More than half of data deficient species predicted to be threatened by extinction

Jan Borgelt, Martin Dorber, Marthe Alnes Høiberg & Francesca Verones

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 7699 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.
Measuring ongoing and anticipating potential threats is vital for preventing damage to the natural world, 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 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 relying for instance on extent of occurrence, area of occupancy, population trends, or population size. However, the sheer amount of known and uncertain species globally, the dynamic nature of threats and trends, and limited human resources for undertaking such Red List assessments 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, unevenly distributed across space, time and taxa. 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”. More specifically Bland et al. identified 8 main justifications as to why species are assessed as DD: uncerfain 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 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. Although DD species are sometimes treated as being not threatened, 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. However, since DD species could belong to any Red List category, they are difficult to handle for practitioners and are therefore generally ignored in studies analysing biodiversity impacts and change. For instance, the Red List Index is built upon well-assessed threat-levels for individual 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 and 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) (Fig. 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, AUROC = 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 (Fig. 1). Binary as well as continuous predictions across marine versus non-marine groups perform well but are not directly comparable.

We further tested our classifier against an IUCN update (Version 2021-2) that was released after our model was trained (Supplementary Fig. 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%).
Fig. 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 (Fig. 2b; Supplementary Fig. 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 Fig. 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 (Fig. 2a; Supplementary Fig. 3).

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 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, reptiles terrestrial mammals, and sharks and rays. Howard and Bickford 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 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. The regions for conservation priorities for both reptiles and amphibians match those previously found, which are congruent with known hotspots for threatened species. A previous assessment for terrestrial mammals 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\textsuperscript{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., 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\textsuperscript{48}, 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 Fig. 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 (Fig. 3a). 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 (Fig. 3a). 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 Figs. 3, 6). In contrast, including DD species did not change or even lowered the average PE score in large parts of international seas (Fig. 3a). Although marine biodiversity as we know it today is richest in coastal waters\textsuperscript{49}, these results should be interpreted with caution because the underlying range maps for many marine species can be too coarse\textsuperscript{50}, which may be especially true for DD species in international seas.

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\textsuperscript{31}. 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 (Fig. 3b). Notably, often only few taxonomic groups accounted for most of the observed increase in average PE score (Supplementary Fig. 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 (Fig. 3b; Supplementary Fig. 6). We further discovered an increase in average PE score in numerous smaller isolated patches distributed

Fig. 2 Potentially threatened fraction of data deficient species. Fraction of Data Deficient species (n = 7699, IUCN Version 2020-3) predicted to be threatened by extinction for marine (a) and non-marine species (b) according to our machine learning classifier.
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, 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 as well as numerically converted threat-levels could be biased. The generated predictions (i.e., PE scores) could facilitate the inclusion of DD species in sustainability-relevant applications and modelling approaches. We encourage the extended use of our algorithm for screening for updates in the status of DS species, as well as large-scale pre-assessments of species not yet evaluated by the IUCN 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, amphitbians, 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) in March 2021, covering 44,924 species in the following taxonomic classes: Actinopterygii, Agaricomycetes, Annelida, Anthozoa, Aves, Bivalvia, Branchiopoda, Bryopsida, Cephalaspidomorphia, Charophyceae, Chondrichthyes, Citellata, Gastropoda, Hydrozoa, Insecta, Jungernanniporida, 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-breeding 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. Habitat preferences were retrieved from the Red List using redlist.r 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) and the Ocean Biodiversity Information System (OBIS) using their corresponding application programming interfaces via the packages rgbif and robis 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...
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 DD species were reclassified into two classes 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\(^{16,38,86,87}\). 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 (marine and non-marin). 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 different strengths and weaknesses\(^{86}\). The best performing data partition (i.e., partition 1; for all species) was selected based on the highest average AUC (see 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\(^ {86} \). 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 imputeR\(^ {84}\), 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\(^ {86}\)). Ultimately, a so-called super-learner\(^ {90}\) was generated using a non-negative generalized linear model with regularization (least absolute shrinkage and selection operator) to produce more sparse ensemble\(^ {90}\), 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 = 6885 species). In addition, we assessed model performance using DD species that have been re-evaluated and assigned a threat category in Red List Version 2023-2 (n = 123 species). In addition, AUC, AUC\(_{PP}\) 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 (Eq. 3) versus false positive rate (Eq. 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\(_{PP}\) as the area under the receiver operating characteristic curve for precision (Eq. 6) versus recall (Eq. 3), is not affected by true negatives (i.e., correctly predicted not-threatened species) but instead affected by how precisely 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 Fig. 7).

Statistics and reproducibility. Analyses were conducted using R version 4.0.3\(^ {91}\) in RStudio version 1.1.453\(^ {92}\). Data were obtained from GBIF, OBIS and IUCN using the packages rgbif, robis, and rredlist\(^ {23,56}\). Handling of spatial and other data was conducted using the R packages carTools, doParallel, exactextact, fastrize, mapTools, parallel, raster, readxl, rgdal, rgeos, sf, stringr, tidyverse, and xlsx\(^ {98–110}\), and in python using the arcpy module from ArcGIS Pro version 2.9\(^ {111}\). Machine learning algorithms were trained and evaluated using the H2O.ai interface (Version 3.36.0.4) for R\(^ {98}\) and caret\(^ {112}\). Figures were created using ggplot2\(^ {113}\), geosphere\(^ {114}\), stringr, tidyr, ggplot2, and base R\(^ {98}\).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability
Previously published and open-access source data were retrieved from ref.\(^ {5,94,115–117}\). All data generated in this study is available without restrictions. The generated predictions for the testing data, complete dataset and updated Data Deficient species are provided as supplementary files (Supplementary Data 1-3). 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 at https://doi.org/10.1386/s42003-022-03638-9. The code is also available in the GitHub repository associated with this paper.
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. Antirr. Rev. 2, 81–98 (2015).

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

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

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

7. IPBES. Summary for policymakers of the global assessment report on biodiversity and ecosystem services. Berlin: IPBES (2019).

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

14. Rondonini, C., Di Marco, M., Visconti, P., Butchart, S. H. M. & Boitani, L. Update or outdated: 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. Towards 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 “Assumed 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., Coughram, C., 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).
