and the future of ecology". In: *Frontiers in Ecology and the Environment* 11.3, pp. 156–162. ISSN: 1540-9309. DOI: 10.1890/120103 (cit. on p. 3).

Han, Barbara A., John Paul Schmidt, Sarah E Bowden, & John M Drake (2015). "Rodent reservoirs of future zoonotic diseases". In: *Proceedings of the National Academy of Sciences* 112.22, pp. 7039–7044 (cit. on pp. 32, 52).

Han, Yuhui, Shikui Dong, Xiaoyu Wu, Shiliang Liu, Xukun Su, Yong Zhang, Haidi Zhao, Xiaolei Zhang, & David Swift (Oct. 1, 2019). "Integrated modeling to identify priority areas for the conservation of the endangered plant species in headwater areas of Asia". In: *Ecological Indicators* 105, pp. 47–56. ISSN: 1470-160X. DOI: 10.1016/j.ecolind.2019.05.064 (cit. on pp. 48, 104).

Harfoot, Michael, Derek P Tittensor, Tim Newbold, Greg McInerny, Matthew J Smith, & Jörn PW Scharlemann (2014). "Integrated assessment models for ecologists: the present and the future". In: *Global Ecology and Biogeography* 23.2, pp. 124–143 (cit. on p. 60).

Harfoot, Michael B. J. et al. (Nov. 2021). "Using the IUCN Red List to map threats to terrestrial vertebrates at global scale". In: *Nature Ecology & Evolution* 5.11, pp. 1510–1519. ISSN: 2397-334X. DOI: 10.1038/s41559-021-01542-9 (cit. on pp. 8, 31).

Harris, David J. (Apr. 2015). "Generating realistic assemblages with a joint species distribution model". In: *Methods in Ecology and Evolution* 6.4. Ed. by David Warton, pp. 465–473. ISSN: 2041-210X, 2041-210X. DOI: 10.1111/2041-210X.12332 (cit. on p. 57).

Hassler, Michael (2023). *World Plants. Synonymic Checklist and Distribution of the World Flora. Version 16.1.* www.worldplants.de (cit. on p. 114).

He, Kate S., Bethany A. Bradley, Anna F. Cord, Duccio Rocchini, Mao-Ning Tuanmu, Sebastian Schmidtlein, Woody Turner, Martin Wegmann, & Nathalie Pettorelli (2015). "Will remote sensing shape the next generation of species distribution models?" In: *Remote Sensing in Ecology and Conservation* 1.1, pp. 4–18. ISSN: 2056-3485. DOI: https://doi.org/10.1002/rse2.7 (cit. on pp. 5, 73, 75, 93, 120).

Heikkinen, Risto K, Mathieu Marmion, & Miska Luoto (2012). "Does the interpolation accuracy of species distribution models come at the expense of transferability?" In: *Ecography* 35.3, pp. 276–288 (cit. on p. 73).

Heinrichs, Julie A, Darren J Bender, David L Gummer, & Nathan H Schumaker (2010). "Assessing critical habitat: evaluating the relative contribution of habitats to population persistence". In: *Biological Conservation* 143.9, pp. 2229–2237 (cit. on p. 48).

Helber, Patrick, Benjamin Bischke, Andreas Dengel, & Damian Borth (Aug. 31, 2017). "EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification". In: DOI: 10.1109/JSTARS.2019.2918242 (cit. on p. 95).

Hidasi-Neto, José, Rafael Loyola, & Marcus V Cianciaruso (2015). "Global and local evolutionary and ecological distinctiveness of terrestrial mammals: identifying priorities across scales". In: *Diversity and Distributions* 21.5, pp. 548–559 (cit. on p. 21).

Hijmans, Robert J., Susan E. Cameron, Juan L. Parra, Peter G. Jones, & Andy Jarvis (2005). "Very high resolution interpolated climate surfaces for global land areas". In: *International*

Journal of Climatology 25.15, pp. 1965–1978. ISSN: 1097-0088. DOI: 10.1002/joc.1276 (cit. on p. 94).

Hill, Mark O (1973). "Diversity and evenness: a unifying notation and its consequences". In: *Ecology* 54.2, pp. 427–432 (cit. on p. 83).

Hobbs, Sarah J & Piran CL White (2012). "Motivations and barriers in relation to community participation in biodiversity recording". In: *Journal for Nature Conservation* 20.6, pp. 364–373 (cit. on p. 38).

Hole, David G., Stephen G. Willis, Deborah J. Pain, Lincoln D. Fishpool, Stuart H. M. Butchart, Yvonne C. Collingham, Carsten Rahbek, & Brian Huntley (2009). "Projected impacts of climate change on a continent-wide protected area network". In: *Ecology Letters* 12.5, pp. 420–431. ISSN: 1461-0248. DOI: 10.1111/j.1461-0248.2009.01297.x (cit. on p. 66).

Hong, Danfeng, Zhu Han, Jing Yao, Lianru Gao, Bing Zhang, Antonio Plaza, & Jocelyn Chanussot (2021). "SpectralFormer: Rethinking Hyperspectral Image Classification with Transformers". In: *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–1. ISSN: 1558-0644. DOI: 10.1109/TGRS.2021.3130716 (cit. on p. 148).

Hooper, David U., E. Carol Adair, Bradley J. Cardinale, Jarrett E. K. Byrnes, Bruce A. Hungate, Kristin L. Matulich, Andrew Gonzalez, J. Emmett Duffy, Lars Gamfeldt, & Mary I. O'Connor (June 2012). "A global synthesis reveals biodiversity loss as a major driver of ecosystem change". In: *Nature* 486.7401, pp. 105–108. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature11118 (cit. on p. 2).

Hornik, Kurt, Maxwell Stinchcombe, & Halbert White (1989). "Multilayer feedforward networks are universal approximators". In: *Neural networks* 2.5, pp. 359–366 (cit. on p. 49).

Hortal, Joaquín, Francesco de Bello, José Alexandre F. Diniz-Filho, Thomas M. Lewinsohn, Jorge M. Lobo, & Richard J. Ladle (2015). "Seven Shortfalls that Beset Large-Scale Knowledge of Biodiversity". In: *Annual Review of Ecology, Evolution, and Systematics* 46.1, pp. 523–549. DOI: 10.1146/annurev-ecolsys-112414-054400 (cit. on pp. 4, 39, 145).

Høye, Toke T., Johanna Ärje, Kim Bjerge, Oskar L. P. Hansen, Alexandros Iosifidis, Florian Leese, Hjalte M. R. Mann, Kristian Meissner, Claus Melvad, & Jenni Raitoharju (Jan. 12, 2021). "Deep learning and computer vision will transform entomology". In: *Proceedings of the National Academy of Sciences* 118.2, e2002545117. DOI: 10.1073/pnas.2002545117 (cit. on p. 61).

Hudson, Lawrence N, Tim Newbold, Sara Contu, Samantha LL Hill, Igor Lysenko, Adriana De Palma, Helen RP Phillips, Rebecca A Senior, Dominic J Bennett, Hollie Booth, et al. (2014). "The PREDICTS database: a global database of how local terrestrial biodiversity responds to human impacts". In: *Ecology and evolution* 4.24, pp. 4701–4735 (cit. on p. 38).

Humboldt, Alexander von & Aimé Bonpland (1807). *Voyage de Humboldt et Bonpland: Essai Sur La Géographie Des Plantes, Accompagné D'Un Tableau Physique Des Régions Équinoxiales: Fondé sur des mesures exécutées, depuis le dixième degré de latitude boréale jusqu'au dixième degré de latitude australe, pendant les années 1799, 1800, 1801, 1802 et 1803/Par Al. De Humboldt Et A. Bonpland; Rédigé Par Alexandre De Humboldt.* Vol. 5. Schoell (cit. on p. 35).

Hurlbert, Allen H & Walter Jetz (2007). "Species richness, hotspots, and the scale dependence of range maps in ecology and conservation". In: *Proceedings of the National Academy of Sciences* 104.33, pp. 13384–13389 (cit. on p. 37).

Hutchinson, G Evelyn (1957). "Cold spring harbor symposia on quantitative biology". In: *(No Title)* 22, p. 415 (cit. on p. 41).

Ientilucci, Emmett J. & Steven Adler-Golden (June 2019). "Atmospheric Compensation of Hyperspectral Data: An Overview and Review of In-Scene and Physics-Based Approaches". In: *IEEE Geoscience and Remote Sensing Magazine* 7.2, pp. 31–50. ISSN: 2168-6831. DOI: 10.1109/MGRS.2019.2904706 (cit. on p. 75).

Ioffe, Sergey & Christian Szegedy (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *International conference on machine learning*. PMLR, pp. 448–456 (cit. on pp. 50, 81).

IPBES (May 4, 2019). *Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services*. Zenodo. DOI: 10.5281/zenodo.6417333 (cit. on pp. 2, 126, 129).

Isaac, Nick JB & William D Pearse (2018). "The use of EDGE (Evolutionary Distinct Globally Endangered) and EDGE-like metrics to evaluate taxa for conservation". In: *Phylogenetic diversity: Applications and challenges in biodiversity science*, pp. 27–39 (cit. on p. 199).

Isaac, Nick JB, Samuel T Turvey, Ben Collen, Carly Waterman, & Jonathan EM Baillie (2007). "Mammals on the EDGE: conservation priorities based on threat and phylogeny". In: *PloS one* 2.3, e296 (cit. on p. 20).

IUCN (2012a). *Guidelines for application of IUCN Red List criteria at regional and national levels : version 4.0*. IUCN. ISBN: 978-2-8317-1247-5 (cit. on p. 17).

— (2012b). *Habitats classification scheme (version 3.1)* (cit. on p. 25).

— (2020). *IUCN Global Ecosystem Typology 2.0*. IUCN. ISBN: 978-2-8317-2077-7. DOI: 10.2305/IUCN.CH.2020.13.en (cit. on p. 26).

— (2022). *Barometer of Life*. https://www.iucnredlist.org/about/barometer-of-life (cit. on p. 104).

Jackson, Stephen T & Jonathan T Overpeck (2000). "Responses of plant populations and communities to environmental changes of the late Quaternary". In: *Paleobiology* 26.S4, pp. 194–220 (cit. on p. 41).

Jarić, Ivan, Ricardo A Correia, Barry W Brook, Jessie C Buettel, Franck Courchamp, Enrico Di Minin, Josh A Firth, Kevin J Gaston, Paul Jepson, Gregor Kalinkat, et al. (2020). "iEcology: harnessing large online resources to generate ecological insights". In: *Trends in Ecology & Evolution* 35.7, pp. 630–639 (cit. on p. 3).

Jarić, Ivan, Franck Courchamp, Jörn Gessner, & David L. Roberts (Sept. 1, 2016). "Potentially threatened: a Data Deficient flag for conservation management". In: *Biodiversity and Conservation* 25.10, pp. 1995–2000. ISSN: 1572-9710. DOI: 10.1007/s10531-016-1164-0 (cit. on p. 19).

Jaynes, Edwin T (1957). "Information theory and statistical mechanics". In: *Physical review* 106.4, p. 620 (cit. on p. 43).

Jetz, Walter & Robert P. Freckleton (Feb. 19, 2015). "Towards a general framework for predicting threat status of data-deficient species from phylogenetic, spatial and environmental information". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1662, p. 20140016. DOI: 10.1098/rstb.2014.0016 (cit. on p. 62).

Jetz, Walter, Melodie A McGeoch, Robert Guralnick, Simon Ferrier, Jan Beck, Mark J. Costello, Miguel Fernandez, Gary N Geller, Petr Keil, Cory Merow, et al. (2019). "Essential biodiversity variables for mapping and monitoring species populations". In: *Nature ecology & evolution* 3.4, pp. 539–551 (cit. on p. 28).

Jetz, Walter, Jennifer McGowan, D. Scott Rinnan, Hugh P. Possingham, Piero Visconti, Brian O'Donnell, & Maria Cecilia Londoño-Murcia (Feb. 2022). "Include biodiversity representation indicators in area-based conservation targets". In: *Nature Ecology & Evolution* 6.2, pp. 123–126. ISSN: 2397-334X. DOI: 10.1038/s41559-021-01620-y (cit. on pp. 2, 33, 34).

Jetz, Walter, Jana M. McPherson, & Robert P. Guralnick (Mar. 1, 2012). "Integrating biodiversity distribution knowledge: toward a global map of life". In: *Trends in Ecology & Evolution* 27.3, pp. 151–159. ISSN: 0169-5347. DOI: 10.1016/j.tree.2011.09.007 (cit. on pp. 28, 151).

Jetz, Walter, Cagan H Sekercioglu, & James EM Watson (2008). "Ecological correlates and conservation implications of overestimating species geographic ranges". In: *Conservation Biology* 22.1, pp. 110–119 (cit. on p. 17).

Ji, Shunping, Chi Zhang, Anjian Xu, Yun Shi, & Yulin Duan (Jan. 2018). "3D Convolutional Neural Networks for Crop Classification with Multi-Temporal Remote Sensing Images". In: *Remote Sensing* 10.1, p. 75. DOI: 10.3390/rs10010075 (cit. on p. 148).

Jiménez-Alfaro, Borja, David Draper, & David Nogués-Bravo (2012). "Modeling the potential area of occupancy at fine resolution may reduce uncertainty in species range estimates". In: *Biological Conservation* 147.1, pp. 190–196 (cit. on p. 47).

Joppa, Lucas N., David L. Roberts, Norman Myers, & Stuart L. Pimm (Aug. 9, 2011). "Biodiversity hotspots house most undiscovered plant species". In: *Proceedings of the National Academy of Sciences* 108.32, pp. 13171–13176. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1109389108 (cit. on p. 94).

Joppa, Lucas N. et al. (2016). «Impact of alternative metrics on estimates of extent of occurrence for extinction risk assessment». In: *Conservation Biology* 30.2, pp. 362–370. ISSN: 1523-1739. DOI: https://doi.org/10.1111/cobi.12591 (cit. on p. 19).

Jørgensen, Sven, Liu Xu, & Robert Costanza (Apr. 19, 2016). *Handbook of Ecological Indicators for Assessment of Ecosystem Health*. CRC Press. 499 pp. ISBN: 978-1-4398-0937-2 (cit. on pp. 11, 121).

Jost, Lou (2006). "Entropy and diversity". In: *Oikos* 113.2, pp. 363–375 (cit. on p. 83).

Juffe-Bignoli, Diego et al. (Aug. 16, 2016). "Assessing the Cost of Global Biodiversity and Conservation Knowledge". In: *PLOS ONE* 11.8, e0160640. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0160640 (cit. on pp. 8, 19).

Jung, Martin, Prabhat Raj Dahal, Stuart H. M. Butchart, Paul F Donald, Xavier De Lamo, Myroslava Lesiv, Valerie Kapos, Carlo Rondinini, & Piero Visconti (2020). "A global map of terrestrial habitat types". In: *Scientific data* 7.1, p. 256 (cit. on p. 26).

Justeau-Allaire, Dimitri, Jeffrey O Hanson, Guillaume Lannuzel, Philippe Vismara, Xavier Lorca, & Philippe Birnbaum (2023). "restoptr: an R package for ecological restoration planning". In: *Restoration Ecology*, e13910 (cit. on p. 67).

Justeau-Allaire, Dimitri, Ghislain Vieilledent, Nicolas Rinck, Philippe Vismara, Xavier Lorca, & Philippe Birnbaum (2021). "Constrained optimization of landscape indices in conservation planning to support ecological restoration in New Caledonia". In: *Journal of Applied Ecology* 58.4, pp. 744–754 (cit. on pp. 67, 150).

Justeau-Allaire, Dimitri, Philippe Vismara, Philippe Birnbaum, & Xavier Lorca (2019). "Systematic Conservation Planning for Sustainable Land-use Policies: A Constrained Partitioning Approach to Reserve Selection and Design." In: *IJCAI 2019-28th International Joint Conference on Artificial Intelligence*. IJCAI, pp. 5902–5908 (cit. on p. 67).

Katal, Negin, Michael Rzanny, Patrick Mäder, & Jana Wäldchen (2022). "Deep learning in plant phenological research: A systematic literature review". In: *Frontiers in Plant Science* 13, p. 805738 (cit. on p. 61).

Kattenborn, Teja, Jana Eichel, Susan Wiser, Larry Burrows, Fabian E. Fassnacht, & Sebastian Schmidtlein (2020). "Convolutional Neural Networks accurately predict cover fractions of plant species and communities in Unmanned Aerial Vehicle imagery". In: *Remote Sensing in Ecology and Conservation* 6.4, pp. 472–486. ISSN: 2056-3485. DOI: https://doi.org/10.1002/rse2.146 (cit. on p. 73).

Keith, David A. et al. (May 8, 2013). "Scientific Foundations for an IUCN Red List of Ecosystems". In: *PLoS ONE* 8.5. Ed. by Matteo Convertino, e62111. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0062111 (cit. on pp. 25, 26, 199).

Keith, David A. et al. (Oct. 2022). "A function-based typology for Earth's ecosystems". In: *Nature* 610.7932, pp. 513–518. ISSN: 1476-4687. DOI: 10.1038/s41586-022-05318-4 (cit. on p. 26).

KEW, Royal Botanic Gardens (2023). *Plants of the Word Online*. https://powo.science.kew.org/results?q=Orchidaceae (cit. on pp. 105, 130).

Kobori, Hiromi et al. (Jan. 1, 2016). "Citizen science: a new approach to advance ecology, education, and conservation". In: *Ecological Research* 31.1, pp. 1–19. ISSN: 1440-1703. DOI: 10.1007/s11284-015-1314-y (cit. on p. 97).

Kokhlikyan, Narine, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. (2020). "Captum: A unified and generic model interpretability library for pytorch". In: *arXiv preprint arXiv:2009.07896* (cit. on p. 147).

König, Christian, Patrick Weigelt, Julian Schrader, Amanda Taylor, Jens Kattge, & Holger Kreft (Mar. 18, 2019). "Biodiversity data integration—the significance of data resolution and domain". In: *PLOS Biology* 17.3, e3000183. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.3000183 (cit. on p. 4).

Kremen, Claire (May 1992). "Assessing the Indicator Properties of Species Assemblages for Natural Areas Monitoring". In: *Ecological Applications* 2.2, pp. 203–217. ISSN: 10510761. DOI: 10.2307/1941776 (cit. on p. 16).

Krizhevsky, Alex, Ilya Sutskever, & Geoffrey E Hinton (2012). "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (cit. on pp. 50, 54, 151).

Kuipers, Koen J. J., Stefanie Hellweg, & Francesca Verones (May 7, 2019). "Potential Consequences of Regional Species Loss for Global Species Richness: A Quantitative Approach for Estimating Global Extinction Probabilities". In: *Environmental Science & Technology* 53.9, pp. 4728–4738. ISSN: 0013-936X. DOI: 10.1021/acs.est.8b06173 (cit. on p. 29).

Kull, Tiiu & Michael J. Hutchings (Apr. 1, 2006). "A comparative analysis of decline in the distribution ranges of orchid species in Estonia and the United Kingdom". In: *Biological Conservation* 129.1, pp. 31–39. ISSN: 0006-3207. DOI: 10.1016/j.biocon.2005.09.046 (cit. on pp. 11, 121).

Lai, Siwei, Liheng Xu, Kang Liu, & Jun Zhao (Feb. 19, 2015). "Recurrent Convolutional Neural Networks for Text Classification". In: *Twenty-Ninth AAAI Conference on Artificial Intelligence*. Twenty-Ninth AAAI Conference on Artificial Intelligence (cit. on p. 98).

Lamba, Aakash, Phillip Cassey, Ramesh Raja Segaran, & Lian Pin Koh (Oct. 2019). "Deep learning for environmental conservation". In: *Current Biology* 29.19, R977–R982. ISSN: 09609822. DOI: 10.1016/j.cub.2019.08.016 (cit. on pp. 59, 129).

Le Lay, Gwenaëlle, Robin Engler, Erika Franc, & Antoine Guisan (2010). "Prospective sampling based on model ensembles improves the detection of rare species". In: *Ecography* 33.6, pp. 1015–1027 (cit. on p. 48).

Leão, Tarciso C. C., Carlos R. Fonseca, Carlos A. Peres, & Marcelo Tabarelli (2014). «Predicting Extinction Risk of Brazilian Atlantic Forest Angiosperms». In: *Conservation Biology* 28.5, pp. 1349–1359. ISSN: 1523-1739. DOI: https://doi.org/10.1111/cobi.12286 (cit. on p. 62).

Leathwick, John R., J Elith, & T Hastie (2006). "Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions". In: *Ecological modelling* 199.2, pp. 188–196 (cit. on p. 53).

Leblanc, César, Alexis Joly, Titouan Lorieul, Maximilien Servajean, & Pierre Bonnet (Sept. 5, 2022). "Species Distribution Modeling based on aerial images and environmental features with

Convolutional Neural Networks". In: CLEF 2022 - Conference and Labs of the Evaluation Forum, p. 2123 (cit. on p. 105).

LeCun, Yann, Bernhard Boser, John Denker, Donnie Henderson, Richard Howard, Wayne Hubbard, & Lawrence Jackel (1989). "Handwritten digit recognition with a back-propagation network". In: *Advances in neural information processing systems* 2 (cit. on p. 54).

LeCun, Yann, Léon Bottou, Yoshua Bengio, & Patrick Haffner (1998). "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11, pp. 2278–2324 (cit. on p. 55).

Lee, Kuang-Huei, Xiaodong He, Lei Zhang, & Linjun Yang (2018). "Cleannet: Transfer learning for scalable image classifier training with label noise". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5447–5456 (cit. on p. 97).

Lembrechts, Jonas J., Ivan Nijs, & Jonathan Lenoir (2019). "Incorporating microclimate into species distribution models". In: *Ecography* 42.7, pp. 1267–1279. ISSN: 1600-0587. DOI: 10.1111/ecog.03947 (cit. on p. 104).

Lenoir, Jonathan, Romain Bertrand, Lise Comte, Luana Bourgeaud, Tarek Hattab, Jérôme Murienne, & Gaël Grenouillet (Aug. 2020). "Species better track climate warming in the oceans than on land". In: *Nature Ecology & Evolution* 4.8, pp. 1044–1059. ISSN: 2397-334X. DOI: 10.1038/s41559-020-1198-2 (cit. on p. 64).

Lenzen, M., D. Moran, K. Kanemoto, B. Foran, L. Lobefaro, & A. Geschke (June 2012). "International trade drives biodiversity threats in developing nations". In: *Nature* 486.7401, pp. 109–112. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature11145 (cit. on pp. 2, 21).

Levin, Michael O., Jared B. Meek, Brian Boom, Sara M. Kross, & Evan A. Eskew (Mar. 2022). "Using publicly available data to conduct rapid assessments of extinction risk". In: *Conservation Science and Practice* 4.3. ISSN: 2578-4854, 2578-4854. DOI: 10.1111/csp2.12628 (cit. on p. 47).

Linardatos, Pantelis, Vasilis Papastefanopoulos, & Sotiris Kotsiantis (Jan. 2021). "Explainable AI: A Review of Machine Learning Interpretability Methods". In: *Entropy* 23.1, p. 18. DOI: 10.3390/e23010018 (cit. on p. 120).

Linden, Sander van der, Anthony Leiserowitz, Seth Rosenthal, & Edward Maibach (2017). "Inoculating the Public against Misinformation about Climate Change". In: *Global Challenges* 1.2, p. 1600008. ISSN: 2056-6646. DOI: 10.1002/gch2.201600008 (cit. on p. 4).

Liu, Canran, Pam M. Berry, Terence P. Dawson, & Richard G. Pearson (2005). "Selecting thresholds of occurrence in the prediction of species distributions". In: *Ecography* 28.3, pp. 385–393. ISSN: 1600-0587. DOI: https://doi.org/10.1111/j.0906-7590.2005.03957.x (cit. on p. 41).

Liu, Canran, Matt White, & Graeme Newell (Apr. 2013). "Selecting thresholds for the prediction of species occurrence with presence-only data". In: *Journal of Biogeography* 40.4. Ed. by Richard Pearson, pp. 778–789. ISSN: 03050270. DOI: 10.1111/jbi.12058 (cit. on p. 41).

Liu, Jianguo et al. (Sept. 2018). "Nexus approaches to global sustainable development". In: *Nature Sustainability* 1.9, pp. 466–476. ISSN: 2398-9629. DOI: 10.1038/s41893-018-0135-8 (cit. on p. 3).

Liu, Ze, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, & Baining Guo (2021). "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows". In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (cit. on p. 148).

Liu, Zhuang, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, & Saining Xie (Mar. 2, 2022). "A ConvNet for the 2020s". In: *arXiv:2201.03545 [cs]* (cit. on p. 148).

Loh, Jonathan, Rhys E Green, Taylor Ricketts, John Lamoreux, Martin Jenkins, Valerie Kapos, & Jorgen Randers (Feb. 28, 2005). "The Living Planet Index: using species population time

series to track trends in biodiversity". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 360.1454, pp. 289–295. DOI: 10.1098/rstb.2004.1584 (cit. on p. 22).
Lomolino, Mark V et al. (2004). "Conservation biogeography". In: *Frontiers of Biogeography: new directions in the geography of nature* 293 (cit. on p. 28).
Lomolino, Mark. V. (2001). "Elevation gradients of species-density: historical and prospective views". In: *Global Ecology and Biogeography* 10.1, pp. 3–13. ISSN: 1466-8238. DOI: 10.1046/j.1466-822x.2001.00229.x (cit. on p. 140).
Lorieul, Titouan (Dec. 2, 2020). "Uncertainty in predictions of Deep Learning models for fine-grained classification". PhD thesis. Université Montpellier (cit. on p. 107).
Lorieul, Titouan, Katelin D Pearson, Elizabeth R Ellwood, Hervé Goëau, Jean-Francois Molino, Patrick W Sweeney, Jennifer M Yost, Joel Sachs, Erick Mata-Montero, Gil Nelson, et al. (2019). "Toward a large-scale and deep phenological stage annotation of herbarium specimens: Case studies from temperate, tropical, and equatorial floras". In: *Applications in Plant Sciences* 7.3, e01233 (cit. on p. 61).
Louis, Jérôme, Vincent Debaecker, Bringfried Pflug, Magdalena Main-Knorn, Jakub Bieniarz, Uwe Mueller-Wilm, Enrico Cadau, & Ferran Gascon (2016). "Sentinel-2 sen2cor: L2a processor for users". In: pp. 1–8 (cit. on p. 75).
Lucas, Tim CD (2020). "A translucent box: interpretable machine learning in ecology". In: *Ecological Monographs* 90.4, e01422 (cit. on p. 52).
Lundberg, Scott M & Su-In Lee (2017). "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30 (cit. on pp. 52, 58, 147).
Luoto, Miska, Raimo Virkkala, & Risto K. Heikkinen (2007). "The role of land cover in bioclimatic models depends on spatial resolution". In: *Global Ecology and Biogeography* 16.1, pp. 34–42. ISSN: 1466-8238. DOI: 10.1111/j.1466-8238.2006.00262.x (cit. on pp. 46, 97).
Lutio, Riccardo de, John Y Park, Kimberly A Watson, Stefano D'Aronco, Jan D Wegner, Jan J Wieringa, Melissa Tulig, Richard L Pyle, Timothy J Gallaher, Gillian Brown, et al. (2022). "The Herbarium 2021 Half–Earth Challenge Dataset and Machine Learning Competition". In: *Frontiers in Plant Science* 12, p. 3320 (cit. on pp. 38, 61).
Mace, Georgina M., Mike Barrett, Neil D. Burgess, Sarah E. Cornell, Robin Freeman, Monique Grooten, & Andy Purvis (Sept. 2018). "Aiming higher to bend the curve of biodiversity loss". In: *Nature Sustainability* 1.9, pp. 448–451. ISSN: 2398-9629. DOI: 10.1038/s41893-018-0130-0 (cit. on pp. 126, 127).
Mace, Georgina M., Nigel J. Collar, Kevin J. Gaston, Craig Hilton-Taylor, H. Resit Akçakaya, Nigel Leader-Williams, E.J. Milner-Gulland, & Simon N. Stuart (Dec. 2008). "Quantification of Extinction Risk: IUCN's System for Classifying Threatened Species". In: *Conservation Biology* 22.6, pp. 1424–1442. ISSN: 08888892, 15231739. DOI: 10.1111/j.1523-1739.2008.01044.x (cit. on pp. 16, 108, 127).
Mace, Georgina M., John L. Gittleman, & Andy Purvis (June 13, 2003). "Preserving the Tree of Life". In: *Science* 300.5626, pp. 1707–1709. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1085510 (cit. on p. 20).
MacKenzie, Darryl I & James D Nichols (2004). "Occupancy as a surrogate for abundance estimation". In: *Animal biodiversity and conservation* 27.1, pp. 461–467 (cit. on p. 38).
Maclean, Ilya M. D. & Robert J. Wilson (July 26, 2011). "Recent ecological responses to climate change support predictions of high extinction risk". In: *Proceedings of the National Academy of Sciences* 108.30, pp. 12337–12342. DOI: 10.1073/pnas.1017352108 (cit. on pp. 65, 127).
Maes, Joachim, Anne Teller, Markus Erhard, C Liquete, L Braat, P Berry, B Egoh, P Puydarrieux, Ch Fiorina, F Santos, et al. (2013). "Mapping and Assessment of Ecosystems and their Services". In: *An analytical framework for ecosystem assessments under action* 5, pp. 1–58 (cit. on p. 26).

Mainali, Kumar, Trevor Hefley, Leslie Ries, & William F Fagan (2020). "Matching expert range maps with species distribution model predictions". In: *Conservation Biology* 34.5, pp. 1292–1304 (cit. on p. 147).

Mair, Louise et al. (June 2021). "A metric for spatially explicit contributions to science-based species targets". In: *Nature Ecology & Evolution* 5.6, pp. 836–844. ISSN: 2397-334X. DOI: 10.1038/s41559-021-01432-0 (cit. on pp. 29, 104).

Malcolm, Jay R., Canran Liu, Ronald P. Neilson, Lara Hansen, & Lee Hannah (Apr. 2006). "Global Warming and Extinctions of Endemic Species from Biodiversity Hotspots". In: *Conservation Biology* 20.2, pp. 538–548. ISSN: 0888-8892, 1523-1739. DOI: 10.1111/j.1523-1739.2006.00364.x (cit. on p. 127).

Mañas, Oscar, Alexandre Lacoste, Xavier Giro-i-Nieto, David Vazquez, & Pau Rodriguez (May 3, 2021). "Seasonal Contrast: Unsupervised Pre-Training from Uncurated Remote Sensing Data". In: *arXiv:2103.16607 [cs]* (cit. on p. 95).

Marcon, Eric (2015). "Mesures de la biodiversité". PhD thesis. AgroParisTech (cit. on pp. 24, 109).

Marconcini, Mattia et al. (July 20, 2020). "Outlining where humans live, the World Settlement Footprint 2015". In: *Scientific Data* 7.1, p. 242. ISSN: 2052-4463. DOI: 10.1038/s41597-020-00580-5 (cit. on p. 46).

Margules, Christopher Robert & Robert L Pressey (2000). "Systematic conservation planning". In: *Nature* 405.6783, pp. 243–253 (cit. on p. 66).

Marmion, Mathieu, Miia Parviainen, Miska Luoto, Risto K Heikkinen, & Wilfried Thuiller (2009). "Evaluation of consensus methods in predictive species distribution modelling". In: *Diversity and distributions* 15.1, pp. 59–69 (cit. on p. 42).

Marsh, Charles J., Yoni Gavish, William E. Kunin, & Neil A. Brummitt (2019). "Mind the gap: Can downscaling Area of Occupancy overcome sampling gaps when assessing IUCN Red List status?" In: *Diversity and Distributions* 25.12, pp. 1832–1845. ISSN: 1472-4642. DOI: 10.1111/ddi.12983 (cit. on p. 19).

Marshall, Erica, Brendan A Wintle, Darren Southwell, & Heini Kujala (2020). "What are we measuring? A review of metrics used to describe biodiversity in offsets exchanges". In: *Biological Conservation* 241, p. 108250 (cit. on p. 146).

Mattila, Niina, Janne S Kotiaho, Veijo Kaitala, & Atte Komonen (2008). "The use of ecological traits in extinction risk assessments: a case study on geometrid moths". In: *Biological Conservation* 141.9, pp. 2322–2328 (cit. on p. 19).

Mauri, Achille, Giovanni Strona, & Jesús San-Miguel-Ayanz (Jan. 5, 2017). "EU-Forest, a high-resolution tree occurrence dataset for Europe". In: *Scientific Data* 4.1, p. 160123. ISSN: 2052-4463. DOI: 10.1038/sdata.2016.123 (cit. on p. 38).

Maxwell, Aaron E., Timothy A. Warner, & Fang Fang (May 3, 2018). "Implementation of machine-learning classification in remote sensing: an applied review". In: *International Journal of Remote Sensing* 39.9, pp. 2784–2817. ISSN: 0143-1161. DOI: 10.1080/01431161.2018.1433343 (cit. on p. 4).

Maxwell, Sean L., Victor Cazalis, Nigel Dudley, Michael Hoffmann, Ana S. L. Rodrigues, Sue Stolton, Piero Visconti, Stephen Woodley, Naomi Kingston, Edward Lewis, et al. (2020). "Area-based conservation in the twenty-first century". In: *Nature* 586.7828, pp. 217–227 (cit. on pp. 33, 66).

McCormick, Melissa K., Dennis F. Whigham, & Armando Canchani-Viruet (2018). "Mycorrhizal fungi affect orchid distribution and population dynamics". In: *New Phytologist* 219.4, pp. 1207–1215. ISSN: 1469-8137. DOI: https://doi.org/10.1111/nph.15223 (cit. on pp. 11, 45, 120).

Medina-Lopez, Encarni (Jan. 2020). "Machine Learning and the End of Atmospheric Corrections: A Comparison between High-Resolution Sea Surface Salinity in Coastal Areas from Top

and Bottom of Atmosphere Sentinel-2 Imagery". In: *Remote Sensing* 12.18, p. 2924. DOI: 10.3390/rs12182924 (cit. on p. 76).

Metzger, Nando, John E. Vargas-Muñoz, Rodrigo C. Daudt, Benjamin Kellenberger, Thao Ton-That Whelan, Ferda Ofli, Muhammad Imran, Konrad Schindler, & Devis Tuia (Nov. 22, 2022). "Fine-grained population mapping from coarse census counts and open geodata". In: *Scientific Reports* 12.1, p. 20085. ISSN: 2045-2322. DOI: 10.1038/s41598-022-24495-w (cit. on p. 62).

Meyer, Carsten, Patrick Weigelt, & Holger Kreft (2016). "Multidimensional biases, gaps and uncertainties in global plant occurrence information". In: *Ecology Letters* 19.8, pp. 992–1006. ISSN: 1461-0248. DOI: https://doi.org/10.1111/ele.12624 (cit. on p. 39).

Michener, William K. & Matthew B. Jones (Feb. 1, 2012). "Ecoinformatics: supporting ecology as a data-intensive science". In: *Trends in Ecology & Evolution* 27.2, pp. 85–93. ISSN: 0169-5347. DOI: 10.1016/j.tree.2011.11.016 (cit. on p. 3).

Mihoub, Jean-Baptiste, Klaus Henle, Nicolas Titeux, Lluis Brotons, Neil A. Brummitt, & Dirk S Schmeller (2017). "Setting temporal baselines for biodiversity: the limits of available monitoring data for capturing the full impact of anthropogenic pressures". In: *Scientific reports* 7.1, p. 41591 (cit. on p. 27).

Miller, David A. W., Krishna Pacifici, Jamie S. Sanderlin, & Brian J. Reich (2019). "The recent past and promising future for data integration methods to estimate species' distributions". In: *Methods in Ecology and Evolution* 10.1, pp. 22–37. ISSN: 2041-210X. DOI: 10.1111/2041-210X.13110 (cit. on p. 37).

Moat, Justin, Tadesse W. Gole, & Aaron P. Davis (2019). "Least concern to endangered: Applying climate change projections profoundly influences the extinction risk assessment for wild Arabica coffee". In: *Global Change Biology* 25.2, pp. 390–403. ISSN: 1365-2486. DOI: 10.1111/gcb.14341 (cit. on pp. 10, 47, 65, 127).

Mod, Heidi K., Daniel Scherrer, Miska Luoto, & Antoine Guisan (2016). "What we use is not what we know: environmental predictors in plant distribution models". In: *Journal of Vegetation Science* 27.6, pp. 1308–1322. ISSN: 1654-1103. DOI: 10.1111/jvs.12444 (cit. on p. 44).

Moilanen, Atte, Aldina M.A Franco, Regan I Early, Richard Fox, Brendan Wintle, & Chris D Thomas (Sept. 22, 2005). "Prioritizing multiple-use landscapes for conservation: methods for large multi-species planning problems". In: *Proceedings of the Royal Society B: Biological Sciences* 272.1575, pp. 1885–1891. ISSN: 0962-8452, 1471-2954. DOI: 10.1098/rspb.2005.3164 (cit. on p. 66).

Moilanen, Atte, Pauli Lehtinen, Ilmari Kohonen, Joel Jalkanen, Elina A. Virtanen, & Heini Kujala (2022). "Novel methods for spatial prioritization with applications in conservation, land use planning and ecological impact avoidance". In: *Methods in Ecology and Evolution* 13.5, pp. 1062–1072. ISSN: 2041-210X. DOI: 10.1111/2041-210X.13819 (cit. on p. 66).

Moilanen, Atte, Kerrie A. Wilson, & Hugh Possingham (2009). *Spatial conservation prioritization: quantitative methods and computational tools*. Oxford University Press (cit. on p. 66).

Monsimet, Jeremy, Olivier Devineau, Julien Petillon, & Denis Lafage (2020). "Explicit integration of dispersal-related metrics improves predictions of SDM in predatory arthropods". In: *Scientific reports* 10.1, pp. 1–12 (cit. on p. 73).

Moreno, Claudia E & Pilar Rodriguez (2010). "A consistent terminology for quantifying species diversity?" In: *Oecologia* 163, pp. 279–282 (cit. on p. 25).

Moret, Pierre, Priscilla Muriel, Ricardo Jaramillo, & Olivier Dangles (June 25, 2019). "Humboldt's Tableau Physique revisited". In: *Proceedings of the National Academy of Sciences* 116.26, pp. 12889–12894. DOI: 10.1073/pnas.1904585116 (cit. on pp. 35, 104).

Mortier, Thomas, Marek Wydmuch, Krzysztof Dembczyński, Eyke Hüllermeier, & Willem Waegeman (2021). "Efficient set-valued prediction in multi-class classification". In: *Data Mining and Knowledge Discovery* 35.4, pp. 1435–1469 (cit. on p. 106).

Mouquet, Nicolas et al. (2015). "REVIEW: Predictive ecology in a changing world". In: *Journal of Applied Ecology* 52.5, pp. 1293–1310. ISSN: 1365-2664. DOI: 10.1111/1365-2664.12482 (cit. on p. 60).

Mouton, Ans M, Bernard De Baets, & Peter LM Goethals (2010). "Ecological relevance of performance criteria for species distribution models". In: *Ecological modelling* 221.16, pp. 1995–2002 (cit. on p. 147).

Mukadam, Meet, Mandhara Jayaram, & Yongfeng Zhang (2020). "A Representation Learning Approach to Animal Biodiversity Conservation". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 294–305. DOI: 10.18653/v1/2020.coling-main.26 (cit. on p. 63).

Munoz, François, Joaquim Estopinan, Ruksan Bose, Raphaël Pélissier, & Ghislain Vieilledent (Nov. 26, 2021). "Future impacts of climate change and deforestation on endemic trees of Western Ghats, South India". In: (cit. on p. vii).

Murdoch, W. James, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, & Bin Yu (Oct. 29, 2019). "Definitions, methods, and applications in interpretable machine learning". In: *Proceedings of the National Academy of Sciences* 116.44, pp. 22071–22080. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1900654116 (cit. on p. 44).

Myers, Norman, Russell A. Mittermeier, Cristina G. Mittermeier, Gustavo A. B. da Fonseca, & Jennifer Kent (Feb. 2000). "Biodiversity hotspots for conservation priorities". In: *Nature* 403.6772, pp. 853–858. ISSN: 1476-4687. DOI: 10.1038/35002501 (cit. on p. 97).

Nair, Vinod & Geoffrey E Hinton (2010). "Rectified linear units improve restricted boltzmann machines". In: *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814 (cit. on p. 50).

Navarro, Laetitia M et al. (Dec. 2017). "Monitoring biodiversity change through effective global coordination". In: *Current Opinion in Environmental Sustainability* 29, pp. 158–169. ISSN: 18773435. DOI: 10.1016/j.cosust.2018.02.005 (cit. on p. 27).

Nelder, John Ashworth & Robert WM Wedderburn (1972). "Generalized linear models". In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 135.3, pp. 370–384 (cit. on p. 43).

Newman, Belinda (2009). "Orchids as Indicators of Ecosystem Health in Urban Bushland Fragments". PhD thesis. Murdoch University (cit. on pp. 11, 16, 74, 106, 121).

Nic Lughadha, Eimear et al. (Jan. 7, 2019). "The use and misuse of herbarium specimens in evaluating plant extinction risks". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 374.1763, p. 20170402. DOI: 10.1098/rstb.2017.0402 (cit. on pp. 63, 104).

Nicholson, Emily, Elizabeth A Fulton, Thomas M Brooks, Ryan Blanchard, Paul Leadley, Jean Paul Metzger, Karel Mokany, Simone Stevenson, Brendan A Wintle, Skipton NC Woolley, et al. (2019). "Scenarios and models to support global conservation targets". In: *Trends in ecology & evolution* 34.1, pp. 57–68 (cit. on p. 142).

Nicholson, Emily et al. (Oct. 2021). "Scientific foundations for an ecosystem goal, milestones and indicators for the post-2020 global biodiversity framework". In: *Nature Ecology & Evolution* 5.10, pp. 1338–1349. ISSN: 2397-334X. DOI: 10.1038/s41559-021-01538-5 (cit. on pp. 25, 103, 199).

Nogués-Bravo, David (2009). "Predicting the past distribution of species climatic niches". In: *Global Ecology and Biogeography* 18.5, pp. 521–531 (cit. on p. 94).

Norberg, Anna, Nerea Abrego, F Guillaume Blanchet, Frederick R Adler, Barbara J Anderson, Jani Anttila, Miguel B Araújo, Tad Dallas, David Dunson, Jane Elith, et al. (2019). "A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels". In: *Ecological monographs* 89.3, e01370 (cit. on p. 57).

Oliver, Tom H. et al. (Nov. 2015). "Biodiversity and Resilience of Ecosystem Functions". In: *Trends in Ecology & Evolution* 30.11, pp. 673–684. ISSN: 01695347. DOI: 10.1016/j.tree.2015.08.009 (cit. on p. 15).

Olson, David M. et al. (Nov. 1, 2001). "Terrestrial Ecoregions of the World: A New Map of Life on EarthA new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity". In: *BioScience* 51.11, pp. 933–938. ISSN: 0006-3568. DOI: 10.1641/0006-3568(2001)051[0933:TEOTWA]2.0.CO;2 (cit. on pp. 46, 87, 98, 203).

Orme, C. David L. et al. (Aug. 2005). "Global hotspots of species richness are not congruent with endemism or threat". In: *Nature* 436.7053, pp. 1016–1019. ISSN: 1476-4687. DOI: 10.1038/nature03850 (cit. on pp. 104, 122).

Özesmi, Stacy L. & Uygar Özesmi (Mar. 1, 1999). "An artificial neural network approach to spatial habitat modelling with interspecific interaction". In: *Ecological Modelling* 116.1, pp. 15–31. ISSN: 0304-3800. DOI: 10.1016/S0304-3800(98)00149-5 (cit. on p. 43).

Pacifici, Michela et al. (Mar. 2015). "Assessing species vulnerability to climate change". In: *Nature Climate Change* 5.3, pp. 215–224. ISSN: 1758-6798. DOI: 10.1038/nclimate2448 (cit. on pp. 2, 64).

Parding, Kajsa M. et al. (Apr. 1, 2020). "GCMeval – An interactive tool for evaluation and selection of climate model ensembles". In: *Climate Services* 18, p. 100167. ISSN: 2405-8807. DOI: 10.1016/j.cliser.2020.100167 (cit. on pp. 128, 135).

Parmesan, Camille (2006). "Ecological and Evolutionary Responses to Recent Climate Change". In: *Annual Review of Ecology, Evolution, and Systematics* 37.1, pp. 637–669. DOI: 10.1146/annurev.ecolsys.37.091305.110100 (cit. on p. 64).

Paukert, Craig P., Kristen L. Pitts, Joanna B. Whittier, & Julian D. Olden (Mar. 1, 2011). "Development and assessment of a landscape-scale ecological threat index for the Lower Colorado River Basin". In: *Ecological Indicators* 11.2, pp. 304–310. ISSN: 1470-160X. DOI: 10.1016/j.ecolind.2010.05.008 (cit. on p. 104).

Pearce, Jennie & Simon Ferrier (2000). "An evaluation of alternative algorithms for fitting species distribution models using logistic regression". In: *Ecological modelling* 128.2-3, pp. 127–147 (cit. on p. 43).

Pecl, Gretta T. et al. (Mar. 31, 2017). "Biodiversity redistribution under climate change: Impacts on ecosystems and human well-being". In: *Science* 355.6332. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aai9214 (cit. on pp. 64, 71).

Pelletier, Tara A., Bryan C. Carstens, David C. Tank, Jack Sullivan, & Anahí Espíndola (Dec. 18, 2018). "Predicting plant conservation priorities on a global scale". In: *Proceedings of the National Academy of Sciences* 115.51, pp. 13027–13032. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1804098115 (cit. on pp. 62, 127).

Pereira, Henrique M, Paul W Leadley, Vânia Proença, Rob Alkemade, Jörn PW Scharlemann, Juan F Fernandez-Manjarrés, Miguel B Araújo, Patricia Balvanera, Reinette Biggs, William WL Cheung, et al. (2010). "Scenarios for global biodiversity in the 21st century". In: *Science* 330.6010, pp. 1496–1501 (cit. on pp. 65, 128).

Pereira, Henrique M. et al. (Jan. 18, 2013). "Essential Biodiversity Variables". In: *Science* 339.6117, pp. 277–278. DOI: 10.1126/science.1229931 (cit. on p. 27).

Pergent-Martini, Christine, Vanina Leoni, Vanina Pasqualini, GD Ardizzone, Elena Balestri, Roberto Bedini, Andrea Belluscio, Thomas Belsher, Joseph Borg, CF Boudouresque, et al. (2005). "Descriptors of Posidonia oceanica meadows: use and application". In: *Ecological Indicators* 5.3, pp. 213–230 (cit. on p. 16).

Persello, Claudio, Jan Dirk Wegner, Ronny Hänsch, Devis Tuia, Pedram Ghamisi, Mila Koeva, & Gustau Camps-Valls (2022). "Deep learning and earth observation to support the sustainable development goals: Current approaches, open challenges, and future opportunities". In: *IEEE Geoscience and Remote Sensing Magazine* 10.2, pp. 172–200 (cit. on p. 62).

Peterson, A Townsend, Victor Sánchez-Cordero, C Ben Beard, & Janine M Ramsey (2002). "Ecologic niche modeling and potential reservoirs for Chagas disease, Mexico." In: *Emerging infectious diseases* 8.7, p. 662 (cit. on p. 48).

Peterson, Garry D, Graeme S Cumming, & Stephen R Carpenter (2003). "Scenario planning: a tool for conservation in an uncertain world". In: *Conservation biology* 17.2, pp. 358–366 (cit. on pp. 65, 128).

Pettorelli, Nathalie, Sadie Ryan, Thomas Mueller, Nils Bunnefeld, Bogumila Jędrzejewska, Mauricio Lima, & Kyrre Kausrud (2011). "The Normalized Difference Vegetation Index (NDVI): unforeseen successes in animal ecology". In: *Climate research* 46.1, pp. 15–27 (cit. on p. 93).

Phillips, Steven J., Robert P Anderson, & Robert E Schapire (2006). "Maximum entropy modeling of species geographic distributions". In: *Ecological modelling* 190.3-4, pp. 231–259 (cit. on pp. 43, 96).

Phillips, Steven J. & Miroslav Dudik (2008). "Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation". In: *Ecography* 31.2, pp. 161–175 (cit. on pp. 43, 71).

Phillips, Steven J., Miroslav Dudík, Jane Elith, Catherine H. Graham, Anthony Lehmann, John R. Leathwick, & Simon Ferrier (2009). "Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data". In: *Ecological Applications* 19.1, pp. 181–197. ISSN: 1939-5582. DOI: 10.1890/07-2153.1 (cit. on pp. 42, 82, 194).

Pichler, Maximilian & Florian Hartig (2023). "Machine learning and deep learning—A review for ecologists". In: *Methods in Ecology and Evolution* n/a (n/a). ISSN: 2041-210X. DOI: 10.1111/2041-210X.14061 (cit. on pp. 44, 48–52, 54, 59, 148, 150).

Pimm, Stuart L., Sky Alibhai, Richard Bergl, Alex Dehgan, Chandra Giri, Zoë Jewell, Lucas Joppa, Roland Kays, & Scott Loarie (Nov. 1, 2015). "Emerging Technologies to Conserve Biodiversity". In: *Trends in Ecology & Evolution* 30.11, pp. 685–696. ISSN: 0169-5347. DOI: 10.1016/j.tree.2015.08.008 (cit. on p. 3).

Pimm, Stuart L., C. N. Jenkins, R. Abell, T. M. Brooks, J. L. Gittleman, L. N. Joppa, P. H. Raven, C. M. Roberts, & J. O. Sexton (May 30, 2014). "The biodiversity of species and their rates of extinction, distribution, and protection". In: *Science* 344.6187. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1246752 (cit. on pp. 2, 3, 19).

Pimm, Stuart L. & Lucas N. Joppa (Mar. 16, 2015). "How Many Plant Species are There, Where are They, and at What Rate are They Going Extinct?" In: *Annals of the Missouri Botanical Garden* 100.3, pp. 170–176. ISSN: 0026-6493, 2162-4372. DOI: 10.3417/2012018 (cit. on p. 127).

Pinborg, Ulla (2002). "An inventory of biodiversity indicators in Europe". In: (cit. on p. 8).

Poggio, Laura, Luis M. de Sousa, Niels H. Batjes, Gerard B. M. Heuvelink, Bas Kempen, Eloi Ribeiro, & David Rossiter (June 14, 2021). "SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty". In: *SOIL* 7.1, pp. 217–240. ISSN: 2199-3971. DOI: 10.5194/soil-7-217-2021 (cit. on pp. 45, 203).

Poggio, Tomaso, Hrushikesh Mhaskar, Lorenzo Rosasco, Brando Miranda, & Qianli Liao (Oct. 1, 2017). "Why and when can deep-but not shallow-networks avoid the curse of dimensionality: A review". In: *International Journal of Automation and Computing* 14.5, pp. 503–519. ISSN: 1751-8520. DOI: 10.1007/s11633-017-1054-2 (cit. on pp. 6, 51, 131).

Pollock, Laura J., Louise M. J. O'Connor, Karel Mokany, Dan F. Rosauer, Matthew V. Talluto, & Wilfried Thuiller (Dec. 1, 2020). "Protecting Biodiversity (in All Its Complexity): New Models and Methods". In: *Trends in Ecology & Evolution* 35.12, pp. 1119–1128. ISSN: 0169-5347. DOI: 10.1016/j.tree.2020.08.015 (cit. on pp. 4, 8, 20, 21, 103, 144–146, 148, 150, 151).

Pollock, Laura J., Wilfried Thuiller, & Walter Jetz (June 2017). "Large conservation gains possible for global biodiversity facets". In: *Nature* 546.7656, pp. 141–144. ISSN: 1476-4687. DOI: 10.1038/nature22368 (cit. on p. 146).

Pollock, Laura J., Reid Tingley, William K. Morris, Nick Golding, Robert B. O'Hara, Kirsten M. Parris, Peter A. Vesk, & Michael A. McCarthy (2014). "Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM)". In: *Methods in Ecology and Evolution* 5.5, pp. 397–406. ISSN: 2041-210X. DOI: 10.1111/2041-210X.12180 (cit. on pp. 42, 43, 104).

Powell-Romero, Francisca, Nicholas M Fountain-Jones, Anna Norberg, & Nicholas J Clark (2023). "Improving the predictability and interpretability of co-occurrence modelling through feature-based joint species distribution ensembles". In: *Methods in Ecology and Evolution* 14.1, pp. 146–161 (cit. on p. 104).

Proença, Vânia, Laura Jane Martin, Henrique M. Pereira, Miguel Fernandez, Louise McRae, Jayne Belnap, Monika Böhm, Neil A. Brummitt, Jaime Garcia-Moreno, Richard D Gregory, et al. (2017). "Global biodiversity monitoring: from data sources to essential biodiversity variables". In: *Biological Conservation* 213, pp. 256–263 (cit. on p. 27).

Puglielli, Giacomo & Meelis Pärtel (2023). "Macroecology of plant diversity across spatial scales". In: *New Phytologist* 237.4, pp. 1074–1077. ISSN: 1469-8137. DOI: 10.1111/nph.18680 (cit. on pp. 104, 120).

Purvis, Andy, John L Gittleman, Guy Cowlishaw, & Georgina M Mace (2000). "Predicting extinction risk in declining species". In: *Proceedings of the royal society of London. Series B: Biological Sciences* 267.1456, pp. 1947–1952 (cit. on p. 62).

Purvis, Andy & Andy Hector (May 2000). "Getting the measure of biodiversity". In: *Nature* 405.6783, pp. 212–219. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/35012221 (cit. on p. 2).

Raghu, Maithra, Ben Poole, Jon Kleinberg, Surya Ganguli, & Jascha Sohl-Dickstein (July 17, 2017). "On the Expressive Power of Deep Neural Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, pp. 2847–2854 (cit. on p. 50).

Raissi, Maziar, Paris Perdikaris, & George Em Karniadakis (Nov. 3, 2018). "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations". In: *Journal of Computational Physics* 378 (C). ISSN: 0021-9991. DOI: 10.1016/j.jcp.2018.10.045 (cit. on pp. 61, 150).

Randin, Christophe F. et al. (Mar. 15, 2020). "Monitoring biodiversity in the Anthropocene using remote sensing in species distribution models". In: *Remote Sensing of Environment* 239, p. 111626. ISSN: 0034-4257. DOI: 10.1016/j.rse.2019.111626 (cit. on pp. 5, 46, 74, 93, 94, 97).

Rapacciuolo, Giovanni (2019). "Strengthening the contribution of macroecological models to conservation practice". In: *Global Ecology and Biogeography* 28.1, pp. 54–60 (cit. on p. 147).

Reeb, Rachel A, Naeem Aziz, Samuel M Lapp, Justin Kitzes, J Mason Heberling, & Sara E Kuebbing (2022). "Using convolutional neural networks to efficiently extract immense

phenological data from community science images". In: *Frontiers in Plant Science* 12, p. 3148 (cit. on p. 61).

Regan, Eugenie C, Luca Santini, Lisa Ingwall-King, Michael Hoffmann, Carlo Rondinini, Andy Symes, Joseph Taylor, & Stuart H. M. Butchart (2015). "Global trends in the status of bird and mammal pollinators". In: *Conservation letters* 8.6, pp. 397–403 (cit. on p. 20).

Reichstein, Markus, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, & Prabhat (Feb. 2019). "Deep learning and process understanding for data-driven Earth system science". In: *Nature* 566.7743, pp. 195–204. ISSN: 1476-4687. DOI: 10.1038/s41586-019-0912-1 (cit. on p. 61).

Remm, Kalle & Liina Remm (2009). "Similarity-based large-scale distribution mapping of orchids". In: *Biodiversity and Conservation* 18.6, pp. 1629–1647 (cit. on p. 73).

Renner, Susanne S. & Constantin M. Zohner (2018). "Climate Change and Phenological Mismatch in Trophic Interactions Among Plants, Insects, and Vertebrates". In: *Annual Review of Ecology, Evolution, and Systematics* 49.1, pp. 165–182. DOI: 10.1146/annurev-ecolsys-110617-062535 (cit. on p. 64).

Reyers, Belinda, Reinette Biggs, Graeme S Cumming, Thomas Elmqvist, Adam P Hejnowicz, & Stephen Polasky (2013). "Getting the measure of ecosystem services: a social–ecological approach". In: *Frontiers in Ecology and the Environment* 11.5, pp. 268–273. ISSN: 1540-9309. DOI: 10.1890/120144 (cit. on p. 126).

Ribeiro, Marco Tulio, Sameer Singh, & Carlos Guestrin (2016). "" Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144 (cit. on p. 147).

Ricotta, Carlo (Apr. 1, 2005). "Through the Jungle of Biological Diversity". In: *Acta Biotheoretica* 53.1, pp. 29–38. ISSN: 1572-8358. DOI: 10.1007/s10441-005-7001-6 (cit. on p. 109).

Ritchie, Hannah, Fiona Spooner, & Max Roser (Dec. 19, 2022). "Biodiversity". In: *Our World in Data* (cit. on pp. 22, 23).

Rivers, Malin C, Neil A. Brummitt, Eimear Nic Lughadha, & Thomas R Meagher (2014). "Do species conservation assessments capture genetic diversity?" In: *Global ecology and conservation* 2, pp. 81–87 (cit. on p. 19).

Roberts, David R, Volker Bahn, Simone Ciuti, Mark S Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, et al. (2017). "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure". In: *Ecography* 40.8, pp. 913–929 (cit. on pp. 81, 107, 191).

Rocchini, Duccio et al. (May 3, 2023). "A quixotic view of spatial bias in modelling the distribution of species and their diversity". In: *npj Biodiversity* 2.1, pp. 1–11. ISSN: 2731-4243. DOI: 10.1038/s44185-023-00014-6 (cit. on p. 129).

Rodrigues, Ana S. L., Thomas M Brooks, Stuart H. M. Butchart, Janice Chanson, Neil Cox, Michael Hoffmann, & Simon N Stuart (2014). "Spatially explicit trends in the global conservation status of vertebrates". In: *Plos one* 9.11, e113934 (cit. on p. 20).

Rodrigues, Ana S. L. & Thomas M. Brooks (2007). "Shortcuts for Biodiversity Conservation Planning: The Effectiveness of Surrogates". In: *Annual Review of Ecology, Evolution, and Systematics* 38.1, pp. 713–737. DOI: 10.1146/annurev.ecolsys.38.091206.095737 (cit. on p. 11).

Rodrigues, Ana S. L. & Victor Cazalis (2020). "The multifaceted challenge of evaluating protected area effectiveness". In: *Nature Communications* 11.1, p. 5147 (cit. on p. 66).

Rodrigues, Ana S. L., J Pilgrim, J Lamoreux, M Hoffmann, & T Brooks (Feb. 2006). "The value of the IUCN Red List for conservation". In: *Trends in Ecology & Evolution* 21.2, pp. 71–76. ISSN: 01695347. DOI: 10.1016/j.tree.2005.10.010 (cit. on p. 16).

Rodrigues, Ana S. L. et al. (2004). "Global Gap Analysis: Priority Regions for Expanding the Global Protected-Area Network". In: *BioScience* 54.12, p. 1092. ISSN: 0006-3568. DOI: 10.1641/0006-3568(2004)054[1092:GGAPRF]2.0.CO;2 (cit. on p. 66).

Rolnick, David, Priya L Donti, Lynn H Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, et al. (2022). "Tackling climate change with machine learning". In: *ACM Computing Surveys (CSUR)* 55.2, pp. 1–96 (cit. on p. 65).

Rolnick, David, Andreas Veit, Serge Belongie, & Nir Shavit (2017). "Deep learning is robust to massive label noise". In: *arXiv preprint arXiv:1705.10694* (cit. on pp. 97, 129).

Rondinini, Carlo, Moreno Di Marco, Piero Visconti, Stuart H. M. Butchart, & Luigi Boitani (2014). "Update or Outdate: Long-Term Viability of the IUCN Red List". In: *Conservation Letters* 7.2, pp. 126–130. ISSN: 1755-263X. DOI: 10.1111/conl.12040 (cit. on pp. 10, 19).

Rossi, Francesca, Peter Van Beek, & Toby Walsh (2006). *Handbook of constraint programming*. Elsevier (cit. on p. 67).

Rounsevell, Mark D. A., Mike Harfoot, Paula A. Harrison, Tim Newbold, Richard D. Gregory, & Georgina M. Mace (June 12, 2020). "A biodiversity target based on species extinctions". In: *Science* 368.6496, pp. 1193–1195. DOI: 10.1126/science.aba6592 (cit. on pp. 34, 126).

Rowland, Jessica A, Lucie M Bland, David A Keith, Diego Juffe-Bignoli, Mark A Burgman, Andres Etter, José Rafael Ferrer-Paris, Rebecca M Miller, Andrew L Skowno, & Emily Nicholson (2020). "Ecosystem indices to support global biodiversity conservation". In: *Conservation Letters* 13.1, e12680 (cit. on p. 26).

Ruder, Sebastian (2016). "An overview of gradient descent optimization algorithms". In: *arXiv preprint arXiv:1609.04747* (cit. on p. 51).

Rußwurm, Marc & Marco Körner (2018). "Multi-temporal land cover classification with sequential recurrent encoders". In: *ISPRS International Journal of Geo-Information* 7.4, p. 129 (cit. on pp. 76, 98, 148).

Ryo, Masahiro, Boyan Angelov, Stefano Mammola, Jamie M. Kass, Blas M. Benito, & Florian Hartig (Feb. 2021). "Explainable artificial intelligence enhances the ecological interpretability of black-box species distribution models". In: *Ecography* 44.2, pp. 199–205. ISSN: 0906-7590, 1600-0587. DOI: 10.1111/ecog.05360 (cit. on pp. 44, 52, 121, 129, 147).

Safi, Kamran, Katrina Armour-Marshall, Jonathan EM Baillie, & Nick JB Isaac (2013). "Global patterns of evolutionary distinct and globally endangered amphibians and mammals". In: *PloS one* 8.5, e63582 (cit. on p. 30).

Saiz, Juan Carlos Moreno, Felipe Dominguez Lozano, Manuel Marrero Gómez, & Ángel Bañares Baudet (2015). "Application of the Red List Index for conservation assessment of Spanish vascular plants". In: *Conservation Biology* 29.3, pp. 910–919 (cit. on p. 20).

Sala, Osvaldo E. et al. (Mar. 10, 2000). "Global Biodiversity Scenarios for the Year 2100". In: *Science* 287.5459, pp. 1770–1774. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.287.5459.1770 (cit. on pp. 2, 65).

Santini, Luca, Stuart H. M. Butchart, Carlo Rondinini, Ana Benítez-López, Jelle P. Hilbers, Aafke M. Schipper, Mirza Cengic, Joseph A. Tobias, & Mark A. J. Huijbregts (2019). "Applying habitat and population-density models to land-cover time series to inform IUCN Red List assessments". In: *Conservation Biology* 33.5, pp. 1084–1093. ISSN: 1523-1739. DOI: https://doi.org/10.1111/cobi.13279 (cit. on p. 47).

Schatz, George E. (Nov. 1, 2009). "Plants on the IUCN Red List: setting priorities to inform conservation". In: *Trends in Plant Science*. Special Issue: Plant science research in botanic gardens 14.11, pp. 638–642. ISSN: 1360-1385. DOI: 10.1016/j.tplants.2009.08.012 (cit. on p. 103).

Scheibenreif, Linus, Joëlle Hanna, Michael Mommert, & Damian Borth (2022). "Self-supervised vision transformers for land-cover segmentation and classification". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1422–1431 (cit. on pp. 46, 150).

Schloss, Carrie A., Tristan A. Nuñez, & Joshua J. Lawler (May 29, 2012). "Dispersal will limit ability of mammals to track climate change in the Western Hemisphere". In: *Proceedings of the National Academy of Sciences* 109.22, pp. 8606–8611. DOI: 10.1073/pnas.1116791109 (cit. on p. 132).

Schmidt, Marco, Thomas Janßen, Stefan Dressler, Karen Hahn, Mipro Hien, Souleymane Konaté, Anne Mette Lykke, Ali Mahamane, Bienvenu Sambou, Brice Sinsin, et al. (2012). "The West African Vegetation Database". In: *Biodiversity and Ecology* 4, pp. 105–110 (cit. on p. 38).

Schmitt, M., L. H. Hughes, C. Qiu, & X. X. Zhu (Sept. 16, 2019). "SEN12MS – A CURATED DATASET OF GEOREFERENCED MULTI-SPECTRAL SENTINEL-1/2 IMAGERY FOR DEEP LEARNING AND DATA FUSION". In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* IV-2/W7, pp. 153–160. ISSN: 2194-9050. DOI: 10.5194/isprs-annals-IV-2-W7-153-2019 (cit. on p. 95).

Schmitt, Sylvain, Robin Pouteau, Dimitri Justeau, Florian de Boissieu, & Philippe Birnbaum (2017). "ssdm: An r package to predict distribution of species richness and composition based on stacked species distribution models". In: *Methods in Ecology and Evolution* 8.12, pp. 1795–1803. ISSN: 2041-210X. DOI: https://doi.org/10.1111/2041-210X.12841 (cit. on p. 42).

Schölkopf, Bernhard & Alexander J. Smola (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press. 658 pp. ISBN: 978-0-262-19475-4 (cit. on p. 43).

Schratz, Patrick, Jannes Muenchow, Eugenia Iturritxa, Jakob Richter, & Alexander Brenning (2019). "Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data". In: *Ecological Modelling* 406, pp. 109–120 (cit. on p. 147).

Schwalm, Christopher R., Spencer Glendon, & Philip B. Duffy (Aug. 18, 2020). "RCP8.5 tracks cumulative CO2 emissions". In: *Proceedings of the National Academy of Sciences* 117.33, pp. 19656–19657. DOI: 10.1073/pnas.2007117117 (cit. on p. 128).

Schwartz, Mark W. (Oct. 2012). "Using niche models with climate projections to inform conservation management decisions". In: *Biological Conservation* 155, pp. 149–156. ISSN: 00063207. DOI: 10.1016/j.biocon.2012.06.011 (cit. on p. 65).

Schweiger, Anna K., Jeannine Cavender-Bares, Philip A. Townsend, Sarah E. Hobbie, Michael D. Madritch, Ran Wang, David Tilman, & John A. Gamon (June 2018). "Plant spectral diversity integrates functional and phylogenetic components of biodiversity and predicts ecosystem function". In: *Nature Ecology & Evolution* 2.6, pp. 976–982. ISSN: 2397-334X. DOI: 10.1038/s41559-018-0551-1 (cit. on p. 4).

Sejnowski, Terrence J (2020). "The unreasonable effectiveness of deep learning in artificial intelligence". In: *Proceedings of the National Academy of Sciences* 117.48, pp. 30033–30038 (cit. on p. 50).

Shannon, C. E. (July 1948). "A mathematical theory of communication". In: *The Bell System Technical Journal* 27.3, pp. 379–423. ISSN: 0005-8580. DOI: 10.1002/j.1538-7305.1948.tb01338.x (cit. on pp. 24, 43, 83, 109).

Shorten, Connor & Taghi M Khoshgoftaar (2019). "A survey on image data augmentation for deep learning". In: *Journal of big data* 6.1, pp. 1–48 (cit. on p. 50).

Silva, Sandro Valerio, Tobias Andermann, Alexander Zizka, Gregor Kozlowski, & Daniele Silvestro (2022). "Global Estimation and Mapping of the Conservation Status of Tree Species

Using Artificial Intelligence." In: *Frontiers in Plant Science* 13. In collab. with Sandro Valerio Silva, Tobias Andermann, Alexander Zizka, Gregor Kozlowski, & Daniele Silvestro, p. 839792. ISSN: 1664-462X (cit. on p. 63).

Silvestro, Daniele, Stefano Goria, Thomas Sterner, & Alexandre Antonelli (May 2022). "Improving biodiversity protection through artificial intelligence". In: *Nature Sustainability* 5.5, pp. 415–424. ISSN: 2398-9629. DOI: 10.1038/s41893-022-00851-6 (cit. on pp. 66, 122).

Soberón, Jorge, J Golubov, & José Sarukhán (2001). "The importance of Opuntia in Mexico and routes of invasion and impact of Cactoblastis cactorum (Lepidoptera: Pyralidae)". In: *Florida Entomologist*, pp. 486–492 (cit. on p. 48).

Soberón, Jorge & Miguel Nakamura (Nov. 17, 2009). "Niches and distributional areas: Concepts, methods, and assumptions". In: *Proceedings of the National Academy of Sciences* 106 (supplement_2), pp. 19644–19650. DOI: 10.1073/pnas.0901637106 (cit. on p. 40).

Sofaer, Helen R et al. (July 1, 2019). "Development and Delivery of Species Distribution Models to Inform Decision-Making". In: *BioScience* 69.7, pp. 544–557. ISSN: 0006-3568. DOI: 10.1093/biosci/biz045 (cit. on pp. 47, 52).

Sokolova, Marina & Guy Lapalme (2009). "A systematic analysis of performance measures for classification tasks". In: *Information processing & management* 45.4, pp. 427–437 (cit. on p. 82).

Song, Conghe, Curtis E. Woodcock, Karen C. Seto, Mary Pax Lenney, & Scott A. Macomber (Feb. 1, 2001). "Classification and Change Detection Using Landsat TM Data: When and How to Correct Atmospheric Effects?" In: *Remote Sensing of Environment* 75.2, pp. 230–244. ISSN: 0034-4257. DOI: 10.1016/S0034-4257(00)00169-3 (cit. on p. 76).

Sonnewald, Maike, Stephanie Dutkiewicz, Christopher Hill, & Gael Forget (2020). "Elucidating ecological complexity: Unsupervised learning determines global marine eco-provinces". In: *Science advances* 6.22, eaay4740 (cit. on p. 150).

Soranno, Patricia A., Kendra S. Cheruvelil, Kevin C. Elliott, & Georgina M. Montgomery (Jan. 1, 2015). "It's Good to Share: Why Environmental Scientists' Ethics Are Out of Date". In: *BioScience* 65.1, pp. 69–73. ISSN: 0006-3568. DOI: 10.1093/biosci/biu169 (cit. on p. 4).

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, & Ruslan Salakhutdinov (2014). "Dropout: a simple way to prevent neural networks from overfitting". In: *The journal of machine learning research* 15.1, pp. 1929–1958 (cit. on pp. 50, 81).

Stanley, Samantha K., Teaghan L. Hogg, Zoe Leviston, & Iain Walker (Mar. 1, 2021). "From anger to action: Differential impacts of eco-anxiety, eco-depression, and eco-anger on climate action and wellbeing". In: *The Journal of Climate Change and Health* 1, p. 100003. ISSN: 2667-2782. DOI: 10.1016/j.joclim.2021.100003 (cit. on p. 2).

Stanton, Jessica C., Kevin T. Shoemaker, Richard G. Pearson, & H. Resit Akçakaya (2015). "Warning times for species extinctions due to climate change". In: *Global Change Biology* 21.3, pp. 1066–1077. ISSN: 1365-2486. DOI: https://doi.org/10.1111/gcb.12721 (cit. on p. 10).

Stévart, T. et al. (Nov. 2019). "A third of the tropical African flora is potentially threatened with extinction". In: *Science Advances* 5.11, eaax9444. ISSN: 2375-2548. DOI: 10.1126/sciadv.aax9444 (cit. on pp. 47, 104, 127).

Stewart, Adam J., Caleb Robinson, Isaac A. Corley, Anthony Ortiz, Juan M. Lavista Ferres, & Arindam Banerjee (Nov. 16, 2021). "TorchGeo: deep learning with geospatial data". In: *arXiv:2111.08872 [cs]* (cit. on pp. 57, 73, 95).

Stirling, Andy (2007). "A general framework for analysing diversity in science, technology and society". In: *Journal of the Royal Society interface* 4.15, pp. 707–719 (cit. on p. 24).

Stupariu, Mihai-Sorin, Samuel A Cushman, Alin-Ionuţ Pleşoianu, Ileana Pătru-Stupariu, & Christine Fuerst (2022). "Machine learning in landscape ecological analysis: a review of recent approaches". In: *Landscape Ecology* 37.5, pp. 1227–1250 (cit. on p. 60).

Sullivan, Brian L, Christopher L Wood, Marshall J Iliff, Rick E Bonney, Daniel Fink, & Steve Kelling (2009). "eBird: A citizen-based bird observation network in the biological sciences". In: *Biological conservation* 142.10, pp. 2282–2292 (cit. on p. 38).

Sumbul, Gencer, Marcela Charfuelan, Begüm Demir, & Volker Markl (July 2019). "Bigearthnet: A Large-Scale Benchmark Archive for Remote Sensing Image Understanding". In: *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*. IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, pp. 5901–5904. DOI: 10.1109/IGARSS.2019.8900532 (cit. on p. 95).

Sundseth, Kerstin & Peter Creed (2008). *Natura 2000: protecting Europe's biodiversity*. European Commission Luxembourg (cit. on p. 27).

Sutton, Richard S & Andrew G Barto (2018). *Reinforcement learning: An introduction*. MIT press (cit. on p. 67).

Svenning, Jens-Christian, Camilla Fløjgaard, Katharine A Marske, David Nógues-Bravo, & Signe Normand (2011). "Applications of species distribution modeling to paleobiology". In: *Quaternary Science Reviews* 30.21-22, pp. 2930–2947 (cit. on p. 94).

Swarts, Nigel D. & Kingsley W. Dixon (Aug. 2009). "Terrestrial orchid conservation in the age of extinction". In: *Annals of Botany* 104.3, pp. 543–556. ISSN: 1095-8290, 0305-7364. DOI: 10.1093/aob/mcp025 (cit. on pp. 11, 121).

Syfert, Mindy M., Lucas Joppa, Matthew J. Smith, David A. Coomes, Steven P. Bachman, & Neil A. Brummitt (Sept. 1, 2014). "Using species distribution models to inform IUCN Red List assessments". In: *Biological Conservation* 177, pp. 174–184. ISSN: 0006-3207. DOI: 10.1016/j.biocon.2014.06.012 (cit. on pp. 47, 104).

Szegedy, Christian, Sergey Ioffe, Vincent Vanhoucke, & Alex Alemi (Aug. 23, 2016a). "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning". In: *arXiv:1602.07261 [cs]* (cit. on p. 131).

Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, & Zbigniew Wojna (June 2016b). "Rethinking the Inception Architecture for Computer Vision". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, pp. 2818–2826. ISBN: 978-1-4673-8851-1. DOI: 10.1109/CVPR.2016.308 (cit. on pp. 80, 107, 191).

Tachikawa, Tetsushi, Masami Hato, Manabu Kaku, & Akira Iwasaki (2011). "Characteristics of ASTER GDEM version 2". In: *2011 IEEE international geoscience and remote sensing symposium*. IEEE, pp. 3657–3660 (cit. on p. 87).

Thomas, Chris D. et al. (Jan. 2004). "Extinction risk from climate change". In: *Nature* 427.6970, pp. 145–148. ISSN: 1476-4687. DOI: 10.1038/nature02121 (cit. on pp. 65, 127, 132).

Thuiller, Wilfried, Guy F. Midgley, Mathieu Rougeti, & Richard M. Cowling (2006a). "Predicting patterns of plant species richness in megadiverse South Africa". In: *Ecography* 29.5, pp. 733–744 (cit. on p. 84).

Thuiller, Wilfried, Bruno Lafourcade, Robin Engler, & Miguel B. Araújo (2009). "BIOMOD – a platform for ensemble forecasting of species distributions". In: *Ecography* 32.3, pp. 369–373. ISSN: 1600-0587. DOI: https://doi.org/10.1111/j.1600-0587.2008.05742.x (cit. on p. 71).

Thuiller, Wilfried, Sébastien Lavergne, Cristina Roquet, Isabelle Boulangeat, Bruno Lafourcade, & Miguel B. Araújo (2011). "Consequences of climate change on the tree of life in Europe". In: *Nature* 470.7335, pp. 531–534 (cit. on p. 65).

Thuiller, Wilfried, Sandra Lavorel, Miguel B. Araújo, Martin T. Sykes, & I. Colin Prentice (June 7, 2005). "Climate change threats to plant diversity in Europe". In: *Proceedings of the National Academy of Sciences* 102.23, pp. 8245–8250. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0409902102 (cit. on p. 65).

Thuiller, Wilfried, Sandra Lavorel, Martin T Sykes, & Miguel B Araújo (2006b). "Using niche-based modelling to assess the impact of climate change on tree functional diversity in Europe". In: *Diversity and Distributions* 12.1, pp. 49–60 (cit. on p. 65).

Todman, Lindsay C, Alex Bush, & Amelia SC Hood (2023). "'Small Data'for big insights in ecology". In: *Trends in Ecology & Evolution* (cit. on pp. 61, 65, 150).

Torrey, Lisa & Jude Shavlik (2010). "Transfer learning". In: *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI global, pp. 242–264 (cit. on p. 96).

Trolliet, Franck, Cédric Vermeulen, Marie-Claude Huynen, & Alain Hambuckers (2014). "Use of camera traps for wildlife studies: a review". In: *Biotechnologie, Agronomie, Société et Environnement* 18.3 (cit. on p. 38).

Troudet, Julien, Philippe Grandcolas, Amandine Blin, Régine Vignes-Lebbe, & Frédéric Legendre (2017). "Taxonomic bias in biodiversity data and societal preferences". In: *Scientific reports* 7.1, p. 9132 (cit. on p. 39).

Tsiftsis, Spyros & Ioannis Tsiripidis (Oct. 1, 2020). "Temporal and spatial patterns of orchid species distribution in Greece: implications for conservation". In: *Biodiversity and Conservation* 29.11, pp. 3461–3489. ISSN: 1572-9710. DOI: 10.1007/s10531-020-02035-0 (cit. on p. 11).

Tuia, Devis et al. (Feb. 9, 2022). "Perspectives in machine learning for wildlife conservation". In: *Nature Communications* 13.1, p. 792. ISSN: 2041-1723. DOI: 10.1038/s41467-022-27980-y (cit. on p. 60).

Tulloch, Vivitskaia JD., Hugh P Possingham, Stacy D Jupiter, Chris Roelfsema, Ayesha IT Tulloch, & Carissa J Klein (2013). "Incorporating uncertainty associated with habitat data in marine reserve design". In: *Biological Conservation* 162, pp. 41–51 (cit. on p. 33).

Tulloch, Vivitskaia JD. et al. (2015). "Why do we map threats? Linking threat mapping with actions to make better conservation decisions". In: *Frontiers in Ecology and the Environment* 13.2, pp. 91–99. ISSN: 1540-9309. DOI: 10.1890/140022 (cit. on pp. 31–33).

Unger, Shem, Mark Rollins, Allison Tietz, & Hailey Dumais (2021). "iNaturalist as an engaging tool for identifying organisms in outdoor activities". In: *Journal of Biological Education* 55.5, pp. 537–547 (cit. on pp. 38, 61).

Urban, Mark C. (May 2015). "Accelerating extinction risk from climate change". In: *Science* 348.6234, pp. 571–573. DOI: 10.1126/science.aaa4984 (cit. on pp. 10, 64, 65, 127, 132, 139).

Valavi, Roozbeh, Jane Elith, José J. Lahoz-Monfort, & Gurutzeta Guillera-Arroita (2023). "Flexible species distribution modelling methods perform well on spatially separated testing data". In: *Global Ecology and Biogeography* n/a (n/a). ISSN: 1466-8238. DOI: 10.1111/geb.13639 (cit. on p. 120).

Valavi, Roozbeh, Gurutzeta Guillera-Arroita, José J. Lahoz-Monfort, & Jane Elith (2022). "Predictive performance of presence-only species distribution models: a benchmark study with reproducible code". In: *Ecological Monographs* 92.1, e01486. ISSN: 1557-7015. DOI: 10.1002/ecm.1486 (cit. on pp. 43, 52, 129).

Valderrábano, Marcos, Cara Nelson, Emily Nicholson, Andrés Etter, Josie Carwardine, James G Hallett, James McBreen, & Emily Botts (2021). "Using ecosystem risk assessment science in ecosystem restoration: a guide to applying the Red List of Ecosystems to ecosystem restoration". In: *IUCN, Gland* (cit. on p. 26).

Vallecillo, Sara, Joachim Maes, Chiara Polce, & Carlo Lavalle (Oct. 1, 2016). "A habitat quality indicator for common birds in Europe based on species distribution models". In: *Ecological Indicators* 69, pp. 488–499. ISSN: 1470-160X. DOI: 10.1016/j.ecolind.2016.05.008 (cit. on p. 16).

Van Vuuren, Detlef P. et al. (Nov. 2011). "The representative concentration pathways: an overview". In: *Climatic Change* 109.1, pp. 5–31. ISSN: 0165-0009, 1573-1480. DOI: 10.1007/s10584-011-0148-z (cit. on p. 128).

Venter, Oscar et al. (Aug. 23, 2016). "Global terrestrial Human Footprint maps for 1993 and 2009". In: *Scientific Data* 3.1, p. 160067. ISSN: 2052-4463. DOI: 10.1038/sdata.2016.67 (cit. on pp. 32, 46, 87, 104, 203, 214).

Verones, Francesca et al. (Sept. 1, 2022). "Global extinction probabilities of terrestrial, freshwater, and marine species groups for use in Life Cycle Assessment". In: *Ecological Indicators* 142, p. 109204. ISSN: 1470-160X. DOI: 10.1016/j.ecolind.2022.109204 (cit. on pp. 29, 104, 199).

Vieilledent, Ghislain, Christelle Vancutsem, Clément Bourgoin, Pierre Ploton, Philippe Verley, & Frédéric Achard (July 23, 2022). *Spatial scenario of tropical deforestation and carbon emissions for the 21st century*. DOI: 10.1101/2022.03.22.485306 (cit. on pp. 142, 149).

Visconti, Piero et al. (2016). "Projecting Global Biodiversity Indicators under Future Development Scenarios". In: *Conservation Letters* 9.1, pp. 5–13. ISSN: 1755-263X. DOI: 10.1111/conl.12159 (cit. on p. 65).

Vitt, Pati, Amanda Taylor, Demetra Rakosy, Holger Kreft, Abby Meyer, Patrick Weigelt, & Tiffany M. Knight (Apr. 25, 2023). "Global conservation prioritization for the Orchidaceae". In: *Scientific Reports* 13.1, p. 6718. ISSN: 2045-2322. DOI: 10.1038/s41598-023-30177-y (cit. on pp. 11, 20, 146, 199).

Vogt-Schilb, Hélène, François Munoz, Franck Richard, & Bertrand Schatz (Oct. 1, 2015). "Recent declines and range changes of orchids in Western Europe (France, Belgium and Luxembourg)". In: *Biological Conservation* 190, pp. 133–141. ISSN: 0006-3207. DOI: 10.1016/j.biocon.2015.05.002 (cit. on p. 12).

Vuuren, Detlef P van, Osvaldo E Sala, & Henrique M Pereira (2006). "The future of vascular plant diversity under four global scenarios". In: *Ecology and Society* 11.2 (cit. on p. 140).

Wagner, Fabien H, Alber Sanchez, Yuliya Tarabalka, Rodolfo G Lotte, Matheus P Ferreira, Marcos PM Aidar, Emanuel Gloor, Oliver L Phillips, & Luiz EOC Aragao (2019). "Using the U-net convolutional network to map forest types and disturbance in the Atlantic rainforest with very high resolution images". In: *Remote Sensing in Ecology and Conservation* 5.4, pp. 360–375 (cit. on p. 73).

Walker, Barnaby E., Tarciso C. C. Leão, Steven P. Bachman, Friederike C. Bolam, & Eimear Nic Lughadha (2020). "Caution Needed When Predicting Species Threat Status for Conservation Prioritization on a Global Scale". In: *Frontiers in Plant Science* 11. ISSN: 1664-462X. DOI: 10.3389/fpls.2020.00520 (cit. on pp. 60, 63, 104).

Walker, Barnaby E., Tarciso C. C. Leão, Steven P. Bachman, Eve Lucas, & Eimear Nic Lughadha (June 4, 2021). *Evidence-based guidelines for developing automated conservation assessment methods*. EcoEvoRxiv. DOI: 10.32942/osf.io/zxq6s (cit. on p. 64).

Wallace, Alfred R (1860). "On the zoological geography of the Malay Archipelago". In: *Zoological Journal of the Linnean Society* 4.16, pp. 172–184 (cit. on p. 36).

Wang, Xudong, Long Lian, Zhongqi Miao, Ziwei Liu, & Stella X. Yu (May 15, 2021). "Long-tailed Recognition by Routing Diverse Distribution-Aware Experts". In: *arXiv:2010.01809 [cs]* (cit. on p. 148).

Watts, Matthew E, Ian R Ball, Romola S Stewart, Carissa J Klein, Kerrie A. Wilson, Charles Steinback, Reinaldo Lourival, Lindsay Kircher, & Hugh P Possingham (2009). "Marxan with

Zones: Software for optimal conservation based land-and sea-use zoning". In: *Environmental Modelling & Software* 24.12, pp. 1513–1521 (cit. on p. 66).
Weigelt, Patrick, Christian König, & Holger Kreft (2020). "GIFT – A Global Inventory of Floras and Traits for macroecology and biogeography". In: *Journal of Biogeography* 47.1, pp. 16–43. ISSN: 1365-2699. DOI: 10.1111/jbi.13623 (cit. on p. 120).
Wheeler, Q. D. et al. (Mar. 2012). "Mapping the biosphere: exploring species to understand the origin, organization and sustainability of biodiversity". In: *Systematics and Biodiversity* 10.1, pp. 1–20. ISSN: 1477-2000, 1478-0933. DOI: 10.1080/14772000.2012.665095 (cit. on p. 2).
Whittaker, R. H. (1960). "Vegetation of the Siskiyou Mountains, Oregon and California". In: *Ecological Monographs* 30.3, pp. 279–338. ISSN: 0012-9615. DOI: 10.2307/1943563 (cit. on pp. 25, 140).
Whittaker, Robert J., Miguel B. Araújo, Paul Jepson, Richard J. Ladle, James E. M. Watson, & Katherine J. Willis (2005). "Conservation Biogeography: assessment and prospect". In: *Diversity and Distributions* 11.1, pp. 3–23. ISSN: 1472-4642. DOI: 10.1111/j.1366-9516.2005.00143.x (cit. on p. 103).
Wilkinson, David P, Nick Golding, Gurutzeta Guillera-Arroita, Reid Tingley, & Michael A McCarthy (2019). "A comparison of joint species distribution models for presence–absence data". In: *Methods in Ecology and Evolution* 10.2, pp. 198–211 (cit. on p. 43).
Williams, John N, Changwan Seo, James Thorne, Julie K Nelson, Susan Erwin, Joshua M O'Brien, & Mark W Schwartz (2009). "Using species distribution models to predict new occurrences for rare plants". In: *Diversity and Distributions* 15.4, pp. 565–576 (cit. on p. 73).
Willig, M.R., D.M. Kaufman, & R.D. Stevens (2003). "Latitudinal Gradients of Biodiversity: Pattern, Process, Scale, and Synthesis". In: *Annual Review of Ecology, Evolution, and Systematics* 34.1, pp. 273–309. DOI: 10.1146/annurev.ecolsys.34.012103.144032 (cit. on p. 140).
Wilson, Kerrie A., Mar Cabeza, & Carissa J Klein (2009). "Fundamental concepts of spatial conservation prioritization". In: *Spatial conservation prioritization: Quantitative methods and computational tools*, pp. 16–27 (cit. on p. 66).
Winter, Lisa, Annekatrin Lehmann, Natalia Finogenova, & Matthias Finkbeiner (Nov. 1, 2017). "Including biodiversity in life cycle assessment – State of the art, gaps and research needs". In: *Environmental Impact Assessment Review* 67, pp. 88–100. ISSN: 0195-9255. DOI: 10.1016/j.eiar.2017.08.006 (cit. on p. 29).
Woodward, F Ian (1990). "The impact of low temperatures in controlling the geographical distribution of plants". In: *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 326.1237, pp. 585–593 (cit. on p. 97).
Wraith, Jenna, Patrick Norman, & Catherine Pickering (Oct. 1, 2020). "Orchid conservation and research: An analysis of gaps and priorities for globally Red Listed species". In: *Ambio* 49.10, pp. 1601–1611. ISSN: 1654-7209. DOI: 10.1007/s13280-019-01306-7 (cit. on pp. 12, 106, 146).
Wraith, Jenna & Catherine Pickering (Apr. 1, 2018). "Quantifying anthropogenic threats to orchids using the IUCN Red List". In: *Ambio* 47.3, pp. 307–317. ISSN: 1654-7209. DOI: 10.1007/s13280-017-0964-0 (cit. on p. 74).
— (2019). "A continental scale analysis of threats to orchids". In: *Biological Conservation* 234, pp. 7–17 (cit. on p. 12).
Wüest, Rafael O. et al. (2020). "Macroecology in the age of Big Data – Where to go from here?" In: *Journal of Biogeography* 47.1, pp. 1–12. ISSN: 1365-2699. DOI: 10.1111/jbi.13633 (cit. on p. 77).
Yee, Thomas W & Neil D Mitchell (1991). "Generalized additive models in plant ecology". In: *Journal of vegetation science* 2.5, pp. 587–602 (cit. on p. 43).

Yousefi, Masoud, Arash Jouladeh-Roudbar, & Anooshe Kafash (May 1, 2020). "Using endemic freshwater fishes as proxies of their ecosystems to identify high priority rivers for conservation under climate change". In: *Ecological Indicators* 112, p. 106137. ISSN: 1470-160X. DOI: 10.1016/j.ecolind.2020.106137 (cit. on pp. 16, 106).

Zarnetske, Phoebe L. et al. (2019). "Towards connecting biodiversity and geodiversity across scales with satellite remote sensing". In: *Global Ecology and Biogeography* 28.5, pp. 548–556. ISSN: 1466-8238. DOI: 10.1111/geb.12887 (cit. on p. 3).

Zhang, Chongliang, Yong Chen, Binduo Xu, Ying Xue, & Yiping Ren (2020). "Improving prediction of rare species' distribution from community data". In: *Scientific reports* 10.1, p. 12230 (cit. on p. 42).

Zhang, Zejin, Yujing Yan, Yu Tian, Junsheng Li, Jin-Sheng He, & Zhiyao Tang (Jan. 2015). "Distribution and conservation of orchid species richness in China". In: *Biological Conservation* 181, pp. 64–72. ISSN: 00063207. DOI: 10.1016/j.biocon.2014.10.026 (cit. on p. 12).

Zhao, Chang-Ming, Wei-Lie Chen, Zi-Qiang Tian, & Zong-Qiang Xie (2005). "Altitudinal Pattern of Plant Species Diversity in Shennongjia Mountains, Central China". In: *Journal of Integrative Plant Biology* 47.12, pp. 1431–1449. ISSN: 1744-7909. DOI: 10.1111/j.1744-7909.2005.00164.x (cit. on p. 140).

Zhao, Qingyuan & Trevor Hastie (2021). "Causal interpretations of black-box models". In: *Journal of Business & Economic Statistics* 39.1, pp. 272–281 (cit. on p. 52).

Zhongming, Zhu, Lu Linong, Yao Xiaona, Zhang Wangqiang, Liu Wei, et al. (2015). "Linking in situ vegetation data to the EUNIS habitat classification: results for forest habitats". In: (cit. on p. 38).

Zhou, Zhi-Hua (2018). "A brief introduction to weakly supervised learning". In: *National science review* 5.1, pp. 44–53 (cit. on p. 150).

Zhu, Xiao Xiang, Sina Montazeri, Mohsin Ali, Yuansheng Hua, Yuanyuan Wang, Lichao Mou, Yilei Shi, Feng Xu, & Richard Bamler (Jan. 5, 2021). "Deep Learning Meets SAR". In: *arXiv:2006.10027 [cs, eess, stat]* (cit. on pp. 62, 149).

Zhu, Xiao Xiang, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, & Friedrich Fraundorfer (Dec. 2017). "Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources". In: *IEEE Geoscience and Remote Sensing Magazine* 5.4, pp. 8–36. ISSN: 2168-6831. DOI: 10.1109/MGRS.2017.2762307 (cit. on pp. 4, 93).

Zizka, Alexander, Tobias Andermann, & Daniele Silvestro (June 18, 2021). "IUCNN - deep learning approaches to approximate species' extinction risk". In: *bioRxiv*, p. 2021.06.17.448832. DOI: 10.1101/2021.06.17.448832 (cit. on pp. 63, 104, 109, 135, 136, 196).

Zizka, Alexander, Daniele Silvestro, Pati Vitt, & Tiffany M. Knight (2020). "Automated conservation assessment of the orchid family with deep learning". In: *Conservation Biology* n/a (n/a). ISSN: 1523-1739. DOI: https://doi.org/10.1111/cobi.13616 (cit. on pp. 47, 63, 72, 74, 103, 111, 127, 130, 199).

Zizka, Alexander et al. (2019). "CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases". In: *Methods in Ecology and Evolution* 10.5, pp. 744–751. ISSN: 2041-210X. DOI: 10.1111/2041-210X.13152 (cit. on pp. 74, 130).

Zou, Hui & Trevor Hastie (2005). "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67.2, pp. 301–320 (cit. on p. 51).

Zurell, Damaris et al. (Sept. 2020). "A standard protocol for reporting species distribution models". In: *Ecography* 43.9, pp. 1261–1277. ISSN: 0906-7590, 1600-0587. DOI: 10.1111/ecog.04960 (cit. on pp. 42, 71).

SUPPLEMENTARY INFORMATION

Chapter 3

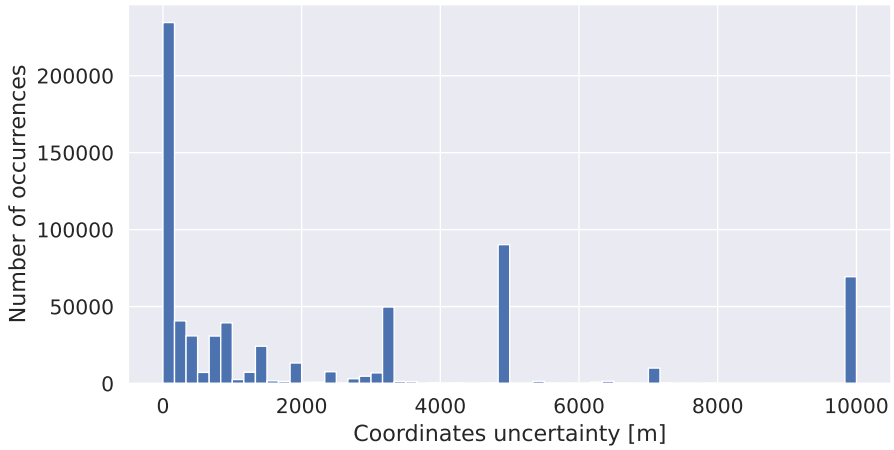


Figure S1: *Histogram of occurrence geolocation uncertainty (60 bins). 31% of the 999,248 occurrences associated with satellite data had no uncertainty provided at all and are not represented in this figure. Uncertainty was limited to 10,000 m on the Figure. First quartile is 100 m, median is 850 m and third quartile is 5,000 m. Recent and citizen science occurrences are usually integrating quite precise geolocation (explaining left peaks accumulation) whereas old observations will be less precise. The peak at 5,000 m certainly witnesses an arbitrary uncertainty value attributed to part of the orchids.*

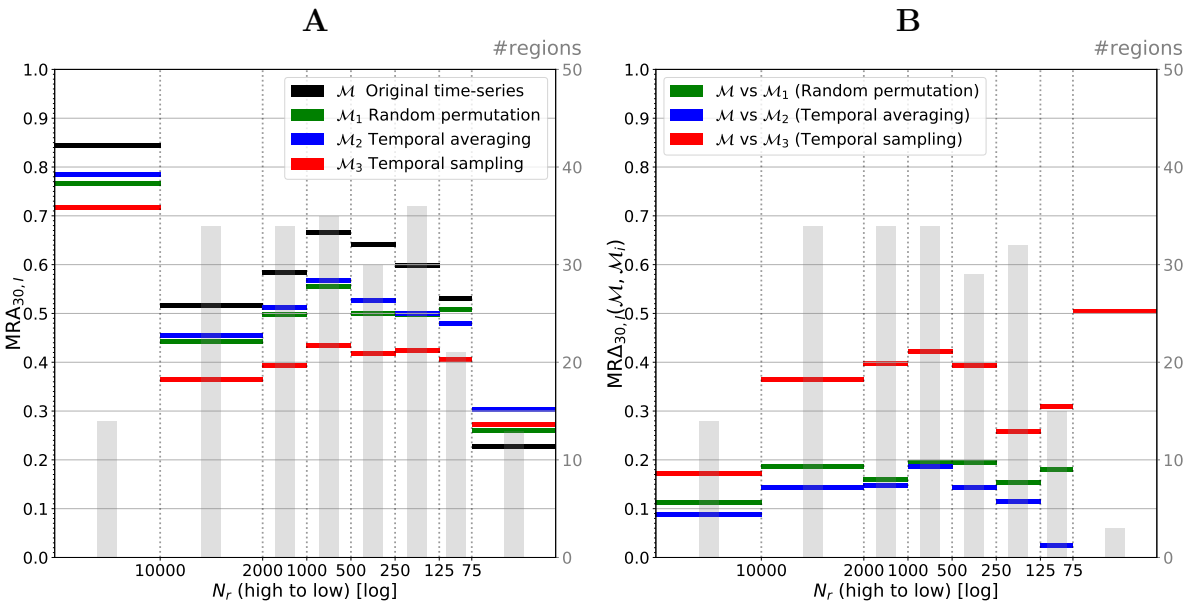


Figure S2: *Region top-30 accuracy* ***(A)*** *and relative top-30 accuracy change* ***(B)*** *averaged per category of number of training occurrences per region N_r. Unlike with categories formed on regions diversity index (Fig. 10A), $MRA_{30,I}(\mathcal{M})$ does not clearly diminish when regions are including less occurrences (this is not because there is less occurrences that the classification task is harder, the few occurrences can be from very common species). $MR\Delta_{30,I}(\mathcal{M}, \mathcal{M}_i)$ is not regularly higher in occurrence-poor regions contrary to diversity-rich regions where the predictive data temporal dimension especially help predictions (Fig. 3.10B).*

Chapter 4

Box A: Deep neural network architecture, dataset spatial split and training procedure

Inception v3 Our backbone model is an adaptation of the Inception v3 (Szegedy et al., 2016b). Initially designed to accept three-channel rgb images, it was modified to deal with a higher number of channels. This convolutional neural network learn patterns from spatialized input predictors. Letting models benefit from the spatial information was shown successful in various literature applications (Botella et al., 2018a; Deneu et al., 2021b). Successive inception modules are composed of convolutional filters of different sizes. This allows the different patch patterns of all sizes to be captured. Convolutional layers reduce the very high input dimension and a final softmax layer outputs the conditional probability distributions. Inputs are concatenated along the channel dimension. It results in $N \times 64 \times 64$ tensors with N the total channel number. Pixel resolution is of 1 km. A large 64×64 km^2 environmental context is therefore provided. The model is also spatially explicit: observation longitude and latitude are supplied in two dedicated channels along with the other predictors.

Deep learning models successfully process large numbers of inputs and classes with few samples. In fact, the modelling paradigm is completely different from combined per-species models. The filters learned during training are applied to all samples, all classes combined. The final softmax layer, which outputs class probabilities conditionally on an observation, is based on a reduced representation space common to all classes. This space has been shown to be structured by the ecological preferences of species in (Deneu et al., 2022). More generally, deep learning classification with strong class imbalance is a very active research avenue. DL outperforms classical approaches to model classes with few samples. In conclusion, our deep-SDM is not affected by the curse of dimensionality.

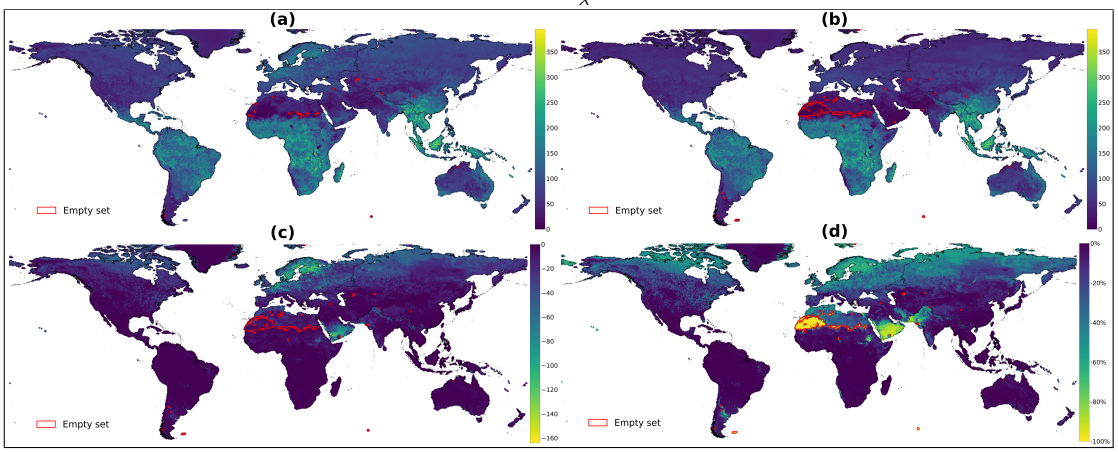
Dataset spatial split A spatial block hold-out validation strategy is employed to limit the effect of the spatial auto-correlation in the data in the evaluation of the model (as suggested in Roberts et al., 2017). 0.025° longitude / latitude blocks were defined worldwide (equivalent to 2.8 km at the equator). A train/validation/test split of 90/5/5 % of the blocks is then applied. The split is further stratified to WGSRPD level 2 zones to ensure a balanced block distribution across vegetation units Brummitt et al., 2001. In order not to over penalise performance, species initially present only in the validation or test sets are transferred to the training set. At the occurrence level, this results in a 902,174 / 46,290 / 50,794 set distribution. At the species level, this leads to a 14,129 / 4,037 / 4,166 split.

Training procedure The deep neural network f_θ is trained with the widely recognised LDAM loss, a modified cross-entropy function giving more emphasis to rare species during the training (Cao et al., 2019). It is a *label-distribution-aware* function specifically designed for strong class-imbalance and multi-class

classification problems. Performances on rare species are pushed upward without deteriorating common species predictions.
Model is fitted on Jean Zay, a supercomputer from the Institute for the Development and Resources in Intensive Scientific Computing (IDRIS). Layer weights are initialized from a truncated normal continuous random variable. Stochastic gradient descent optimizes the parameters on 2 GPUs during 70 epochs. With a batchsize equal to 128 and an initial learning rate of 0.01, the training process took 45 h. Learning rate is decayed at epochs 50 and 65 by a ten factor. A trained model weighs 610 MB. After having validated the model, a new training is lead from scratch on the whole dataset. It is stopped at the best epoch determined beforehand on the validation set. This retraining aims at obtaining the best possible model weights before the global-scale inference.

Box B: Species assemblage post-processing

Species assemblages and their relative probabilities are finally post-processed in two steps. We derive first from the initial occurrence set the inhabited continents of each species (WGSRPD level 1). This allows to filter out species predicted by the model outside their known continents of presence. Filtered assemblages of species are denoted $\hat{S}'_\lambda$. We computed statistics on this filtering step with a geographic prior on a global regular grid with 0.5 decimal degree resolution. The median number of species removed is 6, or 9.1% of the assemblage. Full statistics and map discrepancies are further discussed in Figure S3. Second, kept species conditional probabilities are normalised. Species with a conditional probability of presence smaller than λ are considered absent from the predicted assemblage, as well as species predicted outside their known continents of presence. In both cases, associated conditional probabilities were forced to zero. Normalisation allows to get back to a probability distribution summing to one. For a given input x, final probabilities are obtained with $\hat{\eta}'(x) = \frac{\hat{\eta}(x)}{\sum_{l \in \hat{S}'_\lambda(x)} \hat{\eta}_l(x)}$.



(e)

	Δ species	Relative change [%]
mean	-14.35	-18.21
std	19.37	21.28
min	-164	-100.0
25%	-22	-31.3
50%	-6	-9.1
75%	-1	-0.7
max	0	0.0

FIGURE S3 Maps and statistics illustrating the post-filtering step with the geographic-prior. The support is a global regular grid with 0.5 decimal degree resolution (59,823 points).
(a) $|\hat{S}_\lambda|$, i.e. the species assemblage size *before* the filtering step. Northern latitudes -and especially northern Europe- present abnormally large species assemblages. This is a consequence of the generalisation / over-prediction trade-off described in Discussion. The prediction model is over-confident because of the extensive occurrence training data in northern European countries.
(b) $|\hat{S}'_\lambda|$, i.e. the species assemblage size *after* the filtering step. The over-prediction bias at northern latitudes has been largely compensated. Empty predictions zones (red surrounded) have increased because of the geographic filtering, especially in the Sahara.
(c) $|\hat{S}'_\lambda| - |\hat{S}_\lambda|$, i.e. the absolute size difference of the species assemblage *before/after* the filtering step. Regions having lost the highest number of species are northern European countries and the South Arabian Peninsula.
(d) $\frac{|\hat{S}'_\lambda| - |\hat{S}_\lambda|}{|\hat{S}_\lambda|}$, i.e. the relative change in the species assemblage size *before/after* the filtering step. Regions mentioned in (c) are highlighted again. Saharan regions with empty predictions after geo-filtering do not appear to have lost high species number in (c). However, the clear yellow on map (d) indicates that these regions have lost all of the few species they were predicted to host.
(e) Statistics on the absolute and relative size difference of the species assemblage *before/after* geo-filtering. **Δ species** corresponds to map (c) and **Relative change [%]** corresponds to map (d).

Box C: Model evaluation and calibration

Evaluation of the deep-SDM The deep-SDM model was evaluated on unseen occurrences from the validation spatial blocks. Validation performances set the best epoch choice - the 69th - for final test set metrics to be computed. Selected metrics are the *top-k accuracy* and its per-class counterpart the *top-k accuracy per species*. These set-valued metrics do not require pseudo absences to avoid potential induced bias (Phillips et al., 2009). Top-k accuracy measures if the model returns the correct label among the k most likely classes:

$$\mathrm{A}_k(i) = \begin{cases} 1 & \text{if } \hat{\eta}_{y_i}(x_i) \geq \tilde{\eta}_k(x_i) \\ 0 & \text{otherwise} \end{cases} \tag{s1}$$

with (x_i, y_i) an input/label pair and $\tilde{\eta}$ the permutation of $\hat{\eta}$ sorted in descending order.

The success rate can be calculated for all test set occurrences, all classes combined (*micro-average* denoted A_k) or first for each class individually and then averaged together (*macro-average* denoted MSA_k). The former gives prominence to common species by construction, while the latter depends heavily on rare species performances. Macro-average metrics are suitable for highly imbalanced datasets. Final test set performances at epoch 69 are $\mathrm{A}_{30} = 0.87$ and $\mathrm{MSA}_{30} = 0.48$. This means that i) the correct label is returned among the first 30 species for 87% of the test observations (representative of common species), and ii) when each species in the test set is given the same weight, the correct label is within the first 30 classes returned almost half the time. This second metric may seem low, but it actually measures a particularly difficult task, given that the test set contains 4,166 species ($30/4166 \leq 1\%$). Furthermore, it reflects the performance of the model on rare species, and Figure S5 shows that considering *on average* 124 species significantly improves performance on the validation set, see next paragraph. Finally, training and validation curves show no sign of overfitting, see Figure S4.

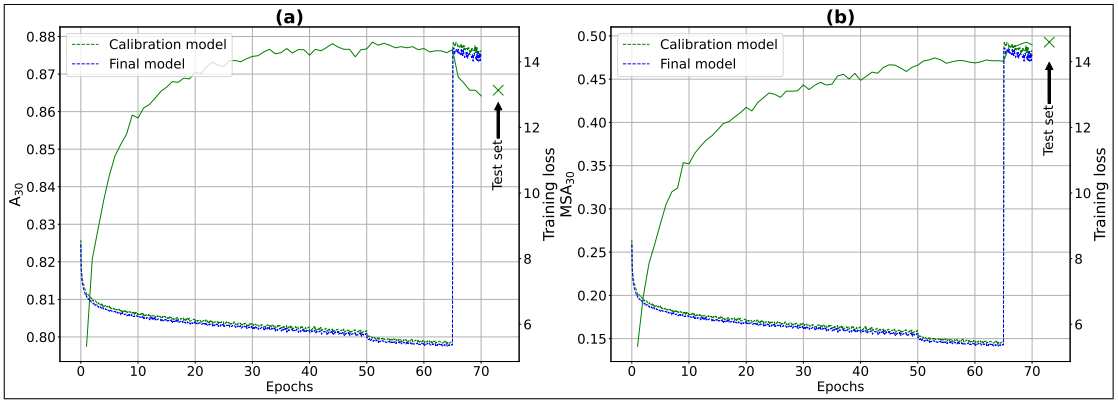


FIGURE S4 *Solid lines*: (a) micro-average and (b) macro-average top-30 accuracy on the validation set over training. *Dotted lines*: training losses of the calibration model (only on the training set) and the final model (full dataset). The first loss jump at epoch 50 corresponds to the first learning rate decay by a tenth factor. The second loss jump at epoch 65 results from the LDAM delayed reweighting

scheme. It gives prominence to the rare species, which is why we observe a decrease in performance for A_{30} and an increase for MSA_{30}. The training and validation curves show no signs of overfitting. The performances of the test sets with the best calibration model (epoch 69) are $A_{30} = 0.865$ and $MSA_{30} = 0.48$.

Calibration of the species assemblage prediction model As discussed earlier, the optimisation of the hyper-parameter λ is done through an average error control method applied on the validation set. In Equation 3, ϵ is set to 0.03. The resulting estimated value for λ is equal to 8.75e−5 (see Figure S5) and the corresponding average size of the predicted species assemblages is equal to 124 species. Reaching 0.97% micro-average accuracy means that the model almost always returns the correct label within the predicted set when a random unseen observation is being provided. The number of observations per class being strongly unbalanced (see Box F Fig. a), the 97% micro-average accuracy is strongly influenced by the performance on common species. Now, when all unseen species are granted the same weight in the average computation (macro-average accuracy), performance is still of 80%. Given how unbalanced the observation dataset is (median occurrence number is four, 25% species have more than 13 occurrences), it becomes clearer that the model's performances are satisfying. Summary statistics on $|\hat{S}_\lambda|$ are reported on Table S1.

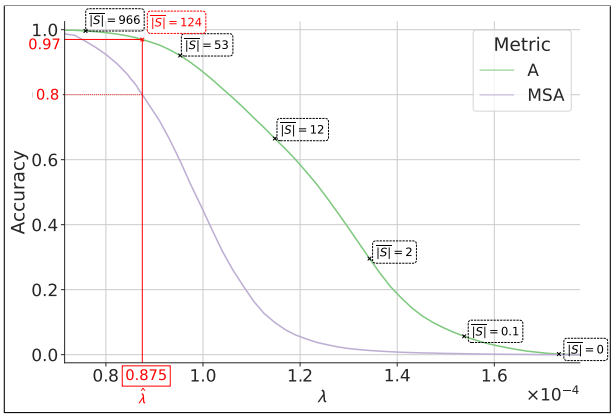


FIGURE S5 Average error control setting on the validation set. Limiting condition on the micro-average accuracy (green curve) is $\epsilon \leq 0.03 \Leftrightarrow A \geq 0.97$. Optimal threshold $\hat{\lambda}$ is highlighted in red while matching macro-average accuracy (grey function) is also reported with a red dashed line. Average set sizes $\overline{|\hat{S}_\lambda|}$ are indicated in dashed boxes (hat and subscript being dropped for readability).

TABLE S1 Validation set statistics on $|\hat{S}_\lambda|$, i.e. the size of the species assemblage after thresholding the conditional probabilities of presence with $\hat{\lambda}$ (46,290 validation points). The minimum number of species retained in the validation set is four. However, on a global scale there are areas with no species above $\hat{\lambda}$, resulting in empty predictions (e.g. western Algeria). It is also very likely that areas are predicted with more than 401 orchids (the maximum on the validation set).

	mean	std	min	25%	50%	75%	max	A_{30}	MSA_{30}
$\hat{\lambda}$	124.067	39.803	4	95	121	150	401	0.970	0.801

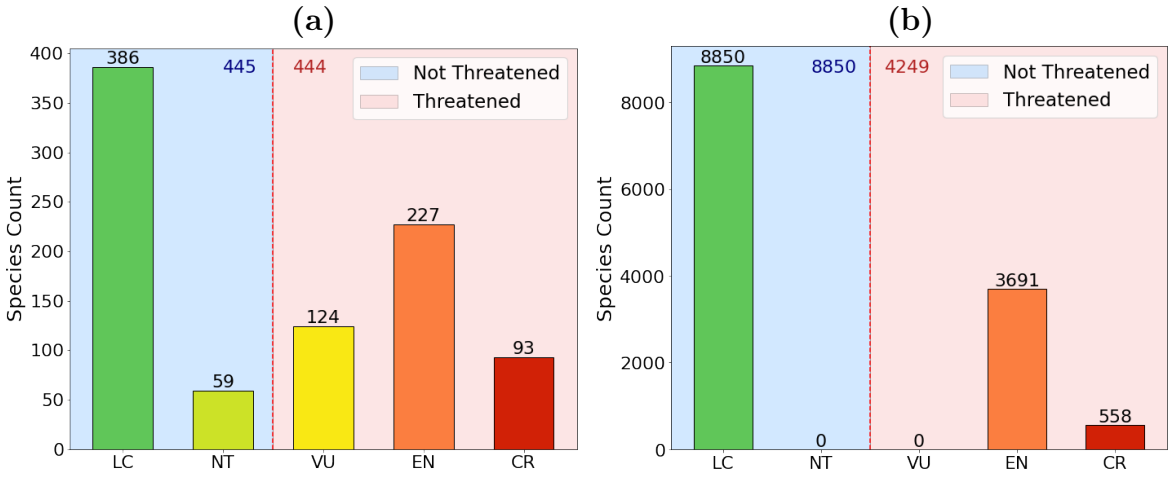


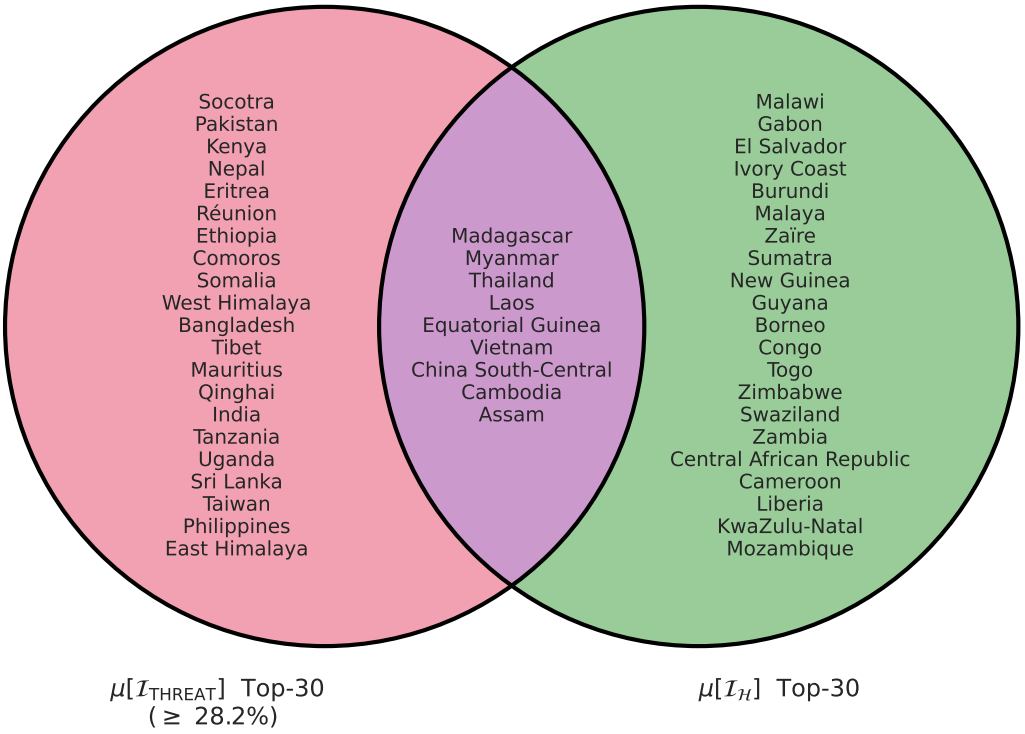
Figure S6: *Distribution of extinction risk status of (a) known IUCN-assessed species and (b) predicted with IUCNN (Zizka et al., 2021). No NT or VU species are predicted. These two classes are by definition quite ambiguous: a species is classified as Near Threatened if the VU thresholds are almost met. When optimising the classifier, predicting NT or VU species results in a high error rate, so the best performance trade-off in this setting is to exclude these classes.*

Box D: Maps batch processing and online access

Batch processing Species assemblages are predicted by 50,176 batches for volume reasons. PyTorch model 512-size predictions are accumulated in a buffer until exceeding the 50,000 limit and are then exported. Raw predicted classes and probabilities are appended in binary files whereas spatial indicators are computed on the fly and saved in distinct .geotiff format. Finally, geotiffs are merged and converted to the Cloud Optimized GeoTIFF format (COG): 256 x 256 blocks are tiling the data and six levels of overviews are added.

Online and interactive map access The web mapping solution used to render the raster data in a web environment is simple. It relies upon a serverless front-end application and a map server able to render OGC (Open Geospatial Consortium) compliant web services : Web Map Service, Web Feature Service, etc. The front-end application is an open-source project called MViewer (https://mviewer.netlify.app/) and is mostly implemented by Brittany region. It parses an .xml configuration file to generate an interactive map. The map server is also an open-source project : GeoServer (https://geoserver.org/). It can read COG data among other geospatial data and serves it as a web service to be displayed by the front-end application. Website is available at https://mapviewer.plantnet.org/?config=apps/store/orchid-status.xml.

(a)



(b)

Continent	Coef.	ρ-value
Africa	0.206716	1.099494e-01
Asia-Temperate	0.499064	2.250013e-04
Asia-Tropical	0.028462	8.925731e-01
Australasia	0.350000	3.558196e-01
Europe	0.650844	5.441973e-06
North America	-0.510027	5.537334e-06
Pacific	-0.300000	6.238377e-01
South America	0.294890	7.228730e-02
Global	0.292719	2.553788e-07

Figure S7: ***(a)*** *Venn diagram between the $\mu[\mathcal{I}_{\mathcal{H}}]$ and $\mu[\mathcal{I}_{THREAT}]$ top-30 countries. With the exception of Madagascar and Equatorial Guinea, all the countries in the intersection are from South and South-East Asia. El Salvador and Guyana are the only countries in the diagram that are not from Africa or Asia.*
(b) *Considering the same two variables, the Spearman correlation and the ρ-value for all countries of the same continent. According to the ρ-values, the correlations are statistically significant only in Asia-Temperate, Europe and North America. Furthermore, looking at the scatter plot, we can see that the European and North American diversity ranges are limited. Finally, it is only in Asia-Temperate that we observe a significant and positive correlation between threat levels and a wide diversity range.*

Box E: Comparison with established indicators

The $\mu[\mathcal{I}_{\text{THREAT}}]$ top countries (i.e. countries with the highest average proportion of predicted threatened species) largely overlap with the countries identified in (Zizka et al., 2020) as having the highest proportion of potentially threatened species.

The Red List of Ecosystems (RLE) is a classification scheme of the risk of ecosystem collapse, with categories and an assessment process that mirror the IUCN Red List of Threatened Species (Keith et al., 2013). This promising indicator currently suffers from poor data coverage, with only 509 assessments registered as of April 2023 (https://assessments.iucnrle.org/, accessed on 26/04/23). This comparison highlights the rare global consistency of our indicators. We could also imagine ecosystem-level indicators that take into account the extinction risk of species assemblages in their construction.

Safeguarding ecosystems within the post-2020 global biodiversity framework requires robust indicators that capture different dimensions: area, integrity and risk of collapse (Nicholson et al., 2021). Among the recommendations for selecting indicators, two are particularly relevant to our work: *4. greater testing and validation of indicators is required to understand their ecosystem relevance, reliability and ease of interpretation* and *5. the connection between global indicators and national or local policy and reporting needs strengthening.* Our indicators meet recommendation five, but suffer from a lack of ground truthing to be confidently applied on the ground, as the fourth recommendation points out.

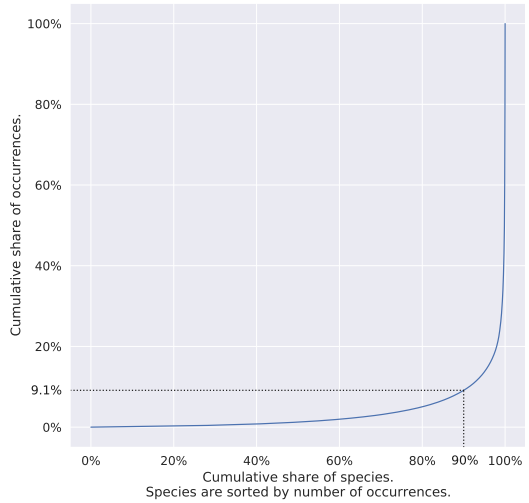
Protecting species for their evolutionary distinctiveness, combined with an IUCN threatened status, is the approach taken by EDGE (Evolutionary Distinct Globally Endangered, Isaac and Pearse, 2018). While EDGE species must be officially listed as threatened by the IUCN in addition to having an above-average ED score (Evolutionary Distinctiveness), Vitt et al., 2023 developed a conservation prioritisation method based on ED and rarity as *the number of occupied regions* or *the area of occupancy*. Here, the spatial ranges considered are compiled from the World Checklist of Selected Plant Families (WCSP) and Global Inventory of Floras and Traits (GIFT) databases. Tropical Africa does not emerge as a clear priority hotspot as our indicators suggest. However, they highlight the Neotropics and Southeast Asia as hotspots of richness, as does our Shannon index indicator. They also identify islands as having particularly high numbers of rare and distinct species. Interestingly, they point out that orchid ED is highly correlated with their richness ($R^2 = 0.87$).

Finally, the closest indicator to date from our work is the global extinction probability of terrestrial vascular plants (Verones et al., 2022). In a given place, this indicator is high if many threatened species are known to occur there and/or if they have very small ranges. However, we defend our kilometre-scale resolution and the novel way in which we calculate $\mathcal{I}_c$. This allows us to weight the contribution of species by their relative probability of occurrence.

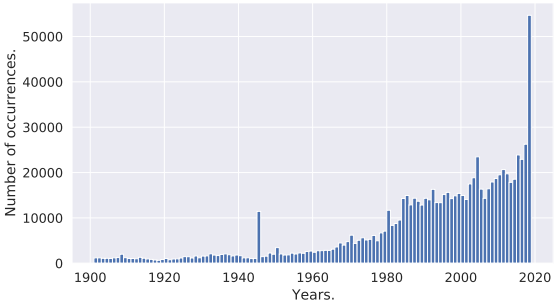
Although the Shannon index measures not only community richness but also its evenness, global vascular plant richness maps such as (Cai et al., 2023) are the closest available point of comparison. Again, both the resolution and construction of our indicator differ from previous work.

Box F: Orchid dataset distributions

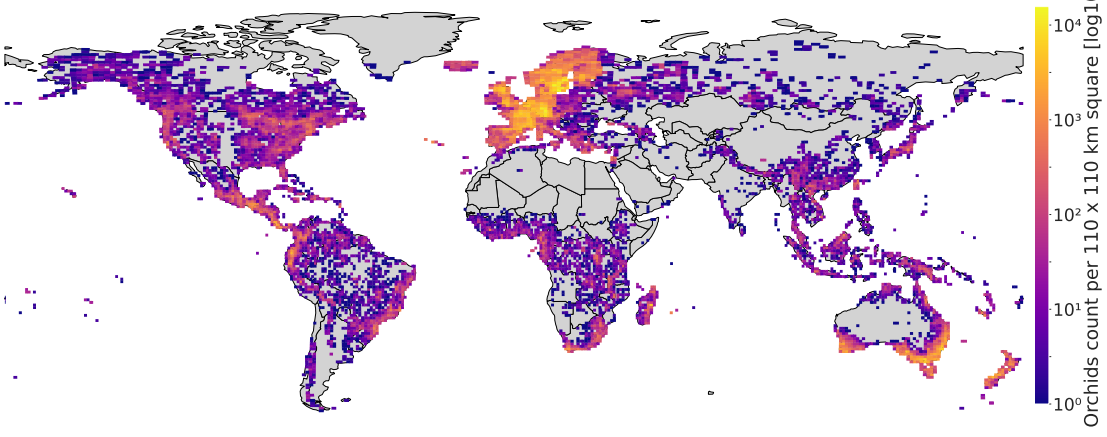
(a)



(b)



(c)



(d)

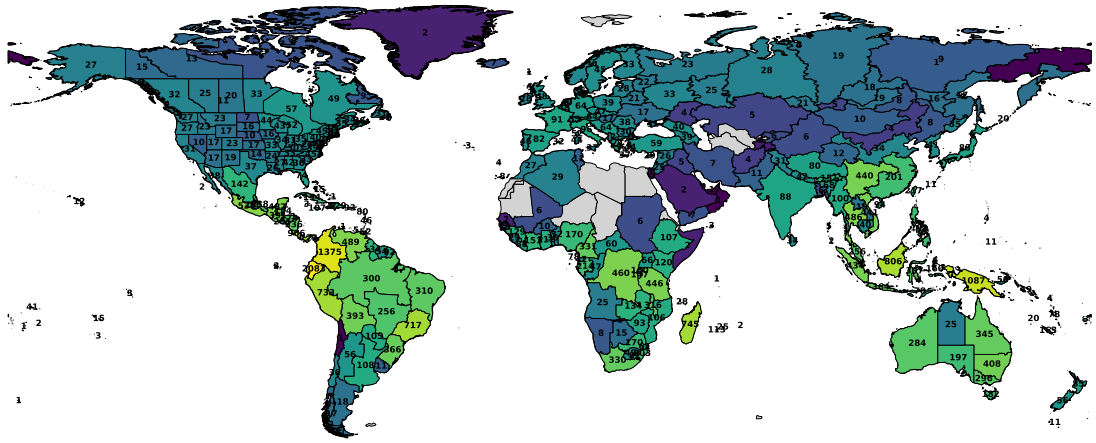


FIGURE S8 (a) Occurrences' distribution. Species are ordered by frequency. The dotted lines are flagging that 90% of the species are only gathering 9.1% of the occurrences. **(b)** Occurrences' temporal distribution. The two previous graphs are based on all dataset's occurrences. **(c)** Observations map coloured by number of records in 110 x 110 km tiles (log10 scale). **(d)** Species richness map stratified by botanical country (WGSRPD level 3). Colours are in log scale.

Box G: Predictive features description

In selecting predictive features, the main limiting criterion was to use only globally available potential drivers of orchid preferences.

WorldClim2 bioclimatic variables The nineteen standard bioclimatic variables from WorldClim version 2 were provided to the model (Fick & Hijmans, 2017). They are historic averages over the 1970-2000 at 30-second resolution. This suits our occurrence date distribution. Variables stem from temperature and precipitation data (https://www.worldclim.org/). They are established indicators of climate annual trends, seasonality and extreme values.

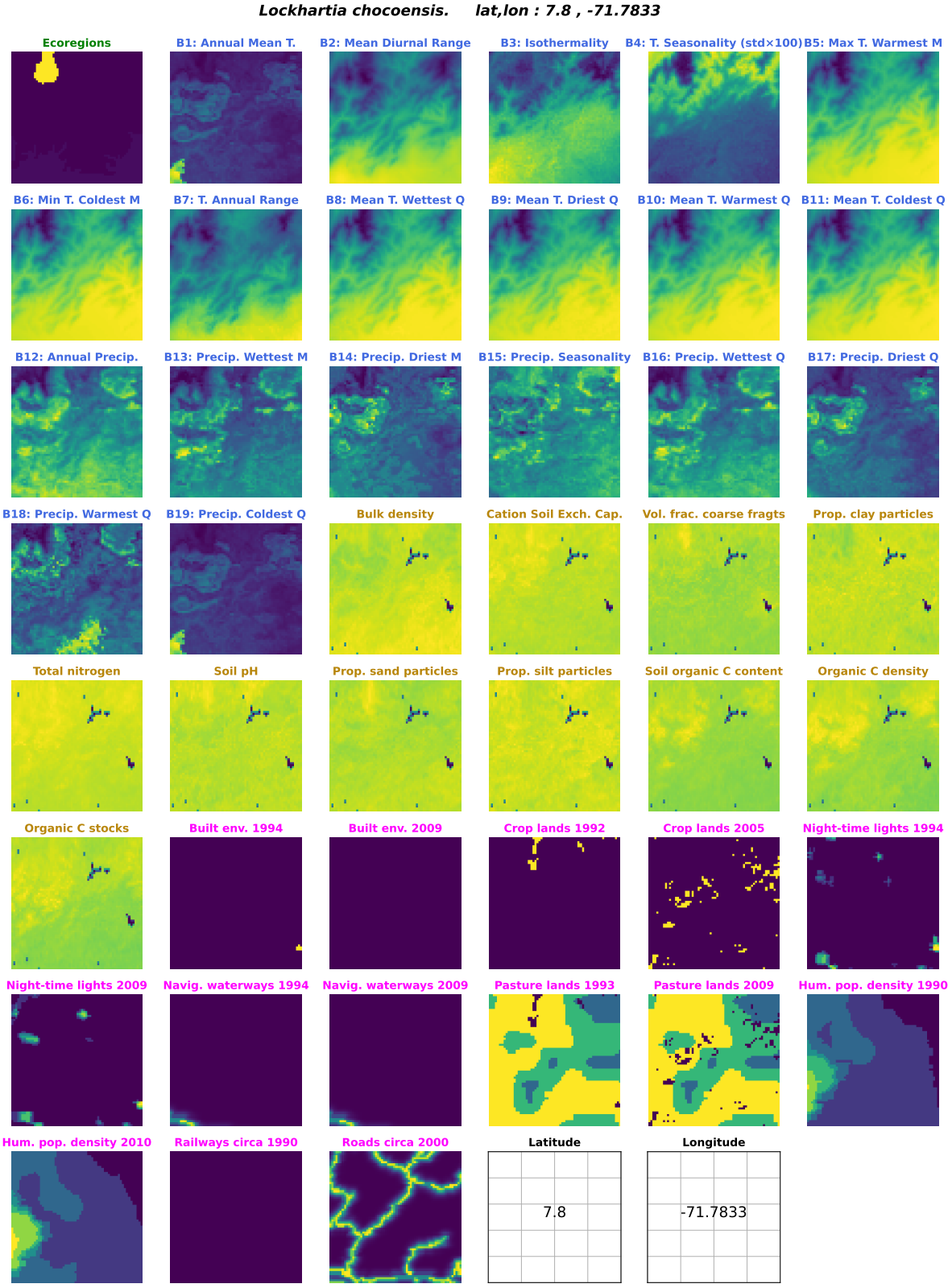
Soilgrids pedological variables Soilgrids is a collection of eleven global soil property and class maps produced by machine learning models (Poggio et al., 2021). They include soil pH, nitrogen concentration annd clay particles proportions among others (more information in the official FAQ: https://www.isric.org/). The exploited statistical models are fitted with 230,000 soil profiles spread wordwide and environmental covariates. We use the 1-kilometre resolution products.

Human footprint detailed rasters Eight variables measure direct and indirect global human pressure: built environments, population density, electric infrastructure, crop lands, pasture lands, roads, railways, and navigable waterways (Venter et al., 2016). They are provided at a 1-kilometre resolution (https://datadryad.org/) and for two distinct years: 1993 and 2009. These rasters spring form both remotely-sensed data and surveys.

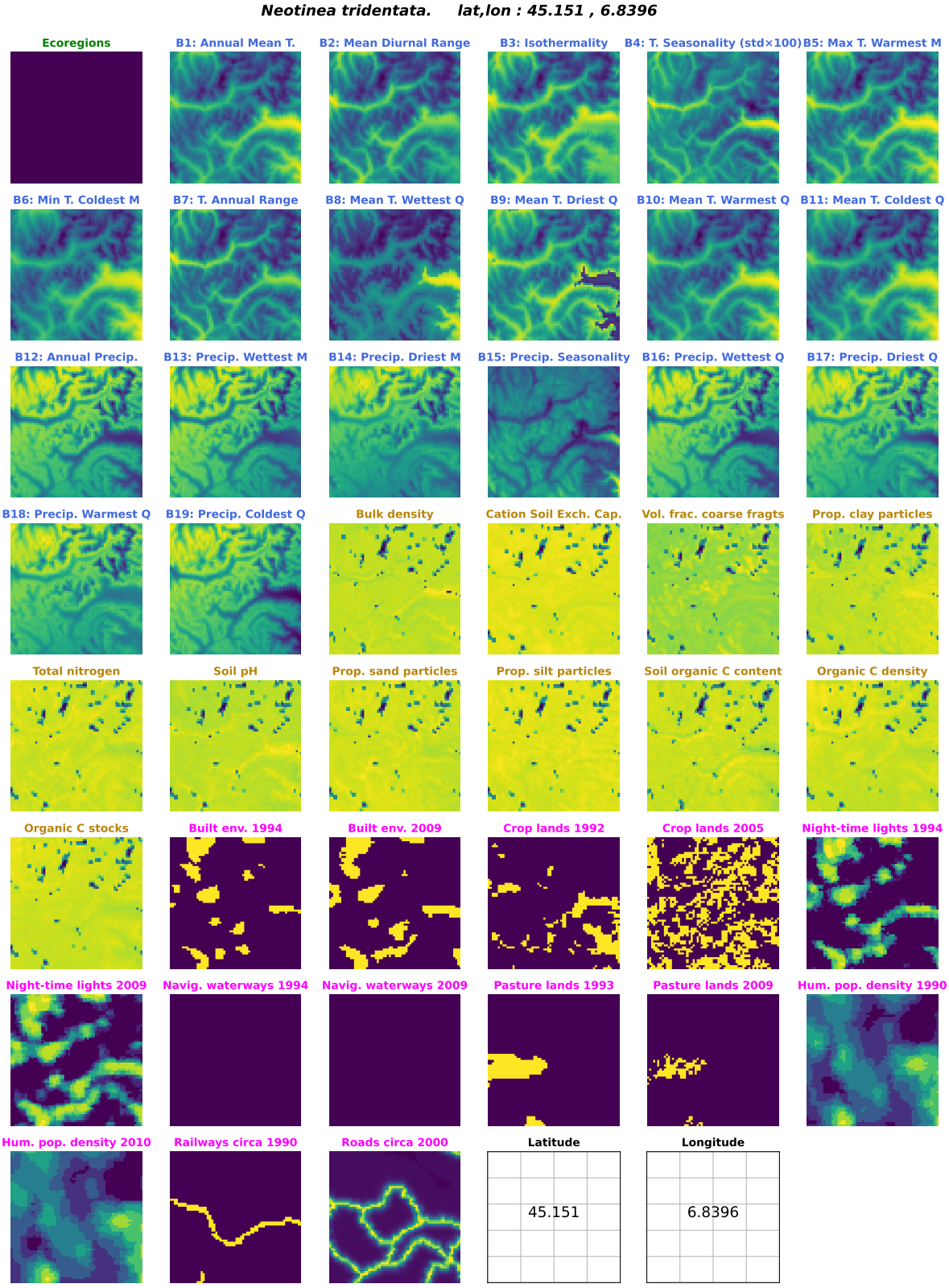
Terrestrial ecoregions of the world This is a biogeographic classification of terrestrial biodiversity. Ecoregions are defined by the authors as *"relatively large units of land containing a distinct assemblage of natural communities and species, with boundaries that approximate the original extent of natural communities prior to major land-use change"* (Olson et al., 2001). 867 ecoregions are gathered into 14 biomes such as boreal forests or deserts. Data (https://www.worldwildlife.org/) was resampled at 30 seconds longitude/latitude resolution.

Location The explicit provision of observation coordinates is a key modelling decision. Both a large regional context and precise location information are provided. The model can make the most of this mixed input. Deep learning models can indeed take advantage of complex combinations of heterogeneous inputs. As longitude and latitude are inputted separately, both indications are processed alongside, can interact, but are also interpreted distinctly.

(a)


Lockhartia chocoensis. lat,lon : 7.8 , -71.7833
Ecoregions
B1: Annual Mean T.
B2: Mean Diurnal Range
B3: Isothermality
B4: T. Seasonality (std×100)
B5: Max T. Warmest M
B6: Min T. Coldest M
B7: T. Annual Range
B8: Mean T. Wettest Q
B9: Mean T. Driest Q
B10: Mean T. Warmest Q
B11: Mean T. Coldest Q
B12: Annual Precip.
B13: Precip. Wettest M
B14: Precip. Driest M
B15: Precip. Seasonality
B16: Precip. Wettest Q
B17: Precip. Driest Q
B18: Precip. Warmest Q
B19: Precip. Coldest Q
Bulk density
Cation Soil Exch. Cap.
Vol. frac. coarse fragts
Prop. clay particles
Total nitrogen
Soil pH
Prop. sand particles
Prop. silt particles
Soil organic C content
Organic C density
Organic C stocks
Built env. 1994
Built env. 2009
Crop lands 1992
Crop lands 2005
Night-time lights 1994
Night-time lights 2009
Navig. waterways 1994
Navig. waterways 2009
Pasture lands 1993
Pasture lands 2009
Hum. pop. density 1990
Hum. pop. density 2010
Railways circa 1990
Roads circa 2000
Latitude
7.8
Longitude
-71.7833

(b)


Neotinea tridentata. lat,lon : 45.151 , 6.8396
Ecoregions
B1: Annual Mean T.
B2: Mean Diurnal Range
B3: Isothermality
B4: T. Seasonality (std×100)
B5: Max T. Warmest M
B6: Min T. Coldest M
B7: T. Annual Range
B8: Mean T. Wettest Q
B9: Mean T. Driest Q
B10: Mean T. Warmest Q
B11: Mean T. Coldest Q
B12: Annual Precip.
B13: Precip. Wettest M
B14: Precip. Driest M
B15: Precip. Seasonality
B16: Precip. Wettest Q
B17: Precip. Driest Q
B18: Precip. Warmest Q
B19: Precip. Coldest Q
Bulk density
Cation Soil Exch. Cap.
Vol. frac. coarse fragts
Prop. clay particles
Total nitrogen
Soil pH
Prop. sand particles
Prop. silt particles
Soil organic C content
Organic C density
Organic C stocks
Built env. 1994
Built env. 2009
Crop lands 1992
Crop lands 2005
Night-time lights 1994
Night-time lights 2009
Navig. waterways 1994
Navig. waterways 2009
Pasture lands 1993
Pasture lands 2009
Hum. pop. density 1990
Hum. pop. density 2010
Railways circa 1990
Roads circa 2000
Latitude
45.151
Longitude
6.8396

(c)

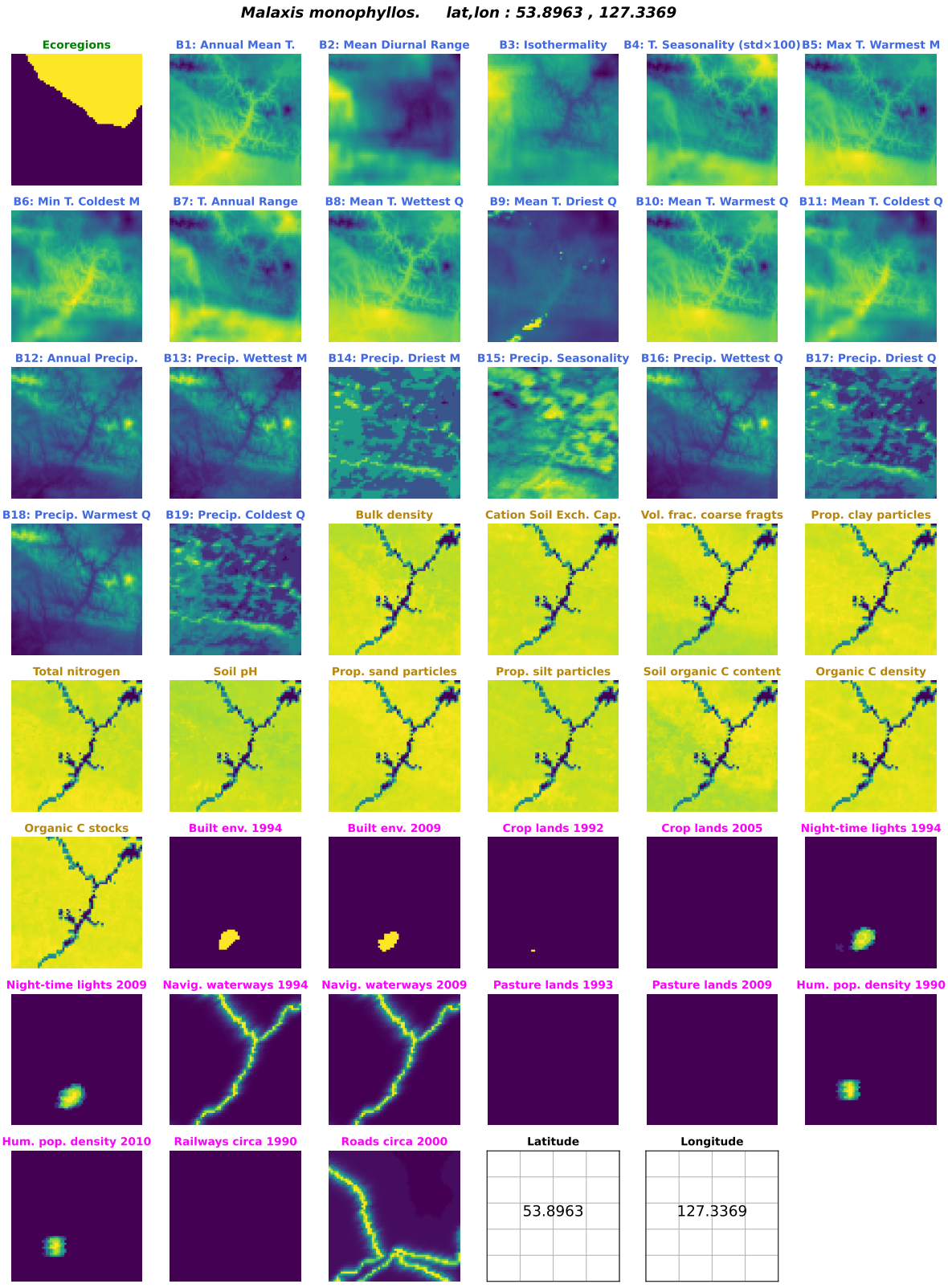


Figure S9: *2D input data associated to three observations located in (a) Venezuela, (b) France and (c) Russia. Feature types are denoted by their title color: (green) ecoregions, (blue) bioclimatic variables, (brown) pedological variables, (pink) human footprint and (black) location.*

Table S2: *List of predictors. They are either categorical or continuous and can be gathered into five groups: the terrestrial ecoregions of the world, the WorldClim2 bioclimatic variables, the Soilgrids pedological variables, the detailed rasters of the human footprint and the location.*

	Group	Name	Type
1	Terrestrial ecoregions of the world	Ecoregions per biome	categorical
2	WorldClim2 bioclimatic variables	BIO1 = Annual Mean Temperature	continuous
3		BIO2 = Mean Diurnal Range (Mean of monthly (max temp - min temp))	continuous
4		BIO3 = Isothermality (BIO2/BIO7) (×100)	continuous
5		BIO4 = Temperature Seasonality (standard deviation ×100)	continuous
6		BIO5 = Max Temperature of Warmest Month	continuous
7		BIO6 = Min Temperature of Coldest Month	continuous
8		BIO7 = Temperature Annual Range (BIO5-BIO6)	continuous
9		BIO8 = Mean Temperature of Wettest Quarter	continuous
10		BIO9 = Mean Temperature of Driest Quarter	continuous
11		BIO10 = Mean Temperature of Warmest Quarter	continuous
12		BIO11 = Mean Temperature of Coldest Quarter	continuous
13		BIO12 = Annual Precipitation	continuous
14		BIO13 = Precipitation of Wettest Month	continuous
15		BIO14 = Precipitation of Driest Month	continuous
16		BIO15 = Precipitation Seasonality (Coefficient of Variation)	continuous
17		BIO16 = Precipitation of Wettest Quarter	continuous
18		BIO17 = Precipitation of Driest Quarter	continuous
19		BIO18 = Precipitation of Warmest Quarter	continuous
20		BIO19 = Precipitation of Coldest Quarter	continuous
21	Soilgrids pedological variables	Bulk density (cg/cm3)	continuous
22		Cation exchange capacity at ph 7 (mmol(c)/kg)	continuous
23		Coarse fragments in cm3/dm3	continuous
24		Clay content in g/kg	continuous
25		Nitrogen in cg/kg	continuous
26		pH water (pH *10)	continuous
27		Sand in g/kg	continuous
28		Silt in g/kg	continuous
29		Soil organic carbon (dg/kg)	continuous
30		Organic carbon density (g/dm3)	continuous
31		Soil organic carbon stock (t/ha)	continuous
32	Human footprint detailed rasters	Individual pressure map of built environments in 1994	categorical
33		Individual pressure map of built environments in 2009	categorical
34		Individual pressure map of crop lands in 1992	categorical
35		Individual pressure map of crop lands in 2005	categorical
36		Individual pressure map of night-time lights in 1994	categorical
37		Individual pressure map of night-time lights in 2009	categorical
38		Individual pressure map of navigable waterways in 1994	continuous
39		Individual pressure map of navigable waterways in 2009	continuous
40		Individual pressure map of pasture lands in 1993	categorical
41		Individual pressure map of pasture lands in 2009	categorical
42		Individual pressure map of human population density in 1990	categorical
43		Individual pressure map of human population density in 2010	categorical
44		Individual pressure map of railways circa 1990	categorical
45		Individual pressure map of roads circa 2000	continuous
46	Location	Longitude (DD)	continuous
47		Latitude (DD)	continuous

Chapter 5

Box H: Description of the *.csv* files containing the status predictions

Description of the status prediction supplementary file *ALL_species_status.csv*. This *.csv* lists the status predicted with our classifier for the 14,129 species of our dataset:

- At two different levels: *broad*, i.e. the binary classification Threatened or not, and *detail*, i.e. the five IUCN categories from LC to CR.
- For two different dispersal scenario: *False* meaning that no dispersal is allowed and species can only be re-predicted only where they once occurred, and *True* meaning that species can potentially be re-predicted on all our dataset points.
- For five different time periods matching the Worldclim 2 bioclimatic projections: *Present, 2021-2040, 2041-2060, 2061-2080* and *2081-2100*.

Here are the meanings of the common fields:

- *species* is the GBIF canonical species name
- *speciesKey* is the GBIF unique key associated to the species
- *lat, lon, HFP09, Elevation* are the species average values across the dataset's observations for respectively *latitude, longitude, 2009 human footprint index* and *elevation*.
- *continents* provides the set of WGSRPD level 3 regions in which the species occurs.
- *IUCNonly* indicates whether the species is currently assessed by the IUCN (886) or not (13,243).
- For IUCN-assessed species, the present categories are the current official red list levels and not predictions.

Description of the Extinct species supplementary file *EXT_species.csv*. This *.csv* lists the 234 species predicted to be extinct by the model. Fields are identical to the *ALL_species_status.csv* file described below, with the noticeable differences being:

- *time period* is the field indicating the first period from which the species is predicted to be extinct.
- All species predicted to be extinct come from the null dispersal scenario.

In addition, one species is predicted to be extinct as of now: *Dendrobium rhytidothece*. This species is not red listed and is known on the GBIF only from six occurrences, five of which date from 1909 and the last one being fuzzy[a]. There are 42 species currently assessed by the IUCN that are predicted to become extinct (and 192 species that have not yet been assessed).

[a] https://www.gbif.org/species/5315531

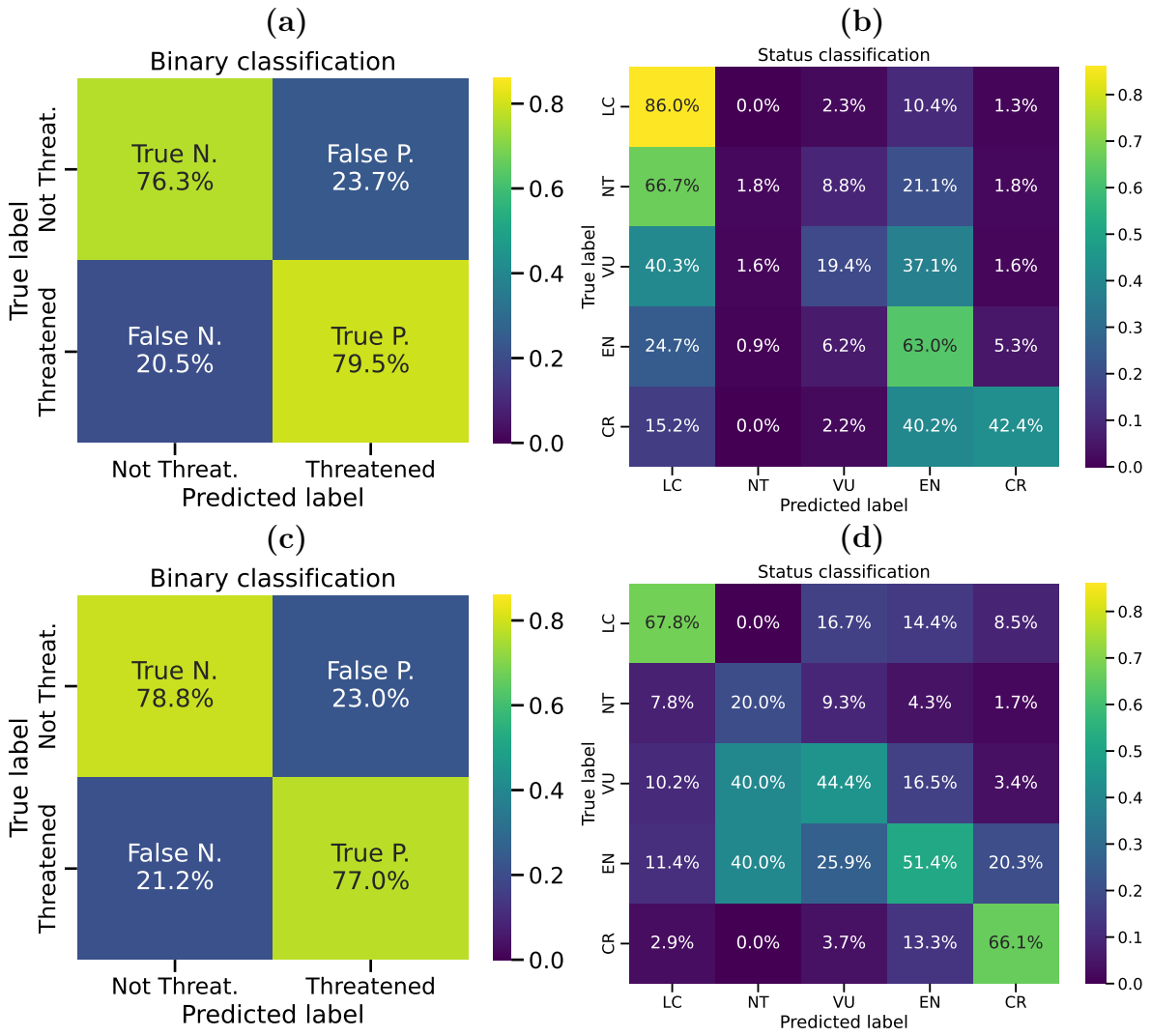


Figure S10: *The first row shows the precision confusion matrices, i.e. with row normalisation. Among the true labels of a given category, we test the proportions predicted in each category. The second row presents recall confusion matrices, i.e. with a normalisation by column. Among the predicted labels of a given category, we test the proportions that actually belong to the predicted category or to others. The first column shows the results for the binary classification and the second column for the status classification.*

(a) *Precision confusion matrix for binary classification. 79.5% of threatened species are correctly classified as threatened by the model.*

(b) *Precision confusion matrix for status classification. 86% of LC species are correctly identified by the model. For NT species, 66.7% are misclassified as LC and 21.1% as EN. 40.3% of the VU species are misclassified as LC, 37.1% as EN and 19.4% are correctly classified. Three quarters of the EN species are classified as either VU, EN or CR. Of the species classified as CR by IUCN, the model predicts 40.2% as EN and 42.4% actually as CR.*

(c) *Recall confusion matrix for binary classification. 78.8% of species predicted as not threatened are actually so.*

(d) *Recall confusion matrix for status classification. Only 67.8% of species predicted as LC are actually Least Concern. Of those predicted to be NT, only a fifth are actually NT, with the remainder split evenly between VU and EN. 44.4% of species predicted to be VU are correctly classified. More than half of the species predicted as EN are actually EN, and more than 80% are from a threatened category. Of the species predicted as CR, 86.4% are actually classified as either CR or EN by IUCN.*

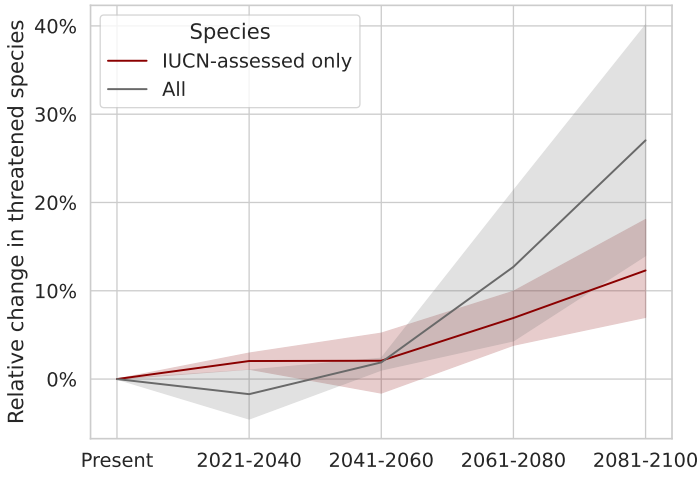


Figure S11: *Relative change in the number of species predicted to be threatened over time. The bands indicate the uncertainty due to the combination of the two dispersal scenarios. Interestingly, for species not yet assessed by the IUCN, the increase in threatened species is predicted to be higher by the end of the century.*

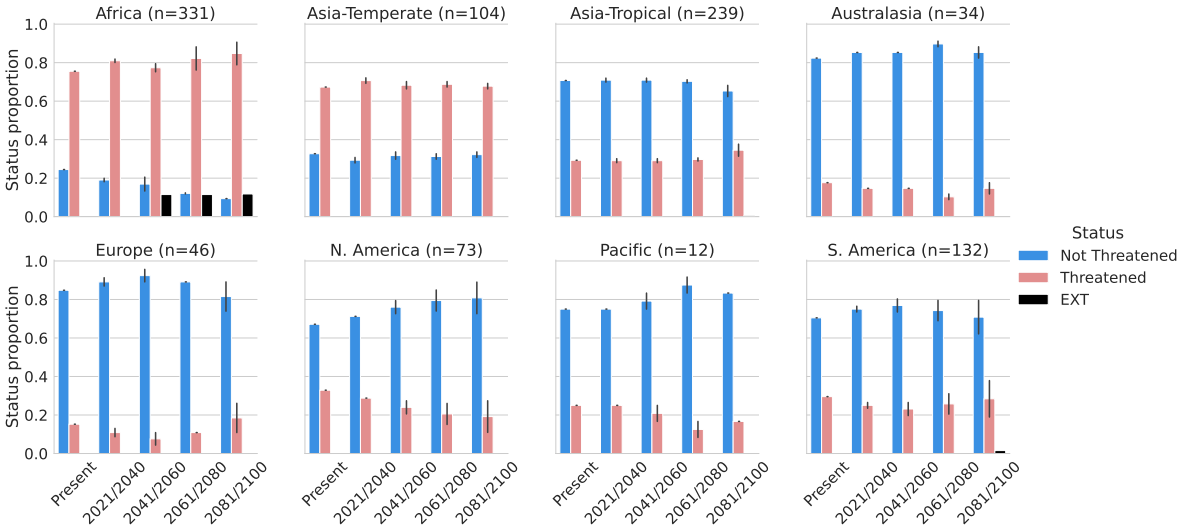


Figure S12: *Binary status proportions per continent and time period. Only IUCN-listed species are included, and their number per continent is given in the subtitle. Error bars account for differences between the two dispersal scenarios. With this restriction, some different trends are apparent. In North America, out of 73 species, the proportion of threatened species is actually predicted to decrease steadily (same trend for the 12 Pacific species, but there are too few species to be considered a robust result). In Africa, the proportion of threatened species is projected to increase to more than 80% by the end of the century. In tropical Asia, the number of threatened species is increasing but is significantly lower than when all species are considered. Levels are also lower in Europe if only IUCN-assessed species are considered. Overall, it is interesting to observe that continental trends and levels can be quite different when only IUCN-assessed species are considered. For us, this is an indication of the model's ability to generalise without simply overfitting and replicating patterns between neighbouring species.*

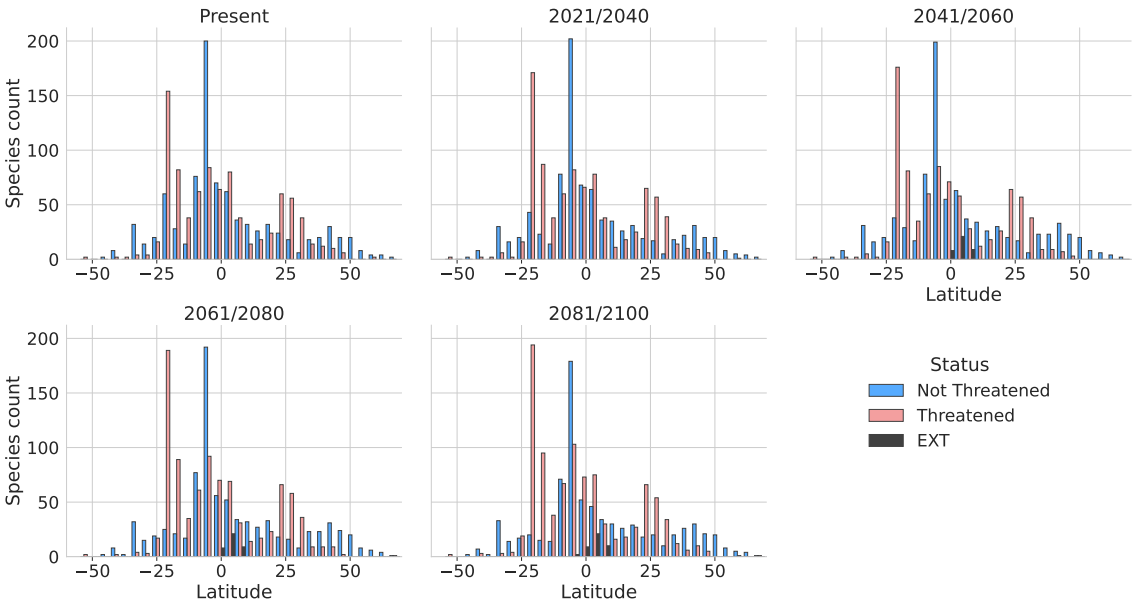


Figure S13: *Species count histograms as a function of average latitude and over time. Only IUCN-listed species are included. The bins cover four degrees of latitude. A trident pattern appears for threatened species, but: i) the highest threatened species peak is at about -20° latitude and ii) high counts of threatened species around the 25° parallel do not increase, unlike the figure for all species. Again, the differences between the two species support confirms that our status classifier relies on specific species-level information and not just spatial information. This figure also confirms that the current IUCN assessment is biased towards the northern hemisphere.*

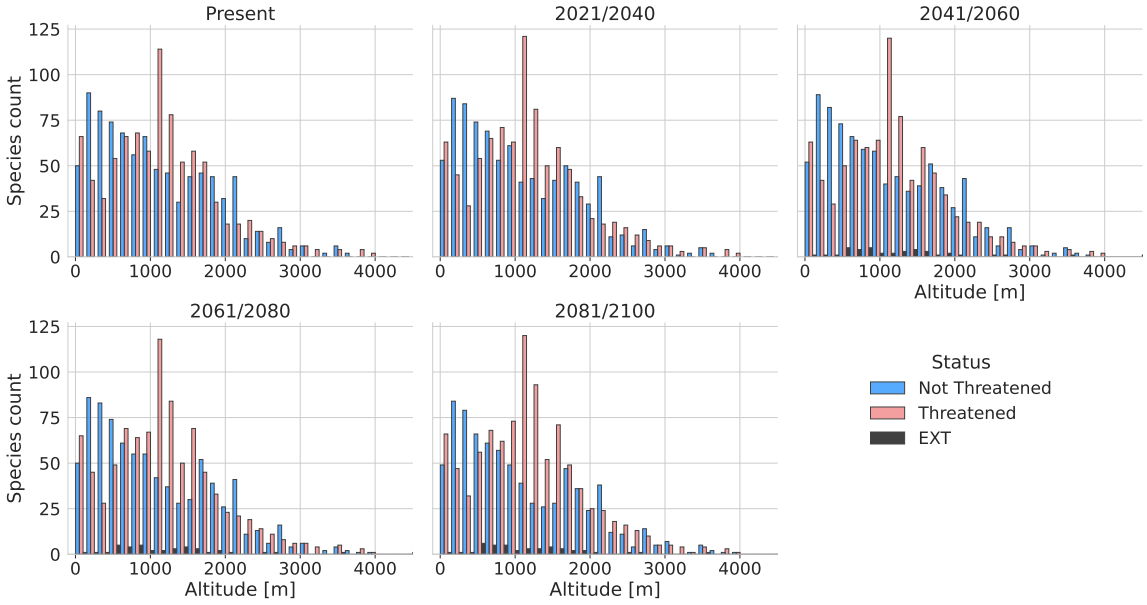


Figure S14: *Species count histograms as a function of average altitude and over time. Only IUCN-listed species are included. Boxes cover 150 metre altitude ranges. Overall, we found the same patterns in this figure as in its all-species counterpart. However, the hump-shaped concentration of threatened species at 800-1,500m is replaced here by a more irregular peak forest. Therefore, in this case, our approach seems to regularise the number of threatened species along the altitudinal gradient. Nevertheless, causality cannot be inferred from our study. A confounding variable such as threat exposure could indeed be at the origin of this pattern. In addition, species already IUCN-assessed and predicted to become extinct seem to be distributed between the lowlands and around 2000 m.*

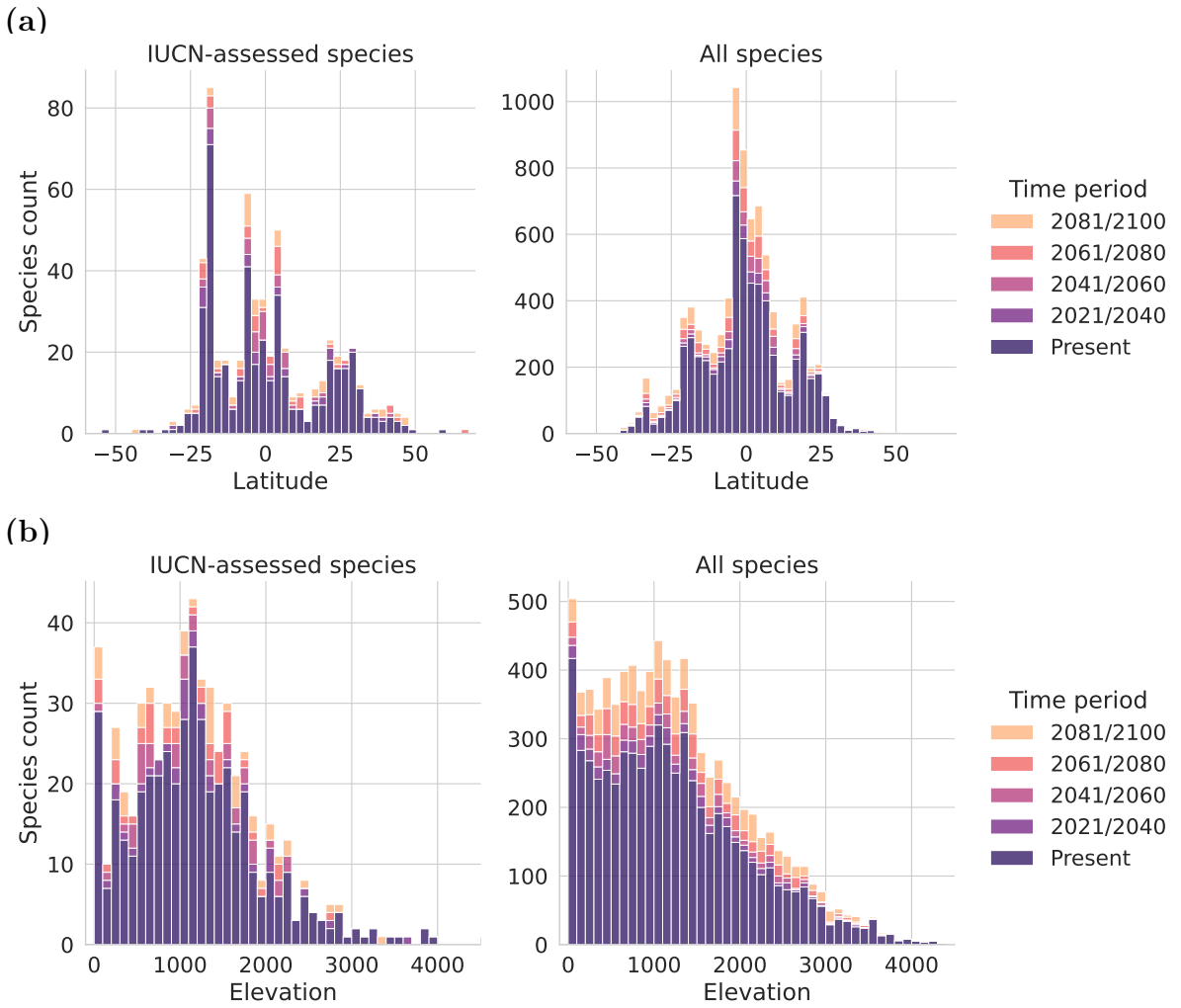


Figure S15: ***(a)*** *Cumulative histogram of the number of threatened species as a function of their average latitude. Bins are 2.5° wide.* ***(b)*** *Cumulative histogram of the number of threatened species as a function of their average altitude. Bins are 100 m wide. This type of histogram implies that a species predicted to be threatened at a given time period cannot later become non-threatened. It must therefore be interpreted with caution, as it ignores possible reclassifications of species as non-threatened.*

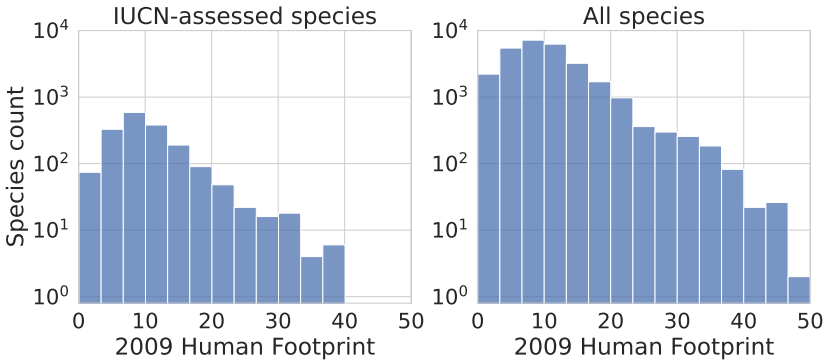


Figure S16: *Histogram of the number of species per 2009 Human Footprint (HFP) bin, considering: (left) IUCN-assessed species only and (right) all species (Venter et al., 2016). The species count axis is* ***log-scale****. The HFP score of a species is averaged over its current occurrences. These histograms allow us to assess the number of species per bin used to calculate the proportions shown in Figure S17. The main message is that high HFP bins contain very few species and the following proportions should then be treated with caution.*

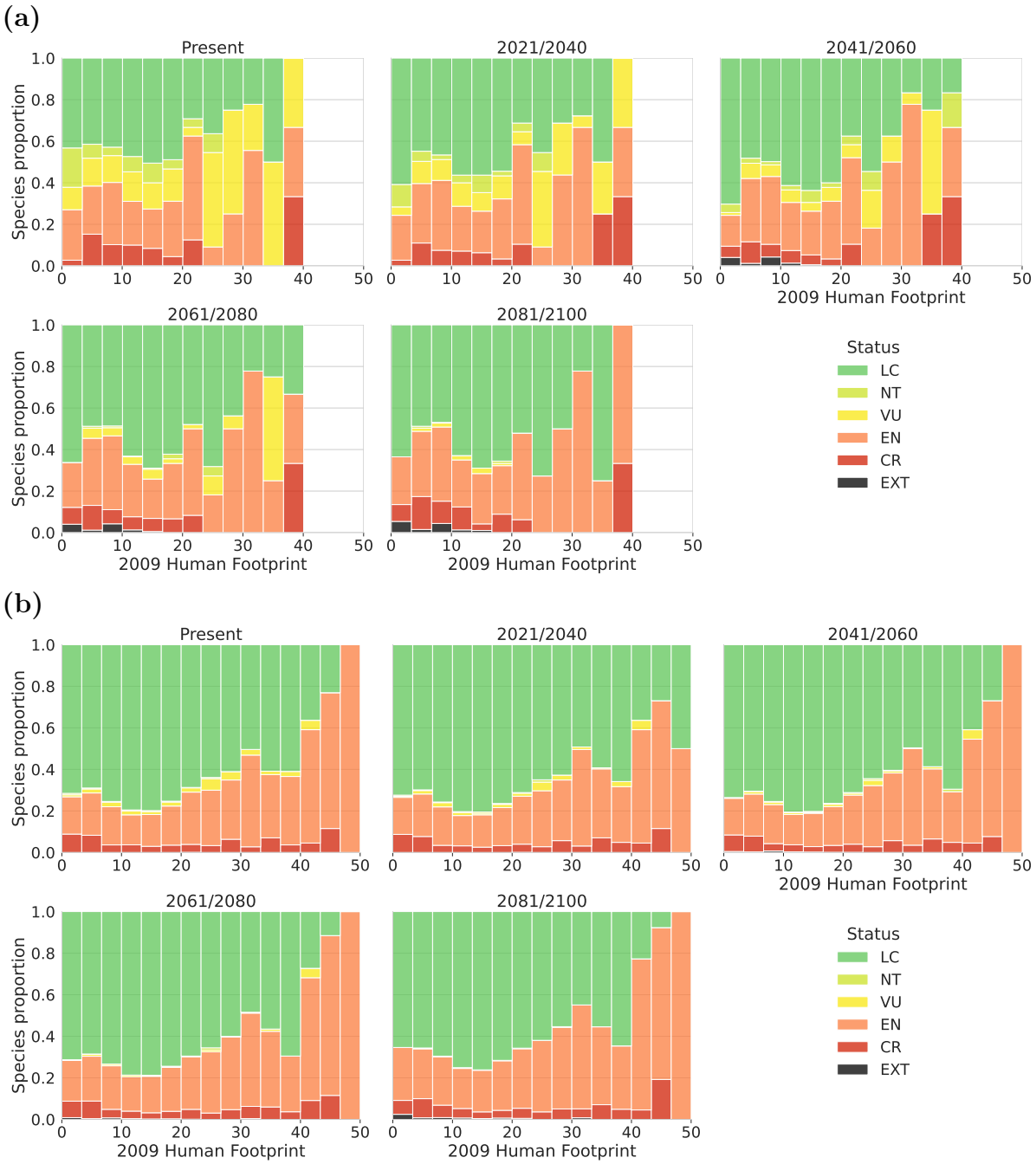


Figure S17: *Predicted status proportions per 2009 Human Footprint (HFP) bin across time periods, considering* ***(a)*** *IUCN-assessed species only or* ***(b)*** *all species in our dataset.* ***(a)*** *Present sub-figure then represents the HFP distribution of currently red listed orchids. We observe that: i) NT and VU proportions are currently significant according to IUCN, but are predicted to disappear over time. This is due to poor classifier performance on these categories and their rather confused definitions, see status confusion matrices Figure S10 and the Discussion. ii) In both cases, the proportions of threatened species appear to be positively correlated with HFP intervals. iii) HFP bins may cover few species (see Figure S16), so the robustness of these results should be further assessed. iv) As with the altitude study, this result shows a correlation, but no causality can be inferred. Again, a confounding variable may be at the origin of this pattern rather than HFP.*