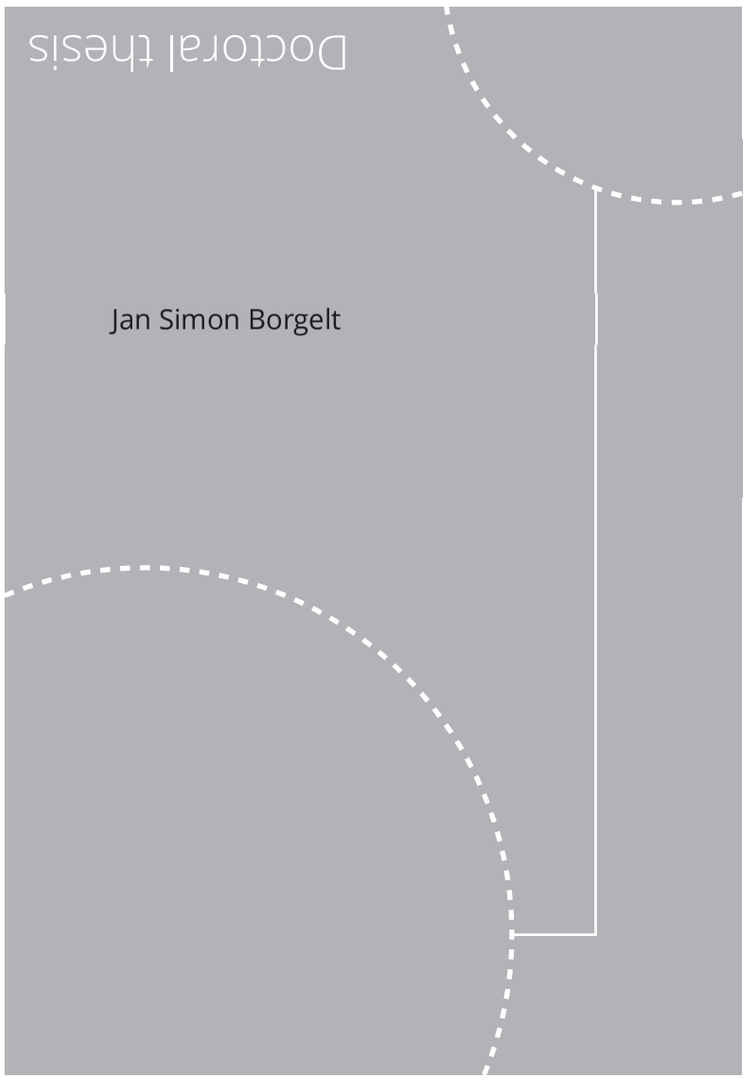
Doctoral theses at NTNU, 2022:270

Jan Simon Borgelt

Potentials of data science for advancing the modelling of biodiversity impacts



ISBN 978-82-326-5878-7 (printed ver.)
ISBN 978-82-326-5348-5 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (electronic ver.)

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of
Science and Technology
Thesis for the degree of
Philosophiae Doctor
Faculty of Engineering
Department of Energy and Process Engineering

Doctoral theses at NTNU, 2022:270

NTNU

Jan Simon Borgelt

Potentials of data science for advancing the modelling of biodiversity impacts

Thesis for the degree of Philosophiae Doctor

Trondheim, September 2022

Norwegian University of Science and Technology
Faculty of Engineering
Department of Energy and Process Engineering

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the degree of Philosophiae Doctor

Faculty of Engineering
Department of Energy and Process Engineering

ISBN 978-82-326-5878-7 (printed ver.)
ISBN 978-82-326-5348-5 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (electronic ver.)

Doctoral theses at NTNU, 2022:270

PRINTED IN NORWAY
NO - 1598

NORDIC SWAN ECOLABEL
Printed matter
2041 0731

Printed by Skipnes Kommunikasjon AS

Preface

The thesis has been submitted to the Faculty of Engineering Science in partial fulfilment of the degree of Philosophiae Doctor. This work was carried out at the Industrial Ecology Programme and the Department of Energy and Process Engineering at the Norwegian University of Science and Technology (NTNU) in Trondheim (Norway) in the period 2018-2022.

This PhD work was part of the *Transforming Citizen Science for Biodiversity* project within NTNU's strategic portfolio of projects to pursue transformative research on the development and application of digital transformation technology.

Jan Simon Borgelt,

Trondheim, June 2022

Acknowledgments
First of all, I want to thank my supervisor Francesca for the positive energy and kindness, for always having an open door, and most of all, for always being supportive and for putting an almost overwhelming level of trust into my ideas. I highly appreciated your support in the past years.

I further would like to thank everyone that I have shared memorable moments with, such as the unforgettable quests for spotting the harlequin duck in Iceland, discovering the vibrant bird life of Runde, and photographing the muskox in Dovrefjell. I am looking forward to more of this in the future!

I especially want to acknowledge the entire LCIA biodiversity group and the digital transformation team. These are namely, Fei, Philip, Ahmed, Yoann, Marthe, Dafna and Martin, and previous colleagues John and Koen. Thank you for creating a warm and welcoming atmosphere in the group. I am looking forward to the next teambuilding trips and activities. Special thanks to Martin for discussing so many ideas and unrelated topics with me! Also many thanks to the digital transformation team, especially Wouter, Caitlin, Jorge, Kwaku, Ben and Philip, with whom is was lucky enough to share ideas, trips to workshops and conferences, and to explore this research field together.

I also appreciated the generous and fruitful work environment at NTNU, and IndEcol in particular. Some names need to be mentioned apart here, these are my previous colleges Eivind and Philomena, the cageball players, the other two (!) Jans, and special thanks to Alex for sharing lots of coffee (and thoughts) with me! Also, many thanks to the Digital Lab, especially Radek for helping to deploy the applications. I further want to highlight the role of my external collaborators, namely Charly, Mark and Olav, whose feedback and contributions were as helpful as essential for this thesis.

Finally, I want to express my gratitude for Kristel, and both Dutch and German families for infinite and unconditional support.

Abstract

The unsuccessful and recently expired Aichi targets triggered a call for measurable, science-based targets in the post-2020 global biodiversity framework. Multiple tools and modelling approaches aim to support the endeavor of halting biodiversity loss by tracking progress of these targets and highlighting trade-offs across impacts. Among these tools are for example Life Cycle Assessment, Ecological Risk Assessment, and Environmentally Extended Input-Output analysis. However, assessing the state of biodiversity is hampered by a range of uncertainties. For instance, the lack of species distribution maps for some taxonomic groups and the incomplete availability of extinction risk categories inhibits attempts to map the location or severity of anthropogenic impacts. In addition, to date, not all significant impact categories are covered in tools such as Life Cycle Assessment because underlying methodologies are either not developed or not operationalized, as is for example the case for impacts from invasive alien species and plastic pollution. However, if not all relevant impacts are considered, or their quantification is inadequate, the inferences of subsequent impact assessments may be incomplete.

The increasing computational power and digital storage capacities, as well as the large-scale data collection by members of the general public, foster the development of new concepts to monitor the state of biodiversity. As part of the *Digital Transformation* initiative in the project *Transforming Citizen Science for Biodiversity*, this PhD thesis aimed to develop tools to advance biodiversity impact assessment modelling by integrating abundantly available open-access biodiversity data within, but not limited to, the framework of Life Cycle Impact Assessment. In particular, the chapters of this thesis contribute toward tackling data limitations, addressing uncertainties, and providing a methodology to consider the previously disregarded impacts of alien species introductions within Life Cycle Assessment.

In chapter 2, a largely automatized workflow was developed to transform open-access biodiversity occurrence data into coarse-scale habitat suitability maps for 27,208 red-listed vascular plant species. This newly generated dataset is already being used in a range of applications, highlighting the utility and need for the generated data.

Chapter 3 highlights potential flaws in previous modelling approaches that require the numeric conversion of species extinction risk categories. Abundantly available data were used in a machine learning classifier to predict the extinction risk of Data Deficient species within relevant taxonomic groups. The predictions suggest that more than half of Data Deficient species may be threatened by extinction and their level of threat is likely to differ within and across taxonomic groups, as well as in space. Besides ranking the species for a future allocation of resources, this study aims to trigger a debate about the appropriate use of Data Deficient species within modelling approaches.

Finally, invasive alien species have long been recognized as a serious threat to global biodiversity. Their current distribution and future spread are powered by worldwide transportation networks that essentially eliminate natural biogeographical barriers. However, while transportation is a fundamental part of any attempt aiming to holistically assess the environmental consequences within Life Cycle Assessment of, for instance, a product, impacts caused by invasive alien species cannot be accounted for due to a lacking impact model. Chapter

4 proposes a globally applicable methodology for the first time, compatible with current developments within the life cycle initiative hosted by UN environment. The quantified impacts are of similar magnitude as climate change impacts, implying that neglecting invasive alien species substantially underestimates the overall environmental consequences of transporting commodities.

The transition into a sustainable world requires innovative approaches that guarantee more accurate, sound, and robust quantitative tools to support policy- and decision-making while keeping up with accelerating global changes. This thesis highlights potentials in data science that can stimulate future advances in next-generation biodiversity impact assessment models, supporting the quest to achieve global biodiversity targets.

Sammendrag
At verden ikke nådde Aichi målene skapte grobunn for å utvikle nye kvantifiserbare og forskningsbaserte mål for det globale post-2020 rammeverket for biologisk mangfold. Flere verktøy og modelleringsmetoder er siktet mot å støtte forsøket på å stoppe tap av biologisk mangfold ved å spore fremskrittene mot målene og å tydeliggjøre eventuelle avveininger på tvers av miljøpåvirkninger. Eksempler på slike verktøy er livsløpsanalyser, økologiske risikovurderinger, og miljøutvidete kryssløpsanalyser. Likevel er slike evalueringer sterkt påvirket av mange usikkerheter. Mangelen på distribusjonskart samt vurdering av risiko for utryddelse av arter på tvers av mange taksonomiske grupper, for eksempel, gjør det utfordrende å beregne påvirkningen av menneskelige aktiviteter. Samtidig er ikke alle signifikante påvirkningskategorier dekket enda, fordi den underliggende metoden enten ikke er utviklet eller operasjonalisert, slik som for eksempel for fremmede arter og plastforurensing. Dersom ikke alle relevante påvirkningskategorier dekkes eller kvantifiseringen av disse ikke er god nok, vil også evalueringer av miljøpåvirkninger lide som følge av dette.

Økende datakraft og digitale lagringsmuligheter, i tillegg til storskala datainnsamling fra borgerforskere skaper muligheter for å utvikle nye konsepter for å overvåke tilstanden til biologisk mangfold. Som en del av prosjektet «*Transforming Citizen Science for Biodiversity»,* en del av *«Digital Transformation»*-initiativet, sikter denne avhandlingen mot å utvikle verktøy for å forbedre modellering av miljøpåvirkninger på biologisk mangfold ved å integrere åpne data på biologisk mangfold innenfor, men ikke begrenset til, livsløpsanalyser. Kapitlene i avhandlingen bidrar til å takle databegrensninger, adressere usikkerhet, og å tilby en metodikk for å vurdere en tidligere oversett miljøpåvirkning innenfor miljøpåvirkningsevalueringer for biologisk mangfold.

I kapittel 2 ble en automatisert arbeidsflyt utviklet for å transformere åpne data på forekomster av biologisk mangfold til grov-skala kart over habitatsegnethet for 27 208 plantearter. Dette nylig genererte datasettet er allerede i bruk i flere applikasjoner, noe som tydeliggjør nytteverdien og behovet for slike data.

Kapittel 3 fremhever potensielle mangler i tidligere modelleringsforsøk som avhenger av numerisk konvertering av kategorier for arters risiko for utryddelse. Rikelig tilgjengelige data ble brukt i kombinasjon med maskinlæring for å forutsi risikoen for utryddelse for arter med datamangel innenfor relevante taksonomiske grupper. Resultatene tyder på at risikoen trolig varierer både innenfor og på tvers av taksonomiske grupper, i tillegg til lokasjon. I tillegg til å rangere arter for fremtidig ressursallokering, tydeliggjør studiet behovet for en debatt rundt riktig modellering av arter med mangelfullt datagrunnlag.

Til slutt rettes fokuset på fremmede arter som lenge er blitt anerkjent som en seriøs trussel mot biologisk mangfold globalt. Deres nåværende distribusjon og fremtidig spredning drives frem av verdensomspennende transportnettverk som visker ut tidligere naturlige biogeografiske barrierer. Mens transport utgjør en grunnleggende del av alle forsøk på helhetlig modellering av miljøkonsekvenser innenfor livsløpsanalyser av, for eksempel, produkter, så er miljøpåvirkninger av fremmede arter ikke mulig å evaluere på grunn av manglende modelleringsmetoder. Kapittel 4 legger frem den første utviklede metodikken for globale evalueringer av miljøpåvirkninger fra fremmede arter som er kompatibel med *«life cycle*

initiativet» av FNs miljøprogram. De kvantifiserte resultatene antyder at å neglisjere fremmede arter vil gi en betydelig underestimering av de totale miljøpåvirkningene på biologisk mangfold fra transporterte produkter.

Overgangen til en bærekraftig verden behøver innovative tilnærminger som kan garantere mer nøyaktige og robuste kvantitative verktøy for å støtte politikk og beslutningstaking, og som samtidig holder tritt med akselererende globale endringer. Denne avhandlingen fremhever potensialet i datavitenskap som kan stimulere fremtidige forbedringer i den neste generasjonen av evalueringsmodeller for miljøpåvirkninger.

Publications

This thesis is based on three articles, listed as primary publications below. One of these articles has been published in *Scientific Data* (Chapter 2), one has been published in *Communications Biology* (Chapter 3), and one of the articles is being prepared for submission at *Environmental Science & Technology* (Chapter 4).

The articles listed in secondary publications are related to, but not included in this thesis. These articles have been published in *Spatial Statistics* and *Biological invasions*.

Primary publications

J Borgelt, J Sicacha-Parada, O Skarpaas, F Verones. 2022. Native Range Estimates for Red-Listed Vascular Plants. Scientific Data 9(1):117. doi: 10.1038/s41597-022-01233-5.
Author contribution: Study design, modelling, writing

J Borgelt, M Dorber, M Alnes Høiberg, F Verones. 2022. More than half of Data Deficient species predicted to be threatened by extinction. Communications Biology 5: 679. doi: 10.1038/s42003-022-03638-9.
Author contribution: Study design, modelling, writing

J Borgelt, M Dorber, C Géron, MAJ Huijbregts, F Verones. In preparation. Terrestrial ecosystem impacts of biological invasions caused by international transportation within the framework of Life Cycle Assessment. To be submitted to Environmental Science & Technology.
Author contribution: Study design, modelling, writing

Secondary publications

C Géron, JJ Lembrechts, **J Borgelt**, J Lenoir, R Hamdi, G Mahy, I Nijs, A Monty. 2021. Urban Alien Plants in Temperate Oceanic Regions of Europe Originate from Warmer Native Ranges. Biological Invasions 23(6):1765–79. doi: 10.1007/s10530-021-02469-9.
Author contribution: Assisted in parts of the modelling

J Sicacha-Parada, I Steinsland, B Cretois, **J Borgelt**. 2021. "Accounting for Spatial Varying Sampling Effort Due to Accessibility in Citizen Science Data: A Case Study of Moose in Norway." Spatial Statistics 42:100446. doi: 10.1016/j.spasta.2020.100446.
Author contribution: Assisted in interpreting and writing

Contents

Chapter 1: Introduction

Chapter 1: Introduction

1.1 The state of planet Earth

The past 10,000 years have been characterized by unusually stable environmental conditions on planet Earth[1]. This stability allowed the transition of humankind from hunters and gatherers to farming communities, setting the corner stone for upcoming civilizations[2]. The invention of the steam engine in the 18th century as well as innovative knowledge for enhancing food production (e.g., the Haber-Bosch process) enabled rapid population growth and increasing wealth, but also heralded the era of mankind[3]. In the so-called "Anthropocene", human actions became the main driver of global changes[4] which threaten the beneficial stability of our planet[5]. Within only a century, the human population has more than tripled to roughly 8 billion individuals[6]. The majority of life-ensuring processes on this planet is affected by anthropogenic activities[4,7] and humanity's safe-operating space is considered at risk for numerous categories (Figure 1). Besides severely impacting vital biochemical flows, humanity induced an ongoing, large-scale loss of biosphere integrity that is expected to cause considerable damage to human well-being[5,8,9].

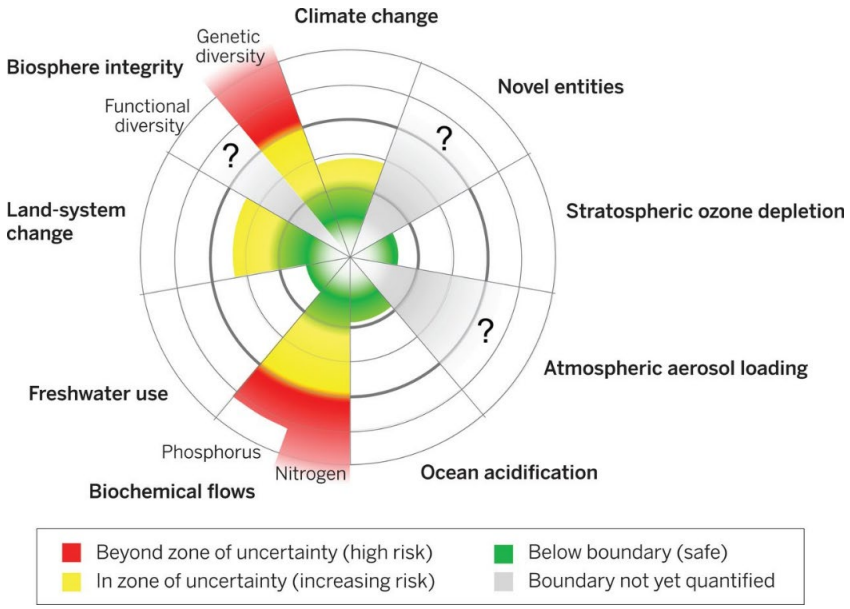


Figure 1: Status for estimated planetary boundaries indicating the safe-operating space for humanity. Source: Steffen et al.[9]

The diversity of life forms that exist today guarantees functioning natural processes and provides essential ecosystem services to humankind[10]. These services include pollination and pest control for food production[11], water purification[12] and maintaining biogeophysical flows such as the carbon cycle[13]. Although today's human well-being was largely built upon a functioning natural world[14], the growing magnitude, intensity, and scale of anthropogenic activities threaten the existence of life on Earth as we know it[9,15,16].

Today, the majority of land and ocean areas is directly affected by human activities[17,18], with habitat loss, invasive species, overexploitation, pollution and climate change forming the main threats to biodiversity[10,19,20]. Approximately 60% of the world's land surface, and 9 out of 14 of the world's terrestrial biomes, have already fallen below a safe planetary boundary threshold[21]. Wilderness areas continue to be lost at alarming rates[22]. One third of terrestrial land area alone is used for growing crops or raising livestock[23]. Globally, a staggering 96% of mammal biomass is comprised of humankind and its livestock[24]. Species such as the Scimitar-Horned Oryx (*Oryx dammah*), once grazing wide areas of Northern Africa's Savannas, currently only survive in fenced and protected areas as a direct result of hunting, competing livestock and habitat loss[25]. In addition, worldwide transport routes promote the introductions of alien species into new environments[26] where they encounter poorly adapted species. In New Zealand for instance, the once widespread Kākāpō (*Strigops habroptila*) (Figure 2) was well adapted against its native predators. When realizing potential threats, the flightless bird would freeze and camouflage against the ground[27]. However, this behavior made the Kākāpō easy prey for introduced mammalian predators, leading to tremendous population declines[27–29] and extinction from its natural habitat, as the species now is confined to islands free of alien predators[27].

Figure 2: Illustration of a Kākāpō (Strigops habroptila) from Birds of New Zealand 1st edition, by Walter Lawry Buller, published in 1873. Source: John G. Keulemans. Minor edits have been made to the original by User: Msikma; Public domain, via Wikimedia Commons.

Moreover, increasing volumes of commodities and people being transported across the oceans also result in more and more ship collisions, one of the major threats[30] to the North Atlantic Right Whale (*Eubalaena glacialis*). The Atlantic Sturgeon (*Acipenser sturio*) is today restricted to only a small portion of its historical range[31] largely due to industrial activities, such as dam construction, pollution, and river regulation[32].

These species are just examples of the globally more than 40,000 species known to be threatened by extinction[33]. Besides the sentimentality of losing unique species, each functional extinction potentially causes cascading effects on those ecosystem services humanity depends on[34]. Concerningly, current species extinction rates are estimated to be 1,000 times higher than historical background extinction rates[20,35], suggesting that planet Earth faces its sixth mass extinction event in the near future if the rapidly increasing human pressures are not diminished[36,37].

1.2 Assessing anthropogenic impacts

Quantitative tools that can map complex and interacting impacts on the environment are vital to reduce human pressures[38]. A variety of tools aim to offer guidance on proactive and reactive impact mitigation. These tools can highlight the most destructive activities within numerous impact pathways on global[39] and regional[40,41] levels. For instance, environmental accounting and its subdisciplines, e.g., ecological footprinting, address practical issues and can be used to calculate measures such as the Earth Overshoot Day, a striking indicator of nowadays' exceeding use of natural resources[42]. Environmentally Extended Input-Output analyses (EEIO) can be used for tracing material flows and their consequential impacts on biodiversity caused by e.g., consumption[43,44] as well as resource use[45]. Another powerful decision-support tool is Life Cycle Assessment (LCA). LCA is used to quantify environmental impacts throughout entire life cycles, e.g., material acquisition, production, use and disposal[46–48], and can be conducted at different levels of detail and complexity, from individual products to entire industries[47]. LCA highlights hotspots of environmental impacts, e.g., within supply chains, technologies, as well as life cycle stages, and, for some impacts, even geographically refined. This enables the comparison of environmental performances between different options and thereby facilitates decision-making toward more sustainable choices[47]. As such, LCA has been applied for assessing trade-offs between pollution, climate change mitigation and biodiversity impacts related to low-carbon electricity[49], as well as suggesting optimal locations for future hydropower plants[39]. Consequently, tools such as LCA not only improve our understanding of anthropogenic impacts on the environment, but also offer effective decision-support by helping to prevent problem shifting from one environmental impact to another[47]. Essential for such assessments are models that quantify the fate of a stressor and its consequential impacts on the environment. These impacts are further aggregated into different areas of protection within LCA, i.e., human health, ecosystem quality and natural resources[50] (Figure 3). Impacts on ecosystem quality are conventionally measured in terms of species loss[51], although other levels of biodiversity can indicate environmental quality as well. However, species loss is the agreed-upon indicator, because of its relatively better data availability, and because one common indicator facilitates comparisons between impacts and across different stressors. Such impact assessment models exist for a range of stressors, including land use[52], climate change[53] and terrestrial acidification[54]. Additional impact models are continuously developed and integrated, such as ocean acidification[55].

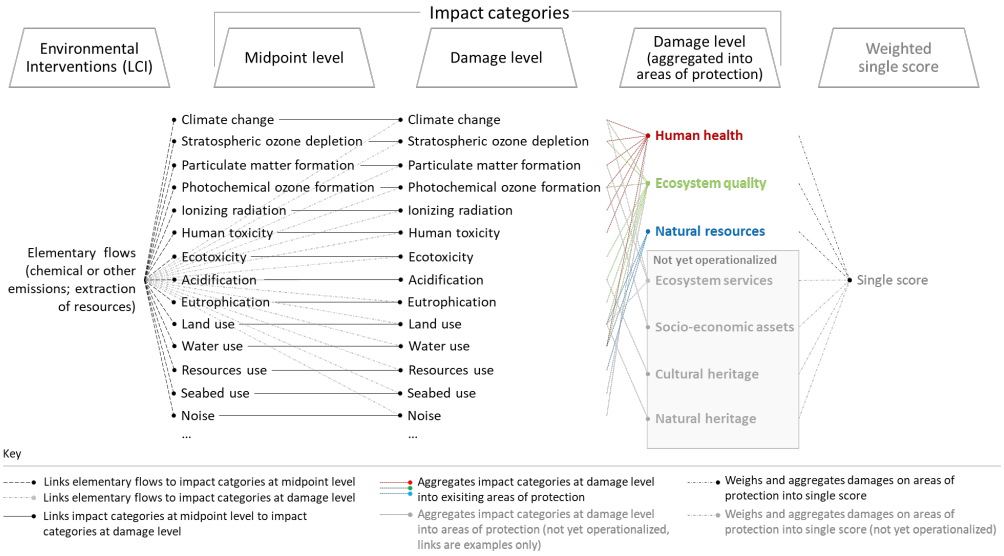


Figure 3: Covered impact categories and their contribution to three areas of protection (i.e., Human health, ecosystem quality, and natural resources) within the Life Cycle Impact Assessment framework. Source: Verones et al.[51]

For instance, models quantifying the impacts of land use changes on ecosystem quality describe the stressor-response dynamic based on the theory of island biogeography[56]. Thereby, the effects of land use changes are calculated by linking the declining area of habitat to species loss following the species-area relationship concept. Notable improvements are the possibility to account for rarity and threat-levels of species[57,58], that some species are better adapted to modified habitats than others[59,60], and that some interventions are more destructive than others, e.g. accounting for the intensity and the fragmentation of land use[52,61]. Recent developments even allow to distinguish between local, regional, and permanent species loss on a global scale[62].

Models for quantifying such impacts often require detailed species-level information. Most fundamentally, this includes data about the species that are likely to be affected by a given stressor, their location, and the expected severity of the impact, e.g., how vulnerable the species are. However, such data are scarce and often not readily available[46,50,63,64]. In general, the availability of species data is biased toward large animals, temperate systems and components of biodiversity used by people[10]. Most commonly utilized databases, such as the International Union for the Conservation of Nature (IUCN) Red List of Threatened Species[33] or the World Wildlife Fund (WWF) WildFinder database[65], only provide data for a selection of species. This selection is subsequently reflected in biodiversity assessments, where only a handful of taxonomic classes are considered as surrogates for entire ecosystems[66–68]. For instance, the estimated ecosystem impacts of water consumption are based on fish species only[69], and, although severely affecting insects[70] and plants[71], land use impacts are frequently assessed for amphibians, birds, mammals and reptiles only[52].

Generating species-level data, e.g., threat-levels, habitat preferences and species distribution maps, is not only time-consuming[72,73], but such data also needs to be updated frequently because the ongoing global changes progress dynamically[74,75]. Advancements in some of the cross-

cutting issues of impact assessment modelling, e.g., increasing the taxonomic coverage[63,64], are therefore unrealistic at the pace at which data are being provided today.

1.3 Open-access data

Yet, the technological prerequisites for monitoring the state of global biodiversity have never been better. The advent of the internet and the recent digital revolution provide new means of data creation, analysis, and sharing. The global digital data storage, as well as computing capacities, are rising exponentially[76], allowing for the implementation of novel, computationally intensive methods. Scientific output is no longer defined by publications only, but just as much by openly sharing findings, analyses and collected data[77].

Openly available data exist across numerous disciplines, ranging from time-series data on bilateral exchange of commodities[78] to large-scale, high-resolution deforestation trackers[79] and numerous environmental parameters on a global scale[80]. The diversity of remote-sensing derived data facilitates the extensive monitoring of species diversity[81], structure[82], as well as ecosystem responses to environmental stress[83]. Today, an unprecedented number of satellites is orbiting Earth[84], providing frequent, high-resolution and openly accessible remote-sensing data[85,86].

In addition, the digital transformation and new ways of communication, empowered the general public to participate in relevant processes, from data collection to decision-making, in numerous disciplines[87,88]. A variety of tools address these new target audiences for crowd-sourcing data on a large scale, such as roughly one hundred thousand internet users that were successfully consolidated to visually classify one million galaxies in *Galaxy Zoo*[89]. The mainstreaming of handheld recording devices (i.e., mobile phones) in the general public further enabled citizen scientists to contribute to large-scale monitoring of the natural world. In recent years, both participation and the number of citizen science projects increased immensely[90]. The value of such data has been successfully shown for different applications both on a local and global scale, e.g., for identifying the source of local freshwater pollution[91], for monitoring plant phenology[92] and highlighting global declines in bee species richness[93]. In the citizen science program *eBird*[94,95], a remarkable number of new data records is being entered each day by bird watchers all over the world, eager to share their observations with the community. Such data become useful at different levels of detail, depending on its quality[96]. For instance, the enthusiastic bird watching community allows for detailed monitoring of individual species, e.g., predicting species abundance across space and time (Figure 4).

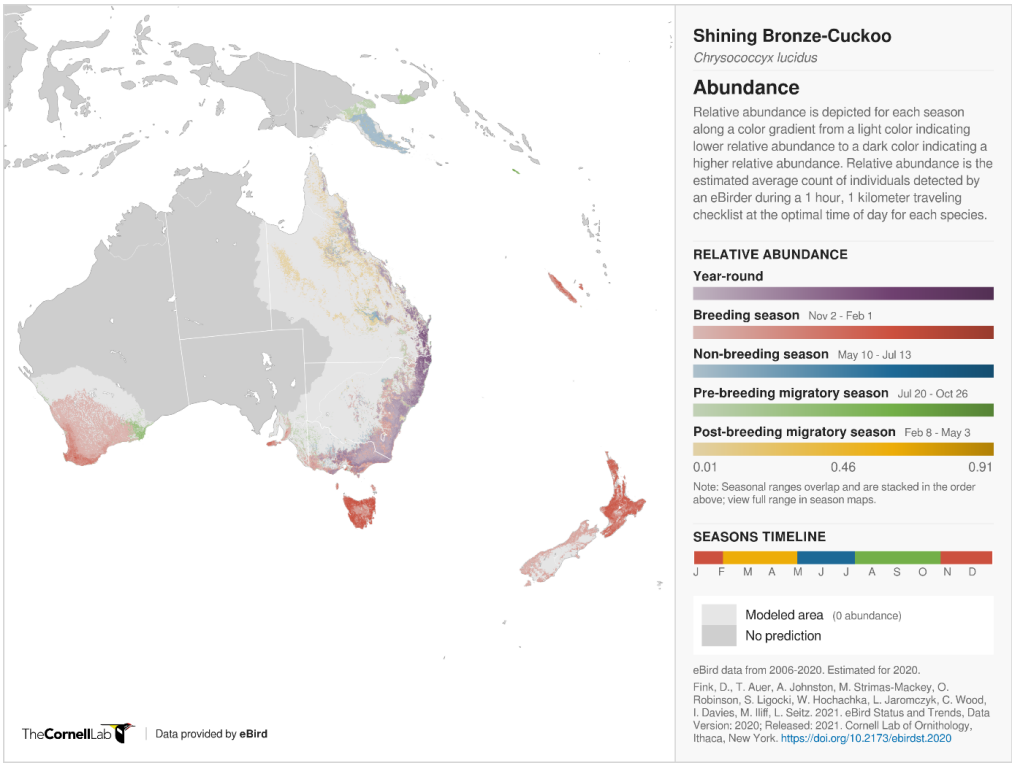


Figure 4: Predicted relative abundance of the Shining Bronze-Cuckoo (Chrysococcyx lucidus) across seasons in Oceania based on data records of the citizen science program eBird. Source: Fink et al.[97]

Consequently, unstructured digital biodiversity data are abundantly available[75,98,99]. Big data infrastructures such as the Global Biodiversity Information Facility (GBIF) aim to channel and provide the data generated by scientists, institutes, and enthusiasts[100,101]. GBIF contains more than 2 billion species records of numerous organism groups from different places across the globe[102]. Such records usually include information on the observed species and their location. Occasionally, features such as the number of individuals or habitat type at the given location are reported. More than 50% of these data have been recorded by citizen scientists[102]. Although for bird species such data are most abundantly available, several other taxa, including different plant groups, are well-represented too[103]. The massive amount of such data stored in open access online databases has a large potential to improve our knowledge on the state of biodiversity. However, currently operationalized impact assessment methods do not utilize the full range of available data sources.

1.4 Thesis contribution

1.4.1 Research gap

The smorgasbord of models aiming to quantify different impacts on the natural world still contain a range of limitations[51,63,64,104,105]. A typical assessment method that relies on diverse and abundant input data is LCA[106]. LCA aims to be standardized and applicable along global value chains[47,107]. The lack of appropriate input data is a major obstacle for advancing the underlying biodiversity impact assessment models. In each impact category, stressors need to

be identified, their effects estimated, and the proportion of affected species needs to be assessed. However, global data coverage is incomplete for many ecological mechanisms and indicators[73,108,109]. The vital modelling of biodiversity impacts is therefore restricted by several data limitations.

Firstly, the availability of species-level data limits biodiversity impact assessment models. For instance, usually only few taxonomic classes (i.e., down to one[110]) are considered that are representative of entire ecosystems. Concerningly, the rationale for not including some taxa reflects the lack of appropriate data rather than the taxa being immune to the assessed stressor[68]. The different impact indicators used today are largely built upon on ready-to-use datasets. Mostly species distribution maps are used to estimate the number of present species, number of affected species, and the relative species loss within a defined geographical unit. This way, the effects of land use have been assessed for amphibians, birds, mammals, and reptiles[52]. Yet, more abundant and functionally important taxa, such as fungi, insects and plants, are neglected. Similarly, the model for tracing biodiversity threats in global supply chains is based on available species distribution maps for the kingdom *Animalia* only[43]. In fact, most spatially explicit impact categories in, for instance, the methodology *LC-impact*[111], are based on such maps. Accordingly, the lack of distribution maps for individual species is one of the most limiting factors for both developing novel and improving existing impact assessment models. As such, some impact categories are either not covered at all in current methodologies, not covered in a spatially-explicit way (e.g., climate change), or had to be improvised due to the lack of distribution data, for e.g., vascular plants. Single attempts exist to use more unstructured data, e.g., available from GBIF. In a remarkable effort, Gade et al.[112] utilized abundant and unstructured data retrieved from GBIF for counting individual vascular plant species across space. However, besides requiring substantial computational resources, such attempts ignore important caveats of the data. For instance, species counts were probably underestimated in some regions because sampling effort and data availability is distributed heterogeneously[113–115], and introduced species are not filtered out in such databases.

Secondly, species-level data uncertainties propagate through biodiversity impact assessment models. Commonly applied data are species-level extinction risk categories retrieved from the IUCN Red List of Threatened Species. These are used to weigh individual taxa according to their conservation relevance. The weights are based on a numerical representation of the species' extinction risk category (i.e., Extinct, Extinct in the Wild, Critically Endangered, Endangered, Vulnerable, Near Threatened, Least Concern and Data Deficient) and follow a linear[57,58,62], categorical[44,116,117] or logarithmic[118] scheme. Hence, the species at a higher IUCN Red List category get assigned a relatively higher weight in the analyses. However, also extinction risk categories are not consistently available (Figure 5) and their assessment is both time-intensive and costly[73,74,119].

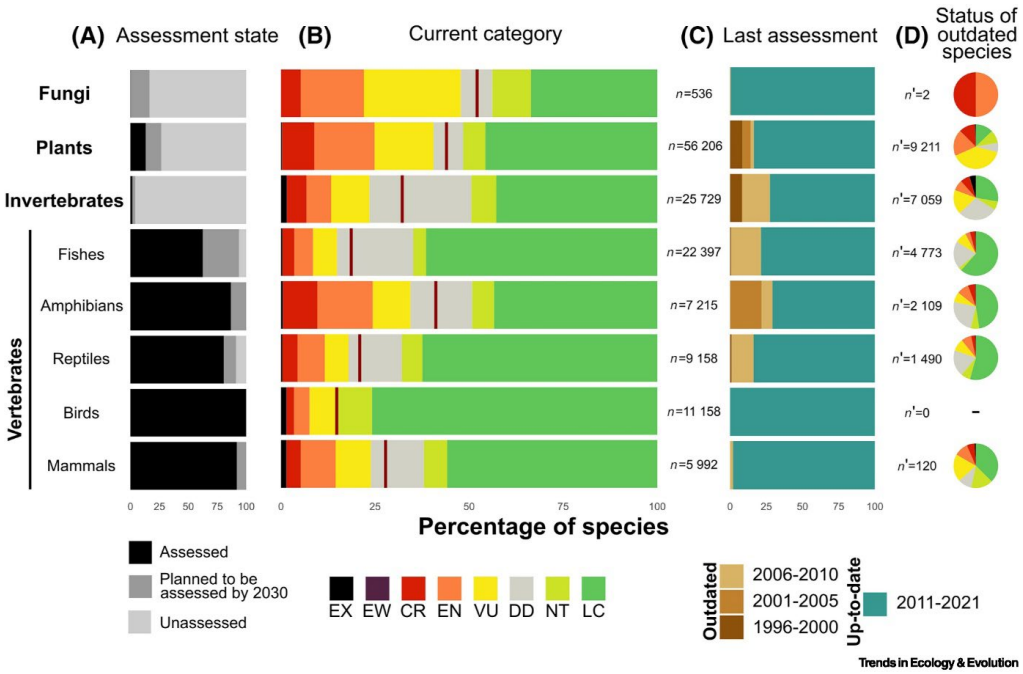


Figure 5: Progress of the IUCN Red List of Threatened Species. A) Proportion of taxonomic class included. B) Percentage of assessed species in the Red List categories Extinct (EX), Extinct in the Wild (EW), Critically Endangered (CR), Endangered (EN), Vulnerable (VU), Data Deficient (DD), Near Threatened (NT), and Least Concern (LC). Percentage of threatened species indicated by red line, assuming that DD species are as threatened as data-sufficient species. C) Timing of Last Assessment. D) Status of outdated species. Source: Cazalis et al.[120]

In addition to only a proportion of global biodiversity being currently assessed for their extinction risk, Red List assessors can assign species as Data Deficient[121]. This label does not indicate at which extinction risk category the species is, leaving the model developer with the difficulty of interpreting Data Deficient species appropriately. This difficulty has led to Data Deficient species having been considered equivalent to species categorized as Critically Endangered[122], Near Threatened[62,117], Least Concern[44] or being completely disregarded[43,57,58,123]. Even though Data Deficient species have been suggested to be likely threatened[124–127] and guidelines for appropriate uses exist[128], the utilization of Data Deficient species is inconsistent within such weighing schemes.

Finally, LCA is restricted by a limited coverage of impact categories. Depending on the method, impact assessment models exist for assessing the consequences of stressors such as eutrophication, acidification, climate change, land stress, water consumption, or toxicity[111]. However, impacts of invasive species are not considered within LCA[64,129], even though they are among the major concerns to global biodiversity[7]. Developing a more encompassing set of impacts on biodiversity is among the priorities for future developments in LCA[64,104], including impacts on marine systems that are rarely being considered so far[105]. Furthermore, additional measures of biodiversity should be operationalized (e.g., ecosystem function)[50], and existing approaches should whenever possible be regionalized[51,130].

1.4.2 Research aim

This thesis explores potentials of data science in combination to relatively novel data sources for biodiversity impact assessment modelling. In particular, the generated data in this thesis aim to assist future assessments of anthropogenic impacts on biodiversity, within LCA and beyond.

The research goals are:

1. Increasing species-level data availability for biodiversity impact assessment models by utilizing open-access databases. Estimate coarse-scale species distribution maps for an important taxonomic group, i.e., vascular plants, and discuss the quality of the generated products (Chapter 2).
2. Addressing species-level data uncertainties in the context of biodiversity impact assessment models. Predict probabilities of being threatened by extinction for Data Deficient species within commonly applied datasets based on novel computational tools and data sources, explore differences across taxa and space, and discuss the performance as well as implications of the approach (Chapter 3).
3. Improving the impact pathway coverage within Life Cycle Assessment. Develop a method to link alien species introductions to the transportation of commodities and generate a set of global characterization factors describing relative impacts on terrestrial ecosystems caused by introduced species across trading partners, while showcasing the utility of the data generated in Chapters 2 & 3 (Chapter 4).
4. Discussing the relevance and applicability of the developed methods in and beyond LCA and make recommendations for future research (Chapter 5).

In chapter 2, species range predictions for vascular plants were generated in an automated approach, allowing the large-scale implementation, and covering the majority of vascular plants of the IUCN Red List Version 2021-1. In chapter 3, the utility of machine learning models to predict the extinction risk status of Data Deficient species, and potential consequences of their incorrect prioritization are discussed. In chapter 4, a set of global characterization factors is derived, based on data created in chapter 2 & 3, and their utility for decision-support discussed. In chapter 5, the advancement of the research field and the applicability of the developed work, as well as their uncertainties and limitations, are discussed.

1.5 References

1. Dansgaard, W. *et al.* Evidence for general instability of past climate from a 250-kyr ice-core record. *Nature* **364**, 218–220 (1993).
2. Bocquet-Appel, J.-P. When the World's Population Took Off: The Springboard of the Neolithic Demographic Transition. *Science (80-.).* **333**, 560–561 (2011).
3. Crutzen, P. J. Geology of mankind. *Nature* **415**, 23–23 (2002).
4. Steffen, W., Crutzen, P. J. & McNeill, J. R. The anthropocene: Are humans now overwhelming the great forces of nature? *Ambio* **36**, 614–621 (2007).
5. Rockström, J. *et al.* A safe operating space for humanity. *Nature* **461**, 472–475 (2009).
6. United Nations. *World Population Prospects 2019: Highlights (ST/ESA/SER.A/423)*. (2019).

7. Butchart, S. H. M. *et al.* Global Biodiversity: Indicators of Recent Declines. *Science (80-.).* **328**, 1164–1168 (2010).

8. Hooper, D. U. *et al.* A global synthesis reveals biodiversity loss as a major driver of ecosystem change. *Nature* **486**, 105–108 (2012).

9. Steffen, W. *et al.* Planetary boundaries: Guiding human development on a changing planet. *Science (80-.).* **347**, (2015).

10. Millennium Ecosystem Assessment. *Ecosystems and Human Well-being: Biodiversity Synthesis*. (World Resources Institute, 2005).

11. Gallai, N., Salles, J., Settele, J. & Vaissière, B. E. Economic valuation of the vulnerability of world agriculture confronted with pollinator decline. *Ecol. Econ.* **68**, 810–821 (2009).

12. Piaggio, M. & Siikamäki, J. The value of forest water purification ecosystem services in Costa Rica. *Sci. Total Environ.* **789**, 147952 (2021).

13. Piao, S. *et al.* The carbon balance of terrestrial ecosystems in China. *Nature* **458**, 1009–1013 (2009).

14. Mace, G. M., Norris, K. & Fitter, A. H. Biodiversity and ecosystem services: a multilayered relationship. *Trends Ecol. Evol.* **27**, 19–26 (2012).

15. Díaz, S. *et al.* Pervasive human-driven decline of life on Earth points to the need for transformative change. *Science (80-.).* **366**, eaax3100 (2019).

16. Steffen, W., Broadgate, W., Deutsch, L., Gaffney, O. & Ludwig, C. The trajectory of the Anthropocene: The Great Acceleration. *Anthr. Rev.* **2**, 81–98 (2015).

17. Allan, J. R., Venter, O. & Watson, J. E. M. Temporally inter-comparable maps of terrestrial wilderness and the Last of the Wild. *Sci. Data* **4**, 170187 (2017).

18. Jones, K. R. *et al.* The Location and Protection Status of Earth's Diminishing Marine Wilderness. *Curr. Biol.* **28**, 2506-2512.e3 (2018).

19. Mace, G. M. Drivers of biodiversity change. in *Trade-offs in Conservation: Deciding What to Save* 349–364 (Wiley, 2010).

20. IPBES. *Summary for policymakers of the global assessment report on biodiversity and ecosystem services*. *Zenodo* (2019) doi:10.5281/zenodo.3831674.

21. Newbold, T. *et al.* Has land use pushed terrestrial biodiversity beyond the planetary boundary? A global assessment. *Science (80-.).* **353**, 288–291 (2016).

22. Watson, J. E. M. *et al.* Catastrophic Declines in Wilderness Areas Undermine Global Environment Targets. *Curr. Biol.* **26**, 2929–2934 (2016).

23. Klein Goldewijk, K., Beusen, A., Van Drecht, G. & De Vos, M. The HYDE 3.1 spatially explicit database of human-induced global land-use change over the past 12,000 years. *Glob. Ecol. Biogeogr.* **20**, 73–86 (2011).

24. Bar-On, Y. M., Phillips, R. & Milo, R. The biomass distribution on Earth. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 6506–6511 (2018).

25. IUCN. SSC Antelope Specialist Group. Oryx dammah. The IUCN Red List of Threatened Species 2016: e.T15568A50191470.

https://dx.doi.org/10.2305/IUCN.UK.2016-2.RLTS.T15568A50191470.en. (2016).

26. Seebens, H. *et al.* No saturation in the accumulation of alien species worldwide. *Nat. Commun.* **8**, 1–9 (2017).

27. Powlesland, R. G., Merton, D. V. & Cockrem, J. F. A parrot apart: The natural history of the kakapo (Strigops habroptilus), and the context of its conservation management. *Notornis* **53**, 3–26 (2006).

28. Bergner, L. M., Dussex, N., Jamieson, I. G. & Robertson, B. C. European Colonization, Not Polynesian Arrival, Impacted Population Size and Genetic Diversity in the Critically Endangered New Zealand Kākāpō. *J. Hered.* **107**, 593–602 (2016).

29. Clout, M. N. & Merton, D. V. Saving the Kakapo: the conservation of the world's most peculiar parrot. *Bird Conserv. Int.* **8**, 281–296 (1998).

30. Van der Hoop, J. M. *et al.* Assessment of Management to Mitigate Anthropogenic Effects on Large Whales. *Conserv. Biol.* **27**, 121–133 (2013).

31. Debus, L. Meristic and morphological features of the Baltic sturgeon (Acipenser sturio L.). *J. Appl. Ichthyol.* **15**, 38–45 (1999).

32. Gesner, J., Williot, P., Rochard, E., Freyhof, J. & Kottelat, M. Acipenser sturio. The IUCN Red List of Threatened Species 2010: e.T230A13040963. https://dx.doi.org/10.2305/IUCN.UK.2010-1.RLTS.T230A13040963.en. (2010).

33. IUCN. The IUCN Red List of Threatened Species. Version 2021-3. https://www.iucnredlist.org (2022).

34. Keyes, A. A., McLaughlin, J. P., Barner, A. K. & Dee, L. E. An ecological network approach to predict ecosystem service vulnerability to species losses. *Nat. Commun.* **12**, 1586 (2021).

35. Pimm, S. L. *et al.* The biodiversity of species and their rates of extinction, distribution, and protection. *Science (80-.).* **344**, 1246752–1246752 (2014).

36. Barnosky, A. D. *et al.* Has the Earth's sixth mass extinction already arrived? *Nature* **471**, 51–57 (2011).

37. Ceballos, G., Ehrlich, P. R. & Raven, P. H. Vertebrates on the brink as indicators of biological annihilation and the sixth mass extinction. *Proc. Natl. Acad. Sci.* **117**, 13596–13602 (2020).

38. Liu, J. *et al.* Nexus approaches to global sustainable development. *Nat. Sustain.* **1**, 466–476 (2018).

39. Dorber, M., Arvesen, A., Gernaat, D. & Verones, F. Controlling biodiversity impacts of future global hydropower reservoirs by strategic site selection. *Sci. Rep.* **10**, 21777 (2020).

40. Verones, F., Bartl, K., Pfister, S., Jiménez Vílchez, R. & Hellweg, S. Modeling the Local Biodiversity Impacts of Agricultural Water Use: Case Study of a Wetland in the Coastal Arid Area of Peru. *Environ. Sci. Technol.* **46**, 4966–4974 (2012).

41. Mutel, C. L., Pfister, S. & Hellweg, S. GIS-Based Regionalized Life Cycle Assessment: How Big Is Small Enough? Methodology and Case Study of Electricity Generation. *Environ. Sci. Technol.* **46**, 13028–13028 (2012).

42. Borucke, M. *et al.* Accounting for demand and supply of the biosphere's regenerative capacity: The National Footprint Accounts' underlying methodology and framework. *Ecol. Indic.* **24**, 518–533 (2013).

43. Moran, D. & Kanemoto, K. Identifying species threat hotspots from global supply chains. *Nat. Ecol. Evol.* **1**, 0023 (2017).

44. Irwin, A. *et al.* Quantifying and categorising national extinction-risk footprints. *Sci. Rep.* **12**, 5861 (2022).

45. Verones, F., Moran, D., Stadler, K., Kanemoto, K. & Wood, R. Resource footprints and their ecosystem consequences. *Sci. Rep.* **7**, 40743 (2017).

46. Finnveden, G. *et al.* Recent developments in Life Cycle Assessment. *J. Environ. Manage.* **91**, 1–21 (2009).

47. Hellweg, S. & Milà i Canals, L. Emerging approaches, challenges and opportunities in life cycle assessment. *Science (80-.).* **344**, 1109–1113 (2014).

48. International Organization for Standardization. ISO 14044: 2018, Environmental Management: Life Cycle Assessment–Requirements and Guidelines. (2018).

49. Gibon, T., Hertwich, E. G., Arvesen, A., Singh, B. & Verones, F. Health benefits, ecological threats of low-carbon electricity. *Environ. Res. Lett.* **12**, 034023 (2017).

50. Maia de Souza, D., Teixeira, R. F. M. & Ostermann, O. P. Assessing biodiversity loss due to land use with Life Cycle Assessment: Are we there yet? *Glob. Chang. Biol.* **21**, 32–47 (2015).

51. Verones, F. *et al.* LCIA framework and cross-cutting issues guidance within the UNEP-SETAC Life Cycle Initiative. *J. Clean. Prod.* **161**, 957–967 (2017).

52. Kuipers, K. J. J., May, R. & Verones, F. Considering habitat conversion and fragmentation in characterisation factors for land-use impacts on vertebrate species richness. *Sci. Total Environ.* **801**, 149737 (2021).

53. De Schryver, A. M., Brakkee, K. W., Goedkoop, M. J. & Huijbregts, M. A. J. Characterization Factors for Global Warming in Life Cycle Assessment Based on Damages to Humans and Ecosystems. *Environ. Sci. Technol.* **43**, 1689–1695 (2009).

54. Azevedo, L. B., van Zelm, R., Hendriks, A. J., Bobbink, R. & Huijbregts, M. A. J. Global assessment of the effects of terrestrial acidification on plant species richness. *Environ. Pollut.* **174**, 10–15 (2013).

55. Scherer, L., Gürdal, İ. & van Bodegom, P. M. Characterization factors for ocean acidification impacts on marine biodiversity. *J. Ind. Ecol.* 1–11 (2022) doi:10.1111/jiec.13274.

56. MacArthur, R. H. & Wilson, E. O. *The theory of island biogeography*. (Princeton University Press, 1967).

57. Verones, F. *et al.* Effects of Consumptive Water Use on Biodiversity in Wetlands of International Importance. *Environ. Sci. Technol.* **47**, 12248–12257 (2013).

58. Chaudhary, A., Verones, F., De Baan, L. & Hellweg, S. Quantifying Land Use Impacts on Biodiversity: Combining Species-Area Models and Vulnerability Indicators. *Environ. Sci. Technol.* **49**, 9987–9995 (2015).

59. Pereira, H. M. & Daily, G. C. Modelling Biodiversity Dynamics in countryside Landscapes. *Ecology* **87**, 1877–1885 (2006).

60. Pereira, H. M., Ziv, G. & Miranda, M. Countryside Species-Area Relationship as a Valid Alternative to the Matrix-Calibrated Species-Area Model. *Conserv. Biol.* **28**, 874–876 (2014).

61. Chaudhary, A. & Brooks, T. M. Land Use Intensity-Specific Global Characterization Factors to Assess Product Biodiversity Footprints. *Environ. Sci. Technol.* **52**, 5094–5104 (2018).

62. Kuipers, K. J. J., Hellweg, S. & Verones, F. Potential Consequences of Regional Species Loss for Global Species Richness: A Quantitative Approach for Estimating Global Extinction Probabilities. *Environ. Sci. Technol.* **53**, 4728–4738 (2019).

63. Curran, M. *et al.* Toward Meaningful End Points of Biodiversity in Life Cycle Assessment. *Environ. Sci. Technol.* **45**, 70–79 (2011).

64. Woods, J. S. *et al.* Ecosystem quality in LCIA: status quo, harmonization, and suggestions for the way forward. *Int. J. Life Cycle Assess.* **23**, 1995–2006 (2018).

65. World Wildlife Fund. *WildFinder: Online database of species distributions*. http://www.worldwildlife.org/WildFinder (2006).

66. Feeley, K. J., Stroud, J. T. & Perez, T. M. Most 'global' reviews of species' responses to climate change are not truly global. *Divers. Distrib.* **23**, 231–234 (2017).

67. Di Marco, M. *et al.* Changing trends and persisting biases in three decades of conservation science. *Glob. Ecol. Conserv.* **10**, 32–42 (2017).

68. Favreau, J. M. *et al.* Recommendations for Assessing the Effectiveness of Surrogate Species Approaches. *Biodivers. Conserv.* **15**, 3949–3969 (2006).

69. Dorber, M., Mattson, K. R., Sandlund, O. T., May, R. & Verones, F. Quantifying net water consumption of Norwegian hydropower reservoirs and related aquatic biodiversity impacts in Life Cycle Assessment. *Environ. Impact Assess. Rev.* **76**, 36–46 (2019).

70. Uhler, J. *et al.* Relationship of insect biomass and richness with land use along a climate gradient. *Nat. Commun.* **12**, 5946 (2021).

71. Buzhdygan, O. Y., Tietjen, B., Rudenko, S. S., Nikorych, V. A. & Petermann, J. S. Direct and indirect effects of land-use intensity on plant communities across elevation in semi-natural grasslands. *PLoS One* **15**, e0231122 (2020).

72. Bachman, S., Moat, J., Hill, A., de la Torre, J. & Scott, B. Supporting Red List threat assessments with GeoCAT: geospatial conservation assessment tool. *Zookeys* **150**, 117–126 (2011).

73. Bachman, S. P. *et al.* Progress, challenges and opportunities for Red Listing. *Biol. Conserv.* **234**, 45–55 (2019).

74. Rondinini, C., Di Marco, M., Visconti, P., Butchart, S. H. M. & Boitani, L. Update or Outdate: Long-Term Viability of the IUCN Red List. *Conserv. Lett.* **7**, 126–130 (2014).

75. La Salle, J., Williams, K. J. & Moritz, C. Biodiversity analysis in the digital era. *Philos. Trans. R. Soc. B Biol. Sci.* **371**, 20150337 (2016).

76. Hilbert, M. & López, P. The World's Technological Capacity to Store, Communicate, and Compute Information. *Science (80-.).* **332**, 60–65 (2011).

77. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).

78. Gaulier, G. & Zignago, S. BACI: International Trade Database at the Product-Level (the 1994-2007 Version). *SSRN Electron. J.* (2010) doi:10.2139/ssrn.1994500.

79. Hansen, M. C. *et al.* High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science (80-.).* **342**, 850–853 (2013).

80. Karger, D. N. *et al.* Climatologies at high resolution for the earth's land surface areas. *Sci. Data* **4**, 170122 (2017).

81. Féret, J.-B. & Asner, G. P. Mapping tropical forest canopy diversity using high-fidelity imaging spectroscopy. *Ecol. Appl.* **24**, 1289–1296 (2014).

82. Schneider, F. D. *et al.* Toward mapping the diversity of canopy structure from space with GEDI. *Environ. Res. Lett.* **15**, 115006 (2020).

83. Stavros, E. N. *et al.* ISS observations offer insights into plant function. *Nat. Ecol. Evol.* **1**, 0194 (2017).

84. Lawrence, A. *et al.* The case for space environmentalism. *Nat. Astron.* **6**, 428–435 (2022).

85. Turner, W. *et al.* Free and open-access satellite data are key to biodiversity conservation. *Biol. Conserv.* **182**, 173–176 (2015).

86. Ustin, S. L. & Middleton, E. M. Current and near-term advances in Earth observation for ecological applications. *Ecol. Process.* **10**, 1 (2021).

87. Silvertown, J. A new dawn for citizen science. *Trends Ecol. Evol.* **24**, 467–471 (2009).

88. Follett, R. & Strezov, V. An Analysis of Citizen Science Based Research: Usage and Publication Patterns. *PLoS One* **10**, e0143687 (2015).

89. Lintott, C. J. *et al.* Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Mon. Not. R. Astron. Soc.* **389**, 1179–1189 (2008).

90. Bonney, R. *et al.* Next Steps for Citizen Science. *Science (80-.).* **343**, 1436–1437 (2014).

91. Middleton, J. V. The Stream Doctor Project : Community-driven stream restoration. *Bioscience* **51**, 293–296 (2001).

92. Fuccillo, K. K., Crimmins, T. M., de Rivera, C. E. & Elder, T. S. Assessing accuracy in citizen science-based plant phenology monitoring. *Int. J. Biometeorol.* **59**, 917–926 (2015).

93. Zattara, E. E. & Aizen, M. A. Worldwide occurrence records suggest a global decline in bee species richness. *One Earth* **4**, 114–123 (2021).

94. Sullivan, B. L. *et al.* eBird: A citizen-based bird observation network in the biological sciences. *Biol. Conserv.* **142**, 2282–2292 (2009).

95. Sullivan, B. L. *et al.* The eBird enterprise: An integrated approach to development and application of citizen science. *Biol. Conserv.* **169**, 31–40 (2014).

96. Dickinson, J. L., Zuckerberg, B. & Bonter, D. N. Citizen Science as an Ecological Research Tool: Challenges and Benefits. *Annu. Rev. Ecol. Evol. Syst.* **41**, 149–172 (2010).

97. Fink, D. *et al.* eBird Status and Trends, Data Version: 2020; Released: 2021. Cornell Lab of Ornithology, Ithaca, New York. https://doi.org/10.2173/ebirdst.2020 (2021) doi:10.2173/ebirdst.2020.

98. Wüest, R. O. *et al.* Macroecology in the age of Big Data – Where to go from here? *J. Biogeogr.* **47**, 1–12 (2020).

99. Hampton, S. E. *et al.* Big data and the future of ecology. *Front. Ecol. Environ.* **11**, 156–162 (2013).

100. Michener, W. K. & Jones, M. B. Ecoinformatics: supporting ecology as a data-intensive science. *Trends Ecol. Evol.* **27**, 85–93 (2012).

101. Peterson, A. T., Soberón, J. & Krishtalka, L. A global perspective on decadal challenges and priorities in biodiversity informatics. *BMC Ecol.* **15**, 15 (2015).

102. GBIF. The Global Biodiversity Information Facility: What is GBIF? https://www.gbif.org/what-is-gbif (2021).

103. Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R. & Legendre, F. Taxonomic bias in biodiversity data and societal preferences. *Sci. Rep.* **7**, 1–14 (2017).

104. Winter, L., Lehmann, A., Finogenova, N. & Finkbeiner, M. Including biodiversity in life cycle assessment – State of the art, gaps and research needs. *Environ. Impact Assess. Rev.* **67**, 88–100 (2017).

105. Woods, J. S., Veltman, K., Huijbregts, M. A. J., Verones, F. & Hertwich, E. G. Toward a meaningful assessment of marine ecological impacts in life cycle assessment (LCA). *Environ. Int.* **89**–**90**, 48–61 (2016).

106. Cooper, J., Noon, M., Jones, C., Kahn, E. & Arbuckle, P. Big Data in Life Cycle Assessment. *J. Ind. Ecol.* **17**, 796–799 (2013).

107. Baitz, M. *et al.* LCA's theory and practice: like ebony and ivory living in perfect harmony? *Int. J. Life Cycle Assess.* **18**, 5–13 (2013).

108. Stuart, S. N., Wilson, E. O., McNeely, J. A., Mittermeier, R. A. & Rodríguez, J. P. The Barometer of Life. *Science (80-.).* **328**, 177–177 (2010).

109. Whittaker, R. J. *et al.* Conservation Biogeography: Assessment and Prospect. *Divers. Distrib.* **11**, 3–23 (2005).

110. Hanafiah, M. M., Xenopoulos, M. A., Pfister, S., Leuven, R. S. E. W. & Huijbregts, M. A. J. Characterization Factors for Water Consumption and Greenhouse Gas Emissions Based on Freshwater Fish Species Extinction. *Environ. Sci. Technol.* **45**, 5272–5278 (2011).

111. Verones, F. *et al.* LC-IMPACT: A regionalized life cycle damage assessment method. *J. Ind. Ecol.* **24**, 1201–1219 (2020).

112. Gade, A. L., Hauschild, M. Z. & Laurent, A. Globally differentiated effect factors for characterising terrestrial acidification in life cycle impact assessment. *Sci. Total Environ.* **761**, 143280 (2021).

113. Rondinini, C., Wilson, K. A., Boitani, L., Grantham, H. & Possingham, H. P. Tradeoffs of different types of species occurrence data for use in systematic conservation planning. *Ecol. Lett.* **9**, 1136–1145 (2006).

114. Amano, T., Lamming, J. D. L. & Sutherland, W. J. Spatial Gaps in Global Biodiversity Information and the Role of Citizen Science. *Bioscience* **66**, 393–400 (2016).

115. Meyer, C., Weigelt, P. & Kreft, H. Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecol. Lett.* **19**, 992–1006 (2016).

116. Butchart, S. H. M. *et al.* Measuring Global Trends in the Status of Biodiversity: Red List Indices for Birds. *PLoS Biol.* **2**, e383 (2004).

117. Montesino Pouzols, F. *et al.* Global protected area expansion is compromised by projected land-use and parochialism. *Nature* **516**, 383–386 (2014).

118. Mooers, A. Ø., Faith, D. P. & Maddison, W. P. Converting Endangered Species Categories to Probabilities of Extinction for Phylogenetic Conservation Prioritization. *PLoS One* **3**, e3700 (2008).

119. Juffe-Bignoli, D. *et al.* Assessing the Cost of Global Biodiversity and Conservation Knowledge. *PLoS One* **11**, e0160640 (2016).

120. Cazalis, V. *et al.* Bridging the research-implementation gap in IUCN Red List assessments. *Trends Ecol. Evol.* **37**, 359–370 (2022).

121. IUCN Standards and Petitions Committee. *Guidelines for using the IUCN Red List Categories and Criteria*. *Prepared by the Standards and Petitions Committee. Downloadable from https://www.iucnredlist.org/documents/RedListGuidelines.pdf* vol. 15 (2022).

122. Verones, F., Pfister, S., van Zelm, R. & Hellweg, S. Biodiversity impacts from water consumption on a global scale for use in life cycle assessment. *Int. J. Life Cycle Assess.* **22**, 1247–1256 (2017).

123. Lenzen, M. *et al.* International trade drives biodiversity threats in developing nations. *Nature* **486**, 109–112 (2012).

124. Roberts, D. L., Taylor, L. & Joppa, L. N. Threatened or Data Deficient: assessing the conservation status of poorly known species. *Divers. Distrib.* **22**, 558–565 (2016).

125. Jarić, I., Courchamp, F., Gessner, J. & Roberts, D. L. Potentially threatened: a Data Deficient flag for conservation management. *Biodivers. Conserv.* **25**, 1995–2000 (2016).

126. Howard, S. D. & Bickford, D. P. Amphibians over the edge: silent extinction risk of Data Deficient species. *Divers. Distrib.* **20**, 837–846 (2014).

127. Jetz, W. & Freckleton, R. P. Toward a general framework for predicting threat status of data-deficient species from phylogenetic, spatial and environmental information. *Philos. Trans. R. Soc. B Biol. Sci.* **370**, 20140016 (2015).

128. IUCN Standards and Petitions Committee. *Guidelines for Reporting on Proportion Threatened (Version 1.2). Annex 1 of the Guidelines for Appropriate Uses of IUCN Red List Data (Version 4.0). Approved by the IUCN Red List Committee in November 2017. Downloadable from: https://www.iucnredlist.org/resources/guidelines-for-appropriate-uses-of-red-list-data.* (2022).

129. Crenna, E., Marques, A., La Notte, A. & Sala, S. Biodiversity Assessment of Value Chains: State of the Art and Emerging Challenges. *Environ. Sci. Technol.* **54**, 9715–9728 (2020).

130. Curran, M. *et al.* How Well Does LCA Model Land Use Impacts on Biodiversity? - A Comparison with Approaches from Ecology and Conservation. *Environ. Sci. Technol.* **50**, 2782–2795 (2016).

Chapter 2: Native Range Estimates for Red-Listed Vascular Plants
Scientific Data (2022) 9(1):117

scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Native range estimates for red-listed vascular plants

Jan Borgelt[1]✉, Jorge Sicacha-Parada[2], Olav Skarpaas[3] & Francesca Verones[1]

Besides being central for understanding both global biodiversity patterns and associated anthropogenic impacts, species range maps are currently only available for a small subset of global biodiversity. Here, we provide a set of assembled spatial data for terrestrial vascular plants listed at the global IUCN red list. The dataset consists of pre-defined native regions for 47,675 species, density of available native occurrence records for 30,906 species, and standardized, large-scale Maxent predictions for 27,208 species, highlighting environmentally suitable areas within species' native regions. The data was generated in an automated approach consisting of data scraping and filtering, variable selection, model calibration and model selection. Generated Maxent predictions were validated by comparing a subset to available expert-drawn range maps from IUCN (n = 4,257), as well as by qualitatively inspecting predictions for randomly selected species. We expect this data to serve as a substitute whenever expert-drawn species range maps are not available for conducting large-scale analyses on biodiversity patterns and associated anthropogenic impacts.

Background & Summary

Life on Earth is essential to human society as it forms the foundation of present welfare[1]. The growing human population, modern lifestyles and associated pressures on the planet have already resulted in a significant loss of natural habitat and are threatening biodiversity[2–6]. Different initiatives promote the protection of biodiversity and aim to halt its loss, such as the UN Sustainable Development Goals[7], the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services[8] and the International Union for the Conservation of Nature (IUCN). Different decision-support tools can contribute to this by assessing environmental performances of products, strategies and policies[2,9–11]. For the development of such tools, but also for the implementation of global conservation strategies and policies itself, spatial data, e.g. in the form of distribution maps of individual species[12], are crucial. However, besides many species remaining undiscovered or undescribed, we still lack spatial information for most of the ones we know[13]. Consequently, comprehensive and ready-to-use datasets for large-scale analyses are only available for a few vertebrate groups[14–16]. This is concerning, as global conservation strategies and biodiversity impact assessments are limited to these groups, while some hyperdiverse species groups, such as plants, are often not considered[17,18].

Here, we provide spatial distribution data for a large fraction of red-listed terrestrial vascular plant species at different levels of spatial detail (Fig. 1), i.e. native regions (n = 47,675), occurrence records (n = 30,906) and modelled range estimates (i.e. a predicted relative environmental suitability[19] within native regions; n = 27,208). The workflow included data scraping and filtering, as well as variable selection, model calibration and model selection, aiming for best practice[20–22] but within the constraints of data limitations and computational feasibility at this scale. Species-specific native regions were retrieved from a scheme specifically developed to challenge the lack of distributional knowledge for plant species[23]. Available native occurrence records were retrieved from the Global Biodiversity Information Facility (GBIF)[24] and subsequently filtered. Range estimates were generated using maximum entropy modelling[19,25–27], and show where environmentally suitable conditions exist within each species' native regions (Fig. 2a–d).

The underlying occurrence data is known to be highly spatiotemporally aggregated and variable across administrative borders for some species[28–31]. We aimed at counteracting a potential sampling bias by using three differently treated occurrence data types (i.e. different degree of spatial filtering: no filter, presence cells, thinned presence cells), and by dividing occurrence data in equally-sized bins during model calibration[32]. Up

[1]Industrial Ecology Programme, Department of Energy and Process Engineering, Norwegian University of Science and Technology (NTNU), Trondheim, Norway. [2]Department of Mathematical Sciences, Norwegian University of Science and Technology (NTNU), Trondheim, Norway. [3]Natural History Museum, University of Oslo, Oslo, Norway. ✉e-mail: jan.borgelt@ntnu.no

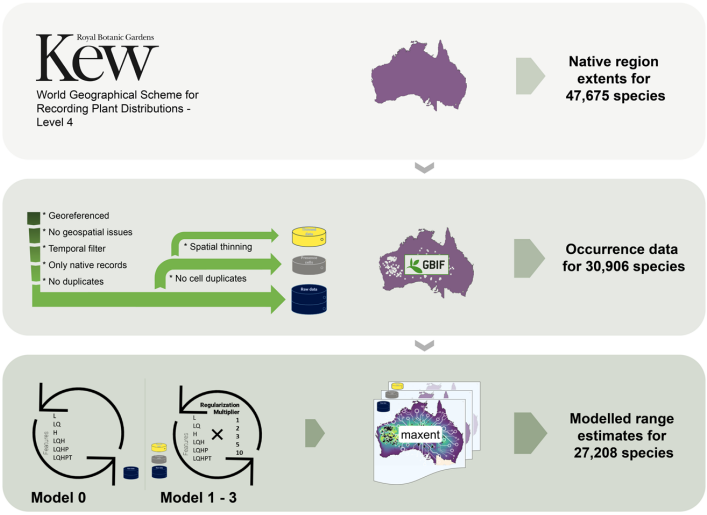


Fig. 1 Schematic summary of the dataset. Top: Native region extents were retrieved from Kew's Plants of the World online. Middle: Occurrence data was retrieved from the Global Biodiversity Information Facility (GBIF)[24] and filtered into three different occurrence data types: raw data (blue), presence cells (grey) and thinned data (yellow). Bottom: The different occurrence data types were used in Maxent models to predict relative environmental suitability indices within native regions (i.e. range estimates). Differences between Model 0 and Model 1 to 3. Model 0 was trained to support variable selection using raw data in k-fold cross validated Maxent models (one model for each combination of feature classes, i.e. linear (L), quadratic (Q), hinge (H), product (P) and threshold (T)). The selected variables and each of the three occurrence data types were used to train a set of separate k-fold cross validated Maxent models (one model for each possible combination of feature classes, regularization multipliers and occurrence data type). The overall best performing model was selected for each species based on performance metrics.

to 96 different models were fitted per species to find optimal variables, model settings and data type. The best prediction was selected for each species based on common performance metrics (i.e. AUC and AUC_{PR}).

However, some predictions will undoubtedly remain flawed by underlying biases. Based on comparisons to expert-drawn range maps available from IUCN (n = 4,257) and qualitative inspection of predictions for randomly selected species, we expect this to mainly influence widespread and common species, and hence, only affect the smallest proportion of global biodiversity[33]. In addition, the species most vital for assessing anthropogenic impacts or for defining conservation priorities, are more likely to be small-ranged and endemic. Although validating each prediction was not feasible, we found most individually inspected predictions to either offer an improvement compared to elsewhere available data or an acceptable substitute, although at a coarser spatial resolution and less detailed.

We want to stress that the presented dataset is generated for the purpose of global spatial screening studies and for building a basis for future, global biodiversity impact assessment models. In concert with powerful, species-specific trait and conservation-related databases, the provided data can benefit future work, such as assessing global extinction probabilities[34], effects of terrestrial acidification[35], drivers of invasion success[36], progress towards reaching global conservation goals[37] and act as pre-assessment prior to expert-based range map generation and red list assessments[38–41]. With a continuously increasing availability of species occurrence records, the presented dataset can be updated frequently to illustrate the state of knowledge at any time. With more data becoming available, precision is likely to increase in the future.

Methods

Taxonomic scope. A species list containing all terrestrial vascular plants (n = 52,372) of the global IUCN red list was retrieved from IUCN in April 2021, IUCN version 2021-1[16]. We retrieved each species' accepted name from Plants of the World Online (POWO)[42] to facilitate communication to various data portals using the package *taxize*[43] in R[44]. Plant family, order and class were retrieved from the Integrated Taxonomic Information System[45] using the package *taxize*[43] in R. Only species outside the IUCN threat categories "Extinct" and "Extinct in the Wild" were kept, and all species considered as subspecies or varieties according to POWO removed. We attempted to assemble spatial data for each of the remaining 48,144 species.

Native regions. Species-specific native regions (Fig. 1) were retrieved from POWO using a customized web-scraper function (see section *Code Availability*) and the packages *taxize*[43] and *rvest*[46] in R. The data follows the World Geographical Scheme for Recording Plant Distributions (WGSRPD)[23] and includes a continental,

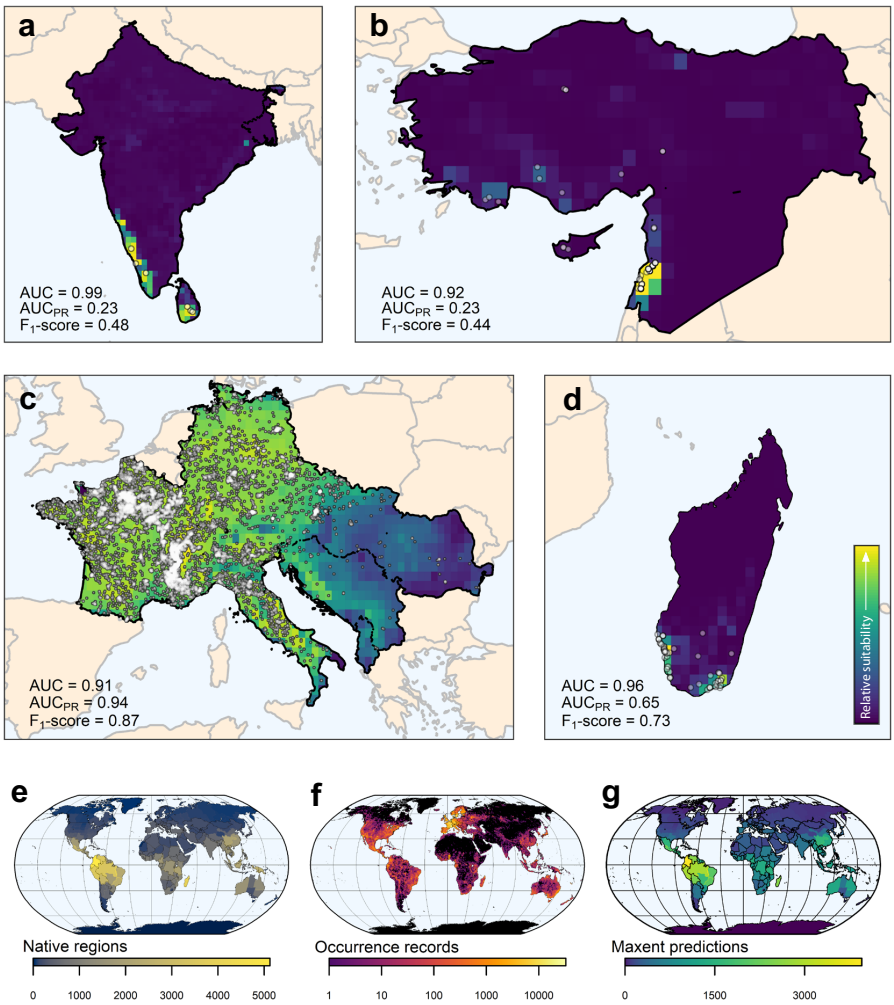


Fig. 2 Data examples for randomly selected species and spatial coverage of the dataset. Best performing Maxent prediction, highlighting environmentally suitable conditions within the species native regions (i.e. modelling extent) along retrieved occurrence records (white points) for (**a**) *Amomum pterocarpum*, (**b**) *Cedrus libani*, (**c**) *Laburnum anagyroides*, (**d**) *Megistostegium nodulosum*. Performance of the shown predictions indicated by maximum F_1-score and the area under the receiver operating characteristics curve for true vs. false positive rate (AUC) and recall vs. precision (AUC_{PR}). Bottom: number of (**e**) retrieved native regions, (**f**) retrieved occurrence records, and (**g**) generated Maxent predictions across the globe.

country and regional level. Retrieved WGSRPD-*regions* were matched to its corresponding shapefile at level 4, available from the Biodiversity Information Standards GitHub repository[47] and rasterized at 30 arc minutes spatial resolution (approximately 56 km at the equator).

Occurrence records. For species with given native extents in POWO, the maximum number of most recent occurrence points (i.e. 100,000) per native WGSRPD-*country* was retrieved from the GBIF application programming interface (API) using the package *rgbif*[48] in R (the equivalent full dataset[49] is available at https://doi.org/10.15468/dl.uvd56q). The considered environmental variables have changed tremendously in the past decades[50,51] and only cover a limited period of time, i.e. the years 1979–2013 and 2015 respectively (see section *Environmental data*). Therefore, only records between the years 2000 and 2020 were considered to temporally align occurrence data to both sets of environmental variables as best as possible. If less than 25 records were available for a given species after the year 2000, no temporal filter was set to maximize data retrieval. GBIF records without specified coordinates and with flagged geospatial issues[48] were not considered. As such, we expect

inaccurate coordinate notations as well as records of specimens preserved in museums or other biodiversity facilities to be typically detected. Only points inside reported native WGSRPD-*regions* were kept and duplicated records were removed (hereafter: raw data). The number of raw data records was counted per cell (30 arc min.) using the package *raster*[52] in R.

Maxent predictions. We generated spatial predictions within species' native WGSRPD-*regions* at 30 arc min. resolution (approximately 56 km at the equator) using maximum entropy modelling (Maxent)[19,26,27], for all species with at least 5 raw data records[53,54] that were distributed across at least 3 cells, and a native region extent of at least 9 cells. Although an arbitrary threshold, we attempted to allocate computational resources to more meaningful predictions, modelled across larger extents. Maxent is a probability density estimation approach widely used for predicting species distributions based on presence-only data[55]. Background information, required to fit response curves[56], was collected from each cell within each species' native regions[57]. For generating models we utilized a high-performance computing infrastructure[58] allowing for parallel computations using the Maxent software[25] via R packages *dismo*[59] and *ENMeval*[60].

Environmental data. We downloaded all CHELSA bioclimatic variables[61,62] ($n = 19$, see Table 1 for full list) in 30 arc seconds resolution and aggregated, for computational efficiency, to the chosen modelling resolution (30 arc min.) by averaging. CHELSA bioclimatic variables are a set of modelled, biologically relevant, climatic variables based on data collected during the years 1979–2013[61]. In addition, fractions for different natural land cover types, including different types and mosaics of forest, shrubland, grassland and sparse vegetation, ($n = 17$, see Table 1 for full list) were calculated based on the European Space Agency's land cover product for the year 2015 in 300 m resolution[63]. Each land cover class was transformed into a binary raster depicting presence (=1) and absence (=0) of the land cover type. The binary raster was then aggregated to modelling resolution by averaging, resulting in one raster for each land cover class, representing the proportion of land covered by that class per pixel.

Occurrence data types. For some species, several raw data records can be in the same cell at the given spatial resolution (30 arc min.). Although pseudo-replication can inflate model performance (here: during model calibration) and, hence, increases the risk of overfitting, we argue that these occurrence points still contain valid information if they are discrete observations and therefore kept this data. However, we henceforth applied two filters to counteract potential spatial biases, as well as pseudo-replication (Fig. 1). We removed all cell-duplicates from the raw data (hereafter: presence cells), and we applied spatial thinning with a minimum distance of two cells on the presence cells (hereafter: thinned data). Occurrence data was spatially filtered using the R package *spThin*[64].

Model training. A set of Maxent models was fitted for each species using the differently treated occurrence data types. All models were calibrated using k-fold cross validation. The employed occurrence data was partitioned into training and testing bins. For species with only few data points ($n < 25$), we used k - 1 Jackknife partitioning ($k = n$)[54]. For species with more data points ($n \geq 25$) we used block partitioning ($k = 4$) to account for spatial autocorrelation of occurrence points in larger datasets[32]. This partitioning splits the occurrence data at a longitudinal and latitudinal line, resulting in approximately equally sized bins[60].

An initial model (Fig. 1; Model 0) was trained to support the selection of uncorrelated environmental variables using the raw data and all environmental variables ($n = 36$) for each species. Separate models, one for each possible combination out of all included feature classes (i.e. environmental variables and transformations thereof), were trained. We included linear (l), quadratic (q), product (p), hinge (h) and threshold (t) transformations, resulting in 6 possible combinations (i.e. l, lq, h, lqh, lqhp, and lqhpt). The best performing model was selected based on the corrected Akaike information criterion (AICc)[65–67]. However, if no model performed best in terms of AICc, or if this metric was unavailable for 50% of fitted models, the average testing area under the receiver operating characteristics curve (AUC; see section *Technical Validation*) during model calibration was used instead. Permutation importance was retrieved for all variables in Model 0. Correlated variables were identified using Spearman's rank correlation coefficient (ρ) and defined as $\rho \geq |\pm 0.7|$. In any set of correlated variables, only the variable with the greatest permutation importance was kept.

The selected environmental variables were used to train separate models for each of the three differently treated occurrence data types: raw data (Model 1), presence cells (Model 2), and thinned data (Model 3). Model 1 was trained if at least 5 raw data records were available, distributed across at least 3 cells (see above). Model 2 and Model 3 were trained if at least 3 records of the corresponding data type were available to avoid computational failure. Although a smaller sample size, we argue that if those models performed better than Model 1, the threshold of 5 records becomes arbitrary and the assessed performance indicators (see section *Technical Validation*) more valuable. The same model architecture as in Model 0 was utilized, including model calibration and selection of the best performing model. However, this time, we added five different regularization multipliers (RM; i.e. 1, 2, 3, 5 and 10; based on previous studies[68–70]) to counteract overfitting[20,56] and for building simpler, ecologically more relevant, models[60]. Hence, separate models for each possible combination out of feature classes and RMs were trained (Fig. 1; Model 1–3), resulting in 30 trained models for each data type and up to 90 models per species.

Metadata. Metadata was assembled for all data and includes general information about species (taxonomy and red list status), provided data type (native regions, occurrence records or Maxent prediction), bounding box of native regions, and if relevant, information about the occurrence data (number of raw data records, Moran's

Variable	Code
Annual Mean Temperature	CHELSA_BIO1
Mean Diurnal Range	CHELSA_BIO2
Isothermality	CHELSA_BIO3
Temperature Seasonality	CHELSA_BIO4
Max Temperature of Warmest Month	CHELSA_BIO5
Min Temperature of Coldest Month	CHELSA_BIO6
Temperature Annual Range	CHELSA_BIO7
Mean Temperature of Wettest Quarter	CHELSA_BIO8
Mean Temperature of Driest Quarter	CHELSA_BIO9
Mean Temperature of Warmest Quarter	CHELSA_BIO10
Mean Temperature of Coldest Quarter	CHELSA_BIO11
Annual Precipitation	CHELSA_BIO12
Precipitation of Wettest Month	CHELSA_BIO13
Precipitation of Driest Month	CHELSA_BIO14
Precipitation Seasonality	CHELSA_BIO15
Precipitation of Wettest Quarter	CHELSA_BIO16
Precipitation of Driest Quarter	CHELSA_BIO17
Precipitation of Warmest Quarter	CHELSA_BIO18
Precipitation of Coldest Quarter	CHELSA_BIO19
Fraction of mosaic cropland/natural vegetation	X30_ESA_CCI
Fraction of mosaic natural vegetation/cropland	X40_ESA_CCI
Fraction of broadleaved evergreen, closed to open, tree cover	X50_ESA_CCI
Fraction of broadleaved deciduous, closed to open, tree cover	X60_ESA_CCI
Fraction of needleleaved evergreen, closed to open, tree cover	X70_ESA_CCI
Fraction of needleleaved deciduous, closed to open, tree cover	X80_ESA_CCI
Fraction of mixed leaf type tree cover	X90_ESA_CCI
Fraction of mosaic tree and shrub/herbaceous cover	X100_ESA_CCI
Fraction of mosaic herbaceous cover/tree and shrub	X110_ESA_CCI
Fraction of shrubland	X120_ESA_CCI
Fraction of grassland	X130_ESA_CCI
Fraction of lichens and mosses	X140_ESA_CCI
Fraction of sparse vegetation	X150_ESA_CCI
Fraction of tree cover, flooded, fresh or brakish water	X160_ESA_CCI
Fraction of tree cover, flooded, saline water	X170_ESA_CCI
Fraction of shrub or herbaceous cover, flooded, fresh/saline/brakish water	X180_ESA_CCI
Fraction of bare areas	X200_ESA_CCI

Table 1. Environmental data used in this study. The layers (n = 36) are based on Karger *et al.*[62] and the European space agency's land cover product[63].

Index[71], calculated as a measure of spatial autocorrelation and based on the number of raw occurrence points obtained per cell), and Maxent metadata: training data (filter treatment, number of training data points), thresholds for converting the prediction into binary range maps[59], model settings (features, parameters, transformations, regularization multiplier, variables) and out of the box[60] model performance, including degree of overfit (DOO) quantified as the difference between calibration and testing AUC during k-fold cross validation[70], as well as self-assessed model performance metrics as described in the section *Technical Validation*.

Data Records

Dataset. The presented dataset is stored in a stable Dryad Digital Repository[72] and can be explored at https://plant-ranges.indecol.no. The dataset includes spatial information for 47,675 species at different levels of detail. In total, range estimates (i.e. relative environmental suitability within native regions) have been predicted for 27,208 species using Maxent, for 30,906 species native occurrence records are provided, and for 47,675 species the spatial extent of its native WGSRPD-*regions* is provided.

All gathered and generated data are stored in netCDF files and can be called by specifying a *varname*. Spatial predictions are provided in Maxent's raw as well as default output (i.e. complementary log-log (cloglog) transformed, but see section *Usage Notes*)[27,59,60]. The suggested data is stored in folder *basic*. These netCDF files (default output and raw output) assemble the best performing Maxent prediction (*varname*: Maxent prediction) for each species selected based on the highest harmonic mean between AUC and AUC_{PR} (see *Technical Validation*), along with number of occurrence records per cell (*varname*: Presence cells) and rasterized native WGSRPD-*regions* (*varname*: Native region).

	Reference		Red list category						
			DD	LC	NT	VU	EN	CR	Total
AUC	Presence - background	Mean	0.939	0.937	0.95	0.96	0.971	0.957	0.945
		Median	0.961	0.951	0.977	0.985	0.994	0.989	0.964
	Reference range	Mean	0.817	0.89	0.927	0.931	0.929	0.915	0.902
		Median	0.852	0.925	0.972	0.974	0.98	0.987	0.943
AUC_{PR}	Presence - background	Mean	0.576	0.529	0.656	0.69	0.749	0.7	0.589
		Median	0.603	0.535	0.717	0.755	0.833	0.797	0.617
	Reference range	Mean	0.516	0.664	0.686	0.653	0.655	0.592	0.658
		Median	0.527	0.702	0.737	0.712	0.699	0.626	0.702

Table 2. Performance of Maxent predictions in the suggested dataset. Mean and median values of area under the receiver operating characteristics curve for true vs. false positive rate (AUC) and recall vs. precision (AUC_{PR}) for all species and across different IUCN threat categories (i.e. data-deficient (DD), least concern (LC), near-threatened (NT), vulnerable (VU), endangered (EN) and critically endangered (CR)). Calculations are based on presence-background data (n = 27,208) and on comparison to expert-based range maps retrieved from IUCN (i.e. reference range, n = 4,257).

The netCDF files in folder *advanced* contain one Maxent prediction for each occurrence data type (*varname*: Model 1, Model 2 or Model 3), instead of best performing Maxent prediction (i.e. *varname* Maxent prediction is not applicable). Number of occurrence records per cell (*varname*: Presence cells) and rasterized native WGSRPD-*regions* (*varname*: Native region) are identical in all netCDF files.

Each band in the netCDF files assembles the mentioned variables for one species. The corresponding bands can be looked up in the metadata (i.e. *speciesID*). Furthermore, the metadata can be used to select appropriate cut-off thresholds for generating binary range maps, filter models based on species, performance, or desired datatypes, and to lookup the relevant study extent for masking individual predictions (see *Usage Notes*).

Technical Validation

Maxent predictions. We calculated performance metrics for model 1 to 3 for each species using its corresponding presence cells to validate the Maxent predictions. Receiver operating characteristic curves and the corresponding area under the curve for *recall* (i.e. *true positive rate, sensitivity*) versus *false positive rate* (AUC) as well as *precision* versus *recall* (AUC_{PR}) were generated using the packages *ROCR*[73] and *PRROC*[74] in R. *Recall* was calculated as the fraction of correctly predicted presence cells compared to all presence cells of the reference (Eq. 1), the *false positive rate* as the fraction of falsely assigned presence cells compared to all true absence cells (Eq. 2), and *precision* as the fraction of correctly assigned presence cells compared to all predicted presence cells (Eq. 3). In addition, F_1-scores (Eq. 4) were calculated as harmonic mean between *recall* and *precision* at all possible cut-off thresholds to transform the Maxent prediction into a binary range map. The maximum obtained F_1-score indicates how well a potential binary range map performs at equal importance of *recall* and *precision*.

$$Recall = \frac{True\ Presence}{True\ Presence + False\ Absence} \tag{1}$$

$$False\ positive\ rate = \frac{False\ Presence}{False\ Presence + True\ Absence} \tag{2}$$

$$Precision = \frac{True\ Presence}{True\ Presence + False\ Presence} \tag{3}$$

$$F_1 = 2\left(\frac{precision \cdot recall}{precision + recall}\right) \tag{4}$$

AUC and AUC_{PR} are threshold-independent performance measures for binary classifiers. An AUC value of 1 indicates a perfect model, an acceptable AUC value (>0.7)[75] indicates the ability to predict many true presences at a low false positive rate, and an AUC value of 0.5 indicates the model performing as good as a random guess. The average AUC obtained across the suggested dataset was 0.95 when comparing predictions to its corresponding presence cells (Table 2), indicating well-performing models for the majority of species. For 26,977 species (99%), at least one Maxent prediction had an AUC value above 0.7[75].

AUC_{PR} is not affected by true negatives (i.e. true absence) which often dominated our dataset. A higher AUC_{PR} value indicates a relatively higher ability to correctly predict a high proportion of presumably true range while maintaining a high precision compared to a lower AUC_{PR}. However, the AUC and AUC_{PR} values, as well as max. F_1-score, described here were calculated based on presence-background data and are highly influenced by class balances. Strictly speaking, both false presences and true absences cannot be determined

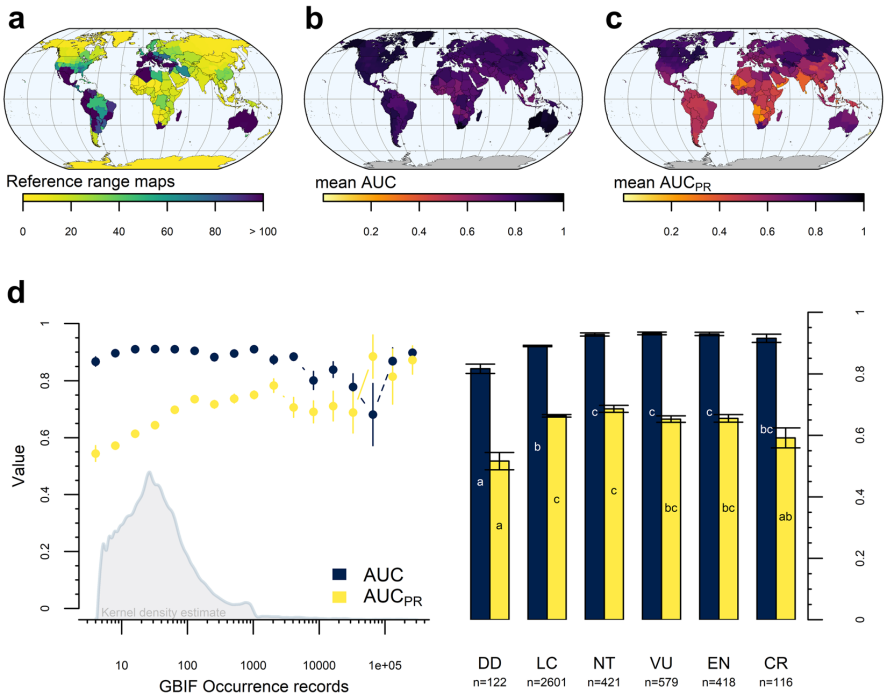


Fig. 3 Performance metrics for the suggested Maxent predictions. (**a**) Number of reference range maps available used for calculating performance metrics. Average values for species native to the corresponding regions of area under the receiver operating characteristics curve for (**b**) true vs. false positive rate (AUC) and (**c**) recall vs. precision (AUC_{PR}). (**d**) Mean and standard deviation of AUC (blue) and AUC_{PR} (yellow) per rounded log-transformed number of raw occurrence data points (left) and for species in different IUCN red list categories (right), i.e. data-deficient (DD), least concern (LC), near-threatened (NT), vulnerable (VU), endangered (EN) and critically endangered (CR). Significant differences across IUCN categories in d are indicated by different letters in bars for AUC (white text) and AUC_{PR} (black text).

with presence-only data. Hence, the performance metrics described here can only be used to compare different models for a given species, but not across different species[76,77].

Therefore, we evaluated the Maxent predictions by comparison to available expert-based range maps, as an additional evaluation dataset[32]. Expert-based range maps were retrieved from IUCN, if available (hereafter: reference ranges). Only reference ranges that were labelled as "native" and "extant (resident)" or "probably extant (resident)" were considered. For 4,257 species of our Maxent predictions, range maps were available at IUCN. These species were unevenly distributed in space (Fig. 3a), across IUCN red list categories (Fig. 3d) as well as the plant classes dicots (Magnoliopsida, n = 3,480), monocots (Liliopsida, n = 731), ferns (Polypodiopsida, n = 27), conifers (Pinopsida, n = 17), and lycopods (Lycopodiopsida, n = 2). Reference ranges were used to calculate the above described performance measures (i.e. max. F_1-score, AUC and AUC_{PR}). However, this time we dealt, presumably, with actual presences and absences of the given species, making the performance metrics comparable across species[76]. Maxent predictions for species classified as "data-deficient" (DD) obtained the lowest, and predictions for species classified as "near-threatened" (NT), "vulnerable" (VU) and "endangered" (EN) the highest AUC values (Fig. 3d). However, these differences were marginal and all average values consistently high across different IUCN categories (mean AUC: 0.9; Table 2) and across the globe (Fig. 3b). Although AUC is a strong indication of model performance[75], the predictions seem to rarely accommodate both a high *recall* and a high *precision* (represented in either max. F1-score or AUC_{PR} value) when compared to reference ranges. However, we found a large variation and no clear trend in AUC_{PR} values for species across different threat-level categories (Fig. 3d), and although the average AUC_{PR} was lowest for species native to parts of central Africa, India and south-eastern Asia (Fig. 3c), we expect these values to be of little explanatory power due to the limited sample sizes in these regions (Fig. 3a). Moreover, AUC_{PR} seems to increase with increasing data availability (Fig. 3d). We assume that low data coverage in sparsely populated areas influenced modelling performance for some, primarily widespread, species, highlighting that sometimes more spatially distributed occurrence data is required for making expert-alike range maps[78].

Furthermore, based on a qualitative assessment of predictions for twelve randomly selected species, we expect uncertainties due to differences in data availability across administrative borders as well as for highly naturalized species. For instance, the clustered occurrence records for *Cedrus libani* in Lebanon (Fig. 2b)

resulted in less precise data than elsewhere available for this species[79], while the prediction for *Laburnum anagyroides* (Fig. 2c) was affected by naturalized occurrence records outside its native origin[80] but still within its native WGSRPD-*regions*. However, this will be most problematic for abundant, widespread, and naturalized species, and hence only relevant for the smallest fraction of global biodiversity[33]. In addition, the predictions for more vulnerable species, presumably small-ranged or endemic, seem to perform better than species in the lowest red list category (i.e. least concern (LC)) in terms of AUC when compared to reference ranges (Fig. 3d).

In fact, the remaining randomly selected predictions were either consistent with point data (e.g. *Terminalia macrostachya*[81]), reflected the current knowledge of elsewhere available data, although at a coarser spatial resolution and less detailed (e.g. *Mammillaria grahamii*[82]), or offered an improvement compared to previously unavailable spatial data (e.g. *Eucalyptus elliptica*[83], *Megistostegium nodulosum*[84] (Fig. 2d), *Memecylon elegantulum*[85], *Psidium salutare*[86,87], *Siparuna conica*[88,89], *Trisetaria dufourei*[90]). However, the prediction of *Pyracantha angustifolia* was difficult to evaluate due to poorly understood range dynamics[91], highlighting the need for more data for vascular plant species.

We want to stress that our predictions indicate environmentally suitable conditions even if isolated from known species occurrence locations. For instance, *Amomum pterocarpum* seems to be restricted to southern India and Sri Lanka[92,93] while our prediction indicates environmentally suitable conditions in north-eastern India (Fig. 2a), which in fact, supports a possible observation nearby[94]. We further detected several expert-based range maps with a substantial mismatch to our data, confirming that some of the expert-based data may be too conservative[95] (e.g. *Magnolia pugana*)[96]. However, we also found expert-based ranges being smaller (e.g. *Vallesia glabra* or *Tetraclinis articulata*)[97,98] than predicted environmental suitability indicates, or being incorrectly georeferenced (e.g. *Corylus cornuta*)[99]. Hence, besides highlighting mismatches to expert-based range maps, we expect this dataset to be of sufficient quality to serve as time- and cost-efficient range map substitutes and pre-assessed range estimates for currently unmapped species.

External data. The retrieved native WGSRPD-*regions* are provided by POWO under a CC BY 3.0 license (https://creativecommons.org/licenses/by/3.0/) and have been checked for consistency to assure proper workflow of data retrieval from POWO and feature matching to the WGSRPD level 4 shapefile. However, the data provider, POWO, cannot warrant the quality or accuracy of the WGSRPD data[42]. In addition, other data (e.g. ecoregions[100]) may ecologically be more relevant than administrative boundaries. However, WGSRPD offers the most detailed data on species' native origins available on a large-scale, to the best of our knowledge. An attempt in matching native WGSRPD-*regions* to ecoregions was discontinued after loss of information due to incompatible geographical boundaries. Hence, we consider the utilized WGSRPD-*regions*, currently, as the best compromise between level of detail and availability of data on species' native origins. Furthermore, spatial inaccuracies and biases in the occurrence data retrieved from GBIF were counteracted by the implemented filtering steps, the coarse spatial resolution, by avoiding non-native occurrence records and the model calibration techniques. However, any unforeseen misclassified or misreported records may flaw predictions for individual species. In addition, data retrieval via GBIF's API was limited to 100,000 occurrence records per request. We extended this limit by sending one request per native country for each species, and hence, expect this issue to be irrelevant for our study. We further want to stress that most of the generated predictions have not been validated individually, and that some predictions may be erroneous either due to data limitations or simply because digitally stored data can contain minor but crucial blunders. For instance, in terms of nomenclature, the red-listed species *Cotoneaster cambricus* is endemic to Wales[101], but also seems to be a synonym for a widespread species according to POWO[42]. Consequently, either our spatial prediction or the expert-based range for this species is incorrect.

Usage Notes

All data handling, modelling and visualization was done using R version 4.0.3[44] in RStudio version 1.4.1103[102]. Handling of all spatial data was done using the R packages *raster*, *rgdal*, *maptools*, *rgeos* and *sp*[52,103–106]. A showcase for opening the different data types for individual species, is available at https://github.com/jannebor/plant_range_estimates. Although functionality of the code may be given at newer, or older, versions, we expect the best user-experience using the versions specified in this descriptor.

Maxent predictions are given as raw and cloglog transformed output. These outputs are related monotonically, meaning that the performance metrics described in this study, as well as a potential binary range map (excluding prevalence dependent thresholds), will be identical for both raw and cloglog output[56]. For users mostly interested in qualitative analyses, both predictions can simply be interpreted as indices of environmental suitability[20]. However, due to rescaling, the exact interpretation and appearance of each output differs. In general, Maxent's output interpretation depends on the underlying data, and differs, in our case between Model 1 (raw data including pseudo-replicates = abundance) compared to Model 2 and 3 (presence), but gives an estimate of the abundance, or presence, of the species in relation to the true modelled quantity (either abundance or presence). Maxent's raw output reflects the exponential Maxent model itself, and can be interpreted as a relative occurrence (or presence) rate summing up to 1[20]. The raw output does not rely on any assumptions[20], however, it may not perform well in visualizing actual differences in suitability[107]. Being rescaled on a more common range from 0 to 1, the cloglog transformation compresses extreme values, and hence facilities visualization and comparison amongst predictions[27]. It can, arguably, be interpreted as a relative probability of presence under certain assumptions[27]. However, as these assumptions are rarely met, we strongly discourage users from this interpretation and suggest interpreting the cloglog output values as an estimate of relative environmental suitability[20] instead.

We further suggest using Maxent predictions with an AUC below 0.7 only in exceptions, and in large-scale studies. In general, our predictions may overestimate true range extents of endemic species and underestimate

ranges of widespread species. However, in worst case, the entire native WGSPRD-*regions* are outlined as being environmentally suitable, which may be acceptable in some cases, but not in others.

In addition, Model 1 has been fitted with the suggested minimum number of records for generating meaningful distributions models[53,54], but Model 2 and 3 were in some cases trained with less records. Whether this low sample size as well as its implied uncertainty is acceptable or not will differ between users and applications and needs to be considered.

The full data, including Maxent predictions (cloglog transformed), underlying occurrence records, native regions and corresponding metadata, can be explored at https://plant-ranges.indecol.no. Here, the predictions based on individual models (Model 1 to 3) as well as a suggested (i.e. best performing) prediction highlight environmentally suitable conditions, if available for the selected species. Predictions can potentially be transformed into a map indicating where the species is most certainly found, as required for local management and conservation actions[95], or into a conservative range map, best suited for analysing global patterns[108] and highlighting where a species is certainly absent[109]. However, the choice of an appropriate cut-off threshold is highly application specific. We outlined "potential range maps" in the data explorer for illustrational purposes only and based on the best performing prediction. We applied different cut-off thresholds to represent different levels of confidence using the R package *dismo*[59]. The threshold at which there was no omission (possibly suitable), the threshold at which the F_1-score is highest (probably suitable) and presence cells (presence).

Code availability

All data and code is available without restrictions under the terms of a Creative Commons Zero (CC0) waiver (https://creativecommons.org/share-your-work/public-domain/cc0/). R code for retrieving and filtering data from POWO and GBIF, and for generating and evaluating Maxent models is available on GitHub (https://github.com/jannebor/plant_range_estimates). Any further requests can be directed to the corresponding author.

Received: 2 June 2021; Accepted: 24 February 2022;
Published online: 29 March 2022

References

1. Millennium Ecosystem Assessment. *Ecosystems and Human Well-being: Biodiversity Synthesis*. (World Resources Institute, 2005).
2. Moran, D. & Kanemoto, K. Identifying species threat hotspots from global supply chains. *Nat. Ecol. Evol.* **1**, 0023 (2017).
3. Newbold, T. Future effects of climate and land-use change on terrestrial vertebrate community diversity under different scenarios. *Proc. R. Soc. B Biol. Sci.* **285**, 20180792 (2018).
4. Newbold, T. *et al*. Global effects of land use on local terrestrial biodiversity. *Nature* **520**, 45–50 (2015).
5. Newbold, T. *et al*. Has land use pushed terrestrial biodiversity beyond the planetary boundary? A global assessment. *Science (80-.).* **353**, 288–291 (2016).
6. Verones, F., Moran, D., Stadler, K., Kanemoto, K. & Wood, R. Resource footprints and their ecosystem consequences. *Sci. Rep.* **7**, 40743 (2017).
7. United Nations. *Transforming our World: the 2030 Agenda for Sustainable Development*. A/RES/70/1 (United Nations, 2015).
8. Díaz, S. *et al*. Pervasive human-driven decline of life on Earth points to the need for transformative change. *Science (80-.).* **366**, eaax3100 (2019).
9. Lenzen, M. *et al*. International trade drives biodiversity threats in developing nations. *Nature* **486**, 109–112 (2012).
10. Hellweg, S. & Milà i Canals, L. Emerging approaches, challenges and opportunities in life cycle assessment. *Science (80-.).* **344**, 1109–1113 (2014).
11. Chaudhary, A. & Brooks, T. M. National Consumption and Global Trade Impacts on Biodiversity. *World Dev.* **121**, 178–187 (2019).
12. Pereira, H. M., Ziv, G. & Miranda, M. Countryside Species-Area Relationship as a Valid Alternative to the Matrix-Calibrated Species-Area Model. *Conserv. Biol.* **28**, 874–876 (2014).
13. Lomolino, M. V & Heaney, L. R. *Frontiers of Biogeography: New Directions in the Geography of Nature*. (Sinauer Associates Inc. Publishers, 2004).
14. World Wildlife Fund. *WildFinder: Online database of species distributions*. http://www.worldwildlife.org/WildFinder (2006).
15. BirdLife International. *IUCN Red List for birds*. http://www.birdlife.org (2019).
16. IUCN. *The IUCN Red List of Threatened Species. Version 2021-1* https://www.iucnredlist.org (2021).
17. Curran, M. *et al*. Toward Meaningful End Points of Biodiversity in Life Cycle Assessment. *Environ. Sci. Technol.* **45**, 70–79 (2011).
18. Woods, J. S. *et al*. Ecosystem quality in LCIA: status quo, harmonization, and suggestions for the way forward. *Int. J. Life Cycle Assess.* **23**, 1995–2006 (2018).
19. Phillips, S. J., Anderson, R. P. & Schapire, R. E. Maximum entropy modeling of species geographic distributions. *Ecol. Modell.* **190**, 231–259 (2006).
20. Merow, C., Smith, M. J. & Silander, J. A. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography (Cop.).* **36**, 1058–1069 (2013).
21. Araújo, M. B. *et al*. Standards for distribution models in biodiversity assessments. *Sci. Adv.* **5**, eaat4858 (2019).
22. Zurell, D. *et al*. A standard protocol for reporting species distribution models. *Ecography (Cop.).* **43**, 1261–1277 (2020).
23. Brummitt, R. K., Pando, F., Hollis, S. & Brummitt, N. A. World Geographical Scheme for Recording Plant Distributions. *International Working Group on Taxonomic Databases (TDWG)* https://www.tdwg.org/standards/wgsrpd/ (2001).
24. GBIF. The Global Biodiversity Information Facility: What is GBIF? https://www.gbif.org/what-is-gbif (2021).
25. Phillips, S. J., Dudík, M. & Schapire, R. E. Maxent software for modeling species niches and distributions (Version 3.4.0). http://biodiversityinformatics.amnh.org/open_source/maxent/ (2016).
26. Phillips, S. J., Dudík, M. & Schapire, R. E. A maximum entropy approach to species distribution modeling. *Proc. Twenty-first Int. Conf. Mach. Learn.* 655–662 (2004).
27. Phillips, S. J., Anderson, R. P., Dudík, M., Schapire, R. E. & Blair, M. E. Opening the black box: an open-source release of Maxent. *Ecography (Cop.).* **40**, 887–893 (2017).
28. Reddy, S. & Dávalos, L. M. Geographical sampling bias and its implications for conservation priorities in Africa. *J. Biogeogr.* **30**, 1719–1727 (2003).
29. Hortal, J., Jiménez-Valverde, A., Gómez, J. F., Lobo, J. M. & Baselga, A. Historical bias in biodiversity inventories affects the observed environmental niche of the species. *Oikos* **117**, 847–858 (2008).
30. Isaac, N. J. B. & Pocock, M. J. O. Bias and information in biological records. *Biol. J. Linn. Soc.* **115**, 522–531 (2015).

31. Feeley, K. J. & Silman, M. R. Keep collecting: accurate species distribution modelling requires more collections than previously thought. *Divers. Distrib.* **17**, 1132–1140 (2011).
32. Radosavljevic, A. & Anderson, R. P. Making better Maxent models of species distributions: complexity, overfitting and evaluation. *J. Biogeogr.* **41**, 629–643 (2014).
33. ter Steege, H. *et al*. Hyperdominance in the Amazonian Tree Flora. *Science (80-.).* **342**, 1243092 (2013).
34. Kuipers, K. J. J., Hellweg, S. & Verones, F. Potential Consequences of Regional Species Loss for Global Species Richness: A Quantitative Approach for Estimating Global Extinction Probabilities. *Environ. Sci. Technol.* **53**, 4728–4738 (2019).
35. Gade, A. L., Hauschild, M. Z. & Laurent, A. Globally differentiated effect factors for characterising terrestrial acidification in life cycle impact assessment. *Sci. Total Environ.* **761**, 143280 (2021).
36. Géron, C. *et al*. Urban alien plants in temperate oceanic regions of Europe originate from warmer native ranges. *Biol. Invasions* **23**, 1765–1779 (2021).
37. Mair, L. *et al*. A metric for spatially explicit contributions to science-based species targets. *Nat. Ecol. Evol.* **5**, 836–844 (2021).
38. Bachman, S., Moat, J., Hill, A., de la Torre, J. & Scott, B. Supporting Red List threat assessments with GeoCAT: geospatial conservation assessment tool. *Zookeys* **150**, 117–126 (2011).
39. Cardoso, P. red - an R package to facilitate species red list assessments according to the IUCN criteria. *Biodivers. Data J.* **5**, e20530 (2017).
40. Lee, C. K. F., Keith, D. A., Nicholson, E. & Murray, N. J. Redlistr: tools for the IUCN Red Lists of ecosystems and threatened species in R. *Ecography (Cop.).* **42**, 1050–1055 (2019).
41. Bachman, S., Walker, B., Barrios, S., Copeland, A. & Moat, J. Rapid Least Concern: towards automating Red List assessments. *Biodivers. Data J.* **8** (2020).
42. POWO. Plants of the World Online. *Facilitated by the Royal Botanic Gardens, Kew*. http://www.plantsoftheworldonline.org/ (2021).
43. Chamberlain, S. *et al. taxize: Taxonomic information from around the web. R package version 0.9.98*. https://github.com/ropensci/taxize (2020).
44. R Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria* https://www.r-project.org/ (2021).
45. ITIS. Integrated Taxonomic Information System. https://www.itis.gov/ (2021).
46. Wickham, H. *rvest: Easily Harvest (Scrape) Web Pages. R package version 0.3.5*. https://cran.r-project.org/package=rvest (2019).
47. Desmet, P. & Page, R. WGSRPD. *GitHub repository* https://github.com/tdwg/wgsrpd (2018).
48. Chamberlain, S. *et al. rgbif: Interface to the Global Biodiversity Information Facility API. R package version 3.6.0*. https://cran.r-project.org/package=rgbif (2021).
49. GBIF. GBIF Occurrence Download. https://doi.org/10.15468/dl.uvd56q (2021).
50. Winkler, K., Fuchs, R., Rounsevell, M. & Herold, M. Global land use changes are four times greater than previously estimated. *Nat. Commun.* **12**, 2501 (2021).
51. Sippel, S., Meinshausen, N., Fischer, E. M., Székely, E. & Knutti, R. Climate change now detectable from any single day of weather at global scale. *Nat. Clim. Chang.* **10**, 35–41 (2020).
52. Hijmans, R. J. *raster: Geographic Data Analysis and Modeling. R package version 3.0-7*. https://cran.r-project.org/package=raster (2019).
53. Hernandez, P. A., Graham, C. H., Master, L. L. & Albert, D. L. The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography (Cop.).* **29**, 773–785 (2006).
54. Pearson, R. G., Raxworthy, C. J., Nakamura, M. & Townsend Peterson, A. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. *J. Biogeogr.* **34**, 102–117 (2006).
55. Phillips, S. J. & Dudík, M. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography (Cop.).* **31**, 161–175 (2008).
56. Elith, J. *et al*. A statistical explanation of MaxEnt for ecologists. *Divers. Distrib.* **17**, 43–57 (2011).
57. Anderson, R. P. & Raza, A. The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: preliminary tests with montane rodents (genus Nephelomys) in Venezuela. *J. Biogeogr.* **37**, 1378–1393 (2010).
58. Själander, M., Jahre, M., Tufte, G. & Reissmann, N. EPIC: An Energy-Efficient, High-Performance GPGPU Computing Research Infrastructure. *arXiv* 1–4 (2019).
59. Hijmans, R. J., Phillips, S., Leathwick, J. & Elith, J. *dismo: Species Distribution Modeling. R package version 1.1-4*. https://cran.r-project.org/package=dismo (2017).
60. Muscarella, R. *et al*. ENMeval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for Maxent ecological niche models. *Methods Ecol. Evol.* **5**, 1198–1205 (2014).
61. Karger, D. N. *et al*. Climatologies at high resolution for the earth's land surface areas. *Sci. Data* **4**, 170122 (2017).
62. Karger, D. N. *et al*. Data from: Climatologies at high resolution for the earth's land surface areas. *Dryad, Dataset* https://doi.org/10.5061/dryad.kd1d4 (2018).
63. ESA. Land Cover CCI Product User Guide Version 2. Tech. Rep. http://maps.elie.ucl.ac.be/CCI/viewer/download.php (2017).
64. Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A., Vilela, B. & Anderson, R. P. spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography (Cop.).* **38**, 541–545 (2015).
65. Akaike, H. Information Theory and an Extension of the Maximum Likelihood Principle. in *2nd International Symposium on Information Theory* (eds. Petrov, B. N. & Csaki, F.) 267–281 (Akademia Kiado, 1973).
66. Hurvich, C. M. & Tsai, C.-L. Regression and time series model selection in small samples. *Biometrika* **76**, 297–307 (1989).
67. Sugiura, N. Further analysts of the data by akaike' s information criterion and the finite corrections. *Commun. Stat. - Theory Methods* **7**, 13–26 (1978).
68. Morales, N. S., Fernández, I. C. & Baca-González, V. MaxEnt's parameter configuration and small samples: are we paying attention to recommendations? A systematic review. *PeerJ* **5**, e3093 (2017).
69. Shcheglovitova, M. & Anderson, R. P. Estimating optimal complexity for ecological niche models: A jackknife approach for species with small sample sizes. *Ecol. Modell.* **269**, 9–17 (2013).
70. Warren, D. L. & Seifert, S. N. Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. *Ecol. Appl.* **21**, 335–342 (2011).
71. Moran, P. A. P. Notes on Continuous Stochastic Phenomena. *Biometrika* **37**, 17 (1950).
72. Borgelt, J., Sicacha-Parada, J., Skarpaas, O. & Verones, F. Native range estimates for red-listed vascular plants. *Dryad, Dataset* https://doi.org/10.5061/dryad.qbzkh18h9 (2022).
73. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics* **21**, 3940–3941 (2005).
74. Grau, J., Grosse, I. & Keilwagen, J. PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics* **31**, 2595–2597 (2015).
75. Hosmer, D. W., Lemeshow, S. & Sturdivant, R. X. *Applied Logistic Regression. The Statistician* **45** (Wiley, 2013).
76. Lobo, J. M., Jiménez-Valverde, A. & Real, R. AUC: a misleading measure of the performance of predictive distribution models. *Glob. Ecol. Biogeogr.* **17**, 145–151 (2008).

77. Sofaer, H. R., Hoeting, J. A. & Jarnevich, C. S. The area under the precision-recall curve as a performance metric for rare binary events. *Methods Ecol. Evol.* **10**, 565–577 (2019).
78. Meyer, C., Weigelt, P. & Kreft, H. Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecol. Lett.* **19**, 992–1006 (2016).
79. Caudullo, G., Welk, E. & San-Miguel-Ayanz, J. Chorological maps for the main European woody species. *Data Br.* **12**, 662–666 (2017).
80. Rivers, M. C. Laburnum anagyroides. *The IUCN Red List of Threatened Species 2017: e.T79919483A79919650* https://doi.org/10.2305/IUCN.UK.2017-3.RLTS.T79919483A79919650.en (2017).
81. Botanic Gardens Conservation International Group & IUCN SSC Global Tree Specialist. Terminalia macrostachya. *The IUCN Red List of Threatened Species 2019: e.T150118895A150118897* https://doi.org/10.2305/IUCN.UK.2019-3.RLTS.T150118895A150118897.en (2019).
82. Heil, K., Terry, M. & Corral-Díaz, R. Mammillaria grahamii (amended version of 2013 assessment). *The IUCN Red List of Threatened Species 2017: e.T152723A121546147* https://doi.org/10.2305/IUCN.UK.2017-3.RLTS.T152723A121546147.en (2017).
83. Brooker, M. & Kleinig, D. *Field Guide to Eucalypts*. (Bloomings Books, 2006).
84. Koopman, M. M. A synopsis of the Malagasy endemic genus Megistostegium Hochr. (Hibisceae, Malvaceae). *Adansonia* **33**, 101–113 (2011).
85. World Conservation Monitoring Centre. Memecylon elegantulum. *The IUCN Red List of Threatened Species 1998: e.T32597A9713234* https://doi.org/10.2305/IUCN.UK.1998.RLTS.T32597A9713234.en (1998).
86. Landrum, L. R. A revision of the Psidium salutare complex (Myrtaceae). *SIDA, Contrib. to Bot.* **20**, 1449–1469 (2003).
87. Tropical Plants Database. Ken Fern. *tropical.theferns.info* https://tropical.theferns.info/viewtropical.php?id=Psidium+salutare (2021).
88. Bernal, R., Gradstein, S. R. & Celis, M. Siparuna conica S.S.Renner & Hausner. *Catálogo de plantas y líquenes de Colombia* http://catalogoplantasdecolombia.unal.edu.co (2015).
89. Renner, S. S. & Hausner, G. New Species of Siparuna (Monimiaceae) II. Seven New Species from Ecuador and Colombia. *Missouri Bot. Gard. Press* **6**, 103–116 (1996).
90. Melendo, M., Giménez, E., Cano, E., Mercado, F. G. & Valle, F. The endemic flora in the south of the Iberian Peninsula: taxonomic composition, biological spectrum, pollination, reproductive mode and dispersal. *Flora - Morphol. Distrib. Funct. Ecol. Plants* **198**, 260–276 (2003).
91. Chari, L. D., Martin, G. D., Steenhuisen, S.-L., Adams, L. D. & Clark, V. R. Biology of Invasive Plants 1. Pyracantha angustifolia (Franch.) C.K. Schneid. *Invasive Plant Sci. Manag.* **13**, 120–142 (2020).
92. Sasidharan, N. Amomum pterocarpum Thwaites. *India Biodiversity Portal* https://indiabiodiversity.org/species/show/258864#habitat-and-distribution (2013).
93. Contu, S. Amomum pterocarpum. *The IUCN Red List of Threatened Species 2013: e.T44393013A44450020* https://doi.org/10.2305/IUCN.UK.2013-1.RLTS.T44393013A44450020.en (2013).
94. Babyrose Devi, N., Das, A. & Singh, P. Amomum Pterocarpum (Zingiberaceae): a new record in the flora of Manipur. *Int. J. Adv. Res.* **6**, 546–549 (2018).
95. Jetz, W., Sekercioglu, C. H. & Watson, J. E. M. Ecological correlates and conservation implications of overestimating species geographic ranges. *Conserv. Biol.* **22**, 110–9 (2008).
96. Gibbs, D. & Khela, S. Magnolia pugana. *The IUCN Red List of Threatened Species 2014: e.T194806A2363344* https://doi.org/10.2305/IUCN.UK.2014-1.RLTS.T194806A2363344.en (2014).
97. Sayer, C. Vallesia glabra. *The IUCN Red List of Threatened Species 2015: e.T62543A72668627* https://doi.org/10.2305/IUCN.UK.2015-2.RLTS.T62543A72668627.en (2015).
98. Sánchez Gómez, P., Stevens, D., Fennane, M., Gardner, M. & Thomas, P. Tetraclinis articulata. *The IUCN Red List of Threatened Species 2011: e.T30318A9534227* https://doi.org/10.2305/IUCN.UK.2011-2.RLTS.T30318A9534227.en (2011).
99. Stritch, L., Roy, S., Shaw, K. & Wilson, B. Corylus cornuta (errata version published in 2017). *The IUCN Red List of Threatened Species 2016: e.T194448A115337731* https://doi.org/10.2305/IUCN.UK.2016-1.RLTS.T194448A2336319.en (2016).
100. Olson, D. M. *et al*. Terrestrial ecoregions of the world: A new map of life on Earth. *Bioscience* **51**, 933–938 (2001).
101. Rivers, M. C. Cotoneaster cambricus. *The IUCN Red List of Threatened Species 2017: e.T102827479A102827485* https://doi.org/10.2305/IUCN.UK.2017-3.RLTS.T102827479A102827485.en (2017).
102. RStudio Team. RStudio: Integrated Development Environment for R. *RStudio, PBC, Boston, MA* http://www.rstudio.com/ (2021).
103. Bivand, R., Keitt, T. & Rowlingson, B. *rgdal: Bindings for the 'Geospatial' Data Abstraction Library*. https://cran.r-project.org/package=rgdal (2019).
104. Bivand, R. & Lewin-Koh, N. *maptools: Tools for Handling Spatial Objects. R package version 0.9-5*. https://cran.r-project.org/package=maptools/ (2019).
105. Bivand, R. & Rundel, C. *rgeos: Interface to Geometry Engine - Open Source ('GEOS'). R package version 0.5-1*. https://cran.r-project.org/package=rgeos (2019).
106. Bivand, R. S., Pebesma, E. & Gómez-Rubio, V. *Applied Spatial Data Analysis with R*. (Springer New York, 2013).
107. Phillips, S. J. & Elith, J. POC plots: calibrating species distribution models with presence-only data. *Ecology* **91**, 2476–2484 (2010).
108. Hurlbert, A. H. & Jetz, W. Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. *Proc. Natl. Acad. Sci.* **104**, 13384–13389 (2007).
109. Jetz, W., McPherson, J. M. & Guralnick, R. P. Integrating biodiversity distribution knowledge: toward a global map of life. *Trends Ecol. Evol.* **27**, 151–159 (2012).

Acknowledgements

We want to thank Radek Lonka and the IndEcol Digital Lab for facilitating the use of the high-performance computing infrastructure and hosting the online application. This study is part of the Transforming Citizen Science for Biodiversity project hosted by the Digital Transformation initiative of the Norwegian University of Science and Technology.

Author contributions

J.B. was responsible for study design, methodologies, code writing, code execution, and writing the manuscript. J.S.P. contributed to methods for technical validation of the data and writing the manuscript. O.S. contributed to methodologies, interpretation of the data, and writing the manuscript. F.V. contributed to study design, interpreting the results, and writing the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Chapter 3: More than half of Data Deficient species predicted to be threatened by extinction

Communications Biology (2022) 5: 679

More than half of Data Deficient species predicted to be threatened by extinction

Jan Borgelt[1], Martin Dorber[1], Marthe Alnes Høiberg[1], Francesca Verones[1]

1. Industrial Ecology Programme, Department of Energy and Process Engineering, Norwegian University of Science and Technology (NTNU), Trondheim, Norway
corresponding author: Jan Borgelt (jan.borgelt@ntnu.no)

Abstract

The IUCN Red List of Threatened Species is essential for practical and theoretical efforts to protect biodiversity. However, species classified as "Data Deficient" (DD) regularly mislead practitioners due to their uncertain extinction risk. Here we present machine learning-derived probabilities of being threatened by extinction for 7,699 DD species, comprising 17% of the entire IUCN spatial datasets. Our predictions suggest that DD species as a group may in fact be more threatened than data-sufficient species. We found that 85% of DD amphibians are likely to be threatened by extinction, as well as more than half of DD species in many other taxonomic groups, such as mammals and reptiles. Consequently, our predictions indicate that, amongst others, the conservation relevance of biodiversity hotspots in South America may be boosted by up to 20% if DD species were acknowledged. The predicted probabilities for DD species are highly variable across taxa and regions, implying current Red List-derived indices and priorities may be biased.

Introduction

Measuring ongoing and anticipating potential threats is vital for preventing damage to the natural world[1–8], which entails detailed knowledge about the current state of biodiversity. A central data resource enabling a multitude of overarching analyses in conservation and sustainability science[9] is the International Union for the Conservation of Nature (IUCN)'s Red List of Threatened Species (hereafter: Red List). The Red List assesses extinction risks and reports Red List categorization for more than 140,000 species based on a set of quantitative criteria[10] relying for instance on extent of occurrence, area of occupancy, population trends, or population size. However, the sheer amount of known and unknown species globally[11,12], the dynamic nature of threats and trends[7], and limited human resources for undertaking such Red List assessments[13,14] turn this critical endeavour into a Sisyphean task.

Consequently, only a small proportion of known species have been assessed for their conservation priority so far[15,16], unevenly distributed across space, time and taxa[13,16]. In addition, numerous assessed species are classified as Data Deficient (DD) even in otherwise comprehensively assessed species groups. A species is considered DD if there is "inadequate information to make a direct, or indirect, assessment of its risk of extinction based on its distribution and/or population status"[17]. More specifically Bland et al.[18] identified 8 main justifications as to why species are assessed as DD: uncertain provenance, type series, few records (< 5), old records (before 1970), uncertain population status or distribution, uncertain threats, new species (discovered in the last 10 years), and taxonomic uncertainty. In parallel, Butchart and Bird[19] stated that the DD category "is probably the most controversial and misunderstood Red List category". One of the main reasons are value choices when dealing with uncertainty and applying the IUCN Guidelines. If, due to uncertain data, a species can be

listed as Critically Endangered (CR) and Least Concern (LC), the species should be listed as DD. However, if the assessor considers a species being not LC but is unsure about its exact threat-level, DD is not the appropriate category. In this case, the assessor needs to decide and assign the species to a category, i.e., risk tolerance. It is important to note that we do not distinguish the DD species according to the reason for their classification as DD.

On average across all taxa and regions, one of six assessed species is classified as DD[15,18,20] (Supplementary Table 1). Although DD species are sometimes treated as being not threatened[21], studies suggest that they are of particular conservation importance because a higher portion of them may be threatened by extinction compared to data-sufficient (DS) species[22–24]. However, since DD species could belong to any Red List category, they are difficult to handle for practitioners[21,25] and are therefore generally ignored in studies analysing biodiversity impacts and change[26,27]. For instance, the Red List Index[27] is built upon well-assessed threat-levels for individual species at several points in time and directly applied in, e.g., sustainable development goals[28] and biodiversity targets[29]. In addition, studies linking biodiversity loss to global trade footprints[30,31] and approaches to transform threat-levels to numerical conservation indicators[32] have ignored DD species. Similarly, the recently suggested metric[26] for measuring success of the post-2020 Global Biodiversity Framework will not be applicable for DD species.

In stark contrast, the continuous growth in knowledge turnover during the digital era has resulted in constant improvement in the availability of global data on biodiversity, human activities, and environmental threats[33]. Statistical tools, such as machine learning (ML), can detect relevant signals in large datasets, thereby offering a time- and cost-effective approach to tackle data deficiency[34–37]. The utility of ML models for predicting species' extinction risk or conservation status was successfully proven for species in single taxonomic groups with great accuracy[24,38–44], regionally as well as globally. However, such predictions are needed consistently for all relevant species to effectively benefit global conservation and sustainability analyses[16].

Here, we present a global multitaxon ML classifier that predicts the probability of being threatened by extinction (hereafter: PE score) based on, amongst others, species taxonomy, range extent, and summarized stressors (min., max., mean and median) within species range maps, as well as species occurrence cells (0.5-degree cells). The classifier was trained and tested on threat levels for 28,363 DS species, drawing on selected features out of more than 400 predictors, human pressures, and environmental stressors (full list in Supplementary Table 2). We applied the classifier to predict PE scores for DD species (n = 7,699) that include range maps of their distribution in their IUCN Red List database record (Version 2020-3)[45,46], to our knowledge the largest data provider of range maps for thousands of species. Since biodiversity varies greatly through space, it is crucial to perform assessments in a spatially explicit way and include their entire spatial extent.

Results & Discussion

Classifier performance

The trained classifier was able to successfully separate between threatened and non-threatened species within a set-aside testing dataset, as well as continuous predictions (i.e., PE scores) (Figure 1). The binary classifier obtained an overall accuracy of 85% (Table 1), being more precise in predicting which species are not threatened by extinction than in predicting which species are threatened. 93% and 92% of species that we predicted to be not threatened were indeed not threatened (for marine and non-marine species respectively). Hence, with only 7-8% of negative predictions (i.e., predicted as not threatened) being incorrect, we are confident that our binary classifier avoids underestimating the conservation status of most taxa. Instead, the binary classifier may be prone to overestimating the status of some taxa; only 60% to 67% of species that we predicted to be threatened are also classified as threatened by the IUCN (for marine and non-marine species respectively). The continuous classifier, however, seems to only underestimate the risk for marine species when directly compared to non-marine species. The relative ranking of continuous predictions within the groups remains valid for all species (AUC = 0.91, AUC_{PR} = 0.80, Gini-Coefficient = 0.82) and across taxonomic classes (Supplementary Table 1). Hence, on average, species being threatened by extinction obtain higher predicted PE scores than not threatened species, for both marine and non-marine species (Figure 1). Binary as well as continuous predictions across marine versus non-marine groups perform well but are not directly comparable.

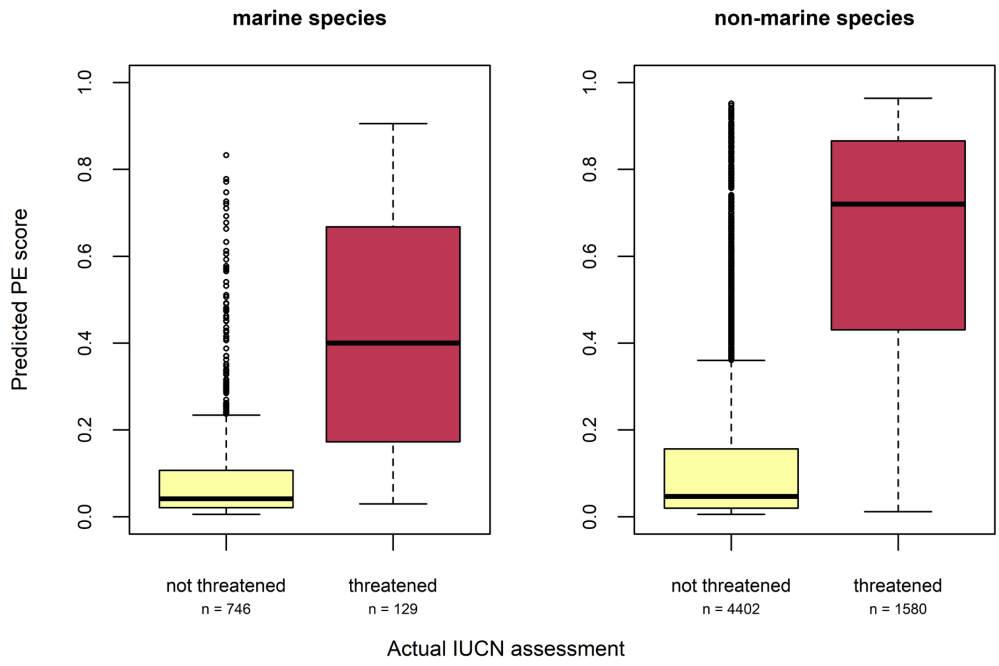


Figure 1: Predicted probability of being threatened by extinction (PE score) across the actual IUCN assessment (not threatened and threatened) for marine (n = 875) and non-marine (n = 5,982) species in the set-aside testing data.

We further tested our classifier against an IUCN update (Version 2021-2)[15] that was released after our model was trained (Supplementary Figure 1). In this update, we found that 123 former DD species from Version 2020-3 were now assigned a threat-level. Our classifier labelled 94 of those species (76%) correctly (Table 1), being equally precise in predicting whether the species was threatened (76%) or not threatened (77%) but more accurate for non-marine (80%) than for marine species (74%).

Table 1: Confusion matrix and resulting performance measures for marine and non-marine species based on the set-aside testing data (25% of the dataset) and based on formerly Data Deficient species (n = 123) in IUCN version 2021-2[15] (in brackets).

	Reference			
Predicted	not threatened	threatened	not threatened	threatened
not threatened	695 (26)	54 (10)	3786 (20)	309 (4)
threatened	51 (8)	75 (25)	616 (7)	1271 (23)
	marine species		**non-marine species**	
Accuracy	0.88 (0.74)		0.85 (0.80)	
Specificity	0.93 (0.76)		0.86 (0.74)	
Sensitivity	0.58 (0.71)		0.80 (0.85)	
Negative Pred. Value	0.93 (0.72)		0.92 (0.83)	
Positive Pred. Value	0.60 (0.76)		0.67 (0.77)	
Balanced Accuracy	0.76 (0.74)		0.83 (0.80)	

Data Deficient species are more threatened by extinction than data-sufficient species

On average we obtained higher PE scores for DD species (43%) than for DS species (26%), resulting in 56% of DD species (n = 4,336) predicted to be threatened by extinction (Supplementary Table 1) versus 28% of DS species[46]. The generated predictions reinforce the concern that DD species are of high conservation interest[21,25] and, given the large variance in predicted probabilities of being threatened (Supplementary Figure 2), highlight the importance of treating DD species individually.

On land, these likely threatened DD species are scattered across all continents and are often geographically restricted to smaller ranges (Figure 2, B; Supplementary Figure 3), such as in central Africa, Madagascar and southern Asia. The greatest number of threatened marine DD species are found in south-eastern Asia, followed by the eastern Atlantic coastline as well as numerous atolls and islands (Supplementary Figure 4). In fact, between a third and half of marine DD species around the world's coastlines were predicted to be threatened by extinction, most notably along the eastern Atlantic coastline including the Mediterranean basin (Figure 2, A; Supplementary Figure 3).

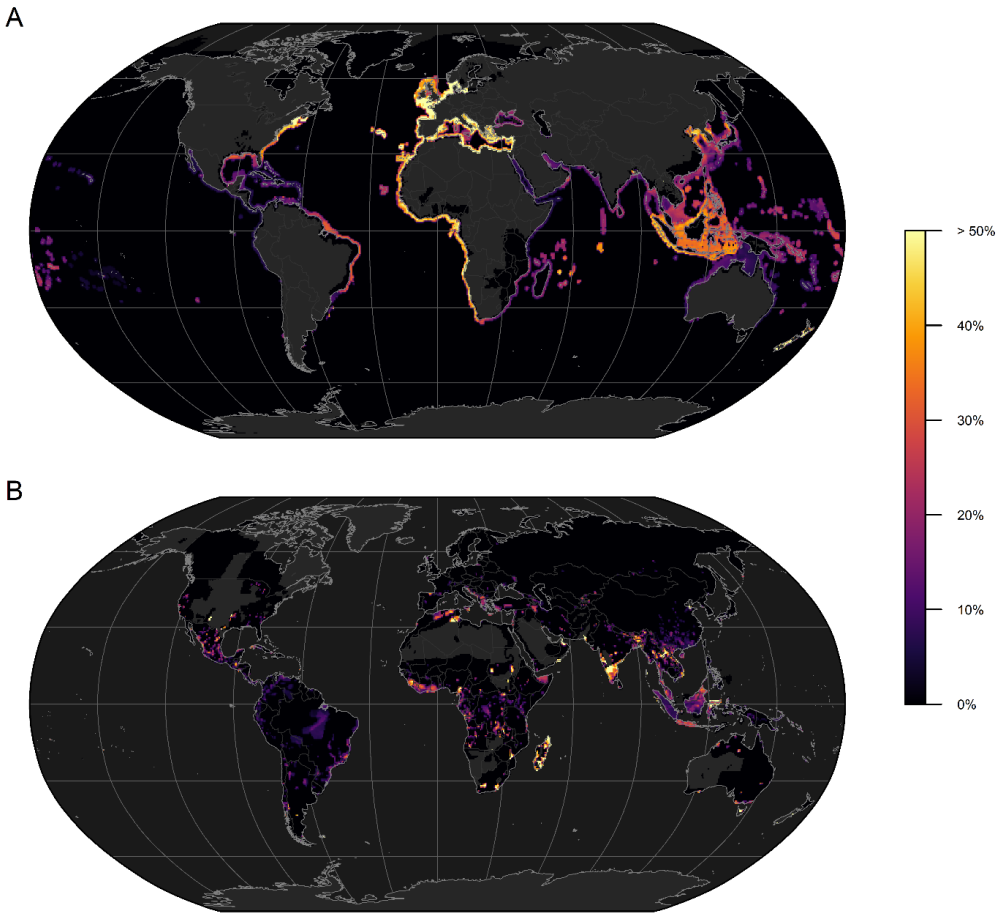


Figure 2: Fraction of Data Deficient species (IUCN Version 2020-3) predicted to be threatened by extinction for marine (A) and non-marine species (B) according to our machine learning classifier.

In addition to roughly 40% of Data Deficient ray-finned fishes (*Actinopterygii*), malacostracans (*Malacostraca*), bivalves, snails and slugs (*Gastropoda*), we found a staggering 960 out of 1130 (85%) Data Deficient amphibians (*Amphibia*), and more than half of Data Deficient anthozoans (*Anthozoa*; marine invertebrates including anemones and corals), insects (*Insecta*), mammals (*Mammalia*) and reptiles (*Reptilia*) likely to be threatened by extinction (Supplementary Table 1).

This is highly relevant for conservation and sustainability analyses, as some of these groups are amongst the most frequently considered ones[7]. More specifically, the classification of DD amphibians, mammals, and reptiles is likely to further increase both the absolute and relative number of species threatened by extinction in these taxonomic groups. For instance, an additional 14% of amphibians were predicted to be threatened by our ML classifier. This would raise the relative number of amphibian species being threatened by extinction from 39% to 47%. Similarly, the fraction of threatened mammals and reptiles likely increases when

accounting for DD species (from 26% to 31% and 19% to 25%, respectively; Supplementary Table 1).

For selected species groups, models that suggest Red List categories or probabilities of being threatened for DD species exist, e.g., for amphibians[24], reptiles[38], terrestrial mammals[39] or sharks and rays[43]. Howard and Bickford[24] found 63% of DD amphibians to be threatened, mostly in South America, central Africa and North Asia, but also state that this is an underestimation. Our model predicts 85% of DD amphibians to be threatened. Bland and Böhm[38] identified 19% out of 292 DD terrestrial reptile species as threatened, while our model identified 59% of reptiles as threatened, but we include over 1000 species and terrestrial, freshwater and marine species, the latter of which are thought to be more likely to be threatened[47]. The regions for conservation priorities for both reptiles and amphibians match those previously found, which are congruent with known hotspots for threatened species[38]. A previous assessment for terrestrial mammals[39] identified 64% of DD terrestrial mammals as threatened, while our model classifies 61% of DD terrestrial and marine mammals as threatened. Sharks and rays in the Mediterranean and North East Atlantic were modelled to contain 62% and 55% threatened species, respectively[44]. On a global scale, we found 26% of DD species in this group to be threatened (Supplementary Table 1). This is concordant with Dulvy et al.[48], which found every fourth species of the ray and shark family to be threatened with extinction and who found the Mediterranean to be a hotspot for extinction, explaining the large discrepancy of the local values to our global one.

Data-deficiency causes regionally biased conservation priorities

The high variance found in the predicted probabilities of being threatened by extinction (i.e., PE scores) at the species level implies that more accurate assessments of DD species could shift regional conservation priorities. We predicted higher PE scores for DD than for DS species in most regions of the world (Supplementary Figure 5), suggesting that current conservation concerns could, in fact, be underestimated. In marine systems, however, this seems to be restricted to coastal waters as well as high latitudes.

DD species in marine systems seem to be most relevant around the world's coastlines, as well as around temperate to tropical islands and atolls, but less relevant in international waters (Figure 3, A). For instance, we found an increase in average PE score by more than 20% once DD were considered alongside DS species in e.g., the Gulf of Mexico, the Caribbean and south America's Atlantic coast (Figure 3, A). Even in biodiversity-rich regions the average PE score increased another 10% to 15% due to the extant DD species, such as in the Gulf of Guinea and South-eastern Asian seas. Here, numerous DD reef forming corals, sharks, rays, chimaeras, and marine fish species seem to be particularly relevant for a timely and expert-based threat assessment (Supplementary Figure 3; Supplementary Figure 6). In contrast, including DD species did not change or even lowered the average PE score in large parts of international seas (Figure 3, A). Although marine biodiversity as we know it today is richest in coastal waters[49], these results should be interpreted with caution because the underlying range maps for many marine species can be too coarse[50], which may be especially true for DD species in international seas.

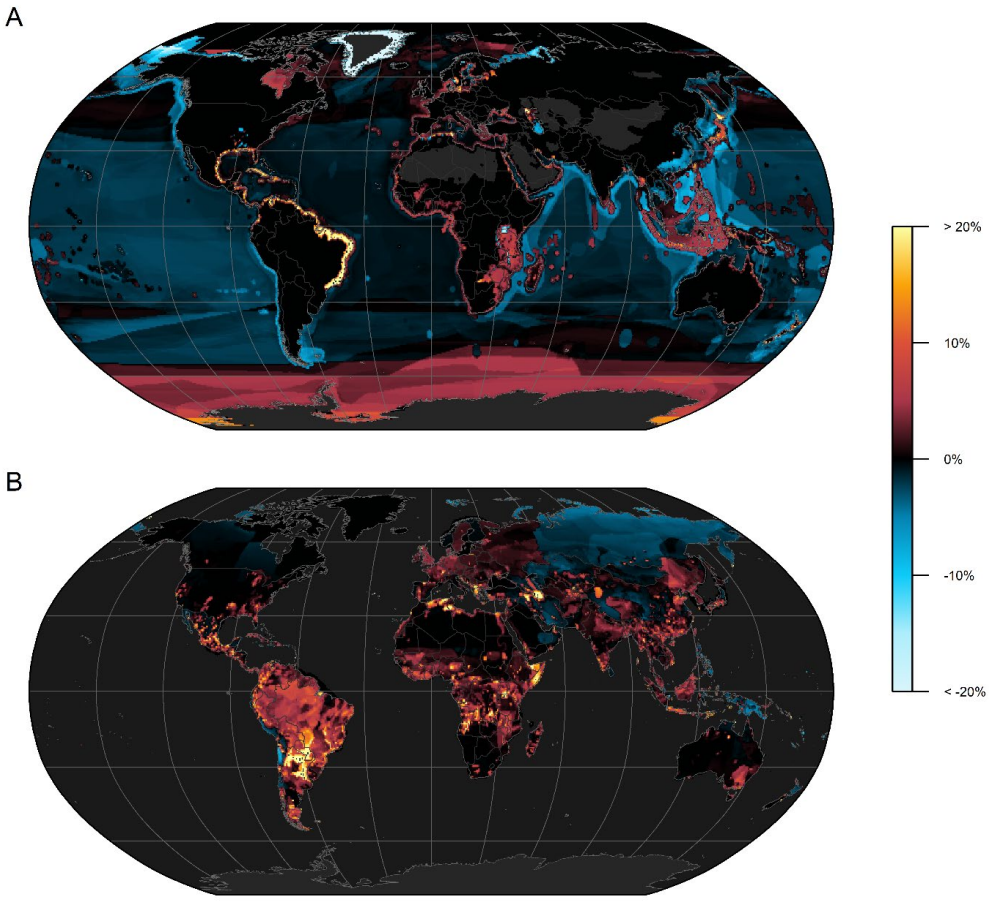


Figure 3: Percent change in average PE score (i.e., predicted probability of being threatened by extinction) for marine (A) and non-marine species (B) following the inclusion of Data Deficient species alongside data-sufficient species.

Furthermore, DD species on land (i.e., strictly non-marine species) seem to have the potential to regionally boost the conservation relevance in most of the world's megadiverse countries[51]. Across Central to South America, we found a widespread increase of 10% to 20% in average PE score when including DD in addition to DS species (Figure 3, B). Notably, often only few taxonomic groups accounted for most of the observed increase in average PE score (Supplementary Figure 6). For instance, the addition of predicted scores for DD amphibians, reptiles, mammals, rays and other freshwater groups in large parts of South America resulted in a widespread increase in average PE score, including for example the Amazon basin, the tropical Andes, the Atlantic Forest and Cerrado. However, these estimates are based on limited taxonomic groups and may be different if spatially explicit range maps for more taxa were available (e.g., plants).

In Africa, DD amphibians, reptiles, mammals, and freshwater ray-finned fishes (*Actinopterygii*) increased the average PE score locally across freshwater systems (e.g., Lake Victoria), tropical rainforests and savannas throughout the continent (Figure 3, B; Supplementary Figure 6). We further discovered an increase in average PE score in numerous

smaller isolated patches distributed around the world once DD extant species' scores were acknowledged, such as in the Northern Territory and the Murray–Darling basin of Australia. Overall, the potential effects on PE score due to DD species were much more restricted to a regional level on land compared to marine systems, presumably due to spatially more explicit, and restricted, range maps for DD species on land.

Conclusion

Previously, the risk of misjudging the importance of individual DD species outweighed the benefits of including them in Red List applications, resulting in regionally biased conservation prioritization. This study suggests that automatized classifiers built on species' range maps and species observations can provide accurate and rapid pre-assessments on a large, global, and multitaxon scale. In contrast to previous approaches, our classifier is able to provide standardized predictions across multiple taxonomic groups[16], making results between taxa directly comparable. The presented results show that DD species vary greatly in probability of being threatened by extinction, indicating a highly heterogenous bias that propagates into consequential Red List applications. As such, inferences built upon Red List-derived numbers of threatened species[30] as well as numerically converted threat-levels[32] could be biased. The generated predictions (i.e., PE scores) could facilitate the inclusion of DD species in sustainability-relevant applications[27] and modelling approaches[26]. We encourage the extended use of our algorithm for screening for updates[14] in the status of DS species, as well as large-scale pre-assessments of species not yet evaluated by the IUCN[42] for a targeted completion of the IUCN Red List of Threatened Species.

Methods

Species data

We retrieved all spatial range map datasets (i.e., mammals, amphibians, reptiles, fish, marine groups, selected vascular plant groups and freshwater groups) available at the IUCN Red List (https://www.iucnredlist.org/resources/spatial-data-download, Version 2020-3)[45,46] in March 2021, covering 44,924 species in the following taxonomic classes: *Actinopterygii*, *Agaricomycetes*, *Amphibia*, *Anthozoa*, *Aves*, *Bivalvia*, *Branchiopoda*, *Bryopsida*, *Cephalaspidomorphi*, *Charophyaceae*, *Chondrichthyes*, *Clitellata*, *Gastropoda*, *Hydrozoa*, *Insecta*, *Jungermanniopsida*, *Lecanoromycetes*, *Liliopsida*, *Lycopodiopsida*, *Magnoliopsida*, *Malacostraca*, *Mammalia*, *Myxini*, *Polypodiopsida*, *Reptilia* and *Sarcopterygii*. Range maps for bird species were not downloaded separately, because of their limited number of DD species. Species taxonomy, native countries, environmental domain (i.e., the occurrence in terrestrial, freshwater, marine systems and combinations thereof) and Red List category were available from IUCN for all species, i.e., Least Concern (LC), Lower Risk/Least Concern (LR/LC), Lower Risk/Conservation Dependent (LR/CD), Near Threatened (NT), Vulnerable (VU), Endangered (EN), Critically Endangered (CR), Extinct (EX), Extinct in the Wild (EW) and Data Deficient (DD). The spatial dataset consists of seasonal range maps (i.e., for each species one or several range maps out of "resident", "breeding season", "non-breading season", "passage", and "seasonal occurrence uncertain" were available). Only those range maps labelled as "native" and "extant" and only species that were not categorized as EW or EX were considered (n = 44,908 species).

Predictor data

The correlate variables are summarized in Supplementary Table 2. Species taxonomy (i.e., taxonomic kingdom, phylum, and class) was included as potential predictor and surrogate for phylogenetic data[42]. Habitat preferences were retrieved from the Red List using *rredlist*[52] in R. Occupied types of habitats as well as the number of different types of habitats, sub-habitats, and habitats of major importance were included as predictor. Occurrence data was retrieved from the Global Biodiversity Information Facility (GBIF)[53] and the Ocean Biodiversity Information System (OBIS)[54] using their corresponding application programming interfaces via the packages *rgbif*[55] and *robis*[56] in R. We only considered occurrence data that were collected between the years 2010 and 2020. For each species, we retrieved the maximum number of occurrence points per native country from GBIF (i.e., 100,000 data points per request), and for marine species, we additionally downloaded all data available from OBIS. The total number of occurrence points as well as the number of occurrence cells in a global grid (0.5-degree cells) was counted.

Because environmental threats can vary considerably across space and we expect the species to be exposed heterogeneously within their ranges, we extracted mean, minimum, maximum, and median values of environmental stressors and features across each species' seasonal range map as well as its occurrence cells.

The included features were representative for the major drivers of biodiversity change, i.e. climate change, habitat change, overexploitation, invasive species and pollution[57]. As climatic dataset we retrieved all CHELSA bioclimatic variables[58,59]. The European Space Agency's land cover product for the year 2015 in 300 m resolution[60] was used to calculate fractions for different natural land cover types (n = 17). One raster was calculated per land cover class, representing the proportion of land covered by that class per cell. As general indicators of anthropogenic land use and land use change we included the global human footprint index[61], including associated stressors such as population density, cropland area and pasture area, human modification index[62], future urban expansion probabilities[63], fraction of land designated to protected areas[64], deforestation rates between the years 2000 and 2019[65], different habitat heterogeneity metrics[66] and cumulative application rates of different pesticides[67]. We counted the number of power plants[68] and dams[69] within each species geographical range, and included country-specific water scarcity estimates[70], annual streamflow[71], stream connectivity indices[72] as well as freshwater environmental variables[73], including eutrophication, pollution and upstream land use fractions, to account for the most severe impacts in freshwater systems[74,75]. Illegal hunting activities remain problematic for many species[76]. Yet, to the best of our knowledge, global poaching data does not exist. Therefore, we included factors that may affect the rate of poaching on a global scale[77,78], i.e., the human development index (HDI) in 2019, the average annual HDI growth between 1990-2019[79] and the corruption perceptions index (CPI) in 2020 at country-level[80]. We further included estimated threats from species invasions, country-specific capacities to respond to invasion[81], a set of modelled impacts on marine ecosystems[82,83] and marine environmental variables[84,85]. All layers were aggregated for computational efficiency by averaging to 0.5-degree cells (approximately 56 km at the equator).

Machine learning classifier

We aimed to estimate the probability of being threatened by extinction (hereafter: PE score) for DD species by training a machine learning classifier, fitted using species with known threat-levels. All DS species were reclassified into two groups based on their IUCN Red List categories: threatened by extinction (i.e., all species in the categories VU, EN, and CR) and not threatened by extinction (i.e., all species in the categories LC, LR/LC, LR/CD and NT). Species classified as DD (n = 7,699) were set aside and not used for training or testing the classifier. All assessments identified by the IUCN as in need of an update were removed[16], with one exception: if fewer than five records remained for a given taxonomic class, outdated assessment were kept to maximize the amount of training data. We used a data split for model validation[16,39,86,87]. Therefore, the remaining dataset (n = 28,363 species) was split into training (75%) and testing (25%) data. During the data split the balance of threat categories were maintained within both taxonomic families and environmental domains, i.e., marine and non-marine.

We used different partitions of the dataset to train ML classifiers in two ways: 1) all species together, and 2) separate classifiers for marine and non-marine species to account for the different spatial extents of the predictor data. For each data partition, we utilized a set of machine learning methods suitable for classification problems, each with its own strengths and weaknesses[88]. The best performing data partition (i.e., partition 1; for all species) was selected based on the highest average AUC (see section *Model evaluation*) across all taxonomic groups. Although irrelevant covariates tend to be automatically ignored in the utilized algorithms[89–92], a smaller set of covariates can improve performance and increase interpretability of the model. Therefore, we performed feature selection on the training data of each partition by using the Boruta algorithm[93]. This algorithm compares the original feature importance to the importance of random shadow features while accounting for possible correlations and interactions. All features considered relevant at the 99% confidence level after 50 runs of the algorithm were kept (i.e., 270 features in partition 1). NA-values were imputed with random values using the package *Hmisc*[94] in R, i.e., about 5% of the values in the remaining features. Optimal model settings and parameters were selected using the AutoML function in H2O.ai[89,90]. We used 10-fold cross validation for calibrating all models (e.g., tuning hyperparameters). In addition, the two classes (i.e., threatened versus not threatened species) were balanced during cross validation by oversampling of the smaller class (i.e., threatened species). In partition 1, a total of 220 models (i.e., base-learners) was trained, including generalized linear models, random forests, gradient boosted classification trees, deep neural networks and an extremely randomized forest (details in reference[90]). Ultimately, a so-called super-learner[95] was generated using a non-negative generalized linear model with regularization (least absolute shrinkage and selection operator) to produce more sparse ensembles[90], combining the best features of the trained base-learners into one superior model. In total, 23 base-learners contributed to the predictions of the super-learner (Supplementary Table 3).

Model evaluation

The performance of all base-learners and the super-learner of the best performing data partition (i.e., partition 1; trained using all species) was assessed using the set aside testing data (n =

6,857 species). In addition, we assessed model performance using DD species that have been re-evaluated and assigned a threat category in Red List Version 2021-2 (n = 123 species)[15].

We calculated accuracy as the fraction of correctly classified species across total number of species (equation 1), specificity as the fraction of not threatened species being correctly classified as not threatened (equation 2), sensitivity (i.e., recall) as the fraction of threatened species being correctly classified as threatened (equation 3), the false positive rate as fraction of not threatened species being classified as threatened (equation 4), the negative predictive value as the fraction of not threatened species across species predicted to be not threatened (equation 5), the positive predictive value (i.e., precision) as the fraction of threatened species across species predicted to be threatened (equation 6) and, balanced accuracy as the average of specificity and sensitivity.

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Negative + True\ Negative + False\ Positive} \quad \text{(equation 1)}$$

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive} \quad \text{(equation 2)}$$

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad \text{(equation 3)}$$

$$False\ positive\ rate = \frac{False\ Positive}{False\ Positive + True\ Negative} \quad \text{(equation 4)}$$

$$Negative\ predictive\ value = \frac{True\ Negative}{True\ Negative + False\ Negative} \quad \text{(equation 5)}$$

$$Positive\ predictive\ value = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad \text{(equation 6)}$$

In addition, AUC, AUC_{PR} and GINI coefficient were calculated[89,90] as threshold-independent performance measures for binary classifiers. A value of 1 depicts the highest performance for all metrics. AUC is the area under the receiver operating characteristic curve for sensitivity (equation 3) versus false positive rate (equation 4). This measure is influenced by correctly assigned species as being not threatened (True Negatives), which is the dominating class in our dataset. In contrast, AUC_{PR}, as the area under the receiver operating characteristic curve for precision (equation 6) versus recall (equation 3), is not affected by true negatives (i.e., correctly predicted not-threatened species) but instead affected by how precise the classifier is in predicting which species are threatened. The GINI coefficient describes the degree of separation between both classes (i.e., threatened versus not threatened), with a value of 1 indicating perfect separation.

Permutation variable importance was calculated as the performance loss (i.e., in AUC) on the testing data before and after a feature was permuted. Features were permuted one at a time in a total of 50 repetitions. In partition 1, the species' taxonomic affiliation, proxies for geographic range size (i.e., number of native countries, species range extent and number of occurrence cells), anthropogenic activities within the species' range (number of dams, road density, number of powerplants, human footprint index), and occupied environmental domains

(combinations of terrestrial, freshwater and marine) are most important for the super-learner in accurately separating not threatened and threatened species (Supplementary Figure 7).

Data handling

All data handling was done using R version 4.0.3[96] in RStudio version 1.4.1103[97]. Data were obtained from GBIF, OBIS and IUCN using the packages *rgbif, robis, and rredlist*[52,55,56]. Handling of spatial and other data was conducted using the R packages *caTools*, *doParallel*, *exactextractr*, *fasterize*, *maptools*, *parallel*, *raster*, *readxl*, *rgdal*, *rgeos*, *sf*, *sp*, *stringr*, *tidyverse*, and *xlsx*[96,98–110], and in python using the arcpy module from ArcGIS Pro version 2.9.0[111]. Machine learning algorithms were trained and evaluated using the H2O.ai interface (Version 3.36.0.4) for R[89] and *caret*[112]. Figures were created using *ggplot*[113], *ggridges*[114], *rnaturalearth*[115], *viridis*[116] and base R[96].

Data availability

Previously published and open-access data were retrieved from refs.[45,46,64–73,53,79–84,54,58–63]. All data generated in this study is available without restrictions. Any further requests can be directed to the corresponding author.

Code availability

All code generated in this study is available without restrictions. R code for preparing the data, for training and testing the ML classifier, as well as applying the algorithm is available on GitHub (https://github.com/jannebor/dd_forecast). The classifier can be applied for single species using our web application (https://ml-extinctionrisk.indecol.no/). Any further requests can be directed to the corresponding author.

References

1. Cardillo, M. & Meijaard, E. Are comparative studies of extinction risk useful for conservation? *Trends Ecol. Evol.* **27**, 167–171 (2012).

2. Mace, G. M., Norris, K. & Fitter, A. H. Biodiversity and ecosystem services: a multilayered relationship. *Trends Ecol. Evol.* **27**, 19–26 (2012).

3. Steffen, W., Broadgate, W., Deutsch, L., Gaffney, O. & Ludwig, C. The trajectory of the Anthropocene: The Great Acceleration. *Anthr. Rev.* **2**, 81–98 (2015).

4. Díaz, S. *et al.* Pervasive human-driven decline of life on Earth points to the need for transformative change. *Science (80-.).* **366**, eaax3100 (2019).

5. Newbold, T. *et al.* Has land use pushed terrestrial biodiversity beyond the planetary boundary? A global assessment. *Science (80-.).* **353**, 288–291 (2016).

6. Pimm, S. L. *et al.* The biodiversity of species and their rates of extinction, distribution, and protection. *Science (80-.).* **344**, 1246752–1246752 (2014).

7. IPBES. *Summary for policymakers of the global assessment report on biodiversity and ecosystem services*. *Zenodo* (2019) doi:10.5281/zenodo.3831674.

8. Barnosky, A. D. *et al.* Has the Earth's sixth mass extinction already arrived? *Nature* **471**, 51–57 (2011).

9. Rodrigues, A., Pilgrim, J., Lamoreux, J., Hoffmann, M. & Brooks, T. The value of the IUCN Red List for conservation. *Trends Ecol. Evol.* **21**, 71–76 (2006).

10. MACE, G. M. *et al.* Quantification of Extinction Risk: IUCN's System for Classifying Threatened Species. *Conserv. Biol.* **22**, 1424–1442 (2008).

11. Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. B. & Worm, B. How Many Species Are There on Earth and in the Ocean? *PLoS Biol.* **9**, e1001127 (2011).

12. Purvis, A. & Hector, A. Getting the measure of biodiversity. *Nature* **405**, 212–219 (2000).

13. Bachman, S. P. *et al.* Progress, challenges and opportunities for Red Listing. *Biol. Conserv.* **234**, 45–55 (2019).

14. Rondinini, C., Di Marco, M., Visconti, P., Butchart, S. H. M. & Boitani, L. Update or Outdate: Long-Term Viability of the IUCN Red List. *Conserv. Lett.* **7**, 126–130 (2014).

15. IUCN. The IUCN Red List of Threatened Species. Version 2021-2. https://www.iucnredlist.org (2021).

16. Cazalis, V. *et al.* Bridging the research-implementation gap in IUCN Red List assessments. *Trends Ecol. Evol.* **37**, 359–370 (2022).

17. IUCN Standards and Petitions Committee. *Guidelines for using the IUCN Red List Categories and Criteria*. *Prepared by the Standards and Petitions Committee. Downloadable from https://www.iucnredlist.org/documents/RedListGuidelines.pdf* vol. 15 (2022).

18. Bland, L. M. *et al.* Toward reassessing data-deficient species. *Conserv. Biol.* **31**, 531–539 (2017).

19. Butchart, S. H. M. & Bird, J. P. Data Deficient birds on the IUCN Red List: What don't we know and why does it matter? *Biol. Conserv.* **143**, 239–247 (2010).

20. Zhao, L. *et al.* Spatial knowledge deficiencies drive taxonomic and geographic selectivity in data deficiency. *Biol. Conserv.* **231**, 174–180 (2019).

21. Parsons, E. C. M. Why IUCN Should Replace "Data Deficient" Conservation Status with a Precautionary "Assume Threatened" Status—A Cetacean Case Study. *Front. Mar. Sci.* **3**, 2015–2017 (2016).

22. Roberts, D. L., Taylor, L. & Joppa, L. N. Threatened or Data Deficient: assessing the conservation status of poorly known species. *Divers. Distrib.* **22**, 558–565 (2016).

23. Jetz, W. & Freckleton, R. P. Towards a general framework for predicting threat status of data-deficient species from phylogenetic, spatial and environmental information. *Philos. Trans. R. Soc. B Biol. Sci.* **370**, 20140016 (2015).

24. Howard, S. D. & Bickford, D. P. Amphibians over the edge: silent extinction risk of Data Deficient species. *Divers. Distrib.* **20**, 837–846 (2014).

25. Jarić, I., Courchamp, F., Gessner, J. & Roberts, D. L. Potentially threatened: a Data Deficient flag for conservation management. *Biodivers. Conserv.* **25**, 1995–2000 (2016).

26. Mair, L. *et al.* A metric for spatially explicit contributions to science-based species targets. *Nat. Ecol. Evol.* **5**, 836–844 (2021).

27. Butchart, S. H. M. *et al.* Measuring Global Trends in the Status of Biodiversity: Red List Indices for Birds. *PLoS Biol.* **2**, e383 (2004).

28. United Nations. *Transforming our World: the 2030 Agenda for Sustainable Development*. A/RES/70/1 (United Nations, 2015).

29. Butchart, S. H. M. *et al.* Using Red List Indices to measure progress towards the 2010 target and beyond. *Philos. Trans. R. Soc. B Biol. Sci.* **360**, 255–268 (2005).

30. Lenzen, M. *et al.* International trade drives biodiversity threats in developing nations. *Nature* **486**, 109–112 (2012).

31. Moran, D. & Kanemoto, K. Identifying species threat hotspots from global supply chains. *Nat. Ecol. Evol.* **1**, 0023 (2017).

32. Mooers, A. Ø., Faith, D. P. & Maddison, W. P. Converting Endangered Species Categories to Probabilities of Extinction for Phylogenetic Conservation Prioritization. *PLoS One* **3**, e3700 (2008).

33. Runting, R. K., Phinn, S., Xie, Z., Venter, O. & Watson, J. E. M. Opportunities for big data in conservation and sustainability. *Nat. Commun.* **11**, 2003 (2020).

34. Hochkirch, A. *et al.* A strategy for the next decade to address data deficiency in neglected biodiversity. *Conserv. Biol.* **35**, 502–509 (2021).

35. Hino, M., Benami, E. & Brooks, N. Machine learning for environmental monitoring. *Nat. Sustain.* **1**, 583–588 (2018).

36. Wearn, O. R., Freeman, R. & Jacoby, D. M. P. Responsible AI for conservation. *Nat. Mach. Intell.* **1**, 72–73 (2019).

37. Bland, L. M. *et al.* Cost-effective assessment of extinction risk with limited information. *J. Appl. Ecol.* **52**, 861–870 (2015).

38. Bland, L. M. & Böhm, M. Overcoming data deficiency in reptiles. *Biol. Conserv.* **204**, 16–22 (2016).

39. Bland, L. M., Collen, B., Orme, C. D. L. & Bielby, J. Predicting the conservation status of data-deficient species. *Conserv. Biol.* **29**, 250–259 (2015).

40. Luiz, O. J., Woods, R. M., Madin, E. M. P. & Madin, J. S. Predicting IUCN Extinction Risk Categories for the World's Data Deficient Groupers (Teleostei: Epinephelidae). *Conserv. Lett.* **9**, 342–350 (2016).

41. Stévart, T. *et al.* A third of the tropical African flora is potentially threatened with extinction. *Sci. Adv.* **5**, eaax9444 (2019).

42. Darrah, S. E., Bland, L. M., Bachman, S. P., Clubbe, C. P. & Trias-Blasi, A. Using coarse-scale species distribution data to predict extinction risk in plants. *Divers. Distrib.* **23**, 435–447 (2017).

43. Walls, R. H. L. & Dulvy, N. K. Tracking the rising extinction risk of sharks and rays in the Northeast Atlantic Ocean and Mediterranean Sea. *Sci. Rep.* **11**, 15397 (2021).

44. Walls, R. H. L. & Dulvy, N. K. Eliminating the dark matter of data deficiency by predicting the conservation status of Northeast Atlantic and Mediterranean Sea sharks and rays. *Biol. Conserv.* **246**, 108459 (2020).

45. IUCN. Species Information Service. Version 2020-3. https://www.iucnredlist.org/resources/spatial-data-download (2021).

46. IUCN. The IUCN Red List of Threatened Species. Version 2020-3. https://www.iucnredlist.org (2020).

47. Böhm, M. *et al.* The conservation status of the world's reptiles. *Biol. Conserv.* **157**, 372–385 (2013).

48. Dulvy, N. K. *et al.* Extinction risk and conservation of the world's sharks and rays. *Elife* **3**, 1–34 (2014).

49. Selig, E. R. *et al.* Global Priorities for Marine Biodiversity Conservation. *PLoS One* **9**, e82898 (2014).

50. O'Hara, C. C., Afflerbach, J. C., Scarborough, C., Kaschner, K. & Halpern, B. S. Aligning marine species range data to better serve science and conservation. *PLoS One* **12**, e0175739 (2017).

51. Mittermeier, R. A., Goetsch Mittermeier, C., Gil, P. R. & Wilson, E. O. Megadiversity: Earth's Biologically Wealthiest Nations. *CEMEX* (2005).

52. Chamberlain, S. rredlist: 'IUCN' Red List Client. R package version 0.7.0. (2020).

53. GBIF. The Global Biodiversity Information Facility: What is GBIF? https://www.gbif.org/what-is-gbif (2021).

54. OBIS. Ocean Biodiversity Information System. Intergovernmental Oceanographic Commission of UNESCO. www.obis.org. (2021).

55. Chamberlain, S. *et al. rgbif: Interface to the Global Biodiversity Information Facility API. R package version 3.6.0*. https://cran.r-project.org/package=rgbif (2021).

56. Provoost, P. & Bosch, S. robis: Ocean Biodiversity Information System (OBIS) Client. R package version 2.3.9. https://CRAN.R-project.org/package=robis. (2020).

57. Pereira, H. M., Navarro, L. M. & Martins, I. S. Global Biodiversity Change: The Bad, the Good, and the Unknown. *Annu. Rev. Environ. Resour.* **37**, 25–50 (2012).

58. Karger, D. N. *et al.* Climatologies at high resolution for the earth's land surface areas. *Sci. Data* **4**, 170122 (2017).

59. Karger, D. N. *et al.* Data from: Climatologies at high resolution for the earth's land surface areas. *Dryad, Dataset* https://doi.org/10.5061/dryad.kd1d4 (2018).

60. ESA. Land Cover CCI Product User Guide Version 2. Tech. Rep. http://maps.elie.ucl.ac.be/CCI/viewer/download.php (2017).

61. Venter, O. *et al.* Global terrestrial Human Footprint maps for 1993 and 2009. *Sci. Data* **3**, 160067 (2016).

62. Kennedy, C. M., Oakleaf, J. R., Theobald, D. M., Baruch-Mordo, S. & Kiesecker, J. Managing the middle: A shift in conservation priorities based on the global human modification gradient. *Glob. Chang. Biol.* **25**, 811–826 (2019).

63. Seto, K. C., Guneralp, B. & Hutyra, L. R. Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools. *Proc. Natl. Acad. Sci.* **109**, 16083–16088 (2012).

64. UNEP-WCMC & IUCN. Protected Planet: The World Database on Protected Areas (WDPA). *Cambridge, UK: UNEP-WCMC and IUCN* www.protectedplanet.net (2021).

65. Hansen, M. C. *et al.* High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science (80-.).* **342**, 850–853 (2013).

66. Tuanmu, M. N. & Jetz, W. A global, remote sensing-based characterization of terrestrial habitat heterogeneity for biodiversity and ecosystem modelling. *Glob. Ecol. Biogeogr.* **24**, 1329–1339 (2015).

67. Maggi, F., Tang, F. H. M., la Cecilia, D. & McBratney, A. PEST-CHEMGRIDS, global gridded maps of the top 20 crop-specific pesticide application rates from 2015 to 2025. *Sci. Data* **6**, 170 (2019).

68. Byers, L. *et al.* A Global Database of Power Plants. *World Resour. Inst.* 1–18 (2019).

69. Mulligan, M., van Soesbergen, A. & Sáenz, L. GOODD, a global dataset of more than 38,000 georeferenced dams. *Sci. Data* **7**, 31 (2020).

70. Boulay, A.-M. *et al.* The WULCA consensus characterization model for water scarcity footprints: assessing impacts of water consumption based on available water remaining (AWARE). *Int. J. Life Cycle Assess.* **23**, 368–378 (2018).

71. Barbarossa, V. *et al.* Erratum: FLO1K, global maps of mean, maximum and minimum annual streamflow at 1 km resolution from 1960 through 2015. *Sci. Data* **5**, 180078 (2018).

72. Barbarossa, V. *et al.* Impacts of current and future large dams on the geographic range connectivity of freshwater fish worldwide. *Proc. Natl. Acad. Sci.* **117**, 3648–3655 (2020).

73. Domisch, S., Amatulli, G. & Jetz, W. Near-global freshwater-specific environmental variables for biodiversity analyses in 1 km resolution. *Sci. Data* **2**, 150073 (2015).

74. Reid, A. J. *et al.* Emerging threats and persistent conservation challenges for freshwater biodiversity. *Biol. Rev.* **94**, 849–873 (2019).

75. Dudgeon, D. *et al.* Freshwater biodiversity: importance, threats, status and conservation challenges. *Biol. Rev.* **81**, 163 (2006).

76. Schlossberg, S., Chase, M. J., Gobush, K. S., Wasser, S. K. & Lindsay, K. State-space models reveal a continuing elephant poaching problem in most of Africa. *Sci. Rep.* **10**, 10166 (2020).

77. Burn, R. W., Underwood, F. M. & Blanc, J. Global Trends and Factors Associated with the Illegal Killing of Elephants: A Hierarchical Bayesian Analysis of Carcass Encounter Data. *PLoS One* **6**, e24165 (2011).

78. Hauenstein, S., Kshatriya, M., Blanc, J., Dormann, C. F. & Beale, C. M. African elephant poaching rates correlate with local poverty, national corruption and global ivory price. *Nat. Commun.* **10**, 2242 (2019).

79. UNDP. *Human Development Report 2020. The Next Frontier: Human Development and the Anthropocene. New York. http://hdr.undp.org/en/content/human-development-report-2020*. (2020).

80. Transparency International. *Corruption Perceptions Index 2020*. (2020).

81. Early, R. *et al.* Global threats from invasive alien species in the twenty-first century and national response capacities. *Nat. Commun.* **7**, 12485 (2016).

82. Halpern, B. S. *et al.* Spatial and temporal changes in cumulative human impacts on the world's ocean. *Nat. Commun.* **6**, 7615 (2015).

83. Halpern, B. S. *et al.* A Global Map of Human Impact on Marine Ecosystems. *Science (80-.).* **319**, 948–952 (2008).

84. Assis, J. *et al.* Bio-ORACLE v2.0: Extending marine data layers for bioclimatic modelling. *Glob. Ecol. Biogeogr.* **27**, 277–284 (2018).

85. Tyberghein, L. *et al.* Bio-ORACLE: a global environmental dataset for marine species distribution modelling. *Glob. Ecol. Biogeogr.* **21**, 272–281 (2012).

86. Zizka, A., Silvestro, D., Vitt, P. & Knight, T. M. Automated conservation assessment of the orchid family with deep learning. *Conserv. Biol.* **35**, 897–908 (2021).

87. Hastie, T., Friedman, J. & Tibshirani, R. *The Elements of Statistical Learning*. *The Elements of Statistical Learning* vol. 27 (Springer New York, 2001).

88. Kampichler, C., Wieland, R., Calmé, S., Weissenberger, H. & Arriaga-Weiss, S. Classification in conservation biology: A comparison of five machine-learning methods. *Ecol. Inform.* **5**, 441–450 (2010).

89. LeDell, E. *et al.* h2o: R Interface for the 'H2O' Scalable Machine Learning Platform. R package version 3.36.0.4. https://github.com/h2oai/h2o-3. (2022).

90. H2O.ai. H2O AutoML. https://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html (2022).

91. Cutler, D. R. *et al.* RANDOM FORESTS FOR CLASSIFICATION IN ECOLOGY. *Ecology* **88**, 2783–2792 (2007).

92. Kuhn, M. Building Predictive Models in R Using the caret Package. *J. Stat. Softw.* **28**, 1–26 (2008).

93. Kursa, M. B. & Rudnicki, W. R. Feature Selection with the Boruta Package. *J. Stat. Softw.* **36**, 1–13 (2010).

94. Harrell Jr, F. E. Hmisc: Harrell Miscellaneous. R package version 4.5-0. (2021).

95. van der Laan, M. J., Polley, E. C. & Hubbard, A. E. Super Learner. *Stat. Appl. Genet. Mol. Biol.* **6**, (2007).

96. R Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria* https://www.r-project.org/ (2021).

97. RStudio Team. RStudio: Integrated Development Environment for R. *RStudio, PBC, Boston, MA* http://www.rstudio.com/ (2021).

98. Hijmans, R. J. *raster: Geographic Data Analysis and Modeling. R package version 3.0-7*. https://cran.r-project.org/package=raster (2019).

99. Tuszynski, J. caTools: Tools: Moving Window Statistics, GIF, Base64, ROC AUC, etc. R package version 1.18.1. https://CRAN.R-project.org/package=caTools. (2021).

100. Wickham, H. *et al.* Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686. (2019).

101. Dragulescu, A. & Arendt, C. xlsx: Read, Write, Format Excel 2007 and Excel 97/2000/XP/2003 Files. R package version 0.6.5. (2020).

102. Wickham, H. & Bryan, J. readxl: Read Excel Files. R package version 1.3.1. https://CRAN.R-project.org/package=readxl. (2019).

103. Bivand, R., Keitt, T. & Rowlingson, B. *rgdal: Bindings for the 'Geospatial' Data Abstraction Library*. https://cran.r-project.org/package=rgdal (2019).

104. Bivand, R. & Lewin-Koh, N. *maptools: Tools for Handling Spatial Objects. R package version 0.9-5*. https://cran.r-project.org/package=maptools/ (2019).

105. Bivand, R. & Rundel, C. *rgeos: Interface to Geometry Engine - Open Source ('GEOS'). R package version 0.5-1*. https://cran.r-project.org/package=rgeos (2019).

106. Bivand, R. S., Pebesma, E. & Gómez-Rubio, V. *Applied Spatial Data Analysis with R*. (Springer New York, 2013).

107. Pebesma, E. Simple Features for R: Standardized Support for Spatial Vector Data. *R J.* **10**, 439 (2018).

108. Ross, N. fasterize: Fast Polygon to Raster Conversion. R package version 1.0.3. https://CRAN.R-project.org/package=fasterize. (2020).

109. Microsoft Corporation & Weston, S. doParallel: Foreach Parallel Adaptor for the 'parallel' Package. R package version 1.0.16. https://CRAN.R-project.org/package=doParallel. (2020).

110. Wickham, H. stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0. https://CRAN.R-project.org/package=stringr. (2019).

111. ESRI. ArcGIS Pro version 2.9.0. https://www.esri.com/en-us/home (2022).

112. Kuhn, M. caret: Classification and Regression Training. R package version 6.0-86. https://CRAN.R-project.org/package=caret. (2020).

113. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. *Springer, NY* (2016).

114. Wilke, C. O. ggridges: Ridgeline Plots in 'ggplot2'. R package version 0.5.3. https://CRAN.R-project.org/package=ggridges. (2021).

115. South, A. rnaturalearth: World Map Data from Natural Earth. R package version 0.1.0. https://CRAN.R-project.org/package=rnaturalearth. (2017).

116. Garnier, S. viridis: Default Color Maps from 'matplotlib'. R package version 0.5.1. https://CRAN.R-project.org/package=viridis. (2018).

Acknowledgements

This study is part of the Digital Transformation initiative of the Norwegian University of Science and Technology. The contribution of M.H., M.D. and F.V. has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 850717). We further thank Daniel Moran and Caitlin Mandeville for valuable feedback during the preparation of this manuscript.

Author contributions

J.B. and M.D. designed the study and gathered predictor data. J.B. performed data analyses, model building and evaluation. J.B., M.D., M.H., and F.V. interpreted the results and wrote the paper.

Competing interest

The authors declare no competing interests.

Chapter 4: Terrestrial ecosystem impacts of biological invasions caused by international transportation within the framework of Life Cycle Assessment

In preparation. To be submitted to Environmental Science & Technology

Chapter 5: Discussion & Conclusion

Chapter 5: Discussion & Conclusion

5.1 Scientific and Practical Relevance

This thesis highlights potentials and presents approaches for utilizing novel data sources for advancing the modelling of biodiversity impacts in the context of LCA. Instead of accepting the restricting limits inherent to time-intensive, expert-based data sources, the findings of this thesis suggest that, in some cases, innovative approaches based on big data and machine learning can provide sufficiently accurate proxy data. This can reduce limitations imposed by data availability. The thesis contributes in four areas to advance the development of biodiversity impact assessments, namely in 1) increasing species data availability, 2) addressing the uncertainties of these data, 3) suggesting a new methodology for quantifying a previously disregarded impact, and 4) providing open, transparent and accessible research.

1) Increasing the availability of species-level data

Species distribution maps are essential for monitoring biodiversity[1,2] and are widely used across various sustainability relevant fields[3–9]. However, such data are not available consistently across the relevant groups or subgroups of species[10]. Although the use of machine learning and big data is being advocated[11–14], large-scale implementations providing comparable data are rare. At the same time, the expert-based generation of distribution maps and a subsequent provision of the data is unfeasible given the large numbers of unmapped species. This collides with the urgent need to develop and advance biodiversity impact assessment modelling for ensuring informed decisions towards reaching the Sustainable Development Goals[15] as well as the goals of the post-2020 global biodiversity framework[16,17]. Chapter 2 presents geographic data for vascular plant species of the IUCN Red List Version 2021-1, which have so far not been mapped. In chapter 2, data at different levels of detail were gathered and coarse-level species range predictions were generated for 27,208 species using automatized maximum entropy models[18]. The accuracy compared to expert-based range maps indicates a sufficient quality to assist future attempts of assessing biodiversity impacts on a global scale. Although the time-consuming efforts of experts remain inevitable for generating more detailed maps, these findings suggest that proxies of species distribution maps can be generated automatically. The study is unprecedented in its scale and the number of species covered. In particular, the presented approach provides a shortcut between data collection and rapid implementation into global impact assessment models. As such, the data are potentially useful for biogeographic analyses on a global scale. Contrary to regional conservation actions, global impact models offer coarse screenings of, e.g., most detrimental actions, and hence, do not require meticulously precise data.

The need for distribution maps of vascular plants is specifically urgent in biodiversity impact assessment modelling in LCA. For instance, previous studies struggled to find the appropriate data and used unstructured data[19] or relied on species richness estimates instead[20–22]. The lack in species-level information, however, impedes subsequent applications attempting to apply species-specific vulnerability weights[8,22] or species characteristics, e.g., dispersal capacities[23].

Consequently, the results of chapter 2 have already been well received by the scientific community. For instance, the workflow was used by Géron et al.[24] for studying determinants of invasion success of introduced plant species. In addition, the modelled species distribution

maps are of substantial interest for the Global Guidance for Life Cycle Impact Assessment Indicators and Methods (GLAM) effort of the life cycle initiative hosted by UN Environment. In its third phase, GLAM aims to provide a consistent and operational life cycle impact assessment method for several impact categories, as well as a metric for translating regional to global species loss within LCA[25], i.e., the global extinction probability (GEP)[8]. The task chairs for developing both land use assessments and GEP reached out and requested the data presented in chapter 2 for subsequent implementation (Figure 1). Moreover, the GEP is recommended to be applied across all available impact categories[26], making all impact calculations that are based on plant species partly based on chapter 2 of this thesis. The practical relevance of the provided distribution data for vascular plant species is thus immediate by directly contributing to improved models for environmental decision-making.

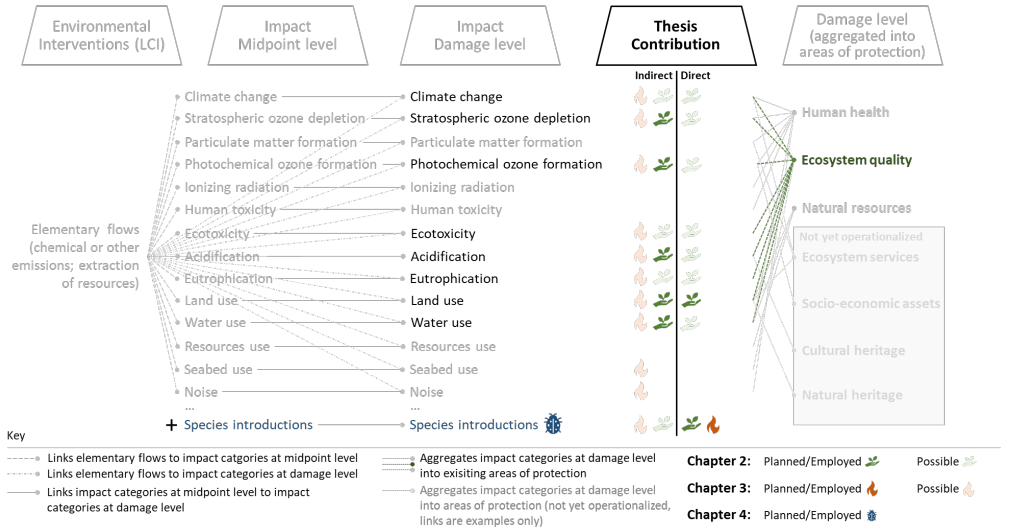


Figure 1: Contribution of chapters 2-4 within the Life Cycle Impact Assessment framework. Direct implementation, i.e., within the impact pathway methodology, and indirect contribution, i.e., via global extincion probabilities[8]. Chapters are indicated by different symbols and planned versus possible implementation by color saturation. Figure adapted from Verones et al.[27]

2) Addressing uncertainties of species-level data

The IUCN Red List of Threatened Species offers valuable information for conservation-relevant assessments, on a regional and global scale[28–30]. However, assessing threat-levels, as well as keeping them up to date, is a time-consuming and costly effort, which impedes the aim of having more than 160,000 species assessed in the near future[31–33]. Global biodiversity impact assessment models frequently rely on species-specific extinction risk categories assessed by the IUCN[8,9,34,35]. The findings of chapter 3 reinforce the concern that species categorized as Data Deficient are problematic for such downstream Red List applications[36]. Depending on the taxonomic group, they hold a considerable share in the number of species[37,38]. Yet, to date, there is no consensus on the appropriate treatment of Data Deficient species[36,39,40], e.g., in approaches aiming to numerically convert a species' extinction risk category. The scientific literature provides numerous applications to support or automatize Red List assessments for individual taxonomic groups[41–44] or specific regions[45–47]. These attempts were mostly targeted towards Red Listing itself. However, for efficiently supporting biodiversity impact assessment

modelling, as well as Red Listing, such data must be available for as many relevant groups and species as possible.

Chapter 3 is the first study to consistently predict probabilities of threat for Data Deficient species across numerous relevant taxonomic groups. The average threat of a Data Deficient species was predicted to be higher compared to their counterparts, supporting earlier studies claiming that they could very well be considered as likely threatened[41,48]. Moreover, Data Deficient species obtained highly variable probabilities of threat within and across taxonomic classes. However, in numerical conversion schemes, they are usually weighed equally within and across taxa, potentially flawing such analyses. In fact, this issue is often neglected in biodiversity impact assessment modelling, e.g., within LCA. The ultimate goal of such studies is to create a ranking based on conservation importance. Indeed, without evidence or further in-depth analyses, the selection of an appropriate weight for a Data Deficient species is largely arbitrary. Though, whether a Data Deficient species is considered to be of equal importance as a Least Concern or as a Critically Endangered species affects this ranking tremendously.

Figure 2A illustrates this dilemma. Here, ecoregions were ranked according to vulnerability weighted species distribution maps for amphibians, mammals, reptiles[49], and birds[50], following the methodologies of Kuipers et al.[8] compared to Verones et al.[22]. Both studies used a linear scheme to numerically convert extinction risk categories, i.e., Least Concern = 0.2, Near Threatened = 0.4, Vulnerable = 0.6, Endangered = 0.8 and Critically Endangered = 1.0. However, Data Deficient species were considered as Least Concern (= 0.2) in Kuipers et al.[8] and as Critically Endangered (= 1.0) in Verones et al.[22]. Naturally, Data Deficient species are systematically underestimated when considered equally important as Least Concern species but overestimated when considered equally important as Critically Endangered species. The difference in ecoregion-rank between these two approaches is striking and may flaw inferences (Figure 2A).

The results of chapter 3 indicate that it could be more appropriate to assign differing numeric weights to Data Deficient species, for instance by splitting them[40] according to whether they were predicted to be threatened (weight of 1.0) or not threatened (weight of 0.2). In fact, the difference in ecoregion-ranking compared to Kuipers et al.[8] highlights specific ecoregions that increase in importance following this approach (Figure 2B), because they contain a considerable number of likely threatened Data Deficient species. At the same time, however, their importance in other ecoregions is not vastly over- or underestimated. In this simplified example, the differences in a hypothetical ecoregion-ranking illustrate the importance to acknowledge threat variations among Data Deficient species, specifically in numeric weighting schemes, but relevant for other approaches too[36,39].

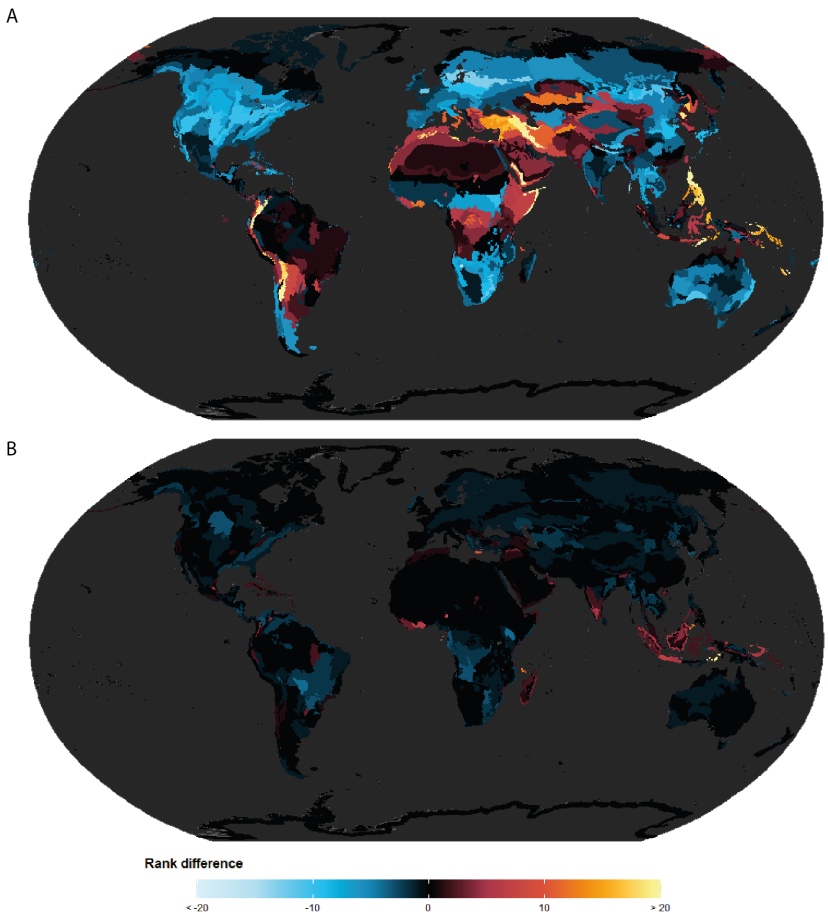


Figure 2: Difference in ranking across terrestrial ecoregions (n = 825) based on vulnerability weighted richness using a linear scheme to numerically convert extinction risk categories, i.e., Least Concern = 0.2, Near Threatened = 0.4, Vulnerable = 0.6, Endangered = 0.8, and Critically Endangered = 1.0. A: Data Deficient species considered as Least Concern in Kuipers et al.[8] compared to Critically Endangered in Verones et al.[22] and B: Data Deficient species considered as Least Concern in Kuipers et al.[8] compared to an evidence-based split of Data Deficient species using predictions of chapter 3, i.e., as Least Concern if predicted to be not threatened, and as Critically Endangered if predicted to be threatened.

Hence, chapter 3 of this thesis contributes to understanding the threat-levels of Data Deficient species and their subsequent consequences in downstream applications. The findings support the concern that the categorization as Data Deficient is highly problematic for conservation prioritization[36] as well as policy-making[39]. Conservation priorities could be consistently underestimated and flawed if Data Deficient species are ignored or misjudged. Besides being applied in biodiversity impact assessment models, extinction risk categories also inform several policy-relevant tools and initiatives, such as the Red List index[51] and the Species Threat Abatement and Restoration (STAR) metric[52]. As such, the presented percentages of threatened species in the global assessment report of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES)[38] are repeatedly too low according to the predictions of chapter 3 (Figure 3). These findings highlight that caution is needed when attempting to employ Data Deficient species in policy-relevant approaches. At best, the predictions generated and provided in chapter 3 can be utilized in future attempts, facilitating a

more evidence-based treatment of Data Deficient species. For instance, these predictions can potentially be implemented into the beforementioned GEP[8] (Figure 1) for ensuring a more appropriate conservation prioritization within LCA. Here, Data Deficient species could be split into potentially threatened and not threatened species based on predicted values of chapter 3, as suggested by Jarić et al.[40]. At the very least, however, studies relying on Data Deficient species should conduct a sensitivity analysis and explore whether and how their results are affected.

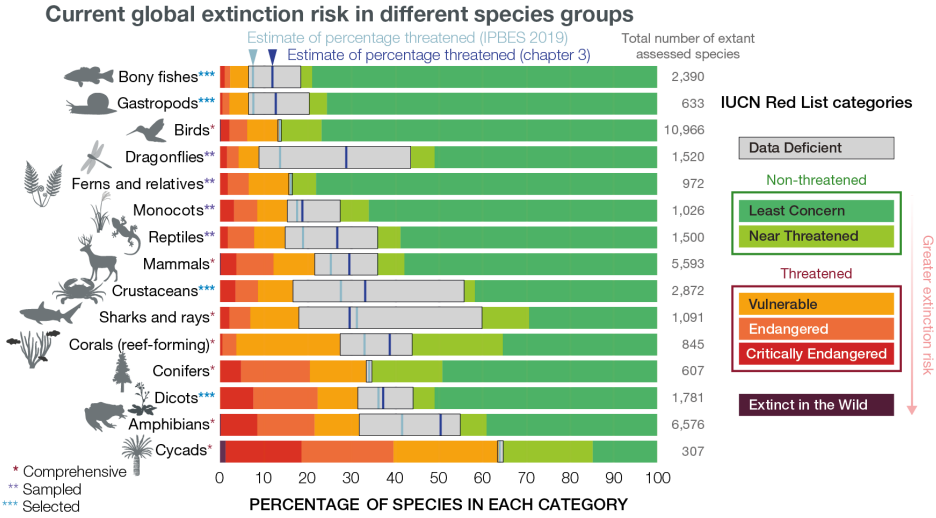


Figure 3: Percentage of species in each IUCN Red List category and estimate for the percentage of species considered threatened across different taxonomic groups assuming that Data Deficient species are as threatened as non-Data Deficient species[38] shown by the vertical light-blue line, and based on predictions for Data Deficient species (chapter 3) shown by the vertical dark-blue lines. Adapted figure from the global assessment report on Biodiversity and Ecosystem Services[38].

3) Improving the coverage of impacts within Life Cycle Assessment

LCA is a policy-relevant tool, often used to support decisions on the corporate level[53]and implemented in several policies by e.g., the European Union[54]. One of its strengths is the possibility to assess performances across numerous environmental impacts[55]. At the same time, however, impacts are overlooked if appropriate impact assessment models are not available[56,57]. Hence, LCA strives for an as complete as possible coverage of different environmental impacts. Yet, although species introductions are to a large extent facilitated by movements of goods and people[58–60], their impacts are neglected in tools aiming to assess environmental consequences e.g., of industrial activities and global supply chains. Until now only one regional approach existed, covering the case of inland shipping[61], but no attempt successfully quantified such impacts on a global scale[56,57,62].

Chapter 4 presents an innovative approach for assessing global impacts from species introductions caused by the transportation of goods in the context of LCA. In line with previous studies, the findings suggest that effects of introduced species can contribute substantially to overall environmental consequences[61]. Hence, neglecting these impacts within and across life cycle stages is especially devastating in increasingly globalized value chains. The proposed

method forms the foundation for further steps towards operationalizing an impact model for including the effects of species introductions alongside other biodiversity impacts (Figure 1). Hence, chapter 4 contributes directly towards complementing the library of existing impact assessment models. Once operationalized, the proposed method will contribute to effective decision-support and allow for a more holistic environmental accounting within LCA and beyond, e.g., in EEIO. In a broader sense, this helps to avoid problem shifting from one environmental impact to another[55] and facilitates better assessment of environmental impacts that aim to ensure responsible consumption and production (Sustainable Development Goal 12), as well as safeguarding life on earth (Sustainable Development Goals 14 & 15)[15].

4) Towards transparent, accessible, and open research

The chapters of this thesis aim to foster advancements in future attempts of biodiversity impact assessment modelling. The utility of the generated data from chapters 2 and 3 was already proven within and beyond this thesis, i.e., through their implementation in chapter 4 and the GLAM framework[25,26]. For effectively contributing to future scientific advancements, however, studies need to go beyond publication and treat their generated data, findings, and related workflows as an enduring product of research[63].

All relevant analysis steps for transparent research and full reproducibility were provided for chapter 2 and 3, and will be provided for chapter 4 upon manuscript submission on GitHub (https://github.com/jannebor). This ensures the validity of the presented and utilized methods, as well as supporting future adaptions and improvements by the scientific community. In addition, a web application[64] was developed and launched upon publication of chapter 2 (Figure 4A). This application aims to increase the interpretability of the data and facilitate its use. Here, the availability, performance, and level of detail of the gathered geographic distribution data can be explored for individual vascular plant species.

Similarly, the utility of the trained machine learning classifier of chapter 3 goes beyond the publication and is extended by another web application (Figure 4B). This beta-version allows for predicting probabilities of extinction threat for single species, facilitating the screening for necessary updates or retrieving estimates for unclassified Data Deficient species. Besides a predicted extinction risk, the stressors that contribute most to this prediction, as well as the species' relative extinction risk in relation to predictions for similar species are shown.

Overall, this thesis contributes to providing open and transparent research for increasing reproducibility, as well as facilitating the re-use of developed methods and data.

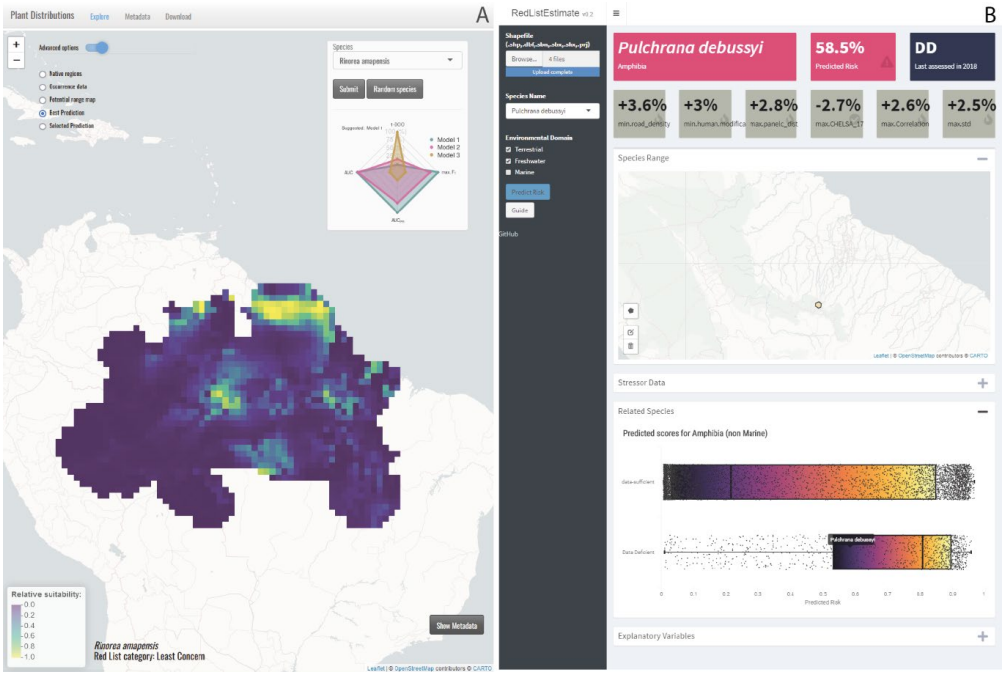


Figure 4: Example screenshots of the developed web applications in this thesis. A: Data explorer for assembled and generated data in chapter 2 (available at https://plant-ranges.indecol.no/), showing a predicted spatial distribution of Rinorea amapensis. B: Interface for applying the trained machine learning classifier of chapter 3 (available at https://ml-extinctionrisk.indecol.no/) with the example of the Data Deficient (DD) species Pulchrana debussyi.

5.2 Limitations and Uncertainty

The findings of this thesis aim to reduce parameter uncertainty (chapters 2 & 3) arising from uncertain or unrepresentative data[65] and facilitate more representative biodiversity impact assessments (chapter 4). However, the underlying methods and data presented in chapters 2-4 contain a range of parameter uncertainties, model uncertainties and subsequent application uncertainties. For instance, the outcomes were generated by models that are data-driven, partly based on modelled third-party data, and require value choices. This and its implications for the utility of the findings of this thesis are discussed below.

5.2.1 Parameter uncertainty

Species occurrence records

Data from the Global Biodiversity Information Facility (GBIF) are considered imperfect[66–70] but were used in all chapters of this thesis. Most of the data are opportunistically recorded sightings of species. Such data are associated with numerous biases and limitations[71]. For instance, typically citizen scientists do not follow any sampling protocol or intend to collect statistically sound data[66]. As such, a species is usually recorded if it was present at a given location and time, but not if it was absent. Although some citizen science projects implemented measures to control flawed parameters such as sampling effort, e.g., *eBird*[72], most GBIF sources provide data that differs fundamentally to data collected based on sampling protocols. Most species' sightings are recorded in accessible and attractive locations[73], as well as for

interesting[66] and easily detectable species[74], implying that the unstructured data are biased in several ways[66,74,75].

In fact, a large part of the uncertainty in predicting geographic distributions of individual vascular plant species (chapter 2) originates in the employed occurrence data. However, we accounted for spatial biases by aggregating to a coarser spatial resolution (i.e., cell size of approximately 56 km at the equator). We expect spatial biases at smaller scales, such as higher observer effort close to infrastructure and accessible attractive places to be filtered out. The most relevant bias for the application in chapter 2 remains the difference across larger spatial scales[69]. These differences are counteracted by block cross-validation which aims to generalize the fitted model. Ultimately this aims to avoid overfitting to regions with many data points[76]. However, there is no assurance that each prediction has been generalized sufficiently. The automatized approach in chapter 2 facilitates the creation of predictions for a large set of species, however, it also makes the validation of each individual prediction unfeasible.

In addition, in chapter 4, the number of species relocations was estimated from species occurrence data retrieved from GBIF. Here, species distribution maps from IUCN[49], BirdLife[50], and chapter 2 were used to generate ecoregion-specific species lists. These lists were contrasted against species lists based on occurrence records from GBIF to identify introduced species and their estimated timing of arrival. However, in countries with a relatively higher sampling effort, those introduced species are more likely to be observed than in underrepresented ones. This primarily affects the fitted curves for estimating both the effect of transport on species introductions and the effect on native biodiversity per introduced species. However, because the estimates of both curves are subsequently combined, this is assumed to not substantially affect the link between transported quantity and biodiversity impacts within a country.

Species-level data

Species distribution maps were used in chapter 2 as validation data, in chapter 3 for retrieving environmental data, and in chapter 4 for counting the number of species in each ecoregion. Range maps are often provided at varying accuracy and can be interpreted as coarse outlines of a species' occurrence[77,78], which are suggested to be most appropriate at coarse spatial resolutions[79]. A disadvantage of using such maps is that it could result in overestimating the species' actual presence in space[79–81].

In addition, species-level extinction risk categories provided by the IUCN[49] were a vital part of chapter 3 for training and testing the machine learning classifier, and were utilized in chapter 4 to estimate the potentially disappeared fraction of species. However, the assessments undertaken by experts contain some level of uncertainty. In chapter 3, outdated assessments had to be used for some species groups to ensure sufficient data availability, which could flaw the trained classifier for some species. The assessments are also unevenly available across space[32]. For some ecoregions, the estimates in chapter 4 are based on relatively few available IUCN assessments, which lowers their specific validity. However, we excluded ecoregions with fewer than 100 assessments to reduce influential effects on the overall outcomes of the study.

Environmental data

Third-party data were used in the generated models of this thesis, such as bioclimatic[82] and land cover variables[83] in chapter 2 and many of the predictors in chapter 3. However, the utilized data sources may contain uncertainties themselves. For instance, Karger et al.[82] rely on in-situ as well as estimated climatic data[84,85] for modelling CHELSA bioclimatic variables, which thus include different levels of uncertainty. Similarly, land cover variables are often based on in-situ observations and predicted into unsampled space using, e.g., random forest algorithms[86]. Such modelled estimates are subsequently used in models of this thesis for making follow-up predictions. Even though the underlying peer-review procedures for publishing the data are assumed to be comprehensive and reliable, this could potentially cause an unnoticed propagation of errors[87], affecting both the interpretability of the models and their predictions.

5.2.2 Methodological uncertainty

The employed models presented in chapter 2 and chapter 3 were calibrated to avoid overfitting[76]. However, the models were only validated under certain assumptions, and are, hence, of limited use for extrapolating into changed conditions, i.e., outside native regions (chapter 2) or for other taxonomic groups (chapter 3). Furthermore, each prediction was not validated individually in chapter 2 due to the automatized workflows with model calibration and selection. The implied uncertainty at the species level must be considered when the data are being used.

Species introductions were predicted using regression techniques in the impact assessment model presented in chapter 4. Thereby, the estimated species introductions were allocated entirely to a specific subset of trade activities that may or may not be complete. For instance, species introductions can be caused by other activities as well, such as tourism[88] or as a result of climate change[89]. For being consistently applicable and comparable to other impact assessment models[27], the impacts had to be described using the agreed-upon indicator, i.e., as potentially disappeared fraction of species (PDF) × years. However, this metric collides with ecological accuracy, because impacts of introduced species do not necessarily cause extinctions only[90,91]. Moreover, the time unit of this metric implies that the invaded ecosystem can return to a natural state once the anthropogenic action stops, which is debatable from an ecological point of view[92,93]. Therefore, it is recommended to apply the estimated impacts within comparative analyses only, not for estimating absolute impacts on species richness, as is usually the case in LCA. In addition, we found the risk and impacts of species introductions to vary per trading partner. However, relative trends could not be validated due to lacking empirical observations or comparable literature. Hence, the impact calculations are data-driven and should be interpreted as such.

5.2.3 Uncertainty in subsequent applications

The presented findings in this thesis are primarily targeted towards biodiversity impact assessments. Potential uncertainties and limitations in this regard are discussed below.

The predicted spatial distributions in chapter 2 could be flawed for individual species. For instance, the validation metrics indicate that the predictions overestimate the species' occurrence for small-ranged species while underestimating it for widespread species, which form the smallest proportion of global biodiversity. In practice this means that impacts are

allocated to locations that are in reality unaffected, if they are associated to an overestimated species distribution map. In turn, if species range maps are underestimated, not all areas of an occurring impact are covered. This implies that impacts will be covered across most regions and for the most important species, i.e., small-ranged, threatened or endemic, when the spatial predictions of chapter 2 are used. Additionally provided data in chapter 2 were native regions. These are provided by third parties and are systematic overestimates of true geographic species ranges. In contrast, species occurrence point localities are systematic underestimates. In cases where such occurrence data have been used to define affected regions, species distribution maps could be more appropriate to fill spatial gaps in comparison to point locality data[81]. Finally, although representing the entire list of vascular plant species of Red List Version 2021-1, the species covered in chapter 2 are only a small proportion of known plant species worldwide. Hence, the dataset is not suited for estimating absolute species richness but rather intends to assist studies that work with relative richness indicators.

The predicted probabilities in chapter 3 were used to assess which Data Deficient species obtained a relatively higher risk of being threatened by extinction. In the context of numerical conversion schemes, the findings stress that the consequences for global-scale modelling approaches may be significant. However, it would be incorrect to use the predicted scores of chapter 3 directly as weights since they do not correspond to actual threat-levels but rather modelled probabilities of being threatened by extinction. Hence, a data split based on the predictions into threatened versus not threatened is the most appropriate way of utilizing the data. At a species-level, however, some predictions may be wrong. Falsely predicting a species to be not threatened is thereby worse than vice versa. In the sense of impact assessment modelling, this means that the extinction risk category of a species is either underestimated (i.e., false negative) or overestimated (i.e., false positive). However, given that in most studies to date, Data Deficient species are typically underestimated (i.e., weighed as Least Concern), the predictions could increase appropriateness.

The characterization factor (CF) presented in chapter 4 indicates the expected damage per kg of transported goods due to species introductions. As a result of limited availability of species-level data the CF is based on a few taxonomic groups, i.e., amphibians, birds, mammals, reptiles, and vascular plants. Some uncertainties of the CF are related to LCA specific limitations. For instance, the pressure-response relationship is assumed to be linear, meaning that impacts increase linearly with increasing activity[94] and ignore vital thresholds. In addition, interactions between stressors are ignored[94] although additive effects could escalate ongoing impacts. However, the CF is primarily intended to be applied in LCA, and possibly EEIO. Such comparative analyses do not look at the absolute impact but investigate which impact is relatively worse or better than the other, aiming to reduce the environmental footprints of products and processes[95]. The comparison to other CFs within the same framework is thereby given if the uncertainty and limitations of the compared impacts are of similar magnitude.

5.3 Conclusion and Outlook

The ever-increasing quantities of available data require innovative approaches for efficiently extracting relevant information. The chapters of this thesis contribute to advancing the field of biodiversity impact assessment modelling by addressing some of the most urgent and fundamental issues in the field[56,57].

Chapters 2 and 3 of this thesis advocate embracing the patterns that emerge from the data, contrary to seeking to confirm expected patterns[12]. Approaches are presented for retrieving knowledge from big data. This aims primarily at reducing data limitations and uncertainties in subsequent applications. Although such data-driven research can flaw inferences on a detailed and local level, it may be sufficient to inform coarse-scale impact assessment models on a global scale. In chapter 2, the availability of geographical distribution data for vascular plant species was increased. The immediate implementation into several studies and frameworks[8,24–26] proves the utility and need for this data. This potentially increases the taxonomic coverage of several impact assessment models[27] (Figure 1) and contributes to a better representation of biodiversity within LCA, and beyond. Chapter 3 stresses limitations in the current paradigm of employing Data Deficient species in impact assessment models. The predicted risks for Data Deficient species suggest potentially flawed biodiversity assessments. However, this issue has been largely neglected until now. The findings of this chapter aim to initiate a debate about the utility and implementation of Data Deficient species within biodiversity models, which hopefully reduces uncertainties in future approaches. A paradigm shift towards purely data-driven science[12] is not advocated[96]. However, automatically generated data-driven proxy data can be a useful complementation to existing expert-based data. This is exemplified in chapter 4, where the data of both preceding chapters, alongside other open data, facilitate the first global impact model for the impacts of species introductions in the context of LCA. Here, frequently updated data from GBIF were utilized for assessing arrivals, locations, and impacts of species introductions. Such data would not exist without the meticulous effort of thousands of diligent and dedicated enthusiasts and experts.

Future improvements of the data generated in chapter 2 should include, but are not limited to, implementing novel approaches for correcting biases in occurrence data[97–101], which could even allow for an increased spatial resolution. In theory, expanding this approach to increase data availability for other taxonomic groups is the logical next step. In practice, however, most taxonomic groups are not sufficiently covered in the relevant databases[68,70] although promising future applications aiming at improving the gathering and use of such data are on the horizon[2,11,102]. In addition, the data should be frequently revised to always represent the current state of knowledge and for following updated species lists of the IUCN Red List. Predictions of chapter 3 could be improved by integrating species-specific data regarding, e.g., species characteristics and threats or estimated population dynamics from species observations[103,104]. However, progress in the Red List is underway[32]. More and more species are being assessed for their extinction risk categories making automatized predictions more robust and feasible for additional taxonomic groups in the future.

In addition, the impact assessment model of chapter 4 is the first of its kind and contains many options for future improvements. Future attempts should prioritize the differentiation between

types of transport vessels and different commodities to facilitate effective decision-making. The effect factor should be replaced as soon as a more sophisticated approach is developed, additional impacts on the environment should be considered and the approach should be adapted to integrate impacts on marine ecosystems as well.

The chapters of this thesis embrace the momentum of the digital revolution. This thesis highlights potentials of data science to facilitate the swift integration of alternative data sources into decision-support tools, vital for keeping up with accelerating environmental threats. The scientific community is encouraged to verify, develop and challenge the presented results. This thesis provides a first attempt to promptly implement unstructured, large-scale databases into policy-relevant tools; it highlights the influential power of Data Deficient species for impact assessment modelling and presents data to tackle this shortcoming; it contributes a first global attempt to quantify the impacts of species introductions in the framework of LCA, and highlights their relevance. These developments represent considerable novelty that clearly advance the modelling of biodiversity impacts beyond the state-of-the-art.

5.4 References

1. Pereira, H. M. *et al.* Essential Biodiversity Variables. *Science (80-.).* **339**, 277–278 (2013).
2. Kissling, W. D. *et al.* Building essential biodiversity variables (EBVs) of species distribution and abundance at a global scale. *Biol. Rev.* **93**, 600–625 (2018).
3. Le Saout, S. *et al.* Protected Areas and Effective Biodiversity Conservation. *Science (80-.).* **342**, 803–805 (2013).
4. Hoffmann, M. *et al.* The Impact of Conservation on the Status of the World's Vertebrates. *Science (80-.).* **330**, 1503–1509 (2010).
5. Stuart, S. N. *et al.* Status and Trends of Amphibian Declines and Extinctions Worldwide. *Science (80-.).* **306**, 1783–1786 (2004).
6. Strassburg, B. B. N. *et al.* Impacts of incentives to reduce emissions from deforestation on global species extinctions. *Nat. Clim. Chang.* **2**, 350–355 (2012).
7. Montesino Pouzols, F. *et al.* Global protected area expansion is compromised by projected land-use and parochialism. *Nature* **516**, 383–386 (2014).
8. Kuipers, K. J. J., Hellweg, S. & Verones, F. Potential Consequences of Regional Species Loss for Global Species Richness: A Quantitative Approach for Estimating Global Extinction Probabilities. *Environ. Sci. Technol.* **53**, 4728–4738 (2019).
9. Lenzen, M. *et al.* International trade drives biodiversity threats in developing nations. *Nature* **486**, 109–112 (2012).
10. Whittaker, R. J. *et al.* Conservation Biogeography: Assessment and Prospect. *Divers. Distrib.* **11**, 3–23 (2005).
11. Wüest, R. O. *et al.* Macroecology in the age of Big Data – Where to go from here? *J. Biogeogr.* **47**, 1–12 (2020).
12. Kelling, S. *et al.* Data-intensive Science: A New Paradigm for Biodiversity Studies. *Bioscience* **59**, 613–620 (2009).

13. Hochachka, W. M. *et al.* Data-intensive science applied to broad-scale citizen science. *Trends Ecol. Evol.* **27**, 130–137 (2012).

14. Sullivan, B. L. *et al.* Using open access observational data for conservation action: A case study for birds. *Biol. Conserv.* **208**, 5–14 (2017).

15. United Nations. *Transforming our World: the 2030 Agenda for Sustainable Development*. A/RES/70/1 (United Nations, 2015).

16. CBD. *Zero draft of the post-2020 global biodiversity framework*. https://www.cbd.int/doc/c/3064/749a/0f65ac7f9def86707f4eaefa/post2020-prep-02-01-en.pdf (2020).

17. Díaz, S. *et al.* Set ambitious goals for biodiversity and sustainability. *Science (80-.).* **370**, 411–413 (2020).

18. Phillips, S. J., Dudík, M. & Schapire, R. E. Maxent software for modeling species niches and distributions (Version 3.4.0). http://biodiversityinformatics.amnh.org/open_source/maxent/ (2016).

19. Gade, A. L., Hauschild, M. Z. & Laurent, A. Globally differentiated effect factors for characterising terrestrial acidification in life cycle impact assessment. *Sci. Total Environ.* **761**, 143280 (2021).

20. De Baan, L., Mutel, C. L., Curran, M., Hellweg, S. & Koellner, T. Land use in life cycle assessment: Global characterization factors based on regional and global potential species extinction. *Environ. Sci. Technol.* **47**, 9281–9290 (2013).

21. Chaudhary, A. & Brooks, T. M. Land Use Intensity-Specific Global Characterization Factors to Assess Product Biodiversity Footprints. *Environ. Sci. Technol.* **52**, 5094–5104 (2018).

22. Verones, F., Pfister, S., van Zelm, R. & Hellweg, S. Biodiversity impacts from water consumption on a global scale for use in life cycle assessment. *Int. J. Life Cycle Assess.* **22**, 1247–1256 (2017).

23. Kuipers, K. J. J., May, R. & Verones, F. Considering habitat conversion and fragmentation in characterisation factors for land-use impacts on vertebrate species richness. *Sci. Total Environ.* **801**, 149737 (2021).

24. Géron, C. *et al.* Urban alien plants in temperate oceanic regions of Europe originate from warmer native ranges. *Biol. Invasions* **23**, 1765–1779 (2021).

25. UNEP/SETAC Life Cycle Initiative. Global LCIA guidance Phase 3 - Creation of a Global Life Cycle Impact Assessment Method: Scoping document. 1–166 (2021).

26. Verones, F. *et al.* Global extinction probabilities of terrestrial, freshwater, and marine species groups for use in Life Cycle Assessment. *Submitted to Ecol. Indic.*

27. Verones, F. *et al.* LCIA framework and cross-cutting issues guidance within the UNEP-SETAC Life Cycle Initiative. *J. Clean. Prod.* **161**, 957–967 (2017).

28. Rodrigues, A., Pilgrim, J., Lamoreux, J., Hoffmann, M. & Brooks, T. The value of the IUCN Red List for conservation. *Trends Ecol. Evol.* **21**, 71–76 (2006).

29. Bennun, L. *et al.* The Value of the IUCN Red List for Business Decision-Making. *Conserv. Lett.* **11**, e12353 (2018).

30. Betts, J. *et al.* A framework for evaluating the impact of the IUCN Red List of threatened species. *Conserv. Biol.* **34**, 632–643 (2020).

31. Stuart, S. N., Wilson, E. O., McNeely, J. A., Mittermeier, R. A. & Rodríguez, J. P. The Barometer of Life. *Science (80-.).* **328**, 177–177 (2010).

32. Bachman, S. P. *et al.* Progress, challenges and opportunities for Red Listing. *Biol. Conserv.* **234**, 45–55 (2019).

33. Rondinini, C., Di Marco, M., Visconti, P., Butchart, S. H. M. & Boitani, L. Update or Outdate: Long-Term Viability of the IUCN Red List. *Conserv. Lett.* **7**, 126–130 (2014).

34. Moran, D. & Kanemoto, K. Identifying species threat hotspots from global supply chains. *Nat. Ecol. Evol.* **1**, 0023 (2017).

35. Irwin, A. *et al.* Quantifying and categorising national extinction-risk footprints. *Sci. Rep.* **12**, 5861 (2022).

36. Bland, L. M., Collen, B., Orme, C. D. L. & Bielby, J. Data uncertainty and the selectivity of extinction risk in freshwater invertebrates. *Divers. Distrib.* **18**, 1211–1220 (2012).

37. Cazalis, V. *et al.* Bridging the research-implementation gap in IUCN Red List assessments. *Trends Ecol. Evol.* **37**, 359–370 (2022).

38. IPBES. Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. (2019) doi:10.5281/ZENODO.5517154.

39. Parsons, E. C. M. Why IUCN Should Replace "Data Deficient" Conservation Status with a Precautionary "Assume Threatened" Status—A Cetacean Case Study. *Front. Mar. Sci.* **3**, 2015–2017 (2016).

40. Jarić, I., Courchamp, F., Gessner, J. & Roberts, D. L. Potentially threatened: a Data Deficient flag for conservation management. *Biodivers. Conserv.* **25**, 1995–2000 (2016).

41. Bland, L. M., Collen, B., Orme, C. D. L. & Bielby, J. Predicting the conservation status of data-deficient species. *Conserv. Biol.* **29**, 250–259 (2015).

42. Bland, L. M. & Böhm, M. Overcoming data deficiency in reptiles. *Biol. Conserv.* **204**, 16–22 (2016).

43. Luiz, O. J., Woods, R. M., Madin, E. M. P. & Madin, J. S. Predicting IUCN Extinction Risk Categories for the World's Data Deficient Groupers (Teleostei: Epinephelidae). *Conserv. Lett.* **9**, 342–350 (2016).

44. Darrah, S. E., Bland, L. M., Bachman, S. P., Clubbe, C. P. & Trias-Blasi, A. Using coarse-scale species distribution data to predict extinction risk in plants. *Divers. Distrib.* **23**, 435–447 (2017).

45. Stévart, T. *et al.* A third of the tropical African flora is potentially threatened with extinction. *Sci. Adv.* **5**, eaax9444 (2019).

46. Walls, R. H. L. & Dulvy, N. K. Eliminating the dark matter of data deficiency by predicting the conservation status of Northeast Atlantic and Mediterranean Sea sharks and rays. *Biol. Conserv.* **246**, 108459 (2020).

47. Walls, R. H. L. & Dulvy, N. K. Tracking the rising extinction risk of sharks and rays in

the Northeast Atlantic Ocean and Mediterranean Sea. *Sci. Rep.* **11**, 15397 (2021).

48. Howard, S. D. & Bickford, D. P. Amphibians over the edge: silent extinction risk of Data Deficient species. *Divers. Distrib.* **20**, 837–846 (2014).

49. IUCN. The IUCN Red List of Threatened Species. Version 2021-3. https://www.iucnredlist.org (2022).

50. BirdLife International. *IUCN Red List for birds*. http://www.birdlife.org (2020).

51. Butchart, S. H. M. *et al.* Using Red List Indices to measure progress towards the 2010 target and beyond. *Philos. Trans. R. Soc. B Biol. Sci.* **360**, 255–268 (2005).

52. Mair, L. *et al.* A metric for spatially explicit contributions to science-based species targets. *Nat. Ecol. Evol.* **5**, 836–844 (2021).

53. Milà i Canals, L. *et al.* Estimating the greenhouse gas footprint of Knorr. *Int. J. Life Cycle Assess.* **16**, 50–58 (2011).

54. Sala, S., Amadei, A. M., Beylot, A. & Ardente, F. The evolution of life cycle assessment in European policies over three decades. *Int. J. Life Cycle Assess.* **26**, 2295–2314 (2021).

55. Hellweg, S. & Milà i Canals, L. Emerging approaches, challenges and opportunities in life cycle assessment. *Science (80-.).* **344**, 1109–1113 (2014).

56. Winter, L., Lehmann, A., Finogenova, N. & Finkbeiner, M. Including biodiversity in life cycle assessment – State of the art, gaps and research needs. *Environ. Impact Assess. Rev.* **67**, 88–100 (2017).

57. Woods, J. S. *et al.* Ecosystem quality in LCIA: status quo, harmonization, and suggestions for the way forward. *Int. J. Life Cycle Assess.* **23**, 1995–2006 (2018).

58. Bellard, C., Leroy, B., Thuiller, W., Rysman, J. F. & Courchamp, F. Major drivers of invasion risks throughout the world. *Ecosphere* **7**, 1–14 (2016).

59. Essl, F. *et al.* Drivers of future alien species impacts: An expert-based assessment. *Glob. Chang. Biol.* **26**, 4880–4893 (2020).

60. Seebens, H. *et al.* Non-native species spread in a complex network: the interaction of global transport and local population dynamics determines invasion success. *Proc. R. Soc. B Biol. Sci.* **286**, 20190036 (2019).

61. Hanafiah, M. M., Leuven, R. S. E. W., Sommerwerk, N., Tockner, K. & Huijbregts, M. A. J. Including the Introduction of Exotic Species in Life Cycle Impact Assessment: The Case of Inland Shipping. *Environ. Sci. Technol.* **47**, 13934–13940 (2013).

62. Crenna, E., Marques, A., La Notte, A. & Sala, S. Biodiversity Assessment of Value Chains: State of the Art and Emerging Challenges. *Environ. Sci. Technol.* **54**, 9715–9728 (2020).

63. Hampton, S. E. *et al.* Big data and the future of ecology. *Front. Ecol. Environ.* **11**, 156–162 (2013).

64. Chang, W. *et al.* shiny: Web Application Framework for R. R package version 1.7.1.9003, https://shiny.rstudio.com/. (2022).

65. van Zelm, R. & Huijbregts, M. A. J. Quantifying the Trade-off between Parameter and Model Structure Uncertainty in Life Cycle Impact Assessment. *Environ. Sci. Technol.*

47, 9274–9280 (2013).

66. Isaac, N. J. B. & Pocock, M. J. O. Bias and information in biological records. *Biol. J. Linn. Soc.* **115**, 522–531 (2015).

67. Amano, T., Lamming, J. D. L. & Sutherland, W. J. Spatial Gaps in Global Biodiversity Information and the Role of Citizen Science. *Bioscience* **66**, 393–400 (2016).

68. Beck, J., Ballesteros-Mejia, L., Nagel, P. & Kitching, I. J. Online solutions and the 'Wallacean shortfall': What does GBIF contribute to our knowledge of species' ranges? *Divers. Distrib.* **19**, 1043–1050 (2013).

69. Meyer, C., Weigelt, P. & Kreft, H. Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecol. Lett.* **19**, 992–1006 (2016).

70. Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R. & Legendre, F. Taxonomic bias in biodiversity data and societal preferences. *Sci. Rep.* **7**, 1–14 (2017).

71. Bird, T. J. *et al.* Statistical solutions for error and bias in global citizen science datasets. *Biol. Conserv.* **173**, 144–154 (2014).

72. Sullivan, B. L. *et al.* eBird: A citizen-based bird observation network in the biological sciences. *Biol. Conserv.* **142**, 2282–2292 (2009).

73. Tiago, P., Ceia-Hasse, A., Marques, T. A., Capinha, C. & Pereira, H. M. Spatial distribution of citizen science casuistic observations for different taxonomic groups. *Sci. Rep.* **7**, 12832 (2017).

74. Isaac, N. J. B., van Strien, A. J., August, T. A., de Zeeuw, M. P. & Roy, D. B. Statistics for citizen science: Extracting signals of change from noisy ecological data. *Methods Ecol. Evol.* **5**, 1052–1060 (2014).

75. Hill, M. O. Local frequency as a key to interpreting species occurrence data when recording effort is not known. *Methods Ecol. Evol.* **3**, 195–205 (2012).

76. Radosavljevic, A. & Anderson, R. P. Making better Maxent models of species distributions: complexity, overfitting and evaluation. *J. Biogeogr.* **41**, 629–643 (2014).

77. Gaston, K. J. & Fuller, R. A. The sizes of species' geographic ranges. *J. Appl. Ecol.* **46**, 1–9 (2009).

78. Brooks, T. M. *et al.* Measuring Terrestrial Area of Habitat (AOH) and Its Utility for the IUCN Red List. *Trends Ecol. Evol.* **34**, 977–986 (2019).

79. Hurlbert, A. H. & Jetz, W. Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. *Proc. Natl. Acad. Sci.* **104**, 13384–13389 (2007).

80. Jetz, W., Sekercioglu, C. H. & Watson, J. E. M. Ecological correlates and conservation implications of overestimating species geographic ranges. *Conserv. Biol.* **22**, 110–9 (2008).

81. Rondinini, C., Wilson, K. A., Boitani, L., Grantham, H. & Possingham, H. P. Tradeoffs of different types of species occurrence data for use in systematic conservation planning. *Ecol. Lett.* **9**, 1136–1145 (2006).

82. Karger, D. N. *et al.* Climatologies at high resolution for the earth's land surface areas. *Sci. Data* **4**, 170122 (2017).

83. ESA. Land Cover CCI Product User Guide Version 2. Tech. Rep. http://maps.elie.ucl.ac.be/CCI/viewer/download.php (2017).

84. Dee, D. P. *et al.* The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.* **137**, 553–597 (2011).

85. Schneider, U. *et al.* GPCC's new land surface precipitation climatology based on quality-controlled in situ data and its role in quantifying the global water cycle. *Theor. Appl. Climatol.* **115**, 15–40 (2014).

86. Inglada, J. *et al.* Operational High Resolution Land Cover Map Production at the Country Scale Using Satellite Image Time Series. *Remote Sens.* **9**, 95 (2017).

87. Meyer, H. & Pebesma, E. Machine learning-based global maps of ecological variables and the challenge of assessing them. *Nat. Commun.* **13**, 2208 (2022).

88. Anderson, L. G., Rocliffe, S., Haddaway, N. R. & Dunn, A. M. The Role of Tourism and Recreation in the Spread of Non-Native Species: A Systematic Review and Meta-Analysis. *PLoS One* **10**, e0140833 (2015).

89. Bellard, C., Jeschke, J. M., Leroy, B. & Mace, G. M. Insights from modeling studies on how climate change affects invasive alien species geography. *Ecol. Evol.* **8**, 5688–5700 (2018).

90. Ehrenfeld, J. G. Ecosystem Consequences of Biological Invasions. *Annu. Rev. Ecol. Evol. Syst.* **41**, 59–80 (2010).

91. Pyšek, P. *et al.* Scientists' warning on invasive alien species. *Biol. Rev.* **95**, 1511–1534 (2020).

92. Rout, T. M., Moore, J. L., Possingham, H. P. & McCarthy, M. A. Allocating biosecurity resources between preventing, detecting, and eradicating island invasions. *Ecol. Econ.* **71**, 54–62 (2011).

93. Roy, E. *et al. Invasive Alien Species - Prioritising prevention efforts through horizon scanning ENV.B.2/ETU/2014/0016*. (2015).

94. Curran, M. *et al.* Toward Meaningful End Points of Biodiversity in Life Cycle Assessment. *Environ. Sci. Technol.* **45**, 70–79 (2011).

95. Hauschild, M. Z. & Huijbregts, M. A. J. *Life Cycle Impact Assessment*. (Springer, 2015). doi:10.1007/BF02978760.

96. Harford, T. Big data: A big mistake? *Significance* **11**, 14–19 (2014).

97. Altwegg, R. & Nichols, J. D. Occupancy models for citizen-science data. *Methods Ecol. Evol.* **10**, 8–21 (2019).

98. Yuan, Y. *et al.* Point process models for spatio-temporal distance sampling data from a large-scale survey of blue whales. *Ann. Appl. Stat.* **11**, 2270–2297 (2017).

99. Johnston, A., Fink, D., Hochachka, W. M. & Kelling, S. Estimates of observer expertise improve species distributions from citizen science data. *Methods Ecol. Evol.* **9**, 88–97 (2018).

100. Sicacha-Parada, J., Steinsland, I., Cretois, B. & Borgelt, J. Accounting for spatial varying sampling effort due to accessibility in Citizen Science data: A case study of moose in

Norway. *Spat. Stat.* **42**, 100446 (2021).

101. Boyd, R. J., Powney, G. D., Carvell, C. & Pescott, O. L. occAssess: An R package for assessing potential biases in species occurrence data. *Ecol. Evol.* **11**, 16177–16187 (2021).

102. Pocock, M. J. O. *et al. A Vision for Global Biodiversity Monitoring With Citizen Science*. *Advances in Ecological Research* vol. 59 (Elsevier Ltd., 2018).

103. Tingley, M. W. & Beissinger, S. R. Detecting range shifts from historical species occurrences: new perspectives on old data. *Trends Ecol. Evol.* **24**, 625–633 (2009).

104. Thomas, J. A. *et al.* Comparative Losses of British Butterflies, Birds, and Plants and the Global Extinction Crisis. *Science (80-.).* **303**, 1879–1881 (2004).

SI1 (Supporting Information for chapter 3): More than half of Data Deficient species predicted to be threatened by extinction

Communications Biology (2022) 5

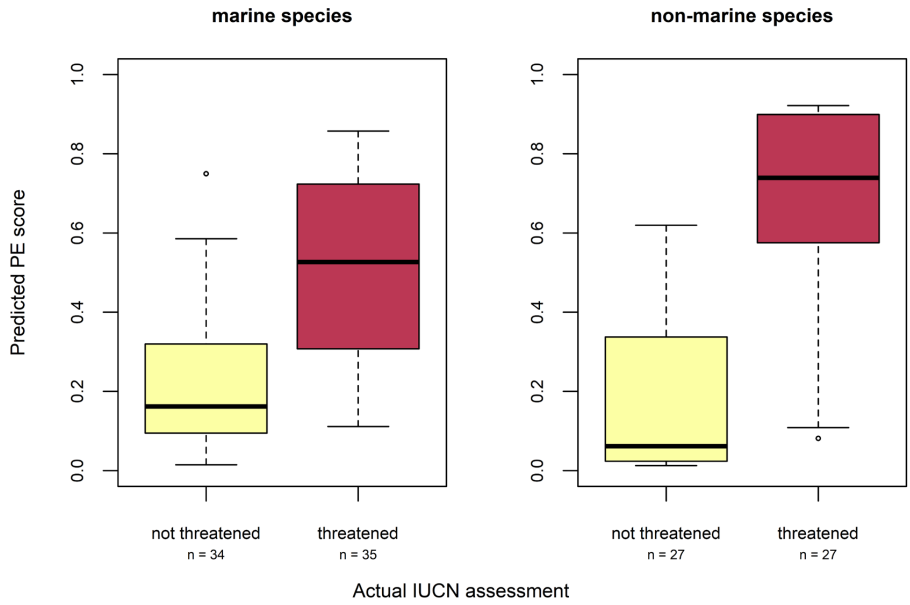
More than half of Data Deficient species predicted to be threatened by extinction

Jan Borgelt[1], Martin Dorber[1], Marthe Alnes Høiberg[1], Francesca Verones[1]

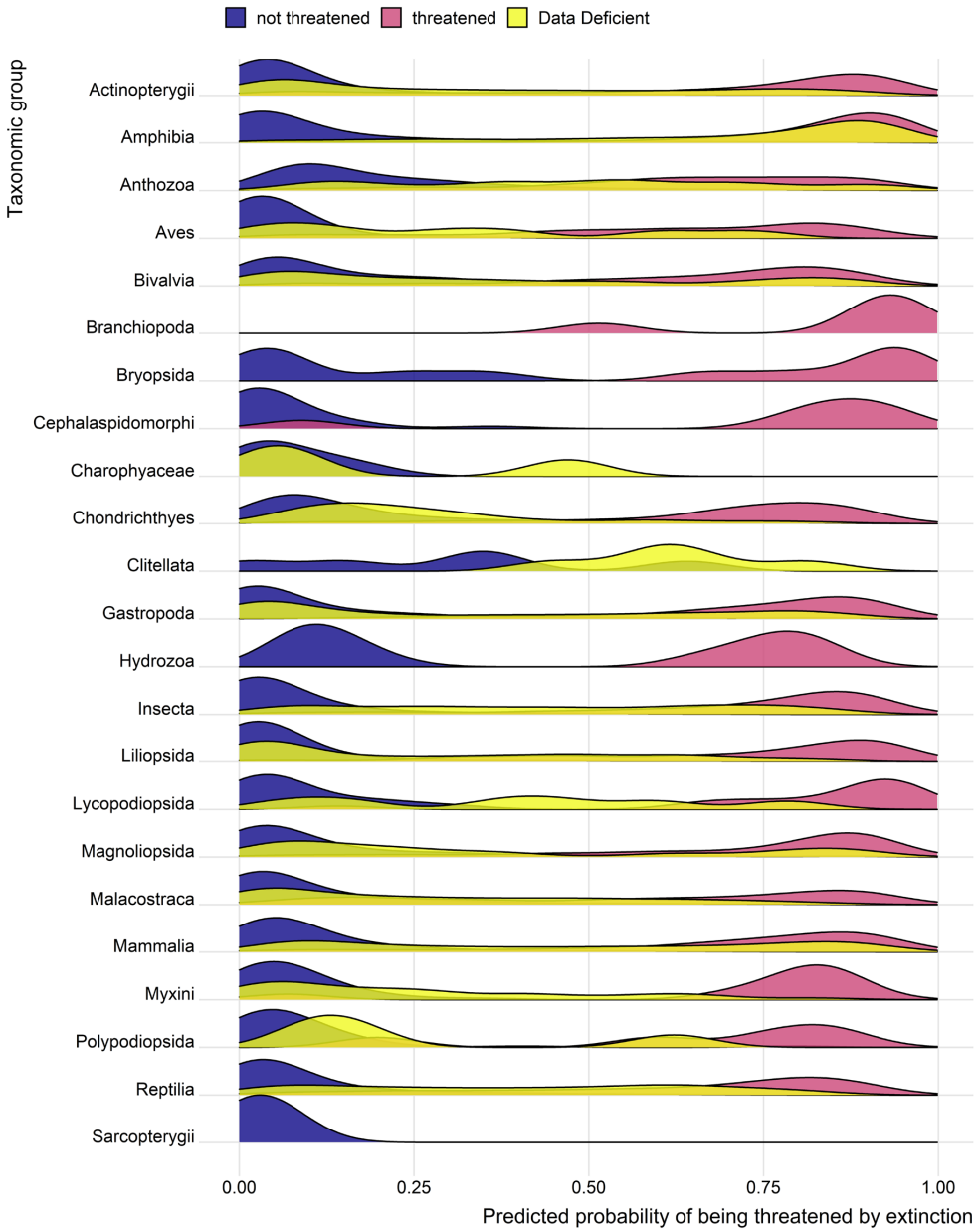
1. Industrial Ecology Programme, Department of Energy and Process Engineering, Norwegian University of Science and Technology (NTNU), Trondheim, Norway
corresponding author: Jan Borgelt (jan.borgelt@ntnu.no)

Supplementary Information

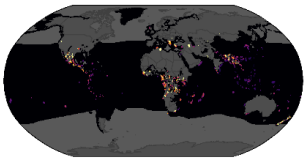
This file includes Supplementary Figures 1-7 and Supplementary Tables 1-3.



Supplementary Figure 1: Boxplot showing the interquartile range (box), median (black line), minimum and maximum values without outliers (error bars), and outliers (points) of predicted probability of being threatened by extinction (i.e., PE score) of formerly Data Deficient species (n = 123 Data Deficient in IUCN Version 2020-3)[1,2] *across updated IUCN assessments (updated in Version 2021-2 to either not threatened or threatened)*[3].

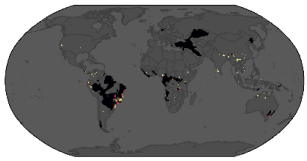


Supplementary Figure 2: Frequency distribution of predicted PE scores for not threatened (blue), threatened (red) and Data Deficient (yellow) species per taxonomic class of the spatial dataset of the IUCN Red List of threatened species (Version 2020-3)[1,2].

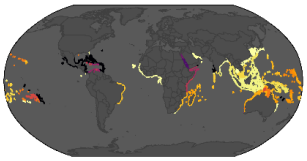


Actinopterygii

Agaricomycetes



Amphibia



Anthozoa

Aves

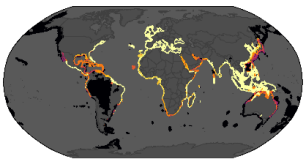
Bivalvia

Branchiopoda

Bryopsida

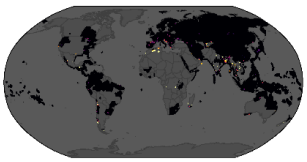
Cephalaspidomorphi

Charophyaceae



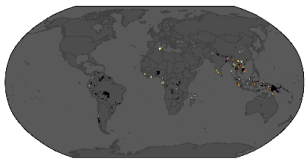
Chondrichthyes

Clitellata



Gastropoda

Hydrozoa



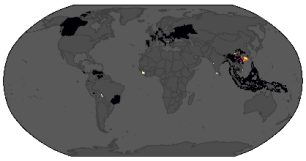
Insecta

Jungermanniopsida

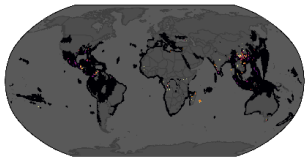
Lecanoromycetes

Liliopsida

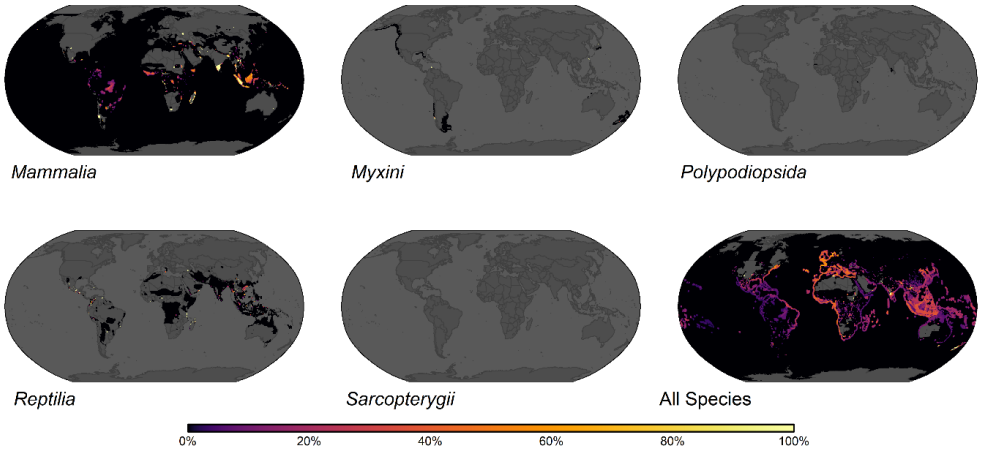
Lycopodiopsida



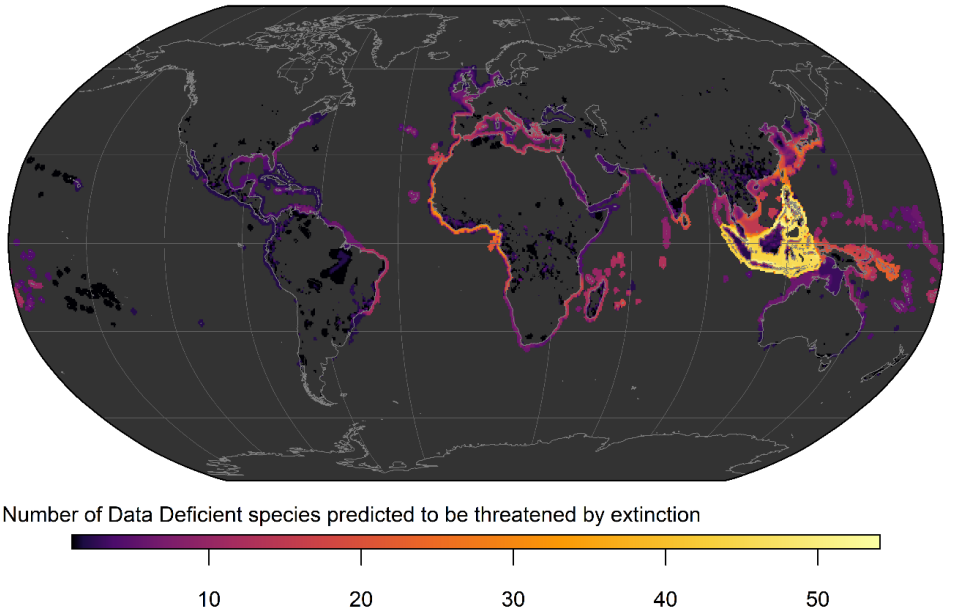
Magnoliopsida



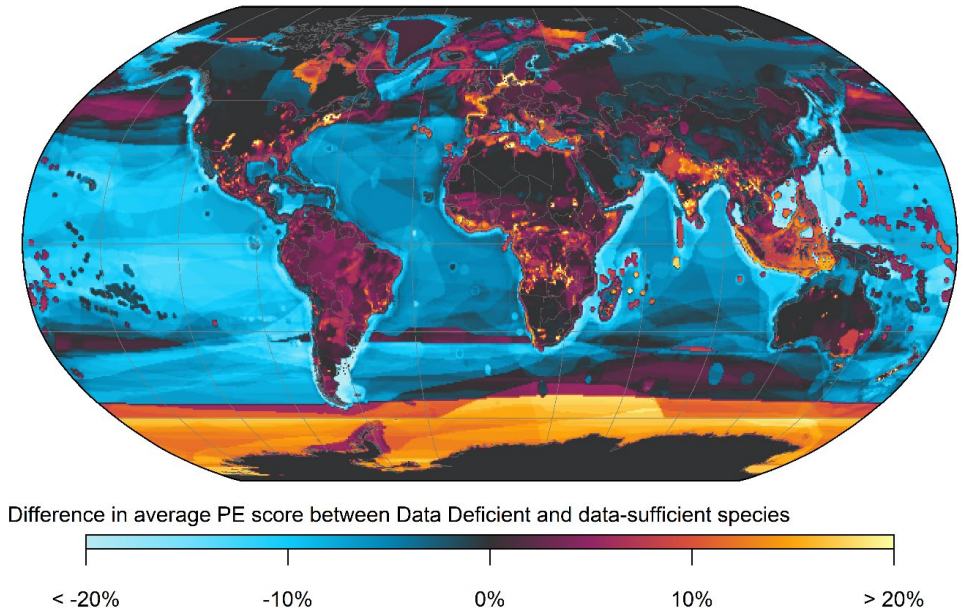
Malacostraca



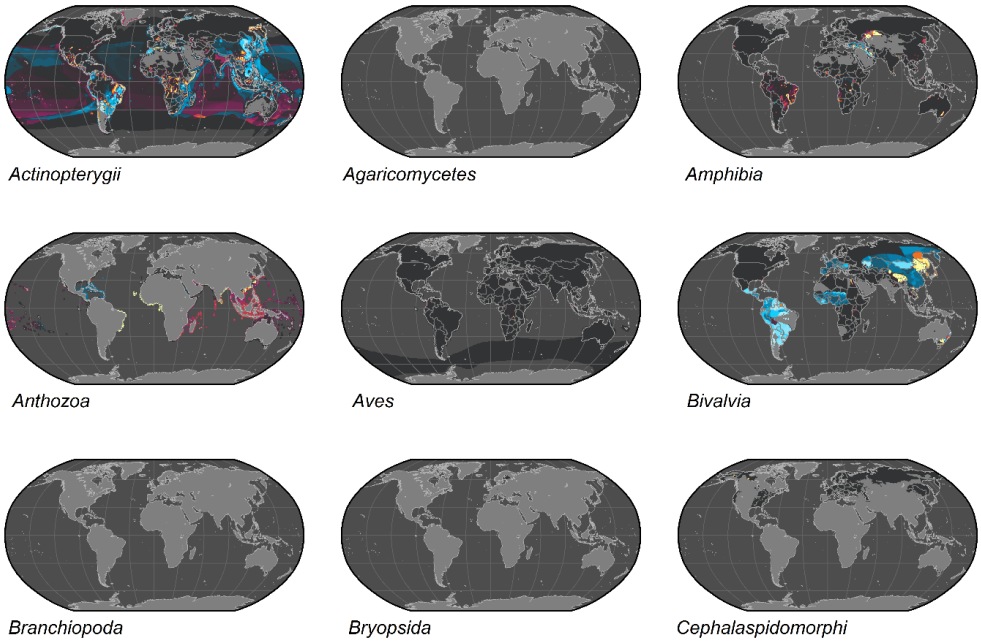
Supplementary Figure 3: Fraction of Data Deficient species predicted to be threatened by extinction per taxonomic class and for all Data Deficient species (n = 7,699) of the spatial dataset of the IUCN Red List of threatened species (Version 2020-3)[1,2].

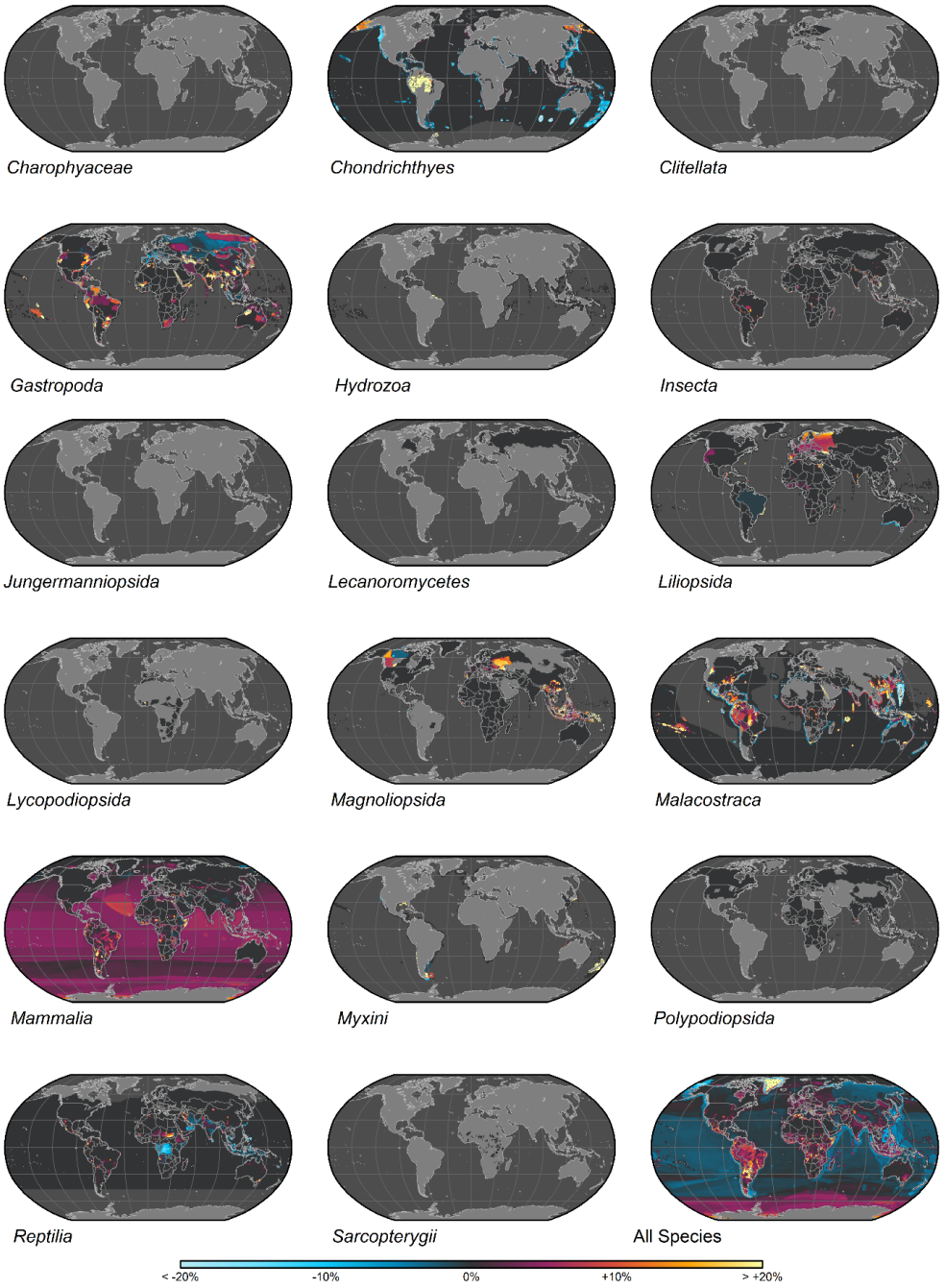


Supplementary Figure 4: Number of Data Deficient species of the spatial dataset of the IUCN Red List of threatened species (Version 2020-3)[1,2] *predicted to be threatened by extinction across the world.*

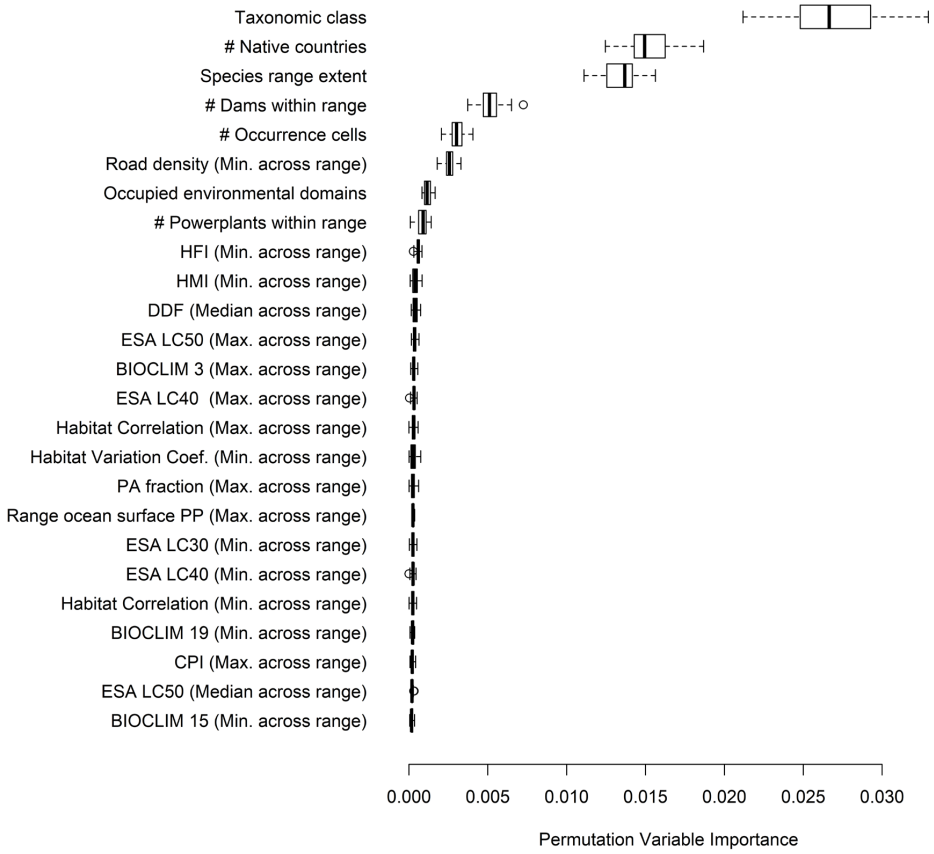


Supplementary Figure 5: Regional differences in average PE score between Data Deficient and data-sufficient species of the spatial dataset of the IUCN Red List of threatened species (Version 2020-3)[1,2].





Supplementary Figure 6: Percent change in average PE score when Data Deficient species are considered along data-sufficient species for all taxonomic groups and all species (n = 44,908) of the spatial dataset of the IUCN Red List of threatened species (Version 2020-3)[1,2].



Supplementary Figure 7: Boxplot showing the interquartile range (box), median (black line), minimum and maximum values without outliers (error bars), and outliers (points) of permutation variable importance for the top 25 variables of the presented classifier, based on performance loss (AUC) during 50 runs of feature permutation. Abbreviations: Human Footprint Index (HFI), Human Modification Index (HMI), demersal destructive fishing (DDF), ESA LC (European Space Agency Land Cover, 30 = Mosaic cropland (>50%) / natural vegetation (tree, shrub, herbaceous cover) (<50%); 40 = Mosaic natural vegetation (tree, shrub, herbaceous cover) (>50%) / cropland (<50%); 50 = Tree cover, broadleaved, evergreen, closed to open (>15%)), BIOCLIM (3 = isothermality; 15 = precipitation seasonality; 19 = mean monthly precipitation amount of the coldest quarter), Protected Area (PA), Primary productivity (PP), Corruption Perception Index (CPI).

*Supplementary Table 1: Predictions across taxonomic groups. Total number of species and number of species (*predicted to be, PE cut-off: 0.388) threatened by extinction across data-sufficient and Data Deficient species for each taxonomic class. Classifier performance (i.e., AUC, as well as Accuracy (Acc.), Sensitivity (Sens.), and Specificity (Spec.)) based on the testing dataset (25%). Note: Performance metrics were calculated for taxonomic classes only if both categories (i.e., threatened vs. not threatened) present in testing data.*

	Data-sufficient		Data Deficient		Performance			
Taxonomic class	total	threatened[2]	total	threatened*	AUC	Acc.	Sens.	Spec.
Actinopterygii	9404	2134 (23%)	1876	880 (47%)	0.93	0.86	0.80	0.88
Amphibia	5801	2268 (39%)	1130	960 (85%)	0.93	0.85	0.93	0.79
Anthozoa	691	227 (33%)	135	98 (73%)	-	-	-	-
Aves	1803	179 (10%)	7	3 (43%)	0.85	0.92	0.40	0.97
Bivalvia	329	72 (22%)	77	37 (48%)	0.83	0.82	0.64	0.87
Branchiopoda	5	5 (100%)	0	0	-	-	-	-
Bryopsida	12	6 (50%)	1	1 (100%)	1.00	0.67	1.00	0.00
Cephalaspidomorphi	27	6 (22%)	1	0 (0%)	0.33	0.75	0.00	1.00
Charophyceae	8	0 (0%)	3	1 (33%)	-	-	-	-
Chondrichthyes	928	309 (33%)	222	58 (26%)	0.85	0.78	0.75	0.79
Clitellata	5	0 (0%)	5	5 (100%)	-	-	-	-
Gastropoda	1839	608 (33%)	551	260 (47%)	0.95	0.90	0.82	0.92
Hydrozoa	14	5 (36%)	2	2 (100%)	-	-	-	-
Insecta	2293	415 (18%)	882	546 (62%)	0.92	0.86	0.85	0.86
Lecanoromycetes	2	1 (50%)	0	0	-	-	-	-
Liliopsida	542	83 (15%)	31	13 (42%)	0.98	0.95	0.93	0.96
Lycopodiopsida	28	15 (54%)	6	4 (67%)	-	-	-	-
Magnoliopsida	542	180 (33%)	55	23 (42%)	0.86	0.79	0.74	0.82
Malacostraca	1380	310 (22%)	832	335 (40%)	0.89	0.75	0.89	0.69
Mammalia	4962	1293 (26%)	818	495 (61%)	0.86	0.83	0.70	0.87
Myxini	46	9 (20%)	29	9 (31%)	-	-	-	-
Polypodiopsida	41	5 (12%)	4	1 (25%)	-	-	-	-
Reptilia	6500	1263 (19%)	1032	605 (59%)	0.89	0.84	0.67	0.87
Sarcopterygii	5	1 (20%)	0	0	-	-	-	-

*Supplementary Table 2: Complete list of data used as correlates in the generated machine learning classifier. *If applicable, variables were generated by retrieving mean, median, minimum, and maximum values of the corresponding layers across species range maps and occurrence cells, or native countries.*

Data	Layers	Variables*
IUCN Red List of threatened species[2]	35	35
Global Biodiversity Information Facility (GBIF)[4] & Ocean Biodiversity Information System (OBIS)[5]	5	5
Climatologies at high resolution for the earth's land surface areas[6]	19	152
ESA Land Cover CCI[7]	17	136
Global terrestrial Human Footprint maps for 1993 and 2009[8]	6	48
Managing the middle: A shift in conservation priorities based on the global human modification gradient[9]	1	8
Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools[10]	1	8
Protected Planet: The World Database on Protected Areas (WDPA)[11]	1	8
High-Resolution Global Maps of 21st-Century Forest Cover Change[12]	1	8
A global, remote sensing-based characterization of terrestrial habitat heterogeneity for biodiversity and ecosystem modelling[13]	14	112
PEST-CHEMGRIDS, global gridded maps of the top 20 crop-specific pesticide application rates from 2015 to 2025[14]	1	8
A Global Database of Power Plants[15]	1	1
GOODD, a global dataset of more than 38,000 georeferenced dams[16]	1	1
The WULCA consensus characterization model for water scarcity footprints: assessing impacts of water consumption based on available water remaining (AWARE)[17]	2	8
FLO1K, global maps of mean, maximum and minimum annual streamflow at 1 km resolution from 1960 through 2015[18]	56	448
Impacts of current and future large dams on the geographic range connectivity of freshwater fish worldwide[19]	2	2
Near-global freshwater-specific environmental variables for biodiversity analyses in 1 km resolution[20]	87	696
The Next Frontier: Human Development and the Anthropocene[21]	2	8
Corruption Perceptions Index[22]	1	4
Global threats from invasive alien species in the twenty-first century and national response capacities[23]	8	56
A Global Map of Human Impact on Marine Ecosystems[24]	18	144
Bio-ORACLE v2.0: Extending marine data layers for bioclimatic modelling[25,26]	161	1288

Supplementary Table 3: Classifier contributions. Weights of all (non-zero) models contributing to the super-learner, i.e., gradient boosted classification trees (GBM) and neural networks (DeepLearning).

Base-learner	Relative Importance	Percentage
GBM_grid_1_AutoML_1_model_45	0.337	0.157
GBM_grid_1_AutoML_1_model_34	0.202	0.094
GBM_grid_1_AutoML_1_model_39	0.194	0.090
GBM_grid_1_AutoML_1_model_25	0.167	0.077
GBM_grid_1_AutoML_1_model_60	0.160	0.075
GBM_1_AutoML_1	0.154	0.071
GBM_grid_1_AutoML_1_model_48	0.140	0.065
GBM_grid_1_AutoML_1_model_31	0.126	0.058
GBM_grid_1_AutoML_1_model_71	0.125	0.058
GBM_grid_1_AutoML_1_model_27	0.100	0.046
GBM_grid_1_AutoML_1_model_5	0.072	0.034
GBM_grid_1_AutoML_1_model_33	0.071	0.033
GBM_grid_1_AutoML_1_model_51	0.063	0.029
GBM_grid_1_AutoML_1_model_10	0.056	0.026
GBM_grid_1_AutoML_1_model_1	0.050	0.023
GBM_grid_1_AutoML_1_model_2	0.038	0.018
DeepLearning_grid_1_AutoML_1_model_35	0.023	0.011
GBM_grid_1_AutoML_1_model_17	0.022	0.010
DeepLearning_grid_1_AutoML_1_model_9	0.017	0.008
GBM_grid_1_AutoML_1_model_4	0.013	0.006
GBM_grid_1_AutoML_1_model_62	0.012	0.006
DeepLearning_grid_1_AutoML_1_model_13	0.009	0.004
GBM_grid_1_AutoML_1_model_46	0.000	0.000

Supplementary References

1. IUCN. Species Information Service. Version 2020-3. https://www.iucnredlist.org/resources/spatial-data-download (2021).

2. IUCN. The IUCN Red List of Threatened Species. Version 2020-3. https://www.iucnredlist.org (2020).

3. IUCN. The IUCN Red List of Threatened Species. Version 2021-2. https://www.iucnredlist.org (2021).

4. Chamberlain, S. *et al. rgbif: Interface to the Global Biodiversity Information Facility API. R package version 3.6.0*. https://cran.r-project.org/package=rgbif (2021).

5. Provoost, P. & Bosch, S. robis: Ocean Biodiversity Information System (OBIS) Client. R package version 2.3.9. https://CRAN.R-project.org/package=robis. (2020).

6. Karger, D. N. *et al.* Data from: Climatologies at high resolution for the earth's land surface areas. *Dryad, Dataset* https://doi.org/10.5061/dryad.kd1d4 (2018).

7. ESA. Land Cover CCI Product User Guide Version 2. Tech. Rep. http://maps.elie.ucl.ac.be/CCI/viewer/download.php (2017).

8. Venter, O. *et al.* Global terrestrial Human Footprint maps for 1993 and 2009. *Sci. Data* **3**, 160067 (2016).

9. Kennedy, C. M., Oakleaf, J. R., Theobald, D. M., Baruch-Mordo, S. & Kiesecker, J. Managing the middle: A shift in conservation priorities based on the global human modification gradient. *Glob. Chang. Biol.* **25**, 811–826 (2019).

10. Seto, K. C., Guneralp, B. & Hutyra, L. R. Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools. *Proc. Natl. Acad. Sci.* **109**, 16083–16088 (2012).

11. UNEP-WCMC & IUCN. Protected Planet: The World Database on Protected Areas (WDPA). *Cambridge, UK: UNEP-WCMC and IUCN* www.protectedplanet.net (2021).

12. Hansen, M. C. *et al.* High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science (80-.).* **342**, 850–853 (2013).

13. Tuanmu, M. N. & Jetz, W. A global, remote sensing-based characterization of terrestrial habitat heterogeneity for biodiversity and ecosystem modelling. *Glob. Ecol. Biogeogr.* **24**, 1329–1339 (2015).

14. Maggi, F., Tang, F. H. M., la Cecilia, D. & McBratney, A. PEST-CHEMGRIDS, global gridded maps of the top 20 crop-specific pesticide application rates from 2015 to 2025. *Sci. Data* **6**, 170 (2019).

15. Byers, L. *et al.* A Global Database of Power Plants. *World Resour. Inst.* 1–18 (2019).

16. Mulligan, M., van Soesbergen, A. & Sáenz, L. GOODD, a global dataset of more than 38,000 georeferenced dams. *Sci. Data* **7**, 31 (2020).

17. Boulay, A.-M. *et al.* The WULCA consensus characterization model for water scarcity footprints: assessing impacts of water consumption based on available water remaining (AWARE). *Int. J. Life Cycle Assess.* **23**, 368–378 (2018).

18. Barbarossa, V. *et al.* Erratum: FLO1K, global maps of mean, maximum and minimum annual streamflow at 1 km resolution from 1960 through 2015. *Sci. Data* **5**, 180078 (2018).

19. Barbarossa, V. *et al.* Impacts of current and future large dams on the geographic range connectivity of freshwater fish worldwide. *Proc. Natl. Acad. Sci.* **117**, 3648–3655 (2020).

20. Domisch, S., Amatulli, G. & Jetz, W. Near-global freshwater-specific environmental variables for biodiversity analyses in 1 km resolution. *Sci. Data* **2**, 150073 (2015).

21. UNDP. *Human Development Report 2020. The Next Frontier: Human Development and the Anthropocene. New York. http://hdr.undp.org/en/content/human-development-report-2020*. (2020).

22. Transparency International. *Corruption Perceptions Index 2020*. (2020).

23. Early, R. *et al.* Global threats from invasive alien species in the twenty-first century and national response capacities. *Nat. Commun.* **7**, 12485 (2016).

24. Halpern, B. S. *et al.* A Global Map of Human Impact on Marine Ecosystems. *Science (80-.).* **319**, 948–952 (2008).

25. Assis, J. *et al.* Bio-ORACLE v2.0: Extending marine data layers for bioclimatic modelling. *Glob. Ecol. Biogeogr.* **27**, 277–284 (2018).

26. Tyberghein, L. *et al.* Bio-ORACLE: a global environmental dataset for marine species distribution modelling. *Glob. Ecol. Biogeogr.* **21**, 272–281 (2012).

SI2 (Supporting Information for chapter 4): Terrestrial ecosystem impacts of biological invasions caused by international transportation within the framework of Life Cycle Assessment

In preparation. To be submitted to Environmental Science & Technology

The main article is awaiting publication and is not included in NTNU Open