HAL open science

Mapping global orchid assemblages with deep learning provides novel conservation insights

Joaquim Estopinan, Maximilien Servajean, Pierre Bonnet, Alexis Joly, François Munoz

▶ **To cite this version:**

Joaquim Estopinan, Maximilien Servajean, Pierre Bonnet, Alexis Joly, François Munoz. Mapping global orchid assemblages with deep learning provides novel conservation insights. Ecological Informatics, 2024, 81, pp.102627. 10.1016/j.ecoinf.2024.102627 . hal-04581266

HAL Id: hal-04581266

https://hal.inrae.fr/hal-04581266

Submitted on 21 May 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Ecological Informatics 81 (2024) 102627

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Mapping global orchid assemblages with deep learning provides novel conservation insights

Joaquim Estopinan [a,b,c,*], Maximilien Servajean [b,d,1], Pierre Bonnet [c,1], Alexis Joly [a,b,1], François Munoz [e,1]

[a] INRIA, Montpellier, France
[b] LIRMM, Univ. Montpellier, CNRS, Montpellier, France
[c] UMR AMAP, CIRAD, Montpellier, France
[d] AMIS, Univ. Paul Valéry Montpellier, Univ. Montpellier, CNRS, Montpellier, France
[e] LIPHY, Univ. Grenoble Alpes, CNRS, Grenoble, France

ARTICLE INFO

Keywords:
Spatial indicator
Species assemblage
Deep learning
Species distribution modelling
IUCN status
Orchids

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 1 M occurrences of 14 K 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 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.

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., 2022; 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

Abbreviations: IUCN, International Union for Conservation of Nature; SDM, Species Distribution Model; GBIF, Global Biodiversity Information Facility; PAs, Protected Areas; SI, Supplementary Information.

* Corresponding author at: ZENITH Team, INRIA, 34095 Montpellier, France.
E-mail address: joaquim.estopinan@inria.fr (J. Estopinan).

[1] Equally contributing senior authors.

https://doi.org/10.1016/j.ecoinf.2024.102627
Received 29 August 2023; Received in revised form 30 April 2024; Accepted 2 May 2024
Available online 10 May 2024
1574-9541/© 2024 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

spatial scale, and the conservation challenges must be addressed at both global scale and fine-resolution (Puglielli and 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 (DL) offers an unprecedented opportunity to characterize complex, scale-dependent relationships between species and their environment (Deneu et al., 2021). 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 and Pärtel, 2023), and Cai et al. (2022) 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 and 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). 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., 2022; González-del Pliego et al., 2019; Nic Lughadha et al., 2019; Zizka et al., 2022). 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 species distribution models (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 and 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., 2022). 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., 2021; 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., 2018a). Spatially Explicit Models (SEMs) 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 and Yurek, 2017).

Our main contribution is to produce kilometre-scale extinction risk maps of species assemblages on a global scale. A species assemblage is 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). In the context of this study, a species assemblage can simply be reformulated as the pool of orchid species with a significant probability of being present at a given location. The first step consisted in training a deep-SDM model on 1 M observations of 14 K species distributed worldwide. We then developed a novel method to estimate species assemblages from the trained deep-SDM. 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?

1.1. 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 and Widmer, 2005; Givnish et al., 2016) and a wealth of observations: 8 M raw GBIF observations, including 6.8 M 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 environment (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).

2. Material and methods

2.1. Data

2.1.1. 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 occurrences. Date distribution summary statistics are min = 1901, Q1 = 1982, med = 1997, 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 SI Box D.

2.1.2. Predictive features

A large environmental context around each observation is collected and provided to the model: 64 × 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 (longitude and latitude), see SI Box E for details. Examples of input are shown in Fig. S5 and the full list of predictors is given in Table S2.

2.2. Species assemblage model

2.2.1. Definition

The objective is to optimise a model returning likely species assemblages worldwide while being learned on a set of presence-only observations. To do so, we optimise a deep species distribution model (Botella et al., 2018b) and further calibrate it to return species assemblages including the initial species observed with very high confidence. This method 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 data, all species combined (multi-species SDM), and is then used to predict a set of labels by thresholding the SDM output categorical probabilities associated to species.

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 by $\mathscr{X} = \{x_1, \ldots, x_n\}$, where n is the total number of samples. The matching species label set is $\mathscr{Y} = \{1, \ldots, C\}$. The objective is to learn a species assemblage predictor on a training dataset composed exclusively of presence-only occurrences $(x_1, y_1), \ldots, (x_n, y_n) \in \mathscr{X} \times \mathscr{Y}$. 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 $\mathscr{X}$, $\mathbb{P}_X$, and the conditional distribution of y given an input x denoted $\eta(x) = (\eta_1(x), \ldots, \eta_C(x))$ and equal to

$$\boldsymbol{\eta}_s(x) = \mathbb{P}_{X,Y}(Y = s | X = x)$$

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

$$S^*_\lambda(x) := \{s \in \mathscr{Y} : \boldsymbol{\eta}_s(x) \geq \lambda\}$$

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 $\widehat{\eta}(x)$ from which we can derive the following *plug-in* estimator of the species assemblage:

$$S_\lambda(x) := \{s \in \mathscr{Y} : \widehat{\eta}_s(x) > \lambda\} \tag{1}$$

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

$$l_{log}(s, \widehat{\eta}) = -\log\ \widehat{\eta}_s(x) \tag{2}$$

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

$$\widehat{\eta}_s(x) = \frac{exp\left(f^s_\theta(x)\right)}{\sum_j exp\left(f^j_\theta(x)\right)}$$

where θ is the set of parameters of the neural network to be optimised by minimizing the loss function of Eq. (2).

Using this very common deep learning framework, it is possible to show that the species assemblage predictor $S_\lambda(x)$ of Eq. (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., 2016). This convolutional neural network learn spatial patterns from two-dimensional predictors (Botella et al., 2018a; Deneu et al., 2021). 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, see Table 1. We use the regions defined by the World geographical scheme for recording plant distribution (WGSRPD) level 2 (Brummitt et al., 2001). Training is done on Jean Zay, a supercomputer from the Institute for Development and Resources in Intensive Scientific Computing (IDRIS). A full description of the model architecture, dataset spatial split and training procedure can be found in supplementary information (SI) Box A. Finally, the settings of our species assemblage model are summarised in the Fig. 1.

2.2.2. Validation

The species assemblage model is calibrated on the occurrences from the validation spatial blocks (see dataset split in SI 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). Furthermore, the SI Box H provides additional results qualifying the model's precision. These include a histogram of the proportion of occupied locations per species and maps comparing individual species predictions with true observations.

Our dataset is highly unbalanced in terms of the number of occurrences per species (see SI Box D). 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 s with a non-null conditional probability $\eta_s(x)$ is potentially present in the assem-

Table 1
Dataset split proportions.

Set	Training	Validation	Test
Proportion	90%	5%	5%
Occurrences	902,174	46,290	50,794
Species	14,129	4037	4166

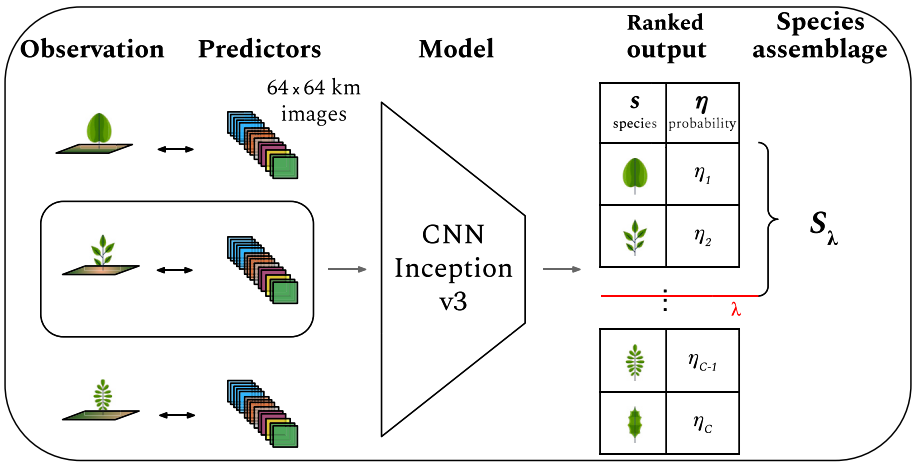

Fig. 1. Method summary scheme. The orchid observation set of 999,258 occurrences from 14,129 species (Section 2.1.1 and Box D) is enriched with predictive features: WorldClim2 bioclimatic variables, Soilgrids pedological variables, Human Footprint rasters, terrestrial ecoregions of the world and location (description Box E, list Table S2 and examples Fig. S5). Next, the Inception v3 model (Szegedy et al., 2016) is first trained with the observation set (Section 3.1.1). Secondly, it is used to predict the most likely species assemblages on a global regular grid of 30-s resolution (Section 2.4.1). The calibration step, which determines the threshold λ on the relative probability of species presence, is described theoretically in Section 2.2.2 and set in Section 3.1.2. When the model is provided with a predictor tensor in inference, the final output is the associated species assemblage S_λ.

blage (i.e. by chosing $\lambda = 0$). However, in practice, the estimator $\widehat{\eta}_s(x)$ 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 the validation set for this calibration step. It allows estimating the average error rate for a given value λ:

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

by computing the percentage of samples x_i in the validation 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:

$$\begin{array}{ll} \underset{\lambda\in[0,1]}{argmin} \ \mathbb{E}[|S_\lambda(X)|] & \\ s.t. \quad \mathscr{E}(S_\lambda) \leq \epsilon & \end{array} \Leftrightarrow \begin{array}{ll} \underset{\lambda\in[0,1]}{max} \ (\lambda) & \\ s.t. \quad \mathscr{E}(S_\lambda) \leq \epsilon & \end{array} \tag{3}$$

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 $1-\epsilon$.

2.3. Conservation indices for species assemblages

2.3.1. Definition of the indices

In addition to the classical Shannon index $\mathscr{I}_{\mathscr{H}}$, we define two novel indices characterizing the extinction risk of a predicted species assemblage, $\mathscr{I}_c$ and $\mathscr{I}_{\mathscr{O}}$. They respectively render the proportion of threatened species in the assemblage and the most critical IUCN status in the assemblage. Let's break down their construction.

2.3.2. 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.

2.3.3. Indicator $\mathscr{I}_{\mathscr{O}}(S)$: 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:

$$\begin{array}{llll} \mathscr{I}_{\mathscr{O}}: & \mathscr{P}(\mathscr{Y}) & \mapsto & E_{\text{status}} \\ & S & \rightarrow & \underset{s_j\in S}{max}\{\varphi(s_j)\} \end{array} \tag{4}$$

2.3.4. Indicator $\mathscr{I}_c(S)$: proportion of species in the assemblage with a given status

Our second indicator $\mathscr{I}_c(S)$ 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). $\mathscr{I}_c$ 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 Eq. (5)). The proportion of critically endangered species is for instance denoted $\mathscr{I}_{\text{CR}}(S)$. And so on for the four other IUCN status in E_{status} and the overall THREAT category.

$$\begin{array}{llll} \mathscr{I}_c: & \mathscr{P}(\mathscr{Y}) \times \mathbb{R}^C & \mapsto & \mathbb{R}^{[0,1]} \\ & (S,\eta) & \rightarrow & \sum_{j\in\varphi^{-1}(c)} \eta_j \end{array} \tag{5}$$

2.3.5. The Shannon index $\mathscr{I}_{\mathscr{H}}(S)$

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 $\mathscr{I}_{\mathscr{H}}$, 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:

$$\mathscr{I}_{\mathscr{H}}(S) = -\sum_{l\in S} \eta_l \cdot log(\eta_l) \tag{6}$$

2.3.6. 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., 2022). The distributions of the

IUCN-assessed and predicted IUCN status are shown in Fig. S3. 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: $\mathscr{I}^{\text{IUCN}}$.

2.4. High-resolution maps construction

2.4.1. 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-s intervals. One second equals 1/3600°, hence $r = 30/3600$ degrees. Let $\mathscr{M} = \{-180, -180 + r, \ldots, 180 - r, 180\}$ and $\mathscr{N}$ be its latitudinal counterpart. The grid support is then obtained by crossing the two sampled axes $\mathscr{M} \times \mathscr{N}$. 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 221 M points. Finally, predictive features are assigned to each land grid position. This results in $\mathscr{G} = \{x_{m,n} \mid m, n \in \mathscr{M} \times \mathscr{N}\}$.

2.4.2. Map definition and construction

Maps are constructed in two steps: First, the species assemblages associated to each $\mathscr{G}$ grid point are predicted by batch with our model: $\widehat{\mathscr{T}}_\lambda(\mathscr{G})$. Second, the spatial indices defined in section 2.3.1 are computed on the predicted assemblages: $\mathscr{I}(\mathscr{G}) = \{\mathscr{I}(S) \mid S \in \widehat{\mathscr{T}}_\lambda(\mathscr{G})\}$. This set of indicators $\{\mathscr{I}(\mathscr{G})\}$ 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:

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

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

2.5. Zonal statistics

Spatial analysis can necesit aggregated regional indicators. With a kilometre scale resolution, $\mathscr{I}_{\mathscr{O}}$ and $\mathscr{I}_c$ 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 2000 km^2 to highlight large area priorities (65 countries out of 369 removed).

2.5.1. Region spatial coverage of the most critical IUCN status

This measure is based on $\mathscr{I}_{\mathscr{O}}$, 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 $\text{Area}_{\%}[\mathscr{I}_{\mathscr{O}}](r, c)$. By default, this variable is computed on the entire species assemblage. Nonetheless, it can also be expressed considering only IUCN-assessed species.

2.5.2. Region average proportions

Second zonal statistic consists in taking $\mathscr{I}_c$ 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 $\mu[\mathscr{I}_c](r, c)$. 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.

3. Results

3.1. Validation of the species assemblage model

3.1.1. Backbone model optimisation

The backbone species distribution model (Inception v3) was first 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:

$$\mathbf{A}_k(i) = \begin{cases} 1 & \text{if } \widehat{\eta}_{y_i}(x_i) \geq \widetilde{\eta}^k(x_i) \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

with (x_i, y_i) an input/label pair, $\widetilde{\eta}$ the permutation of $\widehat{\eta}$ sorted in descending order and $\widehat{\eta}^k$ its component at rank k. The success rate can be calculated for all test set occurrences, all classes combined (*micro-average* denoted $\mathbf{A}_k$) or first for each class individually and then averaged together (*macro-average* denoted $\mathbf{MSA}_k$). 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 $\mathbf{A}_{30} = 0.87$ and $\mathbf{MSA}_{30} = 0.48$. 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 4166 species

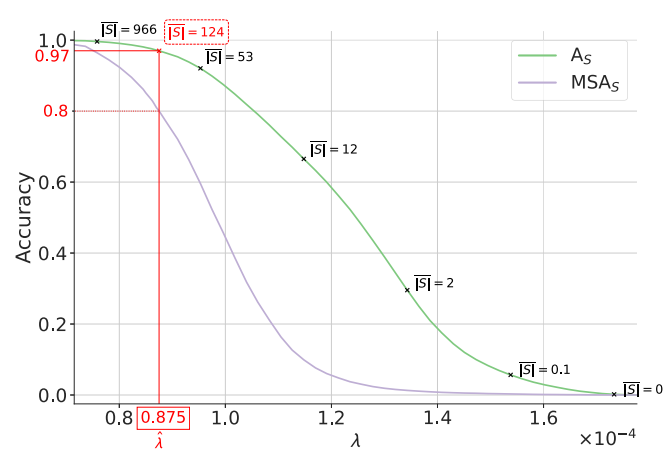

Fig. 2. Average error control setting on the validation set. The accuracy of the model is represented against the threshold on the species conditional probability of presence, denoted by λ. To calibrate the model and reach precautionary species assemblages at any point, we set the limiting condition on the average error to be $\epsilon = 0.03 \Leftrightarrow \mathbf{A}_S \geq 0.97$ (green curve), as shown in Eq. (3). The optimal threshold, denoted by $\widehat{\lambda}$, is highlighted in red. It guarantees that *i)* $\mathbf{A}_S$ is superior or equal to 0.97 while *ii)* the average species assemblage size is as small as possible. Matching macro-average accuracy $\mathbf{MSA}_S$ (grey function) is reported with a red dashed line. Average set sizes are indicated. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

($30/4166 \leq 1\%$). Furthermore, it reflects the performance of the model on rare species, and Fig. 2 shows that considering *on average* 124 species significantly improves performance on the validation set. Training and validation curves show no sign of overfitting, see Fig. S2.

3.1.2. Calibration of the species assemblage predictions

Once the backbone species distribution model optimised, we calibrate the species assemblage predictions. To do so, we control the average error on the validation set when testing if the correct species is returned within the assemblage thresholded with the hyperparameter λ (see method explanation Section 2.2.2). In Eq. (3), ϵ is set to 0.03. As reported on Fig. 2, the resulting value for λ is estimated to be $8.75e-5$. The average size of the predicted species assemblages is equal to 124 species. Summary statistics on $|\widehat{S}_\lambda|$ are reported on Table S1.

$\mathbf{A}_S$ is the micro-average accuracy of the species assemblage model when testing if the observed species is well retained within the predicted assemblage across all validation observations. $\mathbf{MSA}_S$ is its macro-average counterpart: every species contributes therefore equally to this second metric. Reaching $\mathbf{A}_S = 0.97$ accuracy means that, for a random validation point, the model returns the correct label, i.e. the species that has effectively been observed at this point, within the predicted assemblage with 97% confidence. The number of observations per class being strongly unbalanced (see Box D Fig. a), $\mathbf{A}_S$ is strongly influenced by the performance on common species. Now, when all species are granted the same weight with $\mathbf{MSA}_S$, performance is still $\mathbf{MSA}_S = 0.80$, see Fig. 2. Given how unbalanced the observation dataset is (median occurrence number is four, 25% species have more than 13 occurrences), it indicates that the predicted species assemblages are capturing most of the rare validation species as well. The performance at the species level shows the robustness of our assemblages and the performance at the point level its validity in space.

After validating and calibrating the model on the validation set for predicting species assemblages, a new training is started from scratch on the entire dataset. It stops at the best epoch previously determined on the validation set (epoch 69). The aim is to obtain the best possible model weights before global-scale inference. 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 SI Box B for more details and resulting maps.

3.2. $\mathcal{I}_{\mathscr{O}}$ indicator: Most critical status of the species in the assemblage

3.2.1. Global patterns

Considering the worst status of a species assemblage, Fig. 3 compares (a) currently available IUCN information with (b,c) our model results $\mathcal{I}_{\mathscr{O}}^{\text{IUCN}}$ and $\mathcal{I}_{\mathscr{O}}$. 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. 3b.

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. 3c, the value of $\mathcal{I}_{\mathscr{O}}$ 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 Fig. 3c, 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.

3.2.2. Country-level analysis

Table 2 shows the botanical countries with the largest $\mathcal{I}_{\mathscr{O}}$ 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 information T3 for the full table. High on the $\text{Area}_{\%}(\text{CR})$ 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 $\text{Area}_{\%}(\text{EN})$ are different from the first column. European territories such as Corse or Portugal appear in the ranking and Caribbean islands are well represented.

3.3. $\mathcal{I}_c$ indicator: Proportion of species in the assemblage with a given status

3.3.1. Global patterns

Fig. 4a 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 SI Box D 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 $\mathcal{I}_{\mathscr{H}}$ patterns. Colombia orchid richness, estimated for instance at 4327 species according to *World Plants* (Hassler, 2004-2023), is for instance under-represented within our occurrence set with only 1375 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 Fig. 4b $\mathcal{I}_{\text{THREAT}}$. 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 Asia, they are not as high as the region's impressive orchid richness would suggest. This result is quantified on the scatter plot Fig. 5. High diversity does not necessarily imply high threat levels.

On Fig. 3c 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. 4d) 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.

3.3.2. Country-level analysis

In Table 3, the top three botanical countries with the highest average

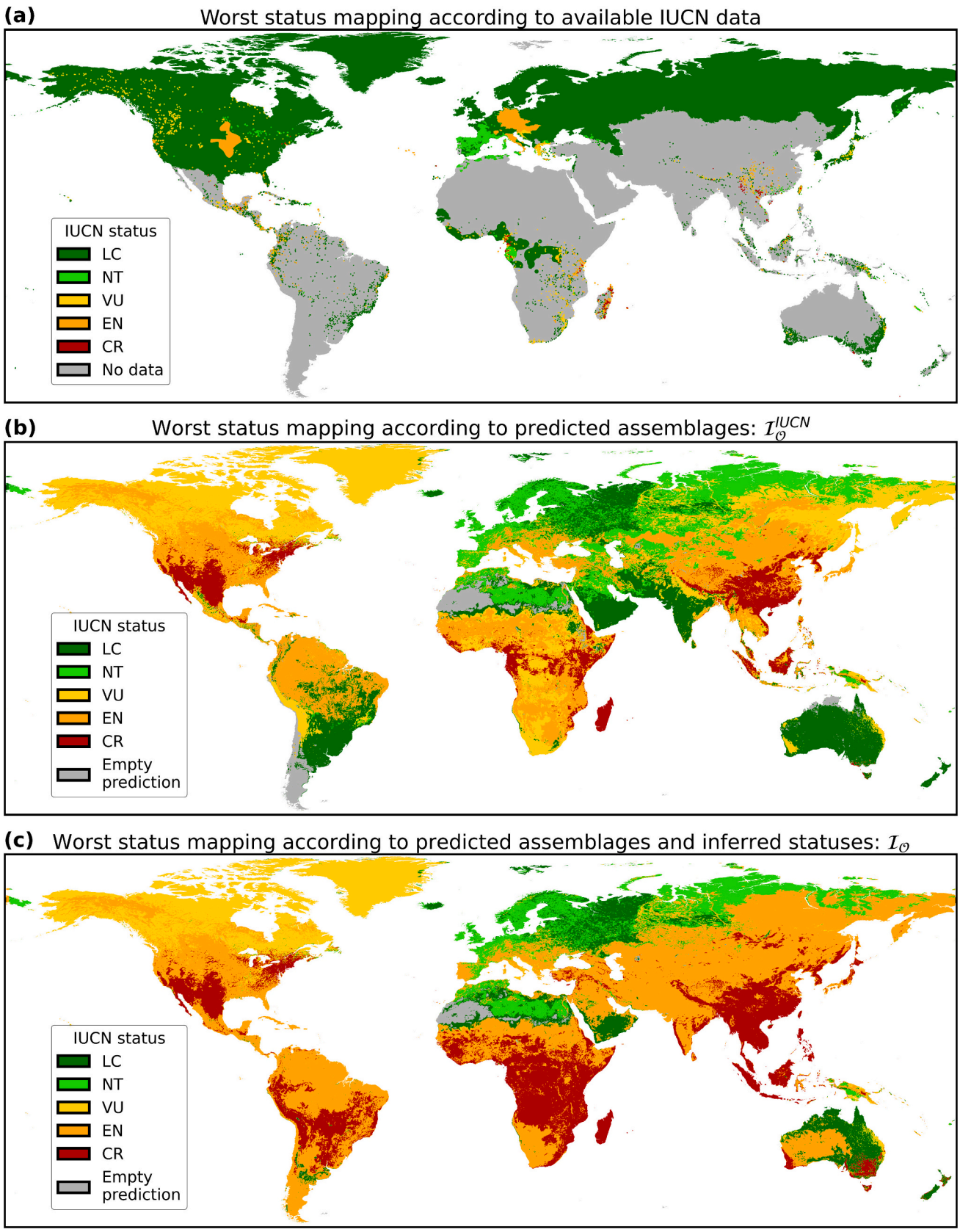

Fig. 3. Global comparison of the most critical IUCN status indicator according to three methods. (a) represents the IUCN information on our dataset: observations and available spatial data (polygons and points from *https://www.iucnredlist.org/resources/spatial-data-download*) taken together. Spatial 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 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].

Table 2
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	$\text{Area}_{\%}$	B. country	$\text{Area}_{\%}$
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).

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 information T3 for the full table.

The scatterplot Fig. 5 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 3 first column) are this time marked with red borders. The top fifteen $\mu[\mathcal{I}_{\mathscr{H}}]$ 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 Fig. 4: this continent is highly diverse with relatively low levels of threat to its species assemblages. A Venn diagram crossing $\mu[\mathcal{I}_{\mathscr{H}}]$ and $\mu[\mathcal{I}_{\text{THREAT}}]$ top-30 countries plus the Spearman correlations per continent are available at Fig. S6.

3.4. Sumatra case study

On the western side of Sumatra, the Barisan Mountains form a sharp relief (see Fig. 6a). The *elevational diversity gradient* theory would suggest that species richness is particularly high along the mountainous area. However, according to the $\mathcal{I}_{\mathscr{H}}$ 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. 4e 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: $\frac{\mu[\mathcal{I}_{\text{CR}}]}{\mu\left[\mathcal{I}_{\text{CR}}^{\text{IUCN}}\right]}(\text{PAs}) = 3.1$. This ratio is even greater when all threatened species are considered together: $\frac{\mu[\mathcal{I}_{\text{THREAT}}]}{\mu\left[\mathcal{I}_{\text{THREAT}}^{\text{IUCN}}\right]}(\text{PAs}) = 7.1$ 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. Discussion

4.1. Modelling choices and considerations on covariates

Our species assemblage predictor has theoretical guarantees that we have validated on a previously unseen observation set (see Section 3.1). 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. Further considerations on the model (on the trade-off between model generalisation and over-prediction, on the difficulty of measuring the precision of the model) are carefully detailed in Box F. In addition, Box H provides additional results on the precision of the model by investigating individual species predictions. A histogram represents the proportion of occupied sites per species for a random sample of two hundred species and prediction maps of a hundred species are compared to the observations in the dataset.

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

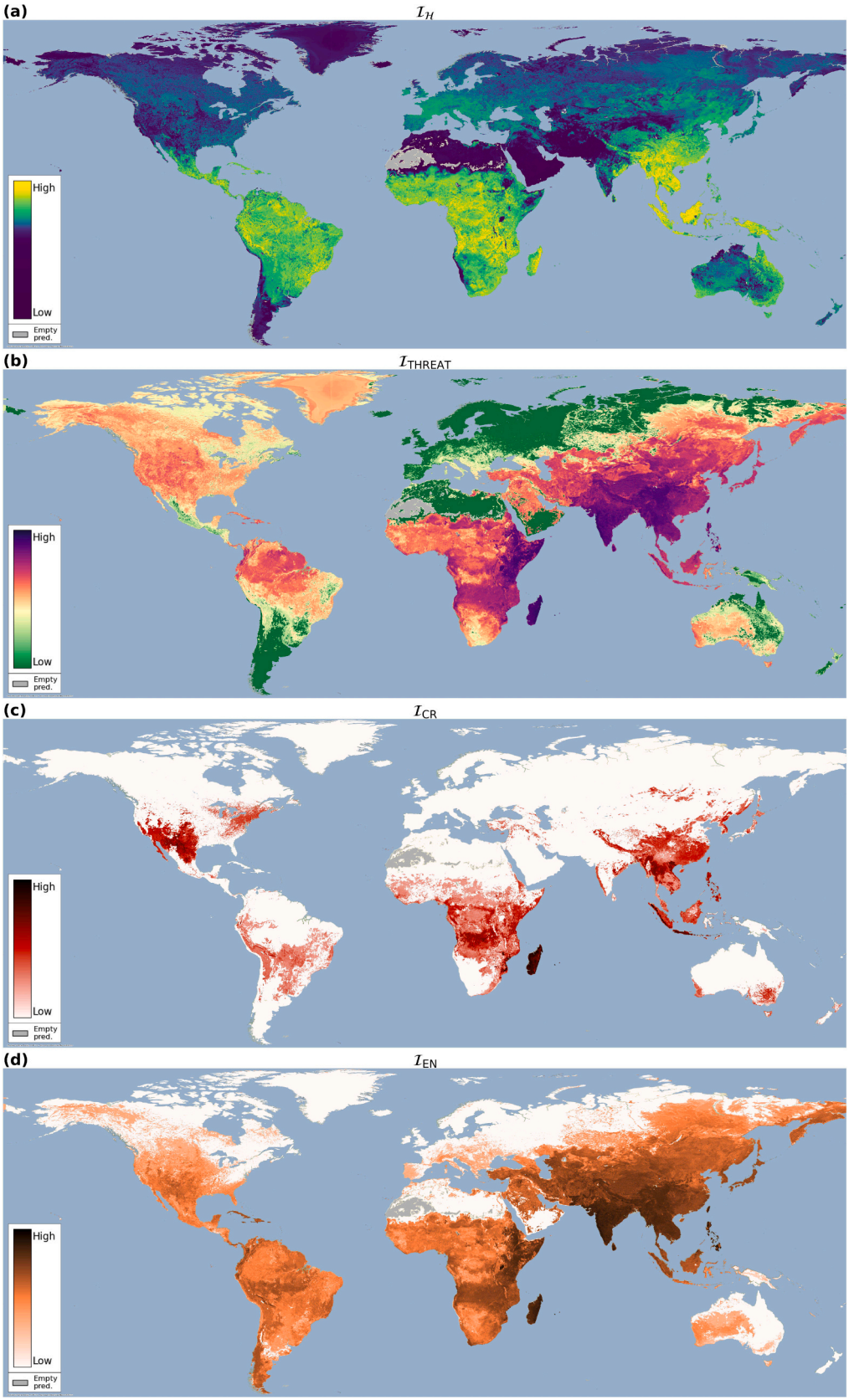

Fig. 4. Four indicators based on species assemblage predictions. (a) $\mathcal{I}_{\mathcal{H}}$ the Shannon index, (b) $\mathcal{I}_{\text{THREAT}}$ the weighted proportion of threatened species, (c) and (d) the weighted proportions of respectively CR species $\mathcal{I}_{\text{CR}}$ and EN species $\mathcal{I}_{\text{EN}}$. [Figure maps are under-sampled, see the website for full-resolution].

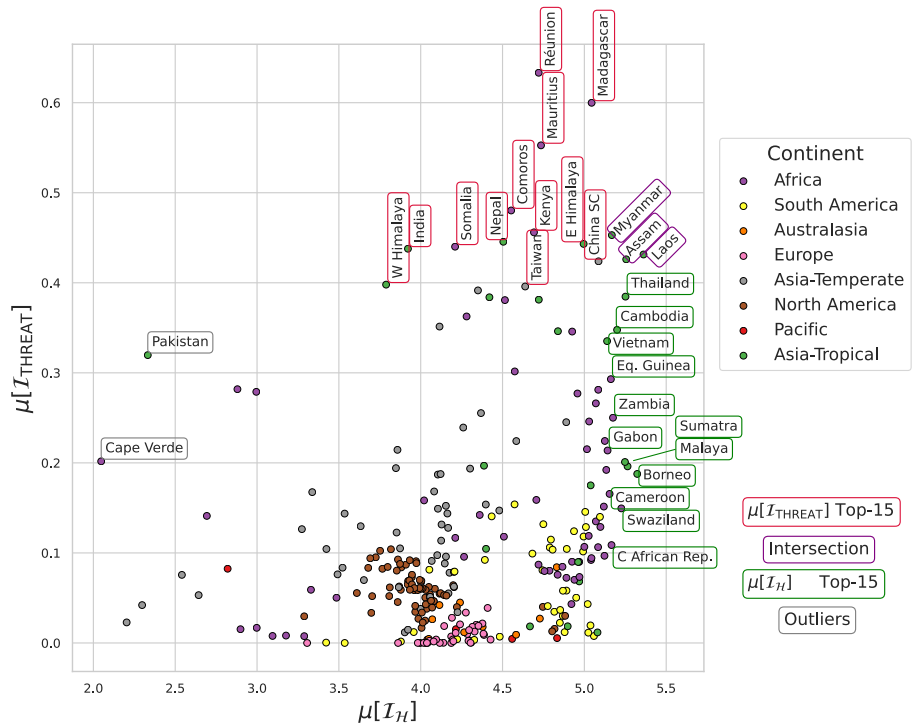

Fig. 5. Average proportion of species predicted as threatened by botanical country (WGSRPD level 3) versus average Shannon index. Countries are coloured in function of their continent (WGSRPD level 1) and top-15 countries of both variables are highlighted. Myanmar, Assam and Laos are the only three regions in the top-15 intersection whereas Pakistan and Cape Verde show especially high threatened species proportions with low diversity indices.

Table 3
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

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 and 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).

4.2. Error and uncertainty quantification

The uncertainty (both aleatoric and epistemic) attached to species assemblage predictions was not estimated using classic methods such as bootstrap, dropout, ensembles or Bayesian neural networks. However, the shortcomings of our method in terms of error and uncertainty are further discussed with: *i)* a map of the average error rate per botanical country when testing whether validated observations are predicted within the assemblages, *ii)* a map estimating the uncertainty associated with the assemblages, and *iii)* a focus on the error propagation of automated IUCN assessments.

First, we introduce the average error rate E_S^R per botanical country of our predicted assemblages, calculated on the validation and test sets. For each observation, E_S measures whether or not the correct label is predicted within the assemblage ($E_S = 1 - A_S$). While the average error rate across the validation set is $E_S = 0.03$, this map shows the spatial variation in the quality of the predicted assemblages, see Fig. S7. Overall, average error rates are higher in the southern hemisphere. Some regions, such as Uruguay or Angola, have no correctly predicted observations within the assemblages, but they host less than five points. India, South East Asia and northern South America have relatively high average error rates. This can be read in parallel with the species richness map per botanical country in Fig. S4. Indeed, it is more difficult to predict the correct species assemblages in rich regions where the potential for confusion between species is greater.

Secondly, we present an attempt to map the uncertainty of predicted assemblages derived from the logit sum of the top-30 predicted species, see Fig. S8. Using the top-30 species is coherent with the validation of the species distribution model. This approach to exploring model uncertainty is based on the assumption that logit levels reflect a degree of model confidence in the predictions. Global uncertainty is high, with the Sahara being the region with the highest uncertainty. France and

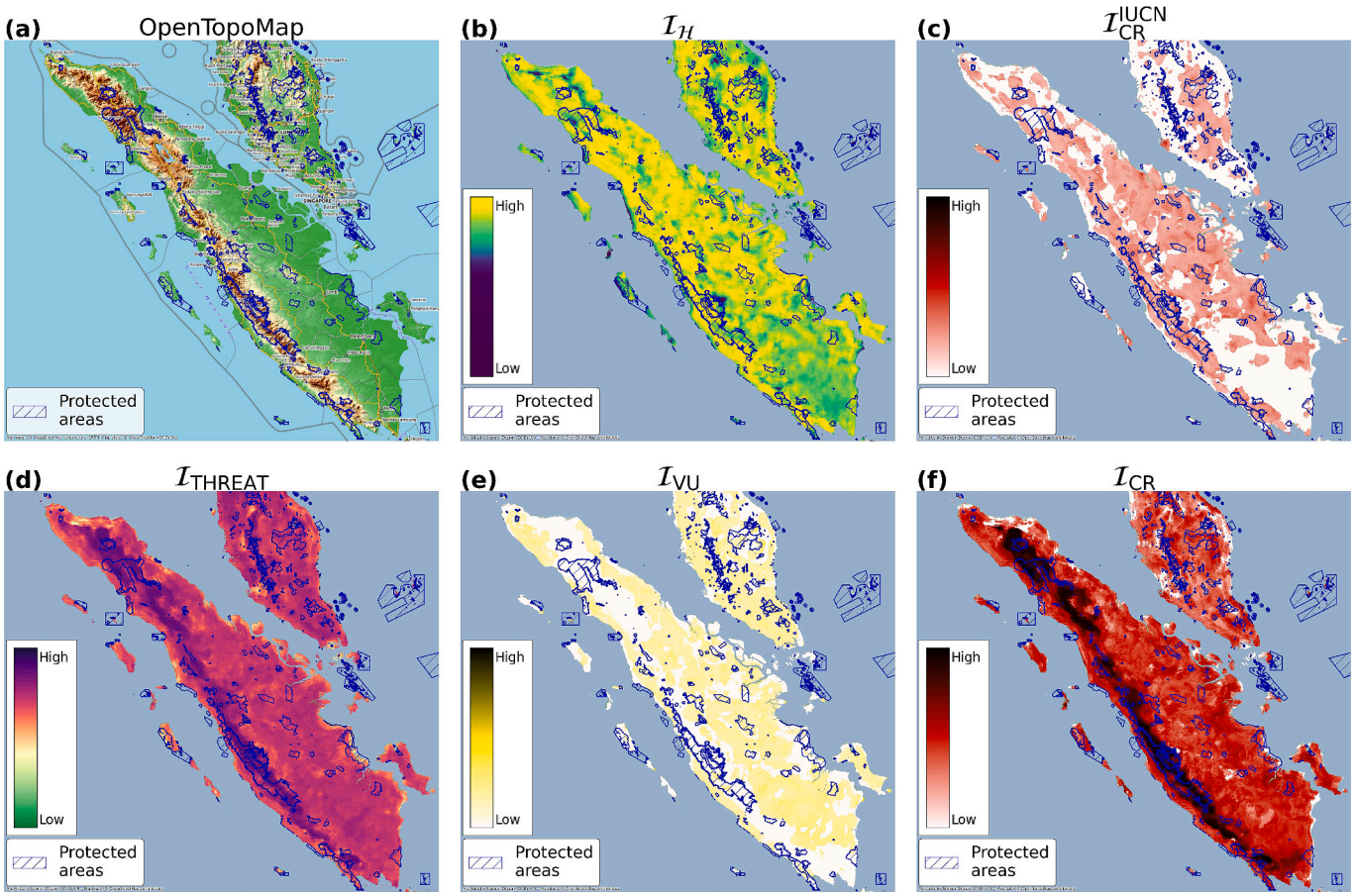

Fig. 6. Five indicators of species assemblage extinction risk applied on Sumatra island. Elevation is also provided and protected areas are hashed in blue (downloaded from *https://www.protectedplanet.net/en*). (a) elevation map, (b) Shannon index, (c) proportion of IUCN-assessed CR species in the predicted 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]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Germany have the lowest uncertainty. The United States and Canada show mixed but overall lower than average uncertainty. The observation map in Fig. S4 explains some of the observed patterns: for example, there are occurrences in southern Chile but not in Patagonia, and no occurrences in the Sahara.

Third, considering not only IUCN-assessed species, but all orchid species in our dataset, 93.7% of the extinction risk information used is estimated using the IUCNN method (Zizka et al., 2020). The corresponding maps and results are therefore dependent on the accuracy of the IUCN automated assessment method. Different IUCN status predictions from IUCNN (or any other status prediction method) would strongly affect our map patterns and zonal statistics when considering all species from the predicted assemblages. The state-of-the-art IUCNN method has an accuracy of 84% for binary prediction and 64% for status prediction. At the status level, the intermediate threat categories NT and VU are particularly difficult to estimate, both because of their close definition and their relatively low representation. Other statuses are well estimated (although CR species are frequently confused with EN), see the confusion matrix in Supplementary Information Table 3 (Zizka et al., 2020). The IUCNN authors also reported that accuracy was highest for species considered threatened by *modification of natural systems*, *energy production* and *mining*, and lowest for species threatened by *human intrusion*, *disturbance* and *pollution*. Indeed, it seems more difficult to detect harmful human disturbance remotely than large-scale habitat modification. Consequently, such a bias will be reflected in the maps produced. In addition, a class of Bayesian neural networks was introduced to quantify prediction uncertainty and possibly retain only high confidence status predictions (Zizka et al., 2022). With this classifier it would be possible to precisely study how status uncertainty propagates into all our extinction risk indicators. However, our aim here was to focus on the proof-of-concept methodological workflow and possible applications, assuming the status information is correct. Furthermore, the quality of automated IUCN assessments will improve with the coverage and consistency of manual assessments, and with increasing access to appropriate environmental predictors. While current automated (and manual) assessments are still plagued by errors, we have taken the position to build on this knowledge and produce ambitious indicators as they continue to improve. Finally, it is partly the uncertainty of status predictions that has led us to design a very precautionary indicator: $\mathcal{I}_{\mathcal{O}}$ the most critical IUCN status from the predicted species assemblages. In doing so, we respect the IUCN principle of using the most critical status when there is uncertainty. In the assessment process, this prevails when different criteria with adequate data lead to different threat categories. Here we apply this principle to all species in a given assemblage. SI Box G proposes a simplified attempt to quantify the probability of underestimating $\mathcal{I}_{\mathcal{O}}$ at a given location.

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. For example, India ranks fourth in terms of its average proportion of CR species (Table 3 last column). Looking at the $\mathcal{I}_{CR}$ 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 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 221 M 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 and Hutchings, 2006), and their interactions with pollinators and mycorrhizal associations contribute to ecosystem functioning (Swarts and 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 $\mathscr{I}_{\mathscr{O}}$ and $\mathscr{I}_c$ 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.

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*. A strength of our indicators is to meet recommendation five. However, a downsize is that they also suffer from a lack of ground truthing to be confidently applied on the ground, as the fourth recommendation points out.

4.4. Comparison with existing indicators

We can further weigh the pros and cons of our method by comparing our results with previous attempts to map orchid extinction risk and diversity.

To start with, we compare our work with (Zizka et al., 2020): The $\mu[\mathscr{I}_{\text{THREAT}}]$ top countries, i.e. the countries with the highest average proportion of predicted threatened species, largely overlap with the countries identified in as having the highest *proportion* (and not the highest *number*) of potentially threatened species. This point of convergence is reassuring since we processed the same species and occurrences.

Protecting species for their evolutionary distinctiveness, combined with an IUCN threatened status, is another 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*. This approach has the advantage of basing conservation priorities on fully available data. It shares the same goal of informing the conservation of data-deficient orchid taxa by highlighting urgent locations. Their analyses and conclusions are carried out at the level of botanical countries. Here, the spatial ranges considered are compiled from the WCSP (World Checklist of Selected Plant Families) and GIFT (Global Inventory of Floras and Traits) databases. Tropical Africa does not emerge as a clear priority hotspot, as our indicators using IUCN information 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 ($R^2 = 0.87$).

On a global scale the speciation rates highlighted by Perez-Escobar et al. (2023) correspond overall to ecoregions identified as highly diverse by the Shannon index and having high threatened species proportions according to $\mathscr{I}_{\text{THREAT}}$. However, this is not reciprocal as many regions predicted with high diversity and threat levels by our indicators do not present high speciation rates, as in Angola and Zambia. Moreover, when we zoom in more details on the Costa Rican ecoregions with the highest speciation rates, some of them, such as the Cordillera de Talamanca, do present very high extinction risk levels, while others, as the Nicoya Peninsula, are projected to be relatively safe from extinction.

The global extinction probability of terrestrial vascular plants from Verones et al. (2022) is an indicator than can be compared to $\mathscr{I}_{\text{THREAT}}$. 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 $\mathscr{I}_c$. 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., 2022) are the closest available point of comparison. Again, both the resolution and construction of our indicator differ from previous work.

4.5. Orchid 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.

5. Conclusions

Based on a deep-SDM architecture, 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 at https://mapviewer.plantnet.org/?config=apps/store/orchid-status.xml#. 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).

Funding information

CACTUS exploration action, INRIA GUARDEN, European Commission, Grant Number: 101060693 MAMBO, European Commission, Grant Number: 101060639.

CRediT authorship contribution statement

Joaquim Estopinan: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation. **Maximilien Servajean:** Writing – review & editing, Visualization, Supervision, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Pierre Bonnet:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization. **Alexis Joly:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization. **François Munoz:** Writing – review & editing, Validation, Supervision, Methodology, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare no conflict of interest.

Data availability

The data and code that support the findings of this study are openly available in github and figshare at doi: https://doi.org/10.6084/m9.figshare.22803431.

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2024.102627.

References

Almpanidou, V., Doxa, A., Mazaris, A.D., 2021. Combining a cumulative risk index and species distribution data to identify priority areas for marine biodiversity conservation in the Black Sea. Ocean Coast. Manag. 213, 105877. URL: https://www.sciencedirect.com/science/article/pii/S0964569121003604.

Borgelt, J., Dorber, M., Høiberg, M.A., Verones, F., 2022. More than half of data deficient species predicted to be threatened by extinction. Commun. Biol. 5, 1–9. URL: https://www.nature.com/articles/s42003-022-03638-9. Number: 1 Publisher: Nature Publishing Group.

Borowiec, M.L., Dikow, R.B., Frandsen, P.B., McKeeken, A., Valentini, G., White, A.E., 2022. Deep learning as a tool for ecology and evolution. Methods Ecol. Evol. https://doi.org/10.1111/2041-210X.13901.

Botella, C., Joly, A., Bonnet, P., Monestiez, P., Munoz, F., 2018a. A deep learning approach to species distribution modelling. In: Multimedia Tools and Applications for Environmental & Biodiversity Informatics, pp. 169–199.

Botella, C., Joly, A., Bonnet, P., Monestiez, P., Munoz, F., 2018b. Species distribution modeling based on the automated identification of citizen observations. Appl. Plant Sci. 6, e1029.

Botella, C., Joly, A., Bonnet, P., Munoz, F., Monestiez, P., 2021. Jointly estimating spatial sampling effort and habitat suitability for multiple species from opportunistic presence-only data. Methods Ecol. Evol. https://doi.org/10.1111/2041-210X.13565.

Bourhis, Y., Bell, J.R., Shortall, C.R., Kunin, W.E., Milne, A.E., 2023. Explainable neural networks for trait-based multispecies distribution modelling–a case study with butterflies and moths. Methods Ecol. Evol. 14, 1531–1542.

Breiner, F.T., Guisan, A., Nobis, M.P., Bergamini, A., 2017. Including environmental niche information to improve IUCN red list assessments. Divers. Distrib. 23, 484–495. https://doi.org/10.1111/ddi.12545.

Brummitt, R.K., Pando, F., Hollis, S., Brummitt, N., 2001. World geographical scheme for recording plant distributions. In: International Working Group on Taxonomic Databases for Plant Sciences, TDWG.

Cai, L., Kreft, H., Taylor, A., Denelle, P., Schrader, J., Essl, F., van Kleunen, M., Pergl, J., Pyšek, P., Stein, A., Winter, M., Barcelona, J.F., Fuentes, N., Inderjit, Karger D.N., Kartesz, J., Kuprijanov, A., Nishino, M., Nickrent, D., Nowak, A., Patzelt, A., Pelser, P.B., Singh, P., Wieringa, J.J., Weigelt, P., 2022. Global models and predictions of plant diversity based on advanced machine learning techniques. New Phytol. https://doi.org/10.1111/nph.18533.

Chzhen, E., Denis, C., Hebiri, M., Lorieul, T., 2021. Set-valued classification–overview via a unified framework. arXiv preprint arXiv:2102.12318.

Collen, B., Ram, M., Zamin, T., McRae, L., 2008. The tropical biodiversity data gap: addressing disparity in global monitoring. Tropic. Conserv. Sci. 1, 75–88.

Cozzolino, S., Widmer, A., 2005. Orchid diversity: an evolutionary consequence of deception? Trends Ecol. Evol. 20, 487–494. URL: https://www.sciencedirect.com/science/article/pii/S0169534705001928.

Cribb, P.J., Kell, S.P., Dixon, K.W., Barrett, R.L., 2003. Orchid conservation: a global perspective. Orchid Conserv. 124.

Dauby, G., Stévart, T., Droissart, V., Cosiaux, A., Deblauwe, V., Simo-Droissart, M., Sosef, M.S.M., Lowry, P.P., Schatz, G.E., Gereau, R.E., Couvreur, T.L.P., 2017. ConR: an R package to assist large-scale multispecies preliminary conservation assessments using distribution data. Ecol. Evol. 7, 11292–11303. https://doi.org/10.1002/ece3.3704.

DeAngelis, D.L., Yurek, S., 2017. Spatially explicit modeling in ecology: a review. Ecosystems 20, 284–300. https://doi.org/10.1007/s10021-016-0066-z.

Deneu, B., Servajean, M., Bonnet, P., Botella, C., Munoz, F., Joly, A., 2021. Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment. PLoS Comput. Biol. 17, e1008856. URL: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008856. Publisher: Public Library of Science.

Deneu, B., Joly, A., Bonnet, P., Servajean, M., Munoz, F., 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? Front. Plant Sci. 13 https://doi.org/10.3389/fpls.2022.839279.

Díaz, S., Settele, J., Brondízio, E.S., Ngo, H.T., Agard, J., Arneth, A., Balvanera, P., Brauman, K.A., Butchart, S.H.M., Chan, K.M.A., Garibaldi, L.A., Ichii, K., Liu, J., Subramanian, S.M., Midgley, G.F., Miloslavich, P., Molnár, Z., Obura, D., Pfaff, A., Polasky, S., Purvis, A., Razzaque, J., Reyers, B., Chowdhury, R.R., Shin, Y.-J., Visseren-Hamakers, I., Willis, K.J., Zayas, C.N., 2019. Pervasive human-driven decline of life on earth points to the need for transformative change. Science 366, eaax3100. https://doi.org/10.1126/science.aax3100. Publisher: American Association for the Advancement of Science.

Domisch, S., Friedrichs, M., Hein, T., Borgwardt, F., Wetzig, A., Jähnig, S.C., Langhans, S. D., 2019. Spatially explicit species distribution models: a missed opportunity in conservation planning? Divers. Distrib. 25, 758–769. https://doi.org/10.1111/ddi.12891.

Elith, J., Leathwick, J.R., 2009. Species distribution models: ecological explanation and prediction across space and time. Annu. Rev. Ecol. Evol. Syst. 40, 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159.

Esri, 2023. World Countries. https://hub.arcgis.com/datasets/esri::world-countries/about. Accessed on 2023-04-20.

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. Front. Plant Sci. 13 https://doi.org/10.3389/fpls.2022.839327.

Fauth, J., Bernardo, J., Camara, M., Resetarits Jr., W., Van Buskirk, J., McCollum, S., 1996. Simplifying the jargon of community ecology: a conceptual approach. Am. Nat. 147, 282–286.

Fay, M.F., 2018. Orchid conservation: how can we meet the challenges in the twenty-first century? Bot. Stud. 59, 16. https://doi.org/10.1186/s40529-018-0232-z.

Fontana, M., Zeni, G., Vantini, S., 2023. Conformal prediction: a unified review of theory and new challenges. Bernoulli 29, 1–23.

Gale, S.W., Fischer, G.A., Cribb, P.J., Fay, M.F., 2018. Orchid conservation: bridging the gap between science and practice. Bot. J. Linn. Soc. 186, 425–434. https://doi.org/10.1093/botlinnean/boy003.

Gaston, K.J., 2000. Global patterns in biodiversity. Nature 405, 220–227. URL: https://www.nature.com/articles/35012228. Number: 6783 Publisher: Nature Publishing Group.

Gaston, K.J., Blackburn, T.M., 1997. The spatial distribution of threatened species: macro-scales and New World birds. Proc. R. Soc. Lond. Ser. B Biol. Sci. 263, 235–240. https://doi.org/10.1098/rspb.1996.0037. Publisher: Royal Society.

GBIF, 2023. Orchidaceae. https://www.gbif.org/species/7689. Accessed on 2023-03-31.

Gillespie, L., Ruffley, M., Exposito-Alonso, M., 2022. An image is worth a thousand species: combining neural networks, citizen science, and remote sensing to map biodiversity. https://doi.org/10.1101/2022.08.16.504150v1. Pages: 2022.08.16.504150 Section: New Results.

Givnish, T.J., Spalink, D., Ames, M., Lyon, S.P., Hunter, S.J., Zuluaga, A., Doucette, A., Caro, G.G., McDaniel, J., Clements, M.A., Arroyo, M.T.K., Endara, L., Kriebel, R., Williams, N.H., Cameron, K.M., 2016. Orchid historical biogeography, diversification, Antarctica and the paradox of orchid dispersal. J. Biogeogr. 43, 1905–1916. https://doi.org/10.1111/jbi.12854.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. J. Am. Stat. Assoc. 102, 359–378.

González-del Pliego, P., Freckleton, R.P., Edwards, D.P., Koo, M.S., Scheffers, B.R., Pyron, R.A., Jetz, W., 2019. Phylogenetic and trait-based prediction of extinction risk for data-deficient amphibians. Curr. Biol. 29, 1557–1563.e3. URL: http://www.sciencedirect.com/science/article/pii/S0960982219304038.

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

Hamilton, H., Smyth, R.L., Young, B.E., Howard, T.G., Tracey, C., Breyer, S., Cameron, D. R., Chazal, A., Conley, A.K., Frye, C., Schloss, C., 2022. Increasing taxonomic diversity and spatial resolution clarifies opportunities for protecting US imperiled species. Ecol. Appl. 32, e2534 https://doi.org/10.1002/eap.2534.

Han, Y., Dong, S., Wu, X., Liu, S., Su, X., Zhang, Y., Zhao, H., Zhang, X., Swift, D., 2019. Integrated modeling to identify priority areas for the conservation of the endangered plant species in headwater areas of Asia. Ecol. Indic. 105, 47–56. URL: https://www.sciencedirect.com/science/article/pii/S1470160X19304091.

Hassler, M., 2004-2023. World Plants. Synonymic Checklist and Distribution of the World Flora. Version 16.1. www.worldplants.de.

He, K.S., Bradley, B.A., Cord, A.F., Rocchini, D., Tuanmu, M.-N., Schmidtlein, S., Turner, W., Wegmann, M., Pettorelli, N., 2015. Will remote sensing shape the next generation of species distribution models? Remote Sens. Ecol. Conserv. 1, 4–18. https://doi.org/10.1002/rse2.7.

Isaac, N.J., Pearse, W.D., 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.

IUCN, 2022. Barometer of Life. https://www.iucnredlist.org/about/barometer-of-life. Accessed: 2023-07-03.

Jørgensen, S., Xu, L., Costanza, R., 2016. Handbook of Ecological Indicators for Assessment of Ecosystem Health. CRC Press.

Kew, R.B.G., 2023. Plants of the Word Online. https://powo.science.kew.org/results?q=Orchidaceae.

Kull, T., Hutchings, M.J., 2006. A comparative analysis of decline in the distribution ranges of orchid species in Estonia and the United Kingdom. Biol. Conserv. 129, 31–39.

Leblanc, C., Joly, A., Lorieul, T., Servajean, M., Bonnet, P., 2022. Species distribution modeling based on aerial images and environmental features with convolutional neural networks. In: Working Notes of CLEF 2022-Conference and Labs of the Evaluation Forum, pp. 2123–2150.

Lembrechts, J.J., Nijs, I., Lenoir, J., 2019. Incorporating microclimate into species distribution models. Ecography 42, 1267–1279. https://doi.org/10.1111/ecog.03947.

Linardatos, P., Papastefanopoulos, V., Kotsiantis, S., 2021. Explainable AI: a review of machine learning interpretability methods. Entropy 23, 18. URL: https://www.mdpi.com/1099-4300/23/1/18. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

Lorieul, T., 2020. Uncertainty in Predictions of Deep Learning Models for Fine-Grained Classification. Ph.D. thesis. Université Montpellier.

Mace, G.M., Collar, N.J., Gaston, K.J., Hilton-Taylor, C., Akçakaya, H.R., Leader-Williams, N., Milner-Gulland, E., Stuart, S.N., 2008. Quantification of extinction risk: IUCN's system for classifying threatened species. Conserv. Biol. 22, 1424–1442. https://doi.org/10.1111/j.1523-1739.2008.01044.x.

Mair, L., Bennun, L.A., Brooks, T.M., Butchart, S.H., Bolam, F.C., Burgess, N.D., Ekstrom, J.M., Milner-Gulland, E., Hoffmann, M., Ma, K., et al., 2021. A metric for spatially explicit contributions to science-based species targets. Nat. Ecol. Evolut. 5, 836–844.

Marcon, E., 2015. Mesures de la Biodiversité. Lecture. AgroParisTech. URL. https://hal-agroparistech.archives-ouvertes.fr/cel-01205813.

McCormick, M.K., Whigham, D.F., Canchani-Viruet, A., 2018. Mycorrhizal fungi affect orchid distribution and population dynamics. New Phytol. 219, 1207–1215. https://doi.org/10.1111/nph.15223.

Moret, P., Muriel, P., Jaramillo, R., Dangles, O., 2019. Humboldt's tableau physique revisited. Proc. Natl. Acad. Sci. 116, 12889–12894. URL: https://www.pnas.org/doi/10.1073/pnas.1904585116. Publisher: Proceedings of the National Academy of Sciences.

Mortier, T., Wydmuch, M., Dembczyński, K., Hüllermeier, E., Waegeman, W., 2021. Efficient set-valued prediction in multi-class classification. Data Min. Knowl. Disc. 35, 1435–1469.

Newman, B., 2009. Orchids as Indicators of Ecosystem Health in Urban Bushland Fragments. Phd, Murdoch University. URL: https://researchrepository.murdoch.edu.au/id/eprint/2374/. Publication Title: Newman, Belinda <https://researchrepository.murdoch.edu.au/view/author/NewmanBelinda.html> (2009) Orchids as Indicators of Ecosystem Health in Urban Bushland Fragments. PhD thesis, Murdoch University.

Nic Lughadha, E., Walker, B.E., Canteiro, C., Chadburn, H., Davis, A.P., Hargreaves, S., Lucas, E.J., Schuiteman, A., Williams, E., Bachman, S.P., Baines, D., Barker, A., Budden, A.P., Carretero, J., Clarkson, J.J., Roberts, A., Rivers, M.C., 2019. The use and misuse of herbarium specimens in evaluating plant extinction risks. Philosoph. Trans. Royal Soc. B: Biol. Sci. 374, 20170402. URL: https://royalsocietypublishing.org/doi/full/10.1098/rstb.2017.0402. Publisher: Royal Society.

Nicholson, E., Watermeyer, K.E., Rowland, J.A., Sato, C.F., Stevenson, S.L., Andrade, A., Brooks, T.M., Burgess, N.D., Cheng, S.-T., Grantham, H.S., et al., 2021. Scientific foundations for an ecosystem goal, milestones and indicators for the post-2020 global biodiversity framework. Nat. Ecol. Evolut. 5, 1338–1349.

Orme, C.D.L., Davies, R.G., Burgess, M., Eigenbrod, F., Pickup, N., Olson, V.A., Webster, A.J., Ding, T.-S., Rasmussen, P.C., Ridgely, R.S., Stattersfield, A.J., Bennett, P.M., Blackburn, T.M., Gaston, K.J., Owens, I.P.F., 2005. Global hotspots of species richness are not congruent with endemism or threat. Nature 436, 1016–1019. URL. https://www.nature.com/articles/nature03850. Number: 7053 Publisher: Nature Publishing Group.

Paukert, C.P., Pitts, K.L., Whittier, J.B., Olden, J.D., 2011. Development and assessment of a landscape-scale ecological threat index for the lower Colorado river basin. Ecol. Indic. 11, 304–310.

Perez-Escobar, O.A., Bogarín, D., Przelomska, N.A., Ackerman, J.D., Balbuena, J.A., Bellot, S., Bühlmann, R.P., Cabrera, B., Cano, J.A., Charitonidou, M., et al., 2023. The origin and speciation of orchids. New Phytol. 242, 700–716. https://doi.org/10.1111/nph.19580.

Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J., Ferrier, S., 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecol. Appl. 19, 181–197.

Pollock, L.J., Tingley, R., Morris, W.K., Golding, N., O'Hara, R.B., Parris, K.M., Vesk, P. A., McCarthy, M.A., 2014. Understanding co-occurrence by modelling species simultaneously with a joint species distribution model (JSDM). Methods Ecol. Evol. 5, 397–406. https://doi.org/10.1111/2041-210X.12180.

Pollock, L.J., O'Connor, L.M.J., Mokany, K., Rosauer, D.F., Talluto, M.V., Thuiller, W., 2020. Protecting biodiversity (in all its complexity): new models and methods. Trends Ecol. Evol. 35, 1119–1128. URL: https://www.sciencedirect.com/science/article/pii/S0169534720302305.

Powell-Romero, F., Fountain-Jones, N.M., Norberg, A., Clark, N.J., 2022. Improving the predictability and interpretability of co-occurrence modelling through feature-based joint species distribution ensembles. Methods Ecol. Evol. https://doi.org/10.1111/2041-210X.13915.

Puglielli, G., Pärtel, M., 2023. Macroecology of plant diversity across spatial scales. New Phytol. 237, 1074–1077. https://doi.org/10.1111/nph.18680.

Ricotta, C., 2005. Through the jungle of biological diversity. Acta Biotheor. 53, 29–38.

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., et al., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40, 913–929.

Ryo, M., Angelov, B., Mammola, S., Kass, J.M., Benito, B.M., Hartig, F., 2021. Explainable artificial intelligence enhances the ecological interpretability of black-box species distribution models. Ecography 44, 199–205. https://doi.org/10.1111/ecog.05360.

Schatz, G.E., 2009. Plants on the iucn red list: setting priorities to inform conservation. Trends Plant Sci. 14, 638–642.

Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27, 379–423.

Silvestro, D., Goria, S., Sterner, T., Antonelli, A., 2022. Improving biodiversity protection through artificial intelligence. Nat. Sustain. 5, 415–424. URL. https://www.nature.com/articles/s41893-022-00851-6. Number: 5 Publisher: Nature Publishing Group.

Stévart, T., Dauby, G., Lowry, P.P., Blach-Overgaard, A., Droissart, V., Harris, D.J., Mackinder, B.A., Schatz, G.E., Sonké, B., Sosef, M.S.M., Svenning, J.-C., Wieringa, J. J., Couvreur, T.L.P., 2019. A third of the tropical African flora is potentially threatened with extinction. Sci. Adv. 5, eaax9444. https://doi.org/10.1126/sciadv.aax9444.

Swarts, N.D., Dixon, K.W., 2009. Terrestrial orchid conservation in the age of extinction. Ann. Bot. 104, 543–556. https://doi.org/10.1093/aob/mcp025.

Syfert, M.M., Joppa, L., Smith, M.J., Coomes, D.A., Bachman, S.P., Brummitt, N.A., 2014. Using species distribution models to inform IUCN red list assessments. Biol. Conserv. 177, 174–184. URL: http://www.sciencedirect.com/science/article/pii/S0006320714002390.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Las Vegas, NV, USA, pp. 2818–2826. URL: http://ieeexplore.ieee.org/document/7780677/.

Venter, O., Sanderson, E.W., Magrach, A., Allan, J.R., Beher, J., Jones, K.R., Possingham, H.P., Laurance, W.F., Wood, P., Fekete, B.M., Levy, M.A., Watson, J.E. M., 2016. Global terrestrial human footprint maps for 1993 and 2009. Sci. Data 3, 160067. URL. https://www.nature.com/articles/sdata201667. Number: 1 Publisher: Nature Publishing Group.

Verones, F., Kuipers, K., Núñez, M., Rosa, F., Scherer, L., Marques, A., Michelsen, O., Barbarossa, V., Jaffe, B., Pfister, S., Dorber, M., 2022. Global extinction probabilities of terrestrial, freshwater, and marine species groups for use in life cycle assessment. Ecol. Indic. 142, 109204. URL: https://www.sciencedirect.com/science/article/pii/S1470160X22006768.

Vitt, P., Taylor, A., Rakosy, D., Kreft, H., Meyer, A., Weigelt, P., Knight, T.M., 2023. Global conservation prioritization for the orchidaceae. Sci. Rep. 13, 6718.

Walker, B.E., Leão, T.C.C., Bachman, S.P., Bolam, F.C., Nic Lughadha, E., 2020. Caution needed when predicting species threat status for conservation prioritization on a global scale. Front. Plant Sci. 11 https://doi.org/10.3389/fpls.2020.00520/full. Publisher: Frontiers.

Weigelt, P., König, C., Kreft, H., 2020. GIFT – a global inventory of floras and traits for macroecology and biogeography. J. Biogeogr. 47, 16–43. https://doi.org/10.1111/jbi.13623.

Whittaker, R.J., Araújo, M.B., Jepson, P., Ladle, R.J., Watson, J.E.M., Willis, K.J., 2005. Conservation biogeography: assessment and prospect. Divers. Distrib. 11, 3–23. https://doi.org/10.1111/j.1366-9516.2005.00143.x.

Wraith, J., Norman, P., Pickering, C., 2020. Orchid conservation and research: an analysis of gaps and priorities for globally red listed species. Ambio 49, 1601–1611. https://doi.org/10.1007/s13280-019-01306-7.

Yousefi, M., Jouladeh-Roudbar, A., Kafash, A., 2020. Using endemic freshwater fishes as proxies of their ecosystems to identify high priority rivers for conservation under climate change. Ecol. Indic. 112, 106137. URL: https://www.sciencedirect.com/science/article/pii/S1470160X20300741.

Zizka, A., Silvestro, D., Vitt, P., Knight, T.M., 2020. Automated conservation assessment of the orchid family with deep learning. Conserv. Biol. https://doi.org/10.1111/cobi.13616.

Zizka, A., Andermann, T., Silvestro, D., 2022. Iucnn–deep learning approaches to approximate species' extinction risk. Divers. Distrib. 28, 227–241.