Understanding the adequacy and representativeness of species distribution data

Alice C. Hughes[1*], Nazli Demirel[2], David K.A. Barnes[3], Ina Helene Ahlquist[4], Ian Ondo[5], Robert Guralnick[6], Tim Hirsch[7], Brian Enquist[8], Cory Merow[8], Kristin Kaschner [9], Gabriel Reygondeau[10], Yulia Egorova[10], Michael Orr[11], Huijie Qiao[12], P.J. Stephenson[13], John Waller[14], Neil D. Burgess[5,15]

1. School of BioSciences, University of Melbourne, Australia. Alice.C.Hughes@unimelb.edu.au
2. Institute of Marine Sciences and Management, Istanbul University, Fatih 34134, Istanbul, Türkiye
3. British Antarctic Survey, High Cross, Madingley Road, Cambridge, CB3 0ET, United Kingdom.
4. SINTEF Ocean, Brattørkaia 17C, Trondheim N-7010, Norway.
5. UN Environment Programme World Conservation Monitoring Center (UNEP-WCMC), 219 Huntington Road, Cambridge, United Kingdom.
6. Florida Museum of Natural History, Gainesville, FL 32611, United States
7. Tim Hirsch Consulting, London, 53 Elstead House, London SW2 3LT, United Kingdom; *(TBC)* Secretariat of the Convention on Biological Diversity, 413 Saint-Jacques Street, Suite 800, Montreal, Quebec, H2Y 1N9, Canada.
8. Eversource Energy Center and Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT 06269, United States
9. Department of Biometry and Environmental System Analysis, University of Freiburg, Freiburg, Germany.
10. University of Miami. Rosenstiel School of Marine, Atmospheric, and Earth Science. Miami, FL 33149. United States.
11. Entomology, State Museum of Natural History, Stuttgart, Germany.
12. State Key Laboratory of Animal Biodiversity Conservation and Integrated Pest Management, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China.
13. IUCN SSC Species Monitoring Specialist Group, Laboratory for Conservation Biology, Department of Ecology & Evolution, University of Lausanne, CH-1015 Lausanne, Switzerland
14. Global Biodiversity Information Facility (GBIF), Copenhagen, Denmark
15. Center for Macroecology, Evolution and Climate, Globe Institute, University of Copenhagen, Copenhagen, Denmark.

Key words: Species distributions, occurrence data, data gaps, Conservation prioritisation, indicators

Abstract

Species occurrence data is the fundamental unit of any species distribution analysis, biodiversity patterns, species extinction vulnerability, and temporal trends. This data is also a critical component of monitoring progress towards global biodiversity targets, such as those included in the Kunming-Montreal Global Biodiversity Framework (GBF) of the Convention

on Biological Diversity (CBD). Recent years have seen massive growth and digitisation of global species occurrence datasets, yet the headline indicators of the GBF's monitoring framework relies on the IUCN RedList index as the species conservation indicator. This paper explores the largest global species distribution databases, and outlines some of the remaining challenges to bringing these data together for enhanced decision making. Countries such as Japan and South Korea have seen dramatic expansions of data coverage, whilst North Africa, Central Asia, and the High Seas have not witnessed comparable growth. In the oceans, expanding geographic coverage partially comes from tracking data from a small number of species. In terms of environmental space, other types of sampling effort are disproportionately concentrated on the temperate continental shelf and slope areas of the north Atlantic. Most of the world's most biodiverse areas, especially in the Tropics, both on land and in the ocean still lack data, and concerted efforts will be needed to improve the coverage of these regions. Furthermore, few long-term monitoring programs exist, and accurately inferring change from small numbers of unstandardised collecting events resulting in a large quantity of uncurated data is challenging. Among other measures, journals should request standardised data be added to the key repositories to address biodiversity data gaps and barriers to usability. Governments should also support the sharing of species occurrence data through standardised data infrastructure, such as GBIF and OBIS, and also ensure support for data curation and quality control to minimize impacts of species misidentification records.

Introduction

We are currently experiencing a global biodiversity crisis, marked by many species declines across habitats and realms (IPBES, 2019). Understanding and attempting to mitigate declines relies heavily on our knowledge of when and where species exist. Species occurrence records represent an essential source of data on species distributions, and may in some cases help establish relative abundances and trends. Such information can be used to guide planning and decision-making across ecological and socio-economic domains, including conservation and sustainable resource management in sectors such as forestry, agriculture, and fisheries. However, the ability to discover, use and re-use datasets is severely compromised by the variety of platforms used to deposit them, a lack of clarity on how data in one system relates to that in another, and the lack of standardisation of data across these platforms (Bayraktarov et al., 2019; Cornford et al., 2022; Marques et al., 2024). Many data sources also fail to define clearly what data are freely and openly available and accessible; the origins of the data, and the length of any time series available, can also be complicated to find on some data platforms (Stephenson & Stengel, 2020). However, dedicated platforms have developed standards to maximise their usability, vastly enhancing the capacity to use and analyse such data, which may not be applied to data stored on more general platforms (Guralnick et al., 2018; Ingenloff et al., 2025). Furthermore, comparable standards for other kinds of biodiversity data are either weakly developed, or absent (see Gonzalez et al., 2025).

Trying to understand the availability of primary biodiversity data (occurrence records and species populations) entails two broad approaches. Firstly it requires analysis of the content of dedicated platforms such as the Global Biodiversity Information Facility (GBIF) and Ocean Biodiversity Information System (OBIS), which themselves aggregate species occurrence data

hosted and shared by a wide variety of institutions, projects and networks. Also, although mainly known for its expert-informed distribution ranges and associated data on threat status and extinction risk, the IUCN Red List database also contains point locality data which is used for some assessments. Secondly, dedicated research efforts that collate studies and programs that analyse biodiversity change over time or space, such as BioTIME, PREDICTS, and the Living Planet Database (LPI) (Dornelas et al., 2025, Hudson et al., 2017; WWF 2024). Understanding the strengths, and current gaps and challenges in using this data may enable the refinement of data and metrics to enhance our ability to monitor diversity at all scales.

Thus, understanding how representative these species occurrence datasets are, identifying their various gaps, and the limitations of their sensible use is critical, not only to inform current analysis, but to prioritise and focus efforts to mobilise new effort to generate data on taxa, regions and ecosystems which are currently least represented. This is especially important when these datasets form the foundation for many species range analyses or modelling outputs, which are used in vast numbers of scientific papers, and also in national to global conservation planning, national to global policy, and work with the business and finance sectors.

Recent years has seen an exponential growth of data, but if this growth overcomes previous biases and gaps (i.e. Hughes et al., 2021; Troudet et al., 2017), and how current data may contribute towards tracking progress to global biodiversity targets remains a question. As data availability has increased exponentially in recent years, understanding where this growth is taking place provides insights on those regions and taxa that are currently under-represented and those that might be expected to sufficiently improve in their data representation and availability based on their present trajectory. This enables two things, firstly the identification of areas where successful growth of data may have transferable lessons for other regions, and secondly the identification of areas, and environmental space, where further efforts are clearly needed to mobilise data, which is critical to provide the basis for proactive conservation planning (i.e see Merow et al., 2025; Feng et al., 2022; Meyer et al., 2015). Furthermore, the representativeness of new data is also a key factor, as single species programs (such as tracking) may fill spatial data gaps, without providing representative data for monitoring.

Here we explore four species occurrence datasets (GBIF, OBIS, BIEN, IUCN), and also include a supplemental analysis of five population or community datasets, and assess how representative they are across space and taxa, and how well they track changes in biodiversity over time. We do not include datasets that include modelling or transformation of the data, species range / polygon data (GARD; Roll et al., 2017, AntMaps; Janicki et al., 2016) as the accuracy of such data is defined by the existence and representativeness of primary species distribution datasets.

We identify some gaps within each dataset, and where those gaps are closing. Furthermore, we assess the relationship between the largest of these databases (GBIF) with basic metrics of intactness and accessibility following the example of Hughes et al. (2021) to assess how representative data is on a gradient of disturbance, highlighting the preponderance of data in highly accessible and more disturbed areas. Finally, we provide recommendations for an improvement of how data is collated to improve the availability of adequate data to track global biodiversity targets.

Methods

We present the major global species occurrence databases in the world, and summarise coverage of data, and how coverage has changed over the last decade (largely based on GBIF data). In addition, we explore how changing the resolution of analysis impacts the perceived regional coverage based on OBIS data. Finally, we conducted additional analysis on spatial biases in GBIF in relation to distance to infrastructure and environmental intactness (based on the human modification index) (Supplementary Methods 1), and taxonomic and spatial representativeness and coverage of other monitoring datasets (PREDICTS, Living Planet Index (LPI), BioTIME and others) (Supplementary Methods 2). All spatial analysis was conducted using an Equal Areas Projection.

Primary distribution datasets

1. GBIF

GBIF-mediated data came directly from the GBIF Secretariat, including a table with the number of occurrence records and species available in a snapshot of GBIF-mediated content for each year (between 2008 and 2025), for each country, and for selected taxa (full list of "taxa", generally at Class level are provided in supplements). Global GeoTIFFS of point density for 2015 and 2025 at a 5km resolution were also provided (GBIF 2025).

Based on the table, summaries of points per year for each group were aggregated for various groupings including by intersecting realms and UN regions (to add refinement to large and heterogeneous regions, such as "Asia" and better reflect both biotic variation and geological regions). The areas are also comparable to IPBES subregions, though some sub-regions have been merged (South and Central America- Latin America; East Africa, Central, Southern Africa, West Africa- Sub-Saharan Africa; North, North-East and Central Asia- Central-North Asia; IPBES 2021).

For mammals, birds, reptiles and amphibians, we also calculated the number of species recorded by IUCN for each of these taxa at a national level, to compare to the number in GBIF, to gain some understanding of "potential completeness" in terms of the number of species recorded in GBIF and the number of species recorded for each country by IUCN (downloaded from the Red List website, selecting the appropriate taxa and downloading the shapefiles, similar to the approach applied by Oliver et al., 2021). Richness from IUCN was calculated using the "count overlapping polygon" toolbox in ArcMap. Taxonomic coverage was also explored using both the table, and directly via the GBIF portal to provide insights when a single year saw a large increase in records in certain regions, and to explore the contribution of citizen science vs traditional forms of data collection.

For areas covered, we used the 5 km resolution GeoTIFFS to assess the percentages of various "zones" with at least one record per 5 km gridcell. This was also repeated using higher numbers of records per cell (10, 25, 50, 100 etc) to gauge the level of coverage for these designations. Designations included countries, biogeographic realms and ecoregions for terrestrial areas, and

for oceans was combined with OBIS data to calculate the percentage of Exclusive Economic Zones (EEZs - these were defined by MRGID designations, as well as ABNJ) and Longhurst regions (Longhurst, 1998) that had data. These designations were selected to provide a useful means to reflect coverage based on both biotic elements (ecoregions/realms, Longhurst regions) and geopolitical entities for management.

Assessments were made based on an equal area projection, the percentage of each designation for 2015 and 2025 which had data (based on the dimensions above) was calculated, and the change between the two mapped. All calculations were conducted in ArcMap 10.8 (ESRI). Ecoregions, realms and biomes were downloaded from the 'Resolve' 2017 dataset (Dinerstein et al., 2017), EEZs were from Flanders Marine Institute (2024), Longhurst regions were from MarineRegions (2025), all spatial data used is listed in Table S1.

2. OBIS

The methods applied to GBIF data were repeated for OBIS data, based on data provided by OBIS in GeoTIFF form. This data was analysed in two ways, firstly it was aggregated to a 5km resolution and combined with GBIF to provide a binary map (data or no data) for the oceans. In addition, analysis was conducted at approximately 1 km, 2 km, 5 km, and 11 km to calculate how the percentage of each "zone" with data varied (based on at least one point per grid) as cell size increased.

Designations (EEZs plus ABNJ, Longhurst regions) were used to assess the percent of each region covered with data (at least one point per cell). In addition, the coverage of coral-reefs with data based on high resolution data (0.01^{o}) was assessed by downloading coral-reef data from UNEP-WCMC (2022), and assessing the coverage of points within reefs (only the highest resolution of data was used for this due to the size and dimensions of reefs). All assessments were made in ArcMap 10.8, based on an equal areas projection. In each case, data was classified to binary (data or no data) and the "tabulate area" tool used to calculate the area covered within each designated zone (EEZs, Longhurst regions etc). In addition, the distributional patterns of various representative taxa (sharks and rays, tunicates, sea cucumbers) were mapped to compare to assumed patterns of diversity based on IUCN Red List of Threatened Species assessments) for those groups. For the comparisons we downloaded shapefiles from each taxa, then used the count overlapping polygon tool in ArcMap 10.8) to map presumed richness vs sampling intensity.

3. BIEN

BIEN focuses solely on plant data and has 269,434,901 samples after cleaning and standardisation, which has been used to map ranges of 289,743 species (though 112,953 are in the Americas) of 350,000 extant plant species. Whilst BIEN does draw on GBIF data (among other sources) it applies its own taxonomic backbone, and does not share data with GBIF (Feng et al., 2022; Enquist et al., in review; Maitner et al., 2018; Feng et al., 2022). The Botanical Information and Ecology Network (BIEN) database integrates over 284 million records for land plants from herbarium specimens, ecological plots, citizen science, and trait databases into

a centralized integrated geospatial database, enabling large-scale biodiversity research. The BIEN workflow has used the primary distribution data to map the distributions of ~290,000 plant species (out of approximately 350,368 species known globally; Antonelli et al., 2023). The core BIEN tools for data cleaning include the Taxonomic Name Resolution Service (TNRS) to standardize synonyms and correct spellings (Boyle et al., 2013); Geographic Name Resolution Service (GNRS) to standardize political names and check whether coordinates fall within the finest political unit specified; Geocoordinate Validation Service (GVS) for removing erroneous coordinates including political centroids and herbaria locations (Boyle et al., 2022); and Native Species Resolver (NSR) to filter records out associated with introduced species, typically based on regional checklists or floras (Boyle et al., 2024). Taken together, this suite of corrections and filters leaves 56% of the entire database, or 159,189,390 unique species occurrence records, available for downstream analyses (Enquist et al., in revision).

4. **World Checklist of Vascular Plants**

Checklist-based distribution datasets, compiled from floras and national/regional checklists, determine species' geographic presence at coarse resolution and provide a complementary source of information to point locality databases for plants. Among these, the World Checklist of Vascular Plants (WCVP) (Govaerts et al 2021), curated by the Royal Botanic Gardens, Kew, records accepted taxonomy and distribution across the 369 'botanical countries' of the World Geographical Scheme for Recording Plant Distributions (WGSRPD) (Brummitt 2001). The resource is disseminated via the Plants of the World Online (POWO) portal (https://powo.science.kew.org/) and as versioned GBIF checklist snapshots (Govaerts et al. 2025). The latest versions available on GBIF indicate ~98% taxonomic alignment between the GBIF and WCVP backbones, suggesting good prospects for reconciling GBIF records with WGSRPD-based ranges. Additional global resources with country/region-level coverage include the Global Inventory of Floras and Traits (GIFT) database (5) (≈5,169 checklists across ≈3,400 regions), which also uses WCVP and can provide less spatially biased regional species lists than GBIF (Weigelt et al. 2020), and the Botanic Gardens Conservation International's (BGCI) GlobalTreeSearch database (https://tools.bgci.org/global_tree_search.php) (≈60,000 tree species with country distributions). Together these sources provide complementary, curated range information where WCVP/WGSRPD might be too coarse or point data are absent.

We have used the point data from BIEN in conjunction with the WCVP botanical country checklists to evaluate the sampling completeness of BIEN occurrence data by assessing the degree to which it aligns with the floristic diversity expected from WCVP, and to help guide collection and digitisation priorities to fill in gaps in plant diversity knowledge. We have not undertaken a comparable analysis between IUCN red list, BIEN and WCVP databases.

We compared species occurrence records compiled and curated by BIEN with species richness estimates derived from the WCVP database. BIEN occurrence records described previously were consolidated into a *parquet* file containing species identity (binomial), latitude, and longitude coordinates. To enable spatial aggregation, we employed the World Geographic Scheme for Recording Plant Distributions (WGSRPD, level 3, hereafter referred to as "botanical countries"), provided through the *rWCVP* R package (Brown et al. 2023). These geographic distribution units were rasterized at a resolution of approximately one kilometre at the equator to facilitate spatial overlay with occurrence records.
All BIEN records were then spatially assigned to botanical countries by intersecting geographic coordinates with the rasterized WGSRPD layer of botanical countries. For each botanical country, we calculated the number of occurrence records and the number of unique species represented in BIEN.

To derive reference estimates of species richness, we used WCVP data as implemented in *rWCVP* and its companion package *rWCVPdata*. Species names were filtered to retain only accepted taxa at the species rank, and distributions were extracted while excluding extinct or doubtful records. Botanical countries were assigned both total species richness and native species richness values, thereby providing reference baselines against which BIEN data could be evaluated.

Comparisons between BIEN and WCVP were conducted at the botanical country level. Differences in richness were quantified both as absolute discrepancies in species counts (expected species richness from WCVP minus species richness obtained from BIEN occurrence records) and as relative deviations from WCVP counts, with relative change expressed as the proportion of BIEN species counts relative to WCVP estimates as follows for botanical country i:

$$\%\Delta i = \frac{BIENcount, i - WCVPcount, i}{WCVPcount, i} \times 100$$

5. IUCN red list of threatened species

The IUCN Red List assessment process requires a map of species distribution as part of the species extinction vulnerability assessment, but data sources used may vary. To analyse this, we downloaded the IUCN Red List data from the online data portal (IUCN 2025). We then assessed the sources of data used to map species ranges, focusing on the availability of point locality data (see Hughes et al., 2024). Summaries were made for each major kingdom where IUCN Red List assessments have been made, in addition to a further analysis of the phylum Chordata given their extensive coverage within the Red List.

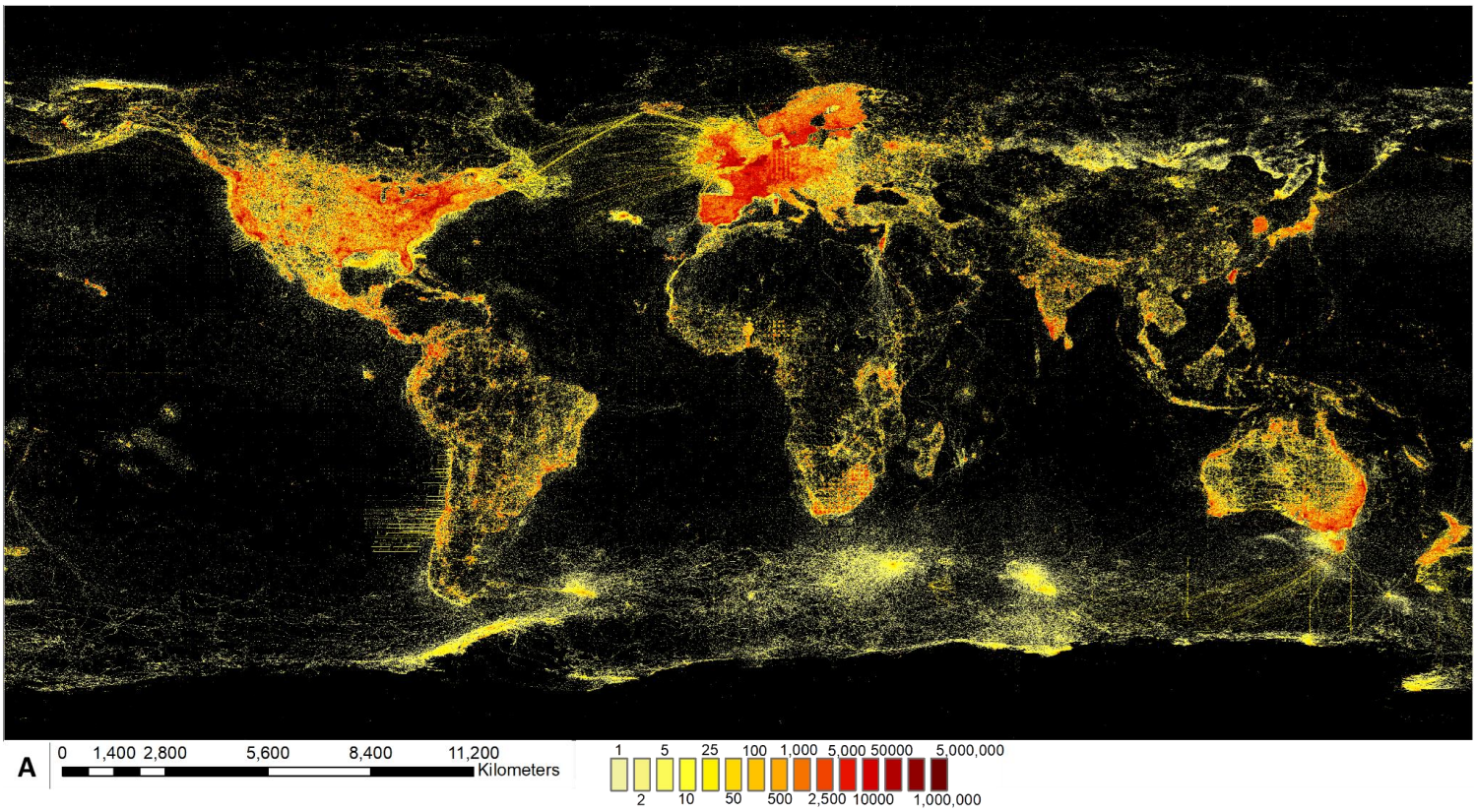
Results

In this paper, we focus our analysis on species occurrence and species diversity and population trends in databases. Additional descriptions of taxonomic biases in GBIF data are provided in Supplementary results S1, spatial biases in Supplementary results S2, and detailed descriptions of other monitoring datasets in Supplementary results S3.

Species occurrence data

GBIF and OBIS

GBIF is the largest global repository of species occurrence data, with 3,133,420,793 records as of July 2025. Of these, 2,013,645,678 are birds (64.3% of all records, due largely to the contribution to GBIF from eBird, which makes up 48.3% of all records, and 75% of all bird records). In terms of coverage, the global spatial coverage has increased dramatically in recent years going from 6.19% of the planet covered in 2015 to 15.52% in 2025 at a 5km resolution (Figure 1, Figure S1A-B). Overall, 78 countries have under 50% of their area represented according to GBIF, and 34 of these have under 25% covered, with only one of these showing a substantial increase (>20% growth) between 2015 and 2025. However, unsurprisingly terrestrial areas were better covered (11.84% to 25.28%) than marine (3.42% to 10.72%) based on GBIF data, though it should be noted that at higher resolutions percentage coverage will decrease (Hughes et al., 2021).



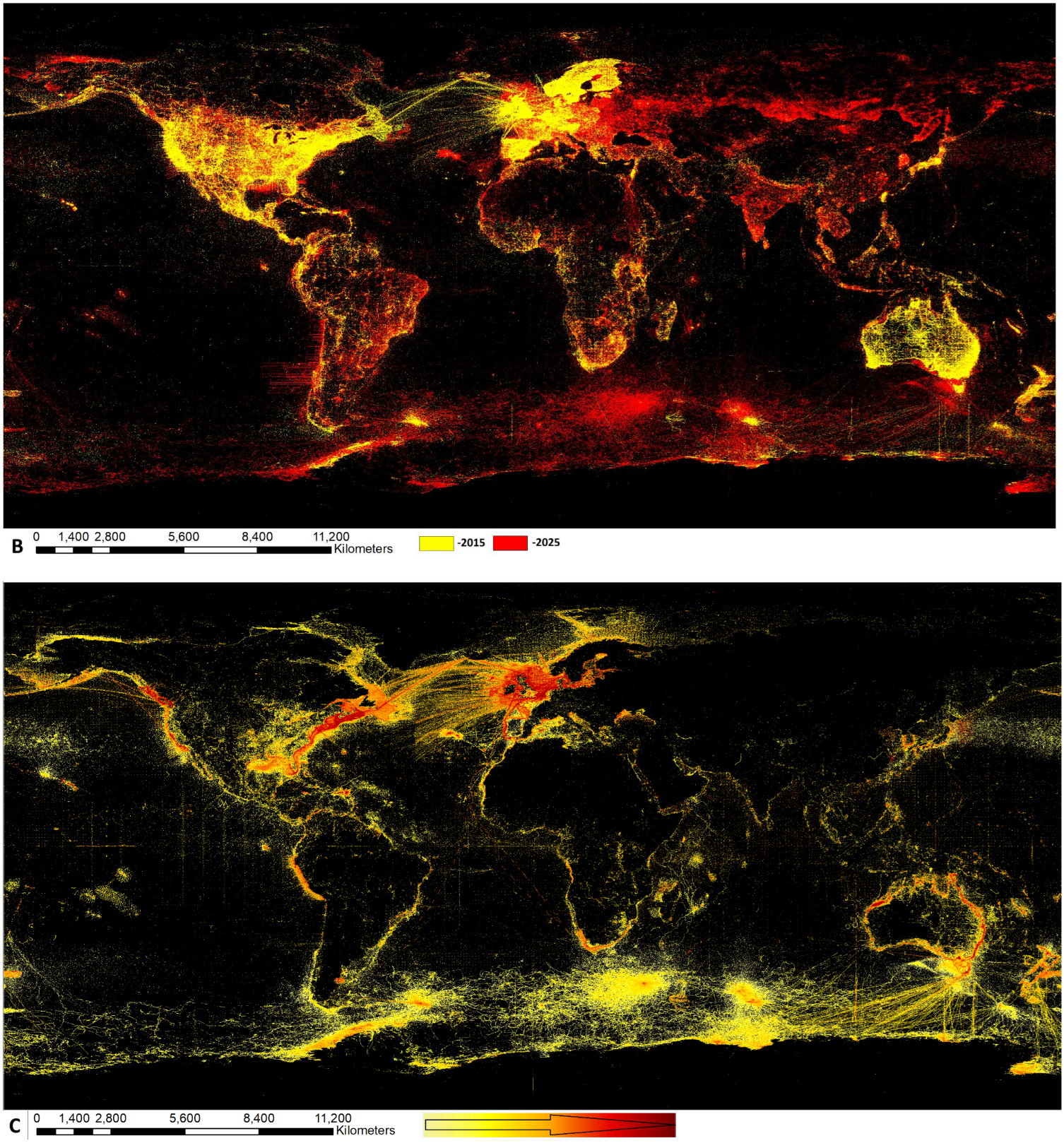


Figure 1. A. *GBIF data density based on data available in 2025, for each year see Figure S1A-B.* ***B.*** Growth of data over a decade in GBIF, Turquoise shows data from 2015, blue shows the additional data growth within GBIF between 2015-2025. For growth of data over time in different regions see Figure S2. **C.** Density of OBIS data at 0.1 degree.

In terms of the national land area covered within GBIF, 62 countries/administrative areas have areas of under one degree (111 km^2) thus assessing the coverage as a percentage may not be indicative. Some parts of Europe and other high-income regions (for example, Scandinavian countries and the UK) had very good coverage (Figure S1C-D), with many reaching 100% by 2025. In addition to looking at absolute coverage, understanding where increases in coverage have occurred in recent years, especially for areas with low coverage, provides an indication of where we may expect to see improvements in the coming years. Some higher-income countries saw dramatic improvement in coverage over the decade: for example, New Zealand increased from 60 to 99%, Japan increased from 44% to 95% coverage, and South Korea increased from only 26% to almost 100% coverage. Conversely, Antarctica has the lowest coverage on the land at 0.11% in 2015, and only expanded to 0.28% by 2025; likewise, Greenland and Libya also have under 5% of their area covered in 2025, with a further eight countries (all in Africa and Central Asia) currently having below 10% covered. It is also important to note that even though the percentage of the area covered for many countries is increasing, 28% of all cells with data (at a 5 km resolution) have only one point, 51% of cells with data have five or fewer points, and 75% have fewer than 50 points per 5 km cell. At a regional level high-income regions such as North America and Europe (in the Neartic, Palaeartic regions) have shown exponential increases in the data included in recent years (Figure S2, Figure S1E-F). The Nearctic has slightly more points overall (but this is primarily driven by birds having the most records), but the Palaearctic has more data for many other taxa.

At the ecoregion level (Figure S1), North America and South-East Australia had the best coverage, followed by European ecoregions (Nearctic 44%-66%, Oceania 71-96%, Australasia 53-65%). Conversely, dry and arid ecoregions in North Africa and Central Asia and ice-bound ecoregions had the lowest, followed by semi-arid and then tropical ecosystems (temperate systems were the best covered; Figure S3). However, changes in coverage are very dynamic,e.g. the Indomalayan region had the second lowest coverage in 2015 (13%) but this had increased to 47% by 2025 (improving its “global rank” relative to many other tropical regions). In contrast, the Afrotropical region was the fourth worst covered in 2015 (19%), but has become the second worst covered in 2025 (37%), whilst Antarctica remained least covered. For biomes, overall tundra regions had the some of the lowest coverage (7%-16%), followed by boreal (8%-21%), then desert and xeric (19-35%) and tropical grasslands (19%-37%), whereas Mediterranean forests had 84% covered by 2025. More broadly, assessment of biomes by region data shows considerable differences in coverage even within a single biome (Figure S3).

Taxonomic coverage

Across taxa, there have been marked increases both in species recorded and in total occurrences. For most regions, the last decade has seen an exponential growth in records across taxa. Yet both the highest total numbers and greatest increases have been in high-income economies; for example, the United States has 26% of all records for reptiles whilst hosting under 5% of described species. Assessing inventory completeness at a national level for different taxa without first filtering point localities to remove non-native species is challenging (due to alien species, captive species, incorrect georeferencing), despite this the taxonomic

completeness of GBIF data varies by country and taxa. However, at least 43 countries have at least 25% of amphibian species not represented by point data (based on comparing counts of species within GBIF with native species recorded according to the IUCN), as well as 34 countries with at least 25% of mammals unrepresented, 17 with reptiles under-represented, but only 5 with birds under represented. In addition, major gaps exist for terrestrial mammals in Asia (at least 55.6% of countries lack records for 25% of species), in Oceania 86% of countries lack records for at least 25% of amphibian species (see Supplementary results 1 provides further details, Data S1). In marine taxa, hotspots for sampling in most groups were also inconsistent with hotspots for species richness (Data S2).

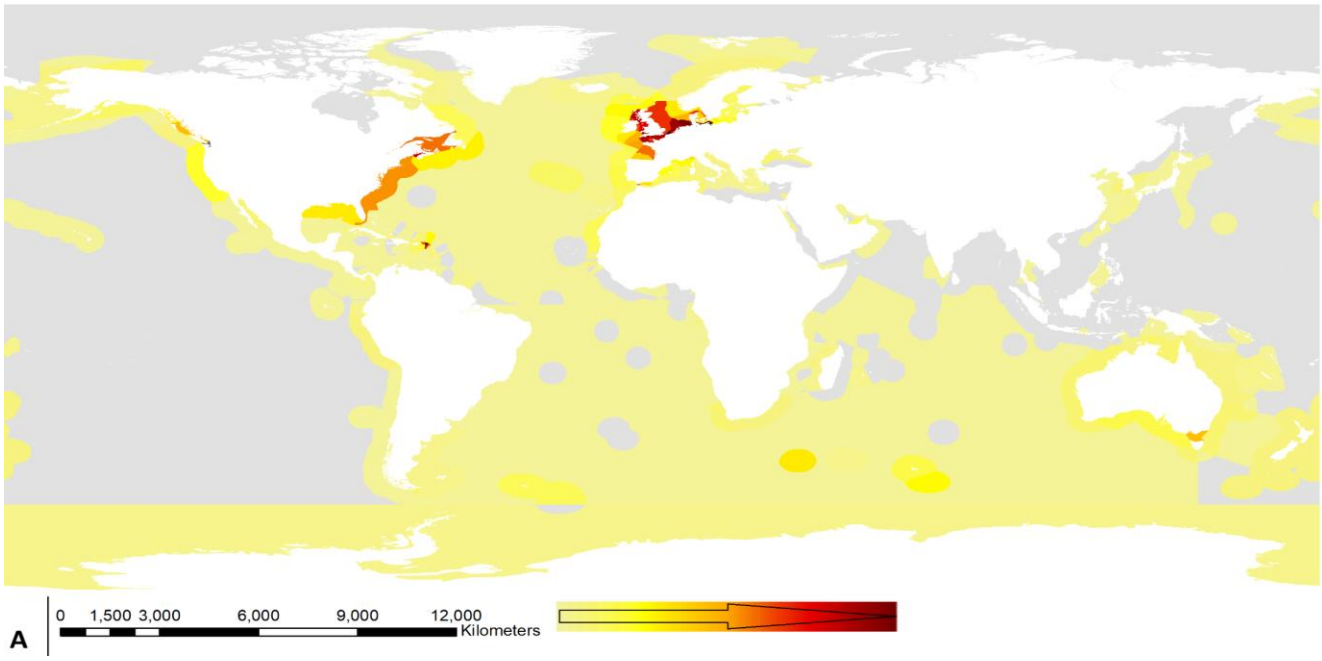
In marine systems, using the AquaMaps 2.0 framework (Reygondeau et al., in review) and World Register of Marine Species (WoRMS) taxonomy, a total of 205,627 marine species were identified along with compiled occurrence data from OBIS, GBIF, and internal AquaMaps sources. Approximately 53% of these species have less than 5 occurrence records. Data gaps are particularly pronounced in groups (in this case phyla) such as Nemertea, Platyhelminthes, and Ctenophora, where about half of the species had no recorded occurrences. In contrast, coverage is much stronger for marine reptiles, fishes, and Porifera, with fewer than 5% of species missing records, and for marine mammals, where all species are represented by at least one occurrence. For marine species with occurrence data, AquaMaps 2.0 pipeline (Reygondeau et al., in prep) also provides quality flags that indicate whether an occurrence point is within the species' range or erroneous. The proportion of verified occurrences is highest in mammals (86%), reptiles (69%), and fishes (62%), moderate in many invertebrate groups (about 25–45%), and very low in nematodes (20%) and platyhelminths (3%).

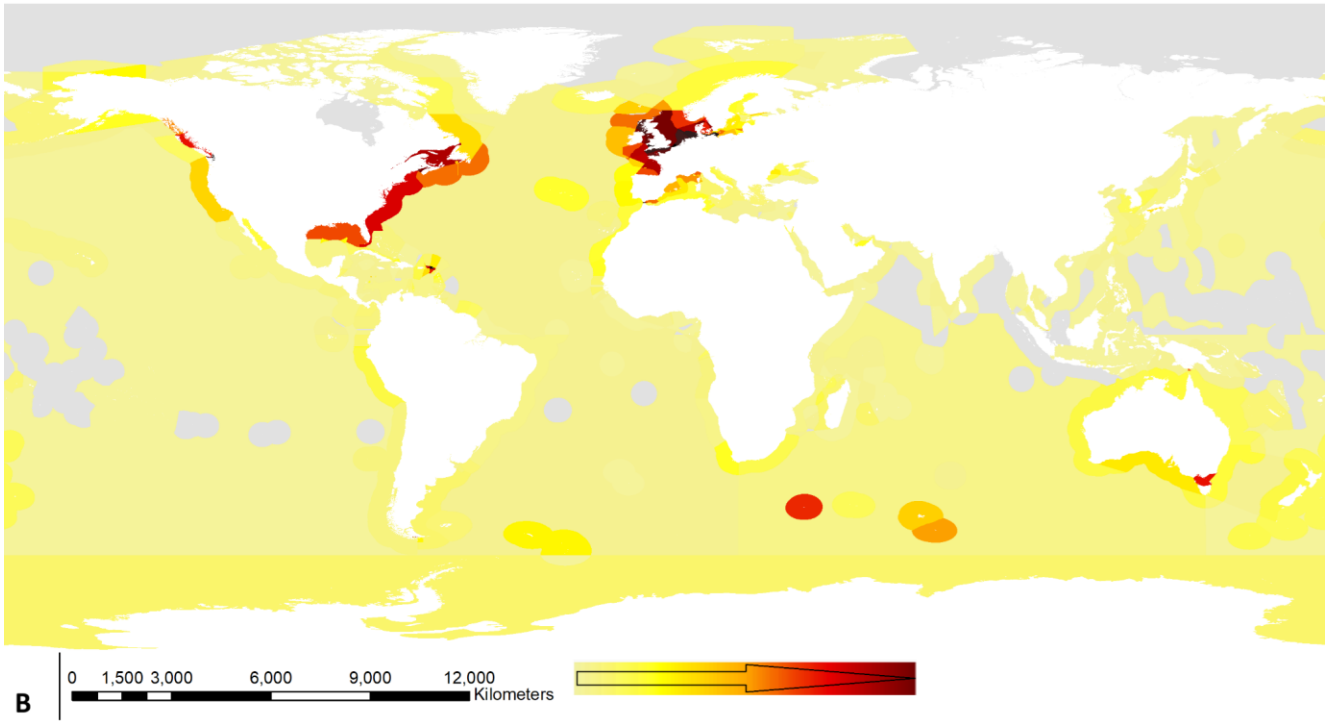
Marine data patterns

To assess the coverage of marine systems GBIF and OBIS data were combined and coverage calculated overall, for EEZs and Longhurst regions (Figure 2, Figure S4). Whilst the data largely overlaps (the overall objective of GBIF-OBIS collaboration is for marine data to be published simultaneously into both networks), differences can arise for a variety of reasons, based on varying workflows and practices among contributing data publishers. Combining OBIS and GBIF data at a 5km resolution covers 19% of the world's surface and 13% of the world's ocean with at least one species occurrence record (Figure 1C). However, the marine component of this data is collected predominantly around the coasts, especially around Europe, North America, and some parts of Australia and New Zealand. Around the European coastline, the Longhurst regions have a coverage of occurrence records up to 87% at a 5km resolution, whilst high-sea regions only have a coverage of 2-4% (for different parts of the high-sea). Hotspots of data density fall within the same regions (Europe, coastal US and Australia), especially to the North of Europe. EEZs have an even higher percentage coverage around certain coasts (especially around Europe), up to 100% in some regions (Figure 2, Figure S4). Conversely, many small island developing states have very little data. For some of the most diverse marine biomes (coral reefs) only 11% of the area had data. In addition, hotspots for

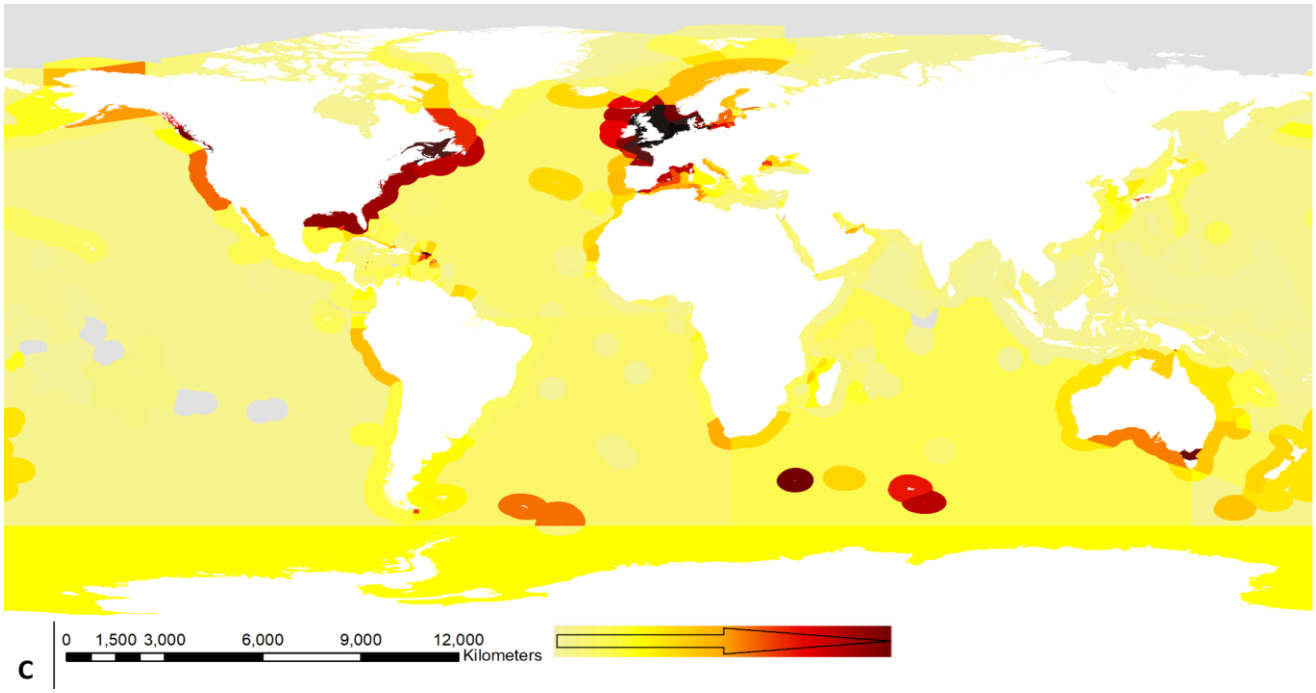
sampling for most groups examined did not correspond to hotspots for species richness (see Data S2).

The proportion of each "zone" with data varied depending on the resolution used. At a 1km resolution, only 1% of the marine areas have data within OBIS. This increases to 3% at a 2 km resolution, 9% at a 5km resolution and 25% at an 11km resolution. Most of the worlds' oceans (especially around Small Island Developing States (SIDS), in tropical regions and the High Seas), have very little data, and EEZs around Southeast Asia had under 1% of data coverage at higher resolutions. Densities of data also matter, and even at 11km around 81% of cells have no data, whilst 6% of cells have only one record (30% of all cells with data), 11% of cells have 1-5 records (60% of cells with data), and 16% have 1-50 points (86% of cells with data, whilst 14% have more). At a higher resolution (1km, which is typically used for distribution modelling), these patterns become more extreme as not only do 99% of cells have no occurrence records, but 51% of cells with data only have one point, and 93% of cells with data have 50 or fewer points. When considering the data needs for single species, this lack of data presents a major challenge to accurate modelling. In the Southern Ocean, whilst areas appear to have high coverage, much of this is tracking data of small numbers of species (mainly Southern elephant seals through Scientific Committee on Antarctic Research, as well as some data on King penguins). Understanding the gaps and representativeness of this data is crucially important for monitoring, especially in the light of the entry into force of the UN Agreement on Marine Biological Diversity of Areas beyond National Jurisdiction (the BBNJ Agreement).




0 1,500 3,000 6,000 9,000 12,000
Kilometers

B


0 1,500 3,000 6,000 9,000 12,000
Kilometers

C

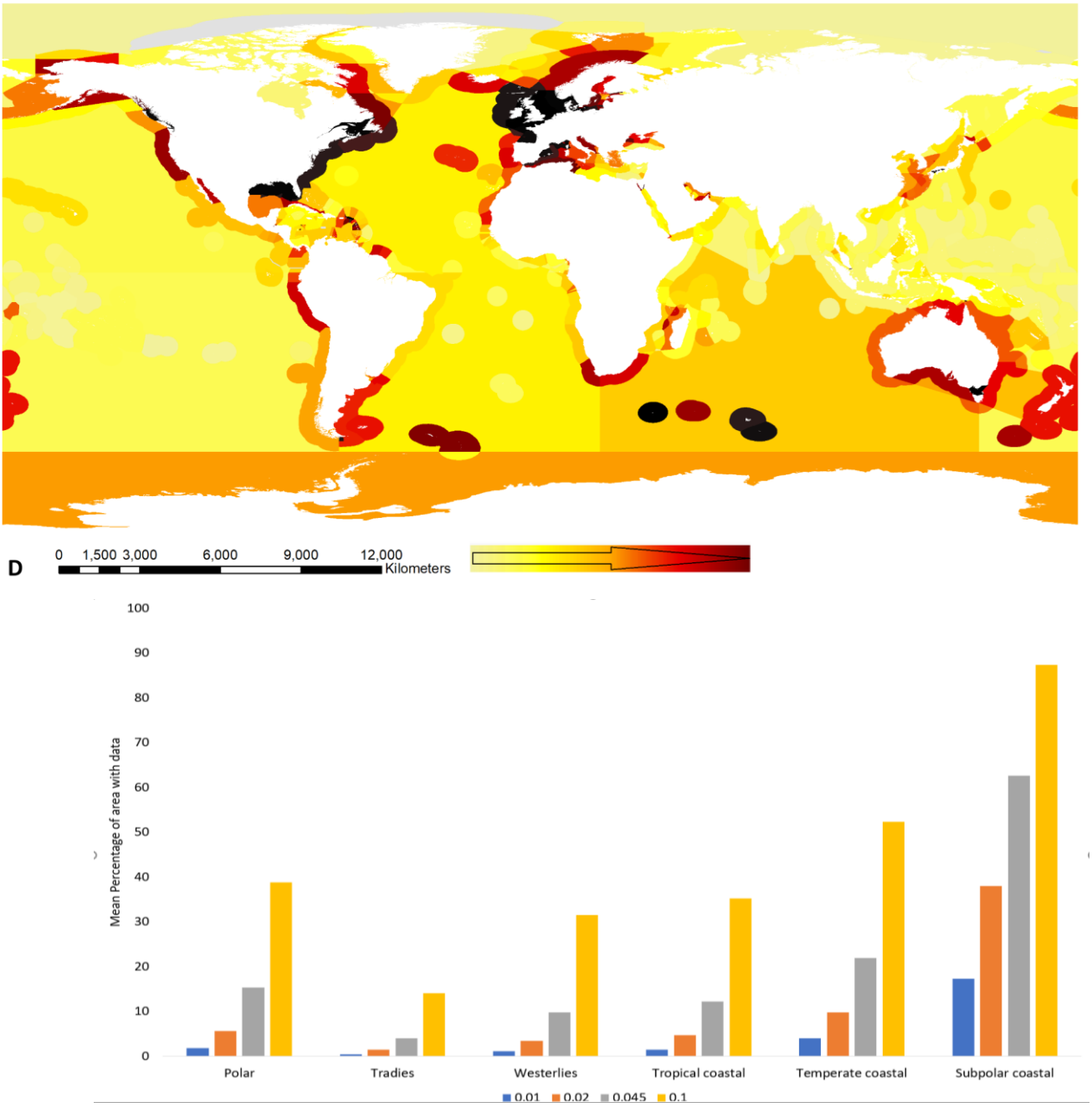


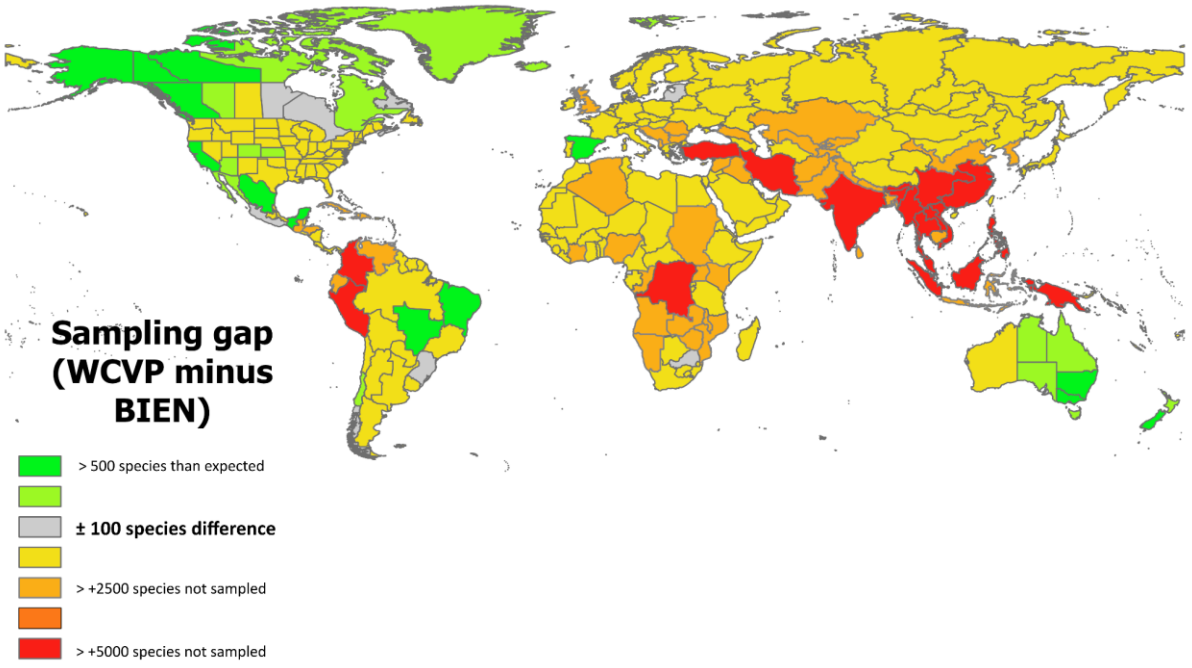
Figure 2. Percent coverage of OBIS point data in the ocean for EEZ (With at least one point per cell), all scales go to 100%. Gray areas have under 1% data coverage. Resolutions A - 0.01^{O}, B - 0.02^{O}, C - 0.045^{O}, D - 0.1^{O}. E. Mean percentage cover per ecotype at each of the four resolutions. See Figure S4 for the equivalent analysis for Longhurst regions, and Figure S1 for terrestrial regions, whilst Figure S3 shows coverage of terrestrial biomes in GBIF.

BIEN (Botanical Information and Ecology Network)

Before a standardised taxonomic backbone was applied, BIEN compiled 1,323,320 unique Land Plant names from 284 million botanical observation records (Figure 3). However, 65-

73% of names have issues requiring correction or cleaning, and the final cleaned set includes 159,189,390 (55.96%) passed all accuracy checks. This includes 323,377 - 404,043 Angiosperms, but far smaller numbers of other plant groups (i.e. 12,102 - 20,921 polypodiophytes [Ferns], 26,808 - 31,556 bryophytes). Furthermore, of plants included, only 36% are assumed native, and at least 26% are introduced.

In terms of spatial representation, similar gaps exist in other data sources, and whilst some diversity hotspots are clear in the Amazon, little data is available from Tropical Southeast Asia, and almost none from Central Asia or North Africa. Interestingly, Angiosperm data and Bryophytes come principally from South America, whereas Fern data has a greater input from China, though largely from South-Eastern regions. Overlays of BIEN and Kew plants data provides some clues on where the greatest spatial disagreements between these data systems are located on land, at the scale of countries (Figure 3). Bias can go in two directions, with apparent underrecording of species by the Kew databases in several parts of North America, some South American countries, and in Spain, Australia, and New Zealand. Conversely, the Kew data seems to show underrecording by the BIEN database in many parts of Africa, Central Europe and the Middle East, and across Asia. The analysis did not reconcile taxonomies between BIEN and WCVP beforehand, which may account for some of the differences observed in species counts across regions. However, the analysis should still effectively reflect broader-scale discrepancies between the two databases. Additionally, our understanding is that BIEN's taxonomic backbone incorporates WCVP, which should limit such inconsistencies, and likely keep them below the 100 species buffer.



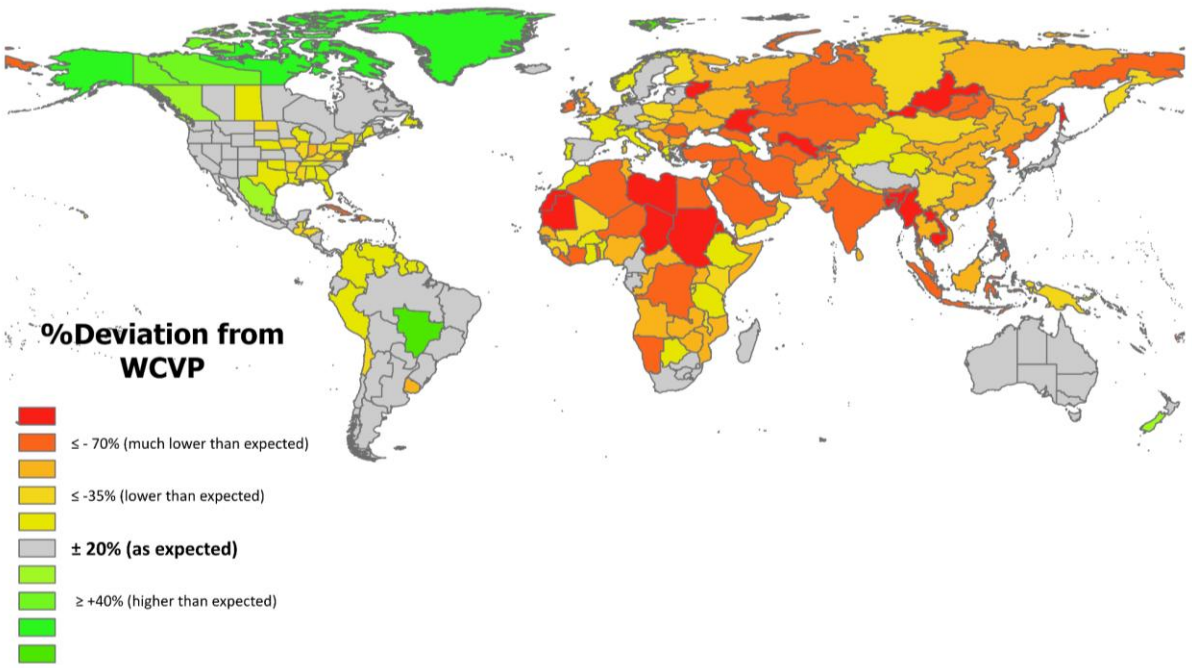


Figure 3: *Top: Absolute species count difference between WCVP and BIEN by botanical country. Bottom: Relative species count difference between WCVP and BIEN by botanical country. Red color gradient indicates regions where BIEN's occurrence sampling contains less species than expected according to WCVP. Green color gradient indicates regions where BIEN's occurrence data contains more species than expected according to WCVP.*

IUCN Red List

The IUCN Red List has been analysed at the taxonomic level of Kingdoms, illustrating the use of point locality data. According to the 2025-1 version of the database, 144,239 species contain some form of map (point/polygon), and 28,713 contain no geographic assessment (Figure 4). Overall, 47,179 species in the Red List consist of point records only, and a further 6,966 species contain point data as well as polygons, whereas 90,094 species only have polygon data (where expert knowledge, and possibly checklists are the main source of information used for range mapping). The source of the point locality data is not always clear, as many assessments are based on expert knowledge and any original data used may not be explicitly made clear (Hughes et al., 2024).

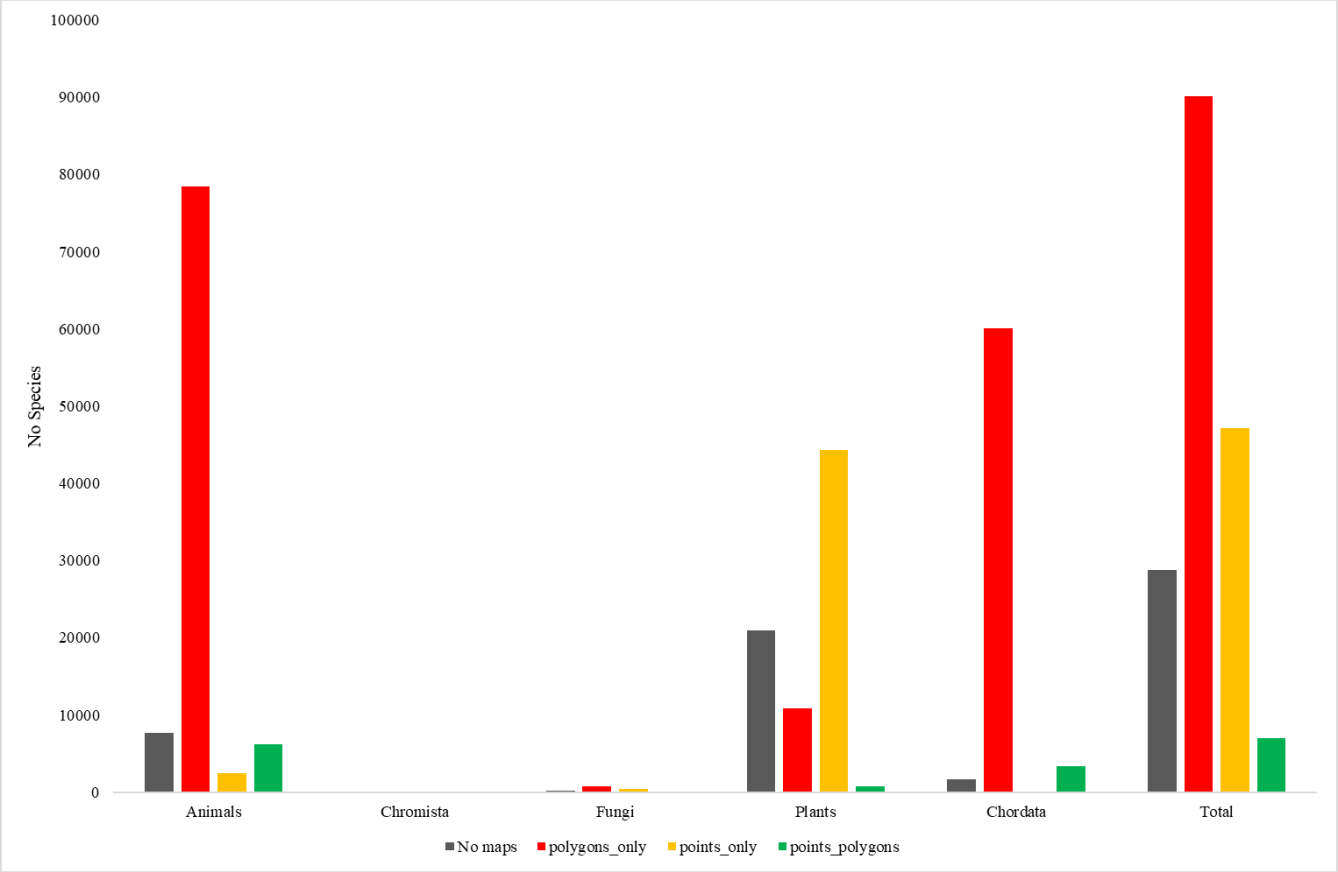


Figure 4. *Numbers of species within the IUCN Red List assessment process that use point locality data only, polygon data only, points and polygons data together, or have no maps associated with their record .*

Sources of data used within the IUCN red list changed between the four Kingdoms (Figure 4). For plants 59% of the species assessed had some form of point data, whereas 14% were only polygons and 27% had no maps. This was reversed for animals (which includes Chordata), where 83% of species only had only polygon maps, whereas 10% had some form of point data (95% and 4% respectively for the phylum Chordata). Thus point maps were largely for non-chordate animals. Fungi and Chromista had too few species assessed to provide further insights.

Discussion

Whilst there has been an exponential growth of data collected and made available over the last decade, this has been accompanied by a fragmentation of data across databases with different schema, making assembling all knowledge of species distributions, even for small regions, potentially challenging (for review see Kemp et al., in press). Understanding trajectories of data growth as well as where persistent gaps remain can help identify areas where extra effort may be needed. Various attempts have been made to explore growth and skew of biodiversity representation in open access databases, typically investigating taxonomic, temporal and geographic bias (see e.g. Petersen et al., 2021). Data may be distributed across multiple

international and national data repositories with varying levels of accessibility and findability (from governmental databases which may not have public access, specialist databases which may be challenging to find and have different data standards, individual project datasets on platforms like Zenodo, to fully open and standardised data on GBIF and OBIS). Whilst new tools are being developed to help aggregate fragmented data from multiple repositories (i.e. Owens et al., 2021), this highlights the expertise needed to even understand aspects of the adequacy of data for any given region, before even considering the standardisation and cleaning typically needed for use (i.e. Orr et al., 2021). This poses several key questions which must be answered to move forwards effectively. Firstly, where (and what taxa) do we have data for? Secondly, how do these data compare to expected diversity patterns within these groups? Further, which locations and taxa show data growth, and for which can we close these gaps?

Assessing patterns of data collection and consistent gaps

Many countries still have low data coverage, for example, 78 countries have under 50% of their area represented in global biodiversity datasets according to GBIF (at a 5km resolution), and 34 of these have under 25% covered, with only one of these showing a substantial increase (>20% growth) between 2015 and 2025. Tropical regions, whilst diverse, are generally undersampled, with the majority of data originating from often higher-income, temperate regions. These biases persist across all databases (Supplementary Results S3). However, despite low sampling, areas such as Brazil and Indonesia still had the highest number of species observed for various taxa (especially when using datasets such as BioTIME; Supplementary results S3A). In marine environments, the High-seas have major datagaps, with much data pertaining to tracking of small numbers of species and individuals, with most data originating from temperate and subpolar coastal regions. Further analysis of these regions would require further data for basic diversity statistics, or more sophisticated approaches (such as modelling).

Conversely, data gaps are so large for some regions as to preclude any form of more advanced analysis of species distributions or population trajectories. For example, based on GBIF data, North Africa and parts of the Middle East (i.e. Libya and Afghanistan, Turkmenistan) show both low coverage and some of the lowest growth at under 4% (Figure S1). Yet the situation for these countries may actually be even worse, for example, much of Afghanistan's data is a non-georeferenced bacteria and fungal assessment from January 2018 with 169,604 samples of bacteria, fungi, archaea etc), meaning virtually no data would be usable for monitoring or modelling. Political and linguistic barriers present a challenge to data growth within these regions, and targeted efforts and partnerships are likely needed to overcome them (see, e.g., Stephenson et al., 2017). Furthermore, within some regions, such as various countries in Africa and Asia, government and ministerial biodiversity data repositories and individual research data may be less likely to be publicly shared, precluding visible data growth despite increasing survey efforts. This all highlights the need for further work to both enhance the findability and access to existing data, in addition to targeting persistent survey gaps.

In addition to geospatial biases in country representation of data, taxonomic coverage is uneven. OBIS and GBIF are built from various databases (eBird, iNaturalist), researchers,

museum collections and other repositories (i.e. Vertnet), and thus reflect global biases in data collection. Thus, whilst the continuous plankton recorder in the ocean will overcome many of the taxonomic biases present in terrestrial regions for the ocean, these issues persist in terrestrial systems. For example, most insect data in GBIF comes from longstanding invertebrate monitoring programs in Europe such as the Swedish Malaise trap program. Similarly, 26% of all insect data in GBIF is from the United Kingdom, and of this 72% comes from three specialised programs on UK moths and butterflies. This shows how effective National Scale efforts can be at mobilising tremendous volumes of data, but few such efforts exist in tropical regions, and even when national monitoring does occur, this data is rarely stored in publicly available repositories (e.g., MyBIS for Malaysia, or Thailand's DNP database). Similarly, based on the BIEN data major gaps exist across Asia and Africa, but have been reconciled by targeted sampling efforts across parts of the Americas.

Whilst most tropical regions globally have seen data-growth this is strongest for birds. For other taxa, the growth of data has been far more variable; Pacific islands and much of Africa have seen little change in the number of species represented in most clades, but the Neotropics and Southeast Asia have seen increases in some groups (such as reptiles, and to a degree amphibians), though rarely for invertebrates. In terms of diversity hotspots experiencing a growth in biodiversity data coverage, the Neotropics is performing well relative to other regions, with some Southeast Asian Nations following this, whilst more arid regions continue to lag. That said, some regions have seen a transformation of their data coverage over the 2015-2025 period. In some cases, such as the case of South Korea, this is in part due to the work of agencies such as the National Institutes for Ecology to digitise collected data in both South Korea and Japan. Furthermore, across sub-Saharan Africa, the Pacific, Caribbean and Asia, the Biodiversity Information for Development (BID), funded by the European Union, and Biodiversity Information Fund for Asia (BIFA), funded by Japan's Ministry of Environment, both initiatives of GBIF, have helped mobilise data. Continued funding will be needed to sustain and expand such efforts into the future, and will likely need to be complemented by national funding to increase inclusion of datasets across global regions.

In addition to geospatial biases in the ecosystems and regions covered, the majority of terrestrial data comes from regions adjacent to roads (over 90% within 2 km; Hughes et al., 2021) and most marine data comes from coastal areas, with all regions but Antarctica showing this pattern (Supplementary results S2). Across all taxa, proximity to a major road is a significant predictor of sampling intensity, with many regions also showing a positive relationship between the degree of human modification (based on the human modification index) and the sampling intensity (Supplementary results S2). Furthermore, most marine data comes from shallow-seas, with little data from 5000-10,000m, and virtually none at greater depths (Bridges & Howell, 2025). Spatial representation of Benthic and Pelagic environments is largely focused around Europe, the West coast of the US and New Zealand, with little data available for other parts of the ocean (Bridges & Howell, 2025). Data showed strong biases towards coastal areas near developed countries, regions with intensive fishing activity, and species of small body size occupying shallow habitats and of commercial or cultural value (Pizarro et al., 2024). However, the coast is very poorly represented around most of the Arctic

(notably in Canada, Greenland, and Russia). In contrast, Antarctica's coast is well represented, yet key habitats are missed, such as unique and globally threatened habitats, under ice shelves, where >70 species may occur in <1 m^2 (Barnes et al., 2021), though notably it is the only area where most data is not immediately adjacent to the coastline. Overall, whilst marine biodiversity data availability has increased exponentially, particularly in the Southern Ocean, Japan and South Korea, clear gaps remain (including most of the high-seas and deep ocean (Figure 1, Figure S1)) (Bridges & Howell, 2025). Compounding these gaps and biases, a global analysis of over four decades of occurrence data found that only 1.14% of available records were suitable for detailed analyses (Pizarro et al., 2024). Notably, these gaps are amplified when higher resolutions of data are assessed, or when data density is considered, highlighting that good data coverage still only occurs for a small part of the world, and whilst data coverage has increased, many gaps still persist. Importantly, perceived increases in data coverage in some ocean regions (such as the Southern Ocean) are in part driven by GPS tracking of Southern Elephant seals (i.e. see Rodríguez et al., 2017), whilst dense collections of data are still primarily centred around Europe. In most regions, it should also be noted that much of the new data is bird data through citizen science initiatives. Whilst such data is invaluable for tracking migration and phenology, it rarely provides data for other taxa, and thus this growth of bird data may not have been mirrored in other taxa. An exception to this is for countries like the UK where multiple National programs have provided a suite of high-quality data for monitoring across taxa, and may have lessons which could be replicated elsewhere.

Taxonomic representation also varies across databases (Supplementary results S3). For example, PREDICTS has a better taxonomic representation than many aggregated datasets, but data is only available for a subset of regions and ecosystems (Supplementary results S3d). The Living Planet database holds considerable data and has been used widely to derive different indicators to be used as policy tools, but there remain areas for improvement for filling geographic and taxonomic data gaps, using it for national monitoring, and developing modelling approaches that better capture uncertainty (McRe et al., 2025). Other areas to address include increasing the use of data that has been standardized for sampling effort before being used to infer population trajectories (Supplementary results S3c). Improving quality by reducing the number of populations through better standardisation may enhance the reliability of assessments (Feng et al., 2022). Monitoring programs also show major regional differences, highlighting both regional differences in focal taxa, and that only a subset of data enters public facing databases (Moussy et al., 2022; Supplemental Results S6). Freshwater systems in particular are some of the most challenging to monitor across scales, as they are not reflected in BioTIME, and RivFishtime shows major geographic biases (Supplementary results S3a-b). Despite their weaknesses, these datasets underpin our understanding of global diversity patterns. Furthermore, the physical dimensions of freshwater systems means that analysis using GBIF data may be challenging, as assessing sample completeness will be difficult at scale, and coordinate imprecision may hinder targeted management or assessments pertaining to stretches of waterways. Whilst these datasets are critical, their biases frequently mirror those of broader biodiversity databases, and the lack of standardised reporting standards can hinder accurate interpretation of the data.

Assessing and addressing sampling representativeness and biases

Understanding the relationship between hotspots of data-collection and hotspots of diversity is critical to ensure that we have adequate sampling in diverse regions. Yet almost all instances of known biodiversity hotspots (based on overlapping species maps from the IUCN) do not coincide with areas of peak data collection in either marine or terrestrial areas. For example, marine data collection intensity does not align with mapped diversity for those taxa (based on species assessments Data S2), especially in the Indo-Malayan region, and high diversity systems such as coral reefs only have species data covering around 11% of their area. For example for marine mammals and marine reptiles, comparisons of occurrence records with IUCN expert range maps reveal substantial coverage gaps (Data S2), with many species having few or no georeferenced records in global repositories, underscoring persistent data mobilisation and accessibility challenges even for well-studied taxa (Moudrý and Devillers, 2020).

Addressing these gaps requires systematic integration of fisheries-dependent and fisheries-independent datasets into global repositories such as GBIF and OBIS. These repositories provide the backbone of marine biodiversity information, but they are still dominated by research surveys and opportunistic observations, which leaves major gaps for exploited taxa and in regions where fisheries provide the most consistent biological records. Linking catch statistics, observer programs, trawl survey data, and acoustic detections with the taxonomic backbones and spatial frameworks of GBIF and OBIS would considerably expand coverage of marine taxa, especially in tropical and subtropical EEZs where biodiversity and fishing pressures coincide. This approach would also reduce duplication between parallel monitoring streams and establish a comprehensive baseline from which cumulative impacts of exploitation, environmental change and conservation interventions could be assessed. Beyond filling gaps, embedding fisheries datasets into these repositories would also promote interoperability across sectors, enabling biodiversity monitoring, conservation planning and fisheries management to draw from a single, coherent evidence base. In doing so, marine biodiversity assessments could reflect both commercially important and non-target taxa, improve comparability across regions, and provide more robust information to support long-term monitoring and reporting under the GBF. This is no small challenge as sharing fisheries data depends on building partnerships, consensus, and capacity across fragmented scientific communities (Maureaud et al., 2025).

Enhancing data usability

In addition to spatial and taxonomic biases and gaps in global datasets, the usability of data is also determined by the accuracy of the records. How many of the records present in a database are robust, that is (to the latest knowledge) truly represent the correct species? While data curation is a key focus for some platforms, which means that uploads are paired with appropriate metadata and species nomenclature is checked as valid and any identification verified by an appropriate taxonomist expert, there are no quality control standards for species records adopted by all platforms. This issue has two steps, one of which can be solved through

cross-referencing with standardised taxonomic backbones (for marine: ITIS, WoRMS], GBIF/Catalogue of Life; for terrestrial plants: World Checklist of Vascular Plants). The other step requires inspecting either the actual specimen or an image. From cleaning based on these taxonomic backbones, large numbers of records are incorrectly listed in databases before cleaning across taxa (i.e. 73% had some issue, which could be corrected in ~⅓ of cases, prior to cleaning in BIEN), with multiple applications requiring such treatment (i.e. wildlife trade data (Marshall et al., 2025)) and pipelines developed for various taxa (i.e. bees Dorey et al., 2023). This is very important, as for most analytical purposes, no data is preferable to 'bad' data. Although any given species name attributed to a record can be checked, it is difficult to ascertain whether the record was correctly identified as this species. Curated datasets such as WoRMS or BOLD can solve this problem but may contain very few records, because of this requirement and cost, and initiatives tend to be taxonomic and regional (i.e. African bats; Monadjem et al., 2024), meaning that the accuracy of most data remains variable, and challenging to quantify. This problem, although rarely quantified, is significant, so much so that often some of the most common and abundant species in a given region (represented in open access databases) are incorrectly identified and do not even occur there (e.g. see Sands et al., 2025). A key unanswered question is thus what proportion of species records data is incorrect and how does this change with taxonomic identity, region and time? Despite such accuracy and other, databasing problems it remains crucial to maximise the potential of what can be gleaned at the moment from existing data to assess the status and change of what we know of global species distributions, especially for rare, endemic, or cryptic species.

Given that global assessment requires standardised and interoperable data from across taxa, sustained efforts are needed to mobilise data which has been collected into a form where it can be accessed, and used to monitor trends in diversity. Notably, IUCN Red List assessments, which often feed into higher reporting, rarely use point data for animal species, whereas this is general practice for plants. Improving spatial data collection and analysis could be a critical component of National Biodiversity Strategic Action plans, as if data is collated and made available based on agreed standards, tools can be applied to the data to facilitate the reporting needed to meet mandated reporting requirements (reducing the effort needed to meet requirements). This requires both political will and sustained funding for monitoring. This would include through agencies such as the GEF (particularly as capacity for monitoring, and systems for data deposit are developed) and at a national level to develop long-term monitoring initiatives, such as those present for insects in the United Kingdom based on the GBIF data).

Moving forwards

The last decade has seen a considerable transformation of biodiversity knowledge accessible via databases for many parts of the world. Yet gaps remain, with Central Asia and Northern Africa on land, and the high-seas, Arctic, and coastlines around the tropics showing little data growth in global databases. Initiatives like BID and BIFA from GBIF have driven improvements in some regions (sub-Saharan Africa, the Pacific, Caribbean and Asia), but there continues to be an urgent need for more capacity building for biodiversity monitoring in low-income high-biodiversity countries (e.g. Schmeller et al., 2017; Stephenon et al., 2017).

Overcoming these challenges will require not only funds, but engagement, especially in areas where data exists but is either increasingly fragmented (e.g. polar regions) or not publicly accessible (e.g. China, Southeast Asia) and areas which may lack capacity or resources. Likewise, intact ecosystems are poorly represented, likely due to the data for these systems coming from Scientists (rather than members of the public) who may be less likely to share data, or tend to publish it within databases which are not connected to, or interoperable with, larger global databases. At present funding from GEF (the Global Environment Facility) cannot be used for monitoring, yet countries cannot plan to manage their biodiversity without either a baseline or a mechanism for standardised monitoring. Thus, dedicated support is clearly needed. Such efforts could be paired with reporting mandates and mechanisms, and by standardising data collection, could also be analysed by standardised tools to reduce reporting burdens and standardise reporting.

There has even been considerable growth of data in some remote regions, such as both Arctic and Antarctic seas, but this is not evident in OBIS or GBIF because much new data is being uploaded to an increasing diversity of specialist and national databases, many of which have little connectivity with global repositories. The proliferation of open access databases for biodiversity records (mainly species-level) keeps increasing, resulting in inevitable problems of data curation, accuracy and fragmentation. Thus although data coming into main repositories such as GBIF and OBIS is increasing, it is also possible for this to be a decreasing proportion of new data. This can be true even if many such new regional or institute databases cascade data to GBIF and OBIS, because of the lag time taken to do this. In the Arctic for example much new data goes to MARBUNN/MAREANO (Institute of Marine Research Norway). ICES (International Council for the Exploration of the Seas), PANGAEA (and other German databases within this e.g. https://critterbase.awi.de/), NOAA and molecular databases (GENBANK, BOLD etc). One of the many problems this generates for biodiversity and conservation evaluation of where species occur, is that key analyses have a reticence to use non-specific databases (GBIF and OBIS) or bypass these completely (e.g. in connectivity studies, Vaughan et al., 2011, O'Hara et al., 2025). Thus it is crucial that action is taken to rectify such curation, accuracy and fragmentation problems to avoid these issues continuing to grow. This likely requires further work to build partnerships (Maureaud et al., 2025), and build trust with communities currently reluctant to share data through central data portals.

In ocean regions, major gaps of easily accessible data in single portals remain, especially in poorly monitored regions such as the high seas, deep ocean, Arctic (but see Greenland data in Zwerschke et al 2025), and many tropical coastlines. The lack of reliable species-level occurrence data makes it difficult to map distributions accurately and to connect biodiversity information with fisheries records, leaving management decisions on uncertain ground. Once fisheries-dependent and independent datasets are embedded within GBIF and OBIS, they can provide the baselines needed for decision-support tools (e.g., MARXAN and Zonation) to inform the design of ecologically representative MPAs, guide spatial fisheries closures, and assess the performance of existing conservation networks. Also, AI-based image and acoustic tools in fisheries along with fishers' experiential knowledge can produce data relevant for enhancing biodiversity surveillance more broadly (Ahlquist et al., 2025; Kelly et al., 2022;

Kuhn et al., 2024). Even in data-rich fisheries (e.g., in Norway or parts of the United States), discrepancies can remain between species distribution information derived from fisheries statistics and scientific surveys and the actual distribution of exploited stocks (Karp et al., 2023; Seljestad et al., 2024). This underscores concerns about the adequacy and representativeness even of available data, even though data on fish stocks and their distribution is becoming increasingly accurate (Sharma et al., 2025).

For offshore and deep-sea regions, integration with information on Vulnerable Marine Ecosystem (VME) indicator taxa and habitat-suitability models can further ensure that conservation priorities capture fragile benthic habitats and associated uncertainty (Burgos et al., 2020). Simultaneously, the development of biodiversity indicators that combine standardised survey observations with aggregated occurrence data allows consistent, policy-relevant measures of ecological condition, making it possible to evaluate protection outcomes across regions (Edgar et al., 2016). Global syntheses show that MPA effectiveness depends not only on coverage but on features such as no-take status, enforcement, connectivity and others, collectively enhancing biodiversity (Edgar et al., 2014). Inclusion of indicator frameworks can further strengthen commitments, for example BBNJ and CBD's 30×30. Linkage of occurrence, VME and fisheries records to planning models, could help translate observations into spatial management actions to balance biodiversity protection with sustainable use. When data are standardised and connected to global infrastructures, e.g. in harmonised bottom-trawl surveys (e.g., FISHGLOB in Europe; Maureaud et al., 2021), such approaches become feasible. Extending such approaches to (lower latitude) biodiversity-rich but data-poor EEZs could increase the efficacy of marine biodiversity data contributions to conservation planning and policy, as terrestrial data is already doing.

However, in addition to basic distribution data, long-term monitoring is also needed, and without government investment and support, many datasets will remain small scale, and short-term For example, long-term monitoring of marine biodiversity settlement and recruitment at an Antarctic coastal using artificial substrata has revealed declines in assemblage-level richness through increasing rarity of many rare species, none of which might be evident through the lens of presence-absence distribution data (Barnes et al., 2025). Such data also reveal 'background' levels of cyclicity at various spatial and temporal scales that must be considered when trying to interpret potential declines and their drivers. Without repeated monitoring, it is very difficult to disentangle responses of assemblages and species to threats (signal) from normal variability and cyclicity in patterns (noise), and thus for example to assess whether and how implemented protection is effective. These types of stochasticity, in addition to seasonal change further hinders the reliability of databases such as the LPI where 20% of populations only have two sampling events, which without careful standardisation of survey effort precludes reliable inference of population trends (supplementary results S3c). In the absence of the standardized, underlying data and the heterogeneity in sampling approaches, hinders the accurate interpretation of indices like these which maximise data intake, possibly at the expense of data quality. National support, and shared common standards would enable more effective monitoring across regions. Emerging technologies like AI and machine learning may offer potential to expand marine monitoring by automating species recognition, catch reporting, and

electronic monitoring, improving both cost-effectiveness and spatiotemporal coverage. Automatic catch registration systems and electronic logbooks are increasingly being trialled in fisheries. With advances in AI, the use of camera traps and acoustic recording devices provides an opportunity to monitor activity and diversity in almost real time of more easily detectable organisms, and such efforts are already implemented in regions such as China based on bioacoustics. These new tools provide a means of scaling up monitoring efforts, and when paired with other methods (e.g. standardised surveys) could greatly enhance our ability to monitor and, therefore, manage biodiversity.

Thus, whilst some of the challenge is scientific (especially around the developing and use of standards, confirming of species level identities, and the sharing of data in some regions), much of the challenge stems from either a lack of government support or a lack of resources. Elements of this are not trivial, confirming robust species identity can often require considerable molecular and micro-morphological expert effort (e.g. see Sands et al., 2025). With various organisations in place to support monitoring efforts, streamlining these activities to make better use of existing data, and to better target data gaps could vastly improve our understanding moving forwards.

However, the use of data from these systems can be challenging, as growing datasets are demanding to download, clean appropriately for the application, and analyse. Thus, the development of GUIs to facilitate these key processes without the need for programming expertise are likely needed both to maximise the impact of this data and their ability to inform monitoring and management, and to incentivise data sharing, by making such tools convenient and accessible. Additionally, given that many of these databases come from distinct research teams (particularly BioTIME and LPI), they are not readily interoperable with other databases. Furthermore, academic journals should not only mandate the public sharing of data (when non-sensitive) for reproducibility, but must specify the repository (such as GBIF or OBIS) to prevent further fragmentation of data. This is in contrast to the situation with genetic sequence data, for which it is standard practice for journals to mandate deposition in one of the repositories within the International Nucleotide Sequence Database Collaboration (INSDC), ensuring that they meet the requirements of the FAIR (findable, accessible, interoperable, reusable) principles. Many ecological journals and molecular communities have already done this, and thus supporting journals to transition would enable the better use (and cleaning) of data that has already been collected. Providing standards for data-upload and enabling upload of data into GBIF and OBIS (as is already often the case for figshare and Zenodo from some journals) would help ensure data remains findable, usable, and interoperable.

Over the last decade, parts of the world that previously had no data have transformed, and we are hopeful that with continued growth the data gaps will continue to narrow. Whilst data growth over the last decade has helped us to understand what we need to study, the insights into trends outside of the small minority of very well-sampled species are more limited. High-resolution data is crucial for effective planning, and thus a key element of metrics for frameworks such as the Kunming-Montreal Global Biodiversity framework. However, at present data-gaps and biases limit the ability to map species in many regions (e.g., central South

Atlantic, central Pacific Ocean and other offshore deep-sea regions in the ocean, and central Asia and North Africa on land; Webb et al., 2010), leading to the use of "expert opinion" as a substitute, and potentially neglecting little known taxa and regions. Looking forward, consolidating existing data before the 2030 CBD and the discussion of future targets would widen the scope for indicator selection, and enhance our ability to detect trends, and adequately conserve wildlife and the habitats it depends on.

Acknowledgements

We thank the many thousands of people who have gathered, made available, organised and made accessible biodiversity occurrence and timeseries data that forms the basis of the databases we summarise here. We would also like to thank the GBIF secretariat team for their inputs, particularly Tim Robertson; and the OBIS secretariat team, particularly Pieter Provoost, Ward Appeltans, Stephen Formel, Silas Principe; Andy Purvis for PREDICTS data; Maria Dornelas for BioTIME data. Corinna Raviollus and Ana Rodrigues from UNEP-WCMC analysed the IUCN red list data with the permission of IUCN. We would also like to acknowledge the work of FISHGLOB who have been working to better collate and centralise data in the face of limited funding support.

References

Ahlquist, I. H., Hatlebrekke, H. H., & Tiller, R. (2025). Fishing for solutions: Norwegian fishers' perspectives on the implementation of automatic catch registration for combating IUU fishing. Marine Policy, 179, 106750. https://doi.org/10.1016/j.marpol.2025.106750

Antonelli, A., Fry, C., & Villaverde, T. (2023). State of the World's Plants and Fungi, 2023.

Barnes, D. K. A., A. Giles, P. Glaz, S. McLoughlin, A. Clement, and S. A. Morley. 2025. " Long Term Marine Biodiversity Monitoring in Coastal Antarctica: Are Fewer Rare Species Recruiting?." Global Change Biology, 31(7), e70341. https://doi.org/10.1111/gcb.70341.

Barnes, D.K.A., Kuhn, G., Hillenbrand, C.D., Gromig, R., Koglin, N., Biskaborn, B.K., Frinault, B.A.V., Klages, J.P., Smith, E.C., Berger, S., Gutt, J., 2021b. Richness, growth, and persistence of life under an Antarctic ice shelf. Current Biology 31, R1566–R1567.

Bayraktarov, E., Ehmke, G., O'Connor, J., Burns, E. L., Nguyen, H. A., McRae, L., ... & Lindenmayer, D. B. (2019). Do big unstructured biodiversity data mean more knowledge?. Frontiers in Ecology and Evolution, 6, 239.

Boyle, B., Hopkins, N., Lu, Z., Raygoza Garay, J. A., Mozzherin, D., Rees, T., ... & Enquist, B. J. (2013). The taxonomic name resolution service: an online tool for automated standardization of plant names. BMC bioinformatics, 14(1), 16.

Boyle, B. L., Maitner, B. S., Barbosa, G. G., Sajja, R. K., Feng, X., Merow, C., ... & Enquist, B. J. (2022). Geographic name resolution service: A tool for the standardization and indexing of world political division names, with applications to species distribution modeling. Plos one, 17(11), e0268162.

Boyle, B. L., Maitner, B., Barbosa, G. C., Rethvick, S. Y. B., & Enquist, B. J. (2024). Native Species Resolver. In Botanical Information and Ecology Network. https://nsr.biendata.org/

Bridges, A. E., & Howell, K. L. (2025). Prioritisation of ocean biodiversity data collection to deliver a sustainable ocean. Communications Earth & Environment, 6(1), 473.

Brown, M. J., Walker, B. E., Black, N., Govaerts, R. H., Ondo, I., Turner, R., & Nic Lughadha, E. (2023). rWCVP: a companion R package for the World Checklist of Vascular Plants. New Phytologist, 240(4), 1355-1365.

Brummitt, R. K., Pando, F., Hollis, S., & Brummitt, N. A. (2001). World geographical scheme for recording plant distributions (Vol. 951, p. 952). Geneva, Switzerland:: International working group on taxonomic databases for plant sciences (TDWG).

Burgess, N., Kemp, H., Hargey, A., Cierna, A., Taylor, C., Bhola, N., ... & Stephenson, P. (2025). The past, present and future of online biodiversity knowledge systems. EcoEvoXIV https://ecoevorxiv.org/repository/view/10289/

Burgos, J. M., Murillo, F. J., Kenchington, E., Sacau, M., & Lirette, C. (2020). Model-based predictions of the distribution of deep-sea Vulnerable Marine Ecosystems in the Northwest Atlantic. *Frontiers in Marine Science, 7*, 131. https://doi.org/10.3389/fmars.2020.00131

Comte, L., Carvajal-Quintero, J., Tedesco, P. A., Giam, X., Brose, U., Erős, T., ... & Olden, J. D. (2021). RivFishTIME: A global database of fish time-series to study global change ecology in riverine systems. Global Ecology and Biogeography, 30(1), 38-50.

Cornford, R., Millard, J., González-Suárez, M., Freeman, R., & Johnson, T. F. (2022). Automated synthesis of biodiversity knowledge requires better tools and standardised research output. Ecography, 2022(3), e06068.

Dinerstein, E., Olson, D., Joshi, A., Vynne, C., Burgess, N. D., Wikramanayake, E., ... & Saleem, M. (2017). An ecoregion-based approach to protecting half the terrestrial realm. BioScience, 67(6), 534-545.

Dorey, J. B., Fischer, E. E., Chesshire, P. R., Nava-Bolaños, A., O'Reilly, R. L., Bossert, S., ... & Cobb, N. S. (2023). A globally synthesised and flagged bee occurrence dataset and cleaning workflow. Scientific Data, 10(1), 747.

Dornelas, M., Antão, L. H., Bates, A. E., Brambilla, V., Chase, J. M., Chow, C. F., ... & Fryxell, J. (2025). BioTIME 2.0: Expanding and improving a database of biodiversity time series. Global Ecology and Biogeography, 34(5), e70003.

Edgar, G. J., Stuart-Smith, R. D., Willis, T. J., Kininmonth, S., Baker, S. C., Banks, S., … Thomson, R. J. (2014). Global conservation outcomes depend on marine protected areas with five key features. *Nature, 506*(7487), 216–220. https://doi.org/10.1038/nature13022

Edgar, G. J., Stuart-Smith, R. D., & others. (2016). New opportunities for conservation of marine biodiversity using standardized global reef fish monitoring. *Annual Review of Marine Science, 8*, 493–519. https://doi.org/10.1146/annurev-marine-122414-033921

Enquist, B., et al., (in review). BIEN: A biodiversity informatics ecosystem advancing open and reproducible workflows for plant observation, plot, and trait data

Feng, X., Enquist, B. J., Park, D. S., Boyle, B., Breshears, D. D., Gallagher, R. V., ... & López-Hoffman, L. (2022). A review of the heterogeneous landscape of biodiversity databases: Opportunities and challenges for a synthesized biodiversity knowledge base. Global Ecology and Biogeography, 31(7), 1242-1260.

Flanders Marine Institute (2024). The intersect of the Exclusive Economic Zones and IHO sea areas, version 5. Available online at https://www.marineregions.org/. https://doi.org/10.14284/699

GBIF.org (01 May 2025) GBIF Occurrence Download https://doi.org/10.15468/dl.a5mqyx

Govaerts, R., Nic Lughadha, E., Black, N., Turner, R., & Paton, A. (2021). The World Checklist of Vascular Plants, a continuously updated resource for exploring global plant diversity. Scientific data, 8(1), 215.

Guralnick, R., Walls, R., & Jetz, W. (2018). Humboldt Core–toward a standardized capture of biological inventories for biodiversity monitoring, modeling and assessment. Ecography, 41(5), 713-725.

Hudson, L.N.; Newbold, T.; Contu, S. et al. (2017). The PREDICTS database: a global database of how local terrestrial biodiversity responds to human impacts [Data set]. Natural History Museum.

Hughes, A. C., Orr, M. C., Yang, Q., & Qiao, H. (2021). Effectively and accurately mapping global biodiversity patterns for different regions and taxa. Global Ecology and Biogeography, 30(7), 1375-1388.

Hughes A., Orr, M.C., Palacio, R.D., Xuan, Y., Qiao, H. (2024). A dire need for 232 better standards of data quality, transparency, and reproducibility in IUCN Red List assessments. Ecoevorxiv

Ingenloff K, Svenningsen C, Earl C, Shimabukuro PHF, Sica Y, Gan Y-M, Kachian ZR, Brenton P, Hochachka W, Wieczorek J, Stevenson R, Kazem A, Baskauf S, Zermoglio PF, Bloom D, Rodrigues A, Gamboa Martínez J & Schigel D. Guide for publishing biological survey and monitoring data to GBIF. GBIF Secretariat: Copenhagen. https://doi.org/10.35035/doc-ynvs-eh84

IPBES (2019). Global assessment report of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services, Brondízio, E. S., Settele, J., Díaz, S., Ngo, H. T. (eds). IPBES secretariat, Bonn, Germany. 1144 pages.

IPBES Technical Support Unit on Knowledge and Data. (2021). IPBES regions and sub-regions (1.2) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5719431

IUCN (2025) The IUCN Red List of Threatened Species. 2025-1. https://www.iucnredlist.org/resources/spatial-data-download Downloaded on 15/05/25.

Janicki, J., Narula, N., Ziegler, M., Guénard, B. Economo, E.P. (2016) Visualizing and interacting with large-volume biodiversity data using client-server web-mapping applications: The design and implementation of antmaps.org. Ecological Informatics 32: 185-193

Karp, M. A., Brodie, S., Smith, J. A., Richerson, K., Selden, R. L., Liu, O. R., Muhling, B. A., Samhouri, J. F., Barnett, L. A. K., Hazen, E. L., Ovando, D., Fiechter, J., Jacox, M. G., & Pozo Buil, M. (2023). Projecting species distributions using fishery-dependent data. Fish and Fisheries, 24, 71–92. https://doi.org/10.1111/faf.12711

Kelly, C., Michelsen, F. A., Reite, K. J., Kolding, J., Varpe, Ø., Berset, A. P., & Alver, M. O. (2022). Capturing big fisheries data: Integrating fishers' knowledge in a web-based decision support tool. Frontiers in Marine Science, 9. https://doi.org/10.3389/fmars.2022.1051879

Kemp, H., Cierna, A., Hughes, A., Arcangeli, A., Soesbergen, A., Hargey, A., Taylor, C., Taylor, C., Degano, M.E., Buschke, F., Tin, F.Y.K:, Sihvonen, H., Weatherdon, L., Miles, L:, Bhola, N., McDermott Long, O., Blanque, V., Sica, Y., Thatey, Z., Demirel, N., Stephenson, P.J., Burgess, N. (2025) The past, present and future of online biodiversity knowledge systems, https://ecoevorxiv.org/repository/view/10289

Kühn, B., Cayetano, A., Fincham, J. I., Moustahfid, H., Sokolova, M., Trifonova, N., … Uusitalo, L. (2024). Machine Learning Applications for Fisheries—At Scales from Genomics to Ecosystems. Reviews in Fisheries Science & Aquaculture, 33(2), 334–357. https://doi.org/10.1080/23308249.2024.2423189

Longhurst, A.R. (1998). Ecological geography of the sea. Academic Press, San Diego, U.S.

Maitner, B. S., Boyle, B., Casler, N., Condit, R., Donoghue, J., Durán, S. M., ... & Enquist, B. J. (2018). The bien r package: A tool to access the Botanical Information and Ecology Network (BIEN) database. Methods in Ecology and Evolution, 9(2), 373-379.

Maitner, B. S., Boyle, B., Casler, N., Condit, R., Donoghue, J., Durán, S. M., ... & Enquist, B. J. (2018). The bien r package: A tool to access the Botanical Information and Ecology Network (BIEN) database. Methods in Ecology and Evolution, 9(2), 373-379.

Marques, N., de Melo Soares, C. D., de Melo Casali, D., Guimarães, E. C., Fava, F. G., da Silva Abreu, J. M., ... & da Cunha Tavares, V. (2024). Retrieving biodiversity data from multiple sources: making secondary data standardised and accessible. Biodiversity Data Journal, 12, e133775.

Marine Regions (2025) Longhurst Provinces Ecological geography of the Sea (Longhurst, 1998) https://www.marineregions.org/gazetteer.php?p=details&id=22538

Marshall, B.M, Alamshah, A. L., Cardoso, P., Cassey, P., Chekunov, S., Eskew, E. A., ... & Hughes, A. C. (2025). The magnitude of legal wildlife trade and implications for species survival. Proceedings of the National Academy of Sciences, 122(2), e2410774121.

Maureaud, A. A., R. Frelat, L. Pécuchet, ... & Thorson, J. T. (2021). Are We Ready to Track Climate-Driven Shifts in Marine Species Across International Boundaries? - A Global Survey of Scientific Bottom Trawl Data. Global Change Biology 27, no. 2: 220–236. https://doi.org/10.1111/gcb.15404.

Maureaud, A. A., Kitchel, Z., Fredston, A., Guralnick, R., Palacios-Abrantes, J., Palomares, M. L., ... & Mérigot, B. (2025). FISHGLOB: A collaborative infrastructure to bridge the gap between scientific monitoring and marine biodiversity conservation. Conservation Science and Practice, 7(6), e70035.

McRae, L., Cornford, R., Marconi, V., Puleston, H., Ledger, S.E., Deinet, S., Oppenheimer, P., Hoffmann, M. and Freeman, R., 2025. The utility of the Living Planet Index as a policy tool and for measuring nature recovery. *Philosophical Transactions B*, *380*(1917), p.20230207.

Merow, C., Maitner, B. S., Schwarz Meyer, A., Pigot, A. L., Serra-Diaz, J. M., & Urban, M. C. (2025). Hottest year in recorded history compounds global biodiversity risks. Proceedings of the National Academy of Sciences, 122(35), e2504945122.

Meyer, C., Kreft, H., Guralnick, R., & Jetz, W. (2015). Global priorities for an effective information basis of biodiversity distributions. Nature communications, 6(1), 1-8.

Monadjem, A., Montauban, C., Webala, P. W., Laverty, T. M., Bakwo-Fils, E. M., Torrent, L., ... & Taylor, P. J. (2024). African bat database: curated data of occurrences, distributions and conservation metrics for sub-Saharan bats. Scientific Data, 11(1), 1309.

Moussy, C., Burfield, I. J., Stephenson, P. J., Newton, A. F., Butchart, S. H., Sutherland, W. J., ... & Donald, P. F. (2022). A quantitative global review of species population monitoring. Conservation Biology, 36(1), e13721.

Moudrý, V., Devillers, R. (2020). Quality and usability challenges of global marine biodiversity databases: An example for marine mammal data. Ecological Indicators, 56, 101051. https://doi.org/10.1016/j.ecolind.2020.101051

Oakleaf, J., Kennedy, C., Wolff, N. H., Terasaki Hart, D. E., Ellis, P., Theobald, D. M., ... & Kiesecker, J. (2024). Mapping global land conversion pressure to support conservation planning. Scientific Data, 11(1), 830.

O'Hara, T.D., Hugall, A.F., Haines, M.L. et al. Spatiotemporal faunal connectivity across global sea floors. Nature 645, 423–428 (2025). https://doi.org/10.1038/s41586-025-09307-1

Oliver, R. Y., Meyer, C., Ranipeta, A., Winner, K., & Jetz, W. (2021). Global and national trends, gaps, and opportunities in documenting and monitoring species distributions. PLoS Biology, 19(8), e3001336.

Orr, M. C., Hughes, A. C., Chesters, D., Pickering, J., Zhu, C. D., & Ascher, J. S. (2021). Global patterns and drivers of bee distribution. Current Biology, 31(3), 451-458.

Owens, H. L., Merow, C., Maitner, B. S., Kass, J. M., Barve, V., & Guralnick, R. P. (2021). occCite: Tools for querying and managing large biodiversity occurrence datasets. Ecography, 44(8), 1228-1235.

Petersen TK, Speed JDM, Grøtan V, Austrheim G. (2021). Species data for understanding biodiversity dynamics: The what, where and when of species occurrence data collection. *Ecol Solut Evidence*. 2021; 2:e12048. https://doi.org/10.1002/2688-8319.12048

Pilotto, F., Kühn, I., Adrian, R., Alber, R., Alignier, A., Andrews, C., ... & Haase, P. (2020). Meta-analysis of multidecadal biodiversity trends in Europe. Nature communications, 11(1), 3486.

Pizarro, O., Castillo, A. G., Piñones, A., Samaniego, H. (2024). Spatial and temporal representation of marine fish occurrences available online. Ecological Indicators, 79, 102403. https://doi.org/10.1016/j.ecoinf.2023.102403

Rodríguez, J. P., Fernández-Gracia, J., Thums, M., Hindell, M. A., Sequeira, A. M., Meekan, M. G., ... & Eguíluz, V. M. (2017). Big data analyses reveal patterns and drivers of the movements of southern elephant seals. Scientific reports, 7(1), 1-10.

Roll, U., Feldman, A., Novosolov, M., Allison, A., Bauer, A. M., Bernard, R., ... & Meiri, S. (2017). The global distribution of tetrapods reveals a need for targeted reptile conservation. Nature ecology & evolution, 1(11), 1677-1682.

Sands C., O'Hara T.D., Guzzi A., Goodall-Copestake W.P. , Convey P., Narayanaswamy B.E., Martín-Ledo R., Stöhr S. (2025) And then there were many: insights from the tangled taxonomy of the Antarctic brittle star *Ophioplinthus gelida* (Echinodermata: Ophiuroidea). Frontiers Marine Science 12 DOI10.3389/fmars.2025.1615695

Schmeller, D.S., Böhm, M., Arvanitidis, C., Barber-Meyer, S., Brummitt, N., Chandler, M., Chatzinikolaou, E., Costello, M.J., Ding, H., García-Moreno, J. and Gill, M., 2017. Building capacity in biodiversity monitoring at the global scale. *Biodiversity and Conservation*, *26*(12), 2765-2790.

Seljestad, G. W., Quintela, M., Bekkevold, D., Pampoulie, C., Farrell, E. D., Kvamme, C., Slotte, A., Dahle, G., Sørvik, A. G., Pettersson, M. E., Andersson, L., Folkvord, A., Glover, K.

A., & Berg, F. (2024). Genetic Stock Identification Reveals Mismatches Between Management Areas and Population Genetic Structure in a Migratory Pelagic Fish. Evolutionary Applications, 17(10), e70030. https://doi.org/10.1111/eva.70030

Sharma, R., Barange, M., Agostini, V., Barros, P., Gutierrez, N.L., Vasconcellos, M., Fernandez Reguera, D., Tiffay, C., & Levontin, P., eds. 2025. Review of the state of world marine fishery resources – 2025. FAO Fisheries and Aquaculture Technical Paper, No. 721. Rome. FAO.

Stephenson, P.J., Bowles-Newark, N., Regan, E., Stanwell-Smith, D., Diagana, M., Hoft, R., Abarchi, H., Abrahamse, T., Akello, C., Allison, H., Banki, O., Batieno, B., Dieme, S., Domingos, A., Galt, R., Githaiga, C.W., Guindol, A.B., Hafashimana, D.L.N., Hirsch, T., Hobern, D., Kaaya, J., Kaggwa, R., Kalemba, M.M., Linjouom, I., Manaka, B., Mbwambo, Z., Musasa, M., Okoree, E., Rwetsiba, A., Siams, A.B. & Thiombiano, A. (2017). Unblocking the flow of biodiversity data for decision-making in Africa. Biological Conservation, 213, 335-340.

Stephenson, P.J. & Stengel, C. (2020). An inventory of biodiversity data sources for conservation monitoring. *PLoS ONE,* 15(12), e0242923. https://doi.org/10.1371/journal.pone.0242923

Theobald, D.M., Oakleaf, J.R., Moncrieff, G., Voigt, M., Kiesecker, J. & Kennedy, C.M. (2025). Global extent and change in human modification of terrestrial ecosystems from 1990 to 2022. *Scientific Data* **12**, 489. [doi:10.1038/s41597-025-04892-2](https://doi.org/10.1038/s41597-025-04892-2)

Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R., & Legendre, F. (2017). Taxonomic bias in biodiversity data and societal preferences. Scientific reports, 7(1), 9132.

UNEP-WCMC (2022) Global Distribution of Coral Reefs https://data-gis.unep-wcmc.org/portal/home/item.html?id=061360436733483686 3f5c0c10e452bf

Vaughan, D.G., Barnes, D.K.A., Fretwell, P.T. & Bingham, R.G. (2011) Potential seaways across West Antarctica. Geochemistry, Geophysics Geosystems, 12, Q10004. https://doi.org/10.1029/2011GC003688

Webb, T. J., Vanden Berghe, E., & O'Dor, R. (2010). Biodiversity's big wet secret: the global distribution of marine biological records reveals chronic under-exploration of the deep pelagic ocean. PloS one, 5(8), e10223.

Weigelt, P., König, C., & Kreft, H. (2020). GIFT–A Global Inventory of Floras and Traits for macroecology and biogeography. Journal of Biogeography, 47(1), 16-43.

WWF (2024) Living Planet Report 2024 – A System in Peril. WWF, Gland, Switzerland.

WWF (2025) Living Planet Index https://www.livingplanetindex.org/data_portal

Zwerschke, N., Arboe, N. H., Behrisch, J., Blicher, M., & Barnes, D. K. (2025). Towards a regional baseline of Greenland's continental shelf seabed biodiversity. Journal of Environmental Management, 382, 125285.

Supplements

1). Supplemental Figures

Figure S1. GBIF data coverage for different designations

Figure S2. Occurrences in GBIF over time for different geographic regions.

Figure S3. Mean coverage of each biome within each realm based on ecoregional units

2). Supplementary Text

Supplementary Methods 1. Spatial biases in GBIF data

Supplementary Methods 2. Species population trends monitoring datasets

A. BioTIME
B. Living Planet Index and other monitoring databases

Supplementary Results S1. Taxonomic biases and patterns in GBIF data

Supplementary Results S2. Geospatial biases in GBIF data

Supplementary Results S3. Biases in monitoring databases

A. BioTIME
B. RivFishTIME
C. Living Planet Index
D. PREDICTS
E. Species monitoring programs

3). Supplemental Tables

Table S1. Data types used for analysis

4). Supplemental Data

(https://drive.google.com/drive/folders/1q-MVM7EJXm4oEKxCDm7NlonhwzJ_g5Ui?usp=sharing)

Data S1. GBIF data growth per taxa over time including number of species and samples annually.

Data S2. Example maps of data coverage and hotspots (based on IUCN data) for various marine taxa.

Data S3. BioTIME maps of species and sampling for each zone and taxonomic group.

Data S4. Living Planet data summary and analysis

Data S5. PREDICTS summary maps for each region and taxa

Data S6. Summary of the monitoring programs from Moussy et al.

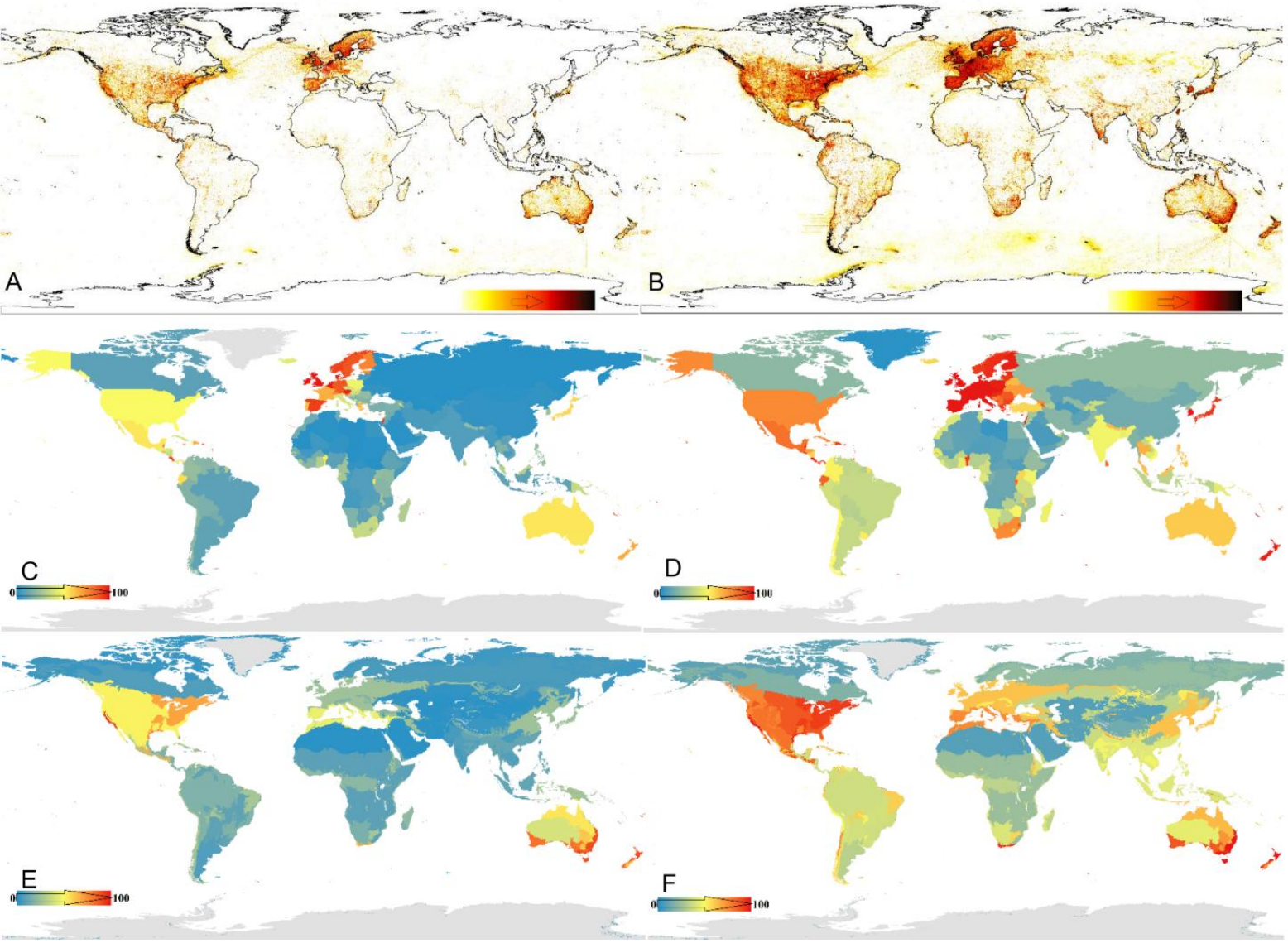


Figure S1. GBIF data coverage. Maps on the left (A, C, E) show data for 2015, and maps on the right (B, D, F) provide data for 2025. The spatial scales are Top (A-B) data plotted at a 5 km resolution, middle (C-D) data at national coverage (with at least 1 point per 5km cell) and bottom (E-F) data at ecoregion scale.

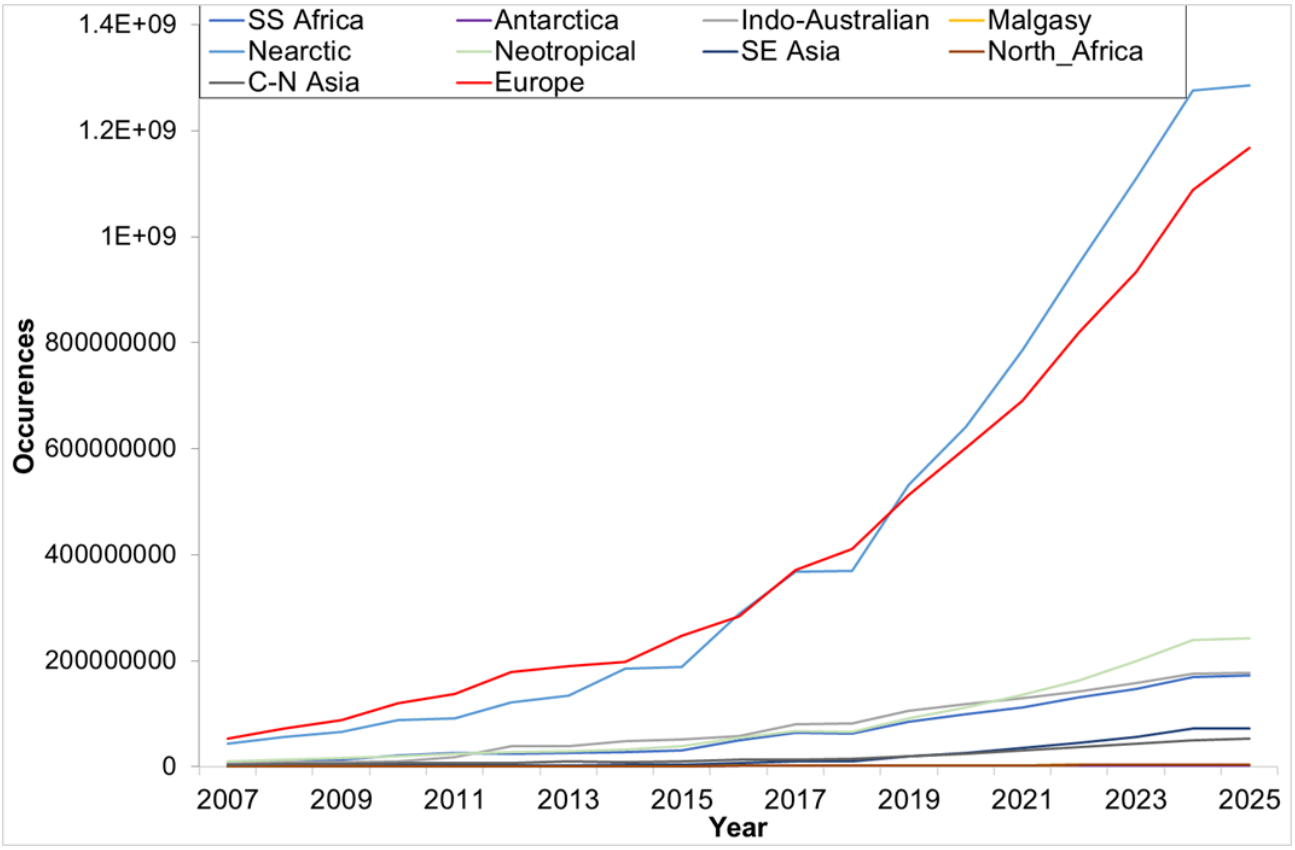


Figure S2. Records in GBIF over time for different geographic regions.

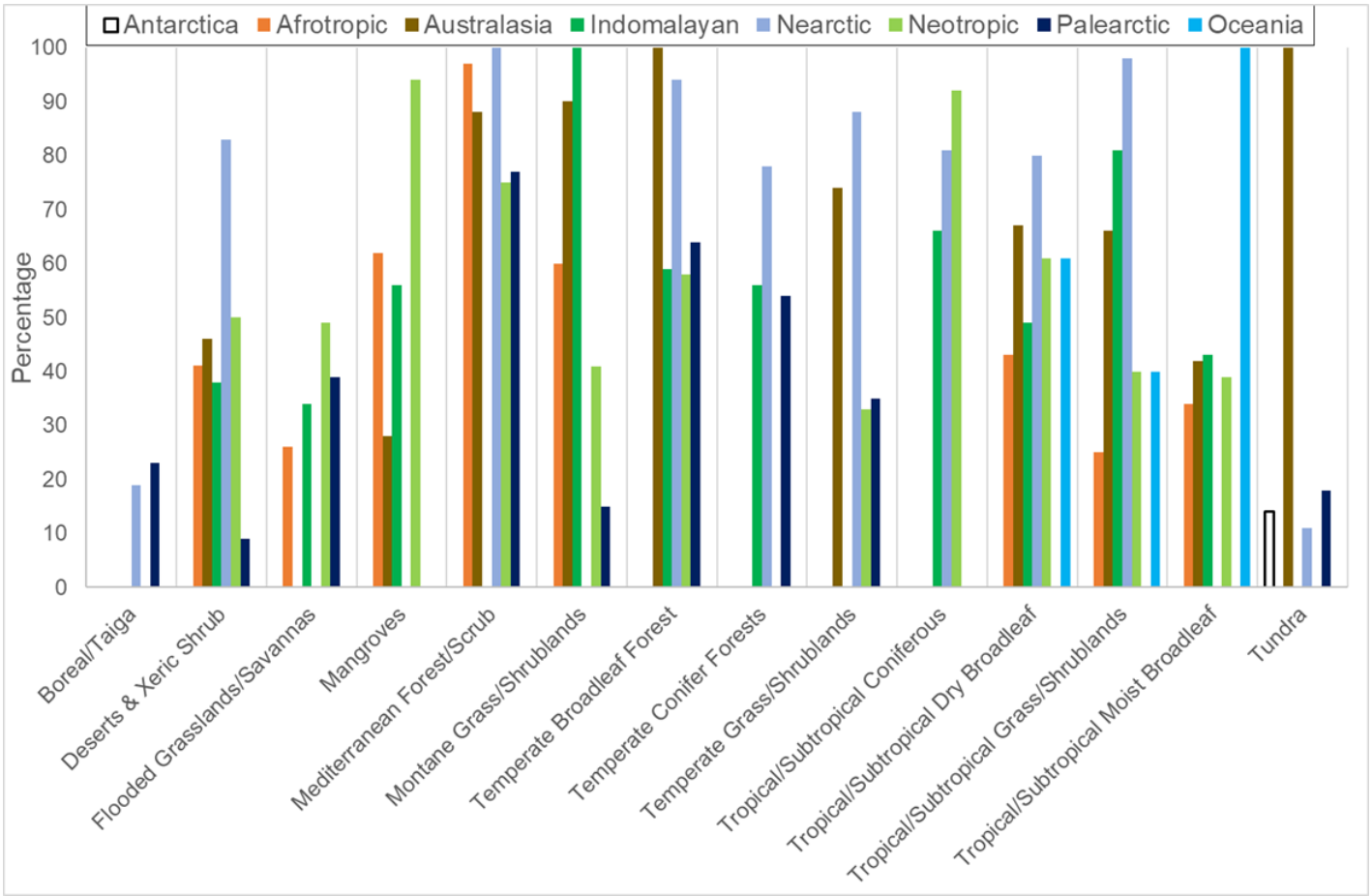
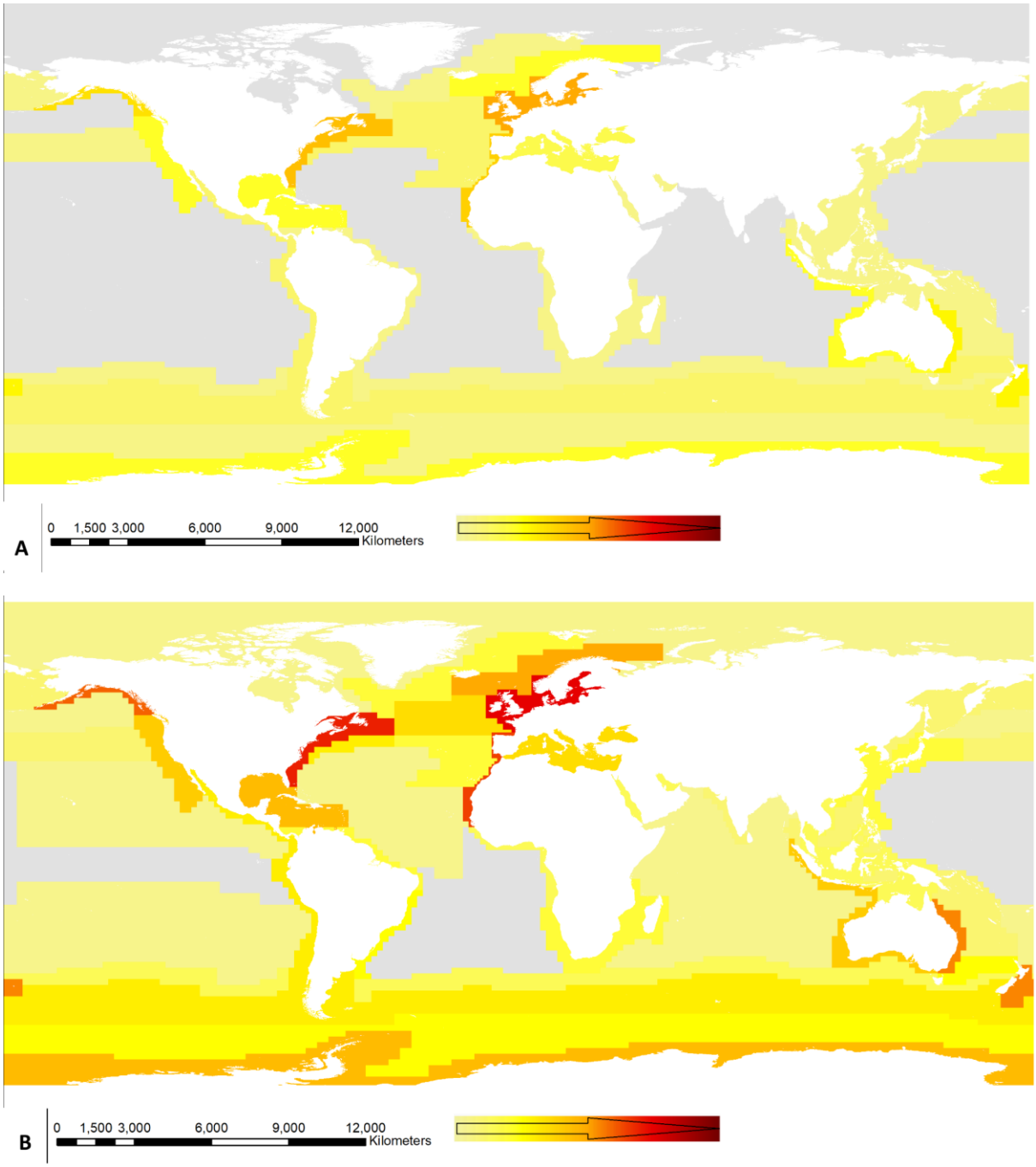


Figure S3. Mean coverage of each biome within each realm based on ecoregional units.



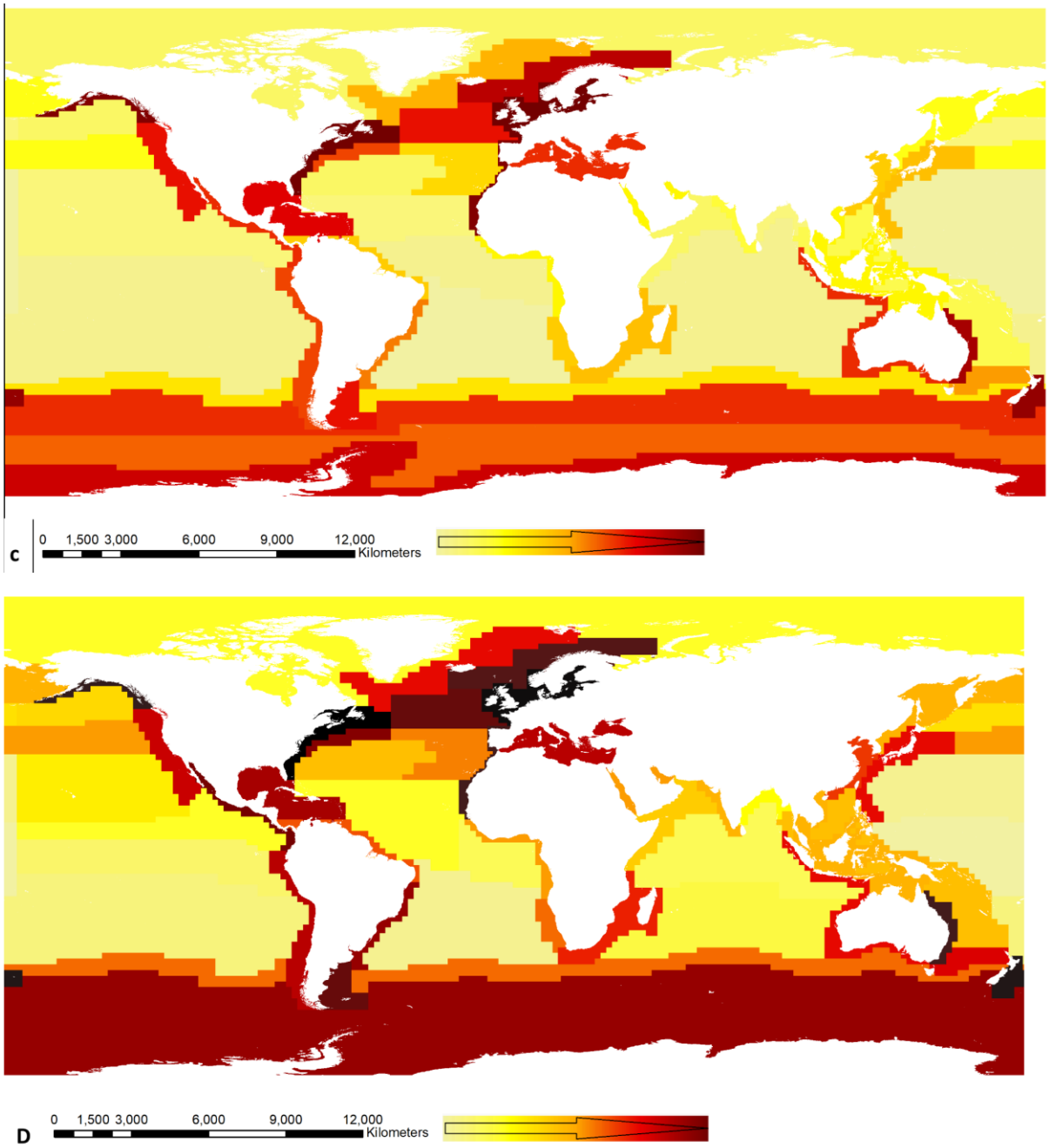


Figure S4. *Percent coverage of OBIS point data in the ocean (With at least one point per cell) for Longhurst Zones. Gray areas have under 1% data coverage. Longhurst regions. Resolutions A - 0.01^{O}, B - 0.02^{O}, C - 0.045^{O}, D - 0.1^{O}. See Figure 2 for the equivalent analysis for EEZs, and Figure S1 for terrestrial regions.*

Supplemental text

Supplementary Methods

Supplementary Methods 1: Spatial biases in GBIF data

Analysing other biases in primary observation data

Given that GBIF was the largest dataset examined, and largely draws data from, or provides data to the other databases, bias analysis here was focused on GBIF data. Based on the Hughes et al 2021 analysis we also assessed the relationship between sampling intensity and distance from road, elevation, human modification (Theobold et al., 2024) and conversion pressure, as well as distance from the coast in marine systems. Firstly all road datasets were downloaded from OSM (Open Street Map), for those where downloadable data was unavailable Hotosim (humanitarian data) was used, and the UK was obtained from Ordnance Survey. To remove tracks and local roads we only selected major roads (road numbers starting with 511 or 513 categories within the OSM system). Obviously, many observations may be from smaller roads, but this should be reflected by the modification indices, and we wanted to not include too generous a classification of roads due to the variation in smaller roads and their classification in different regions. Roads were converted to rasters with a 1km resolution then reclassified to remove smaller road categories. We then used the Euclidean distance tool to measure the distance to the nearest road at a 1km resolution for each country. Maximum distance was set to the equivalent of 18 degrees. The Mosaic to new raster tool was then used to assemble maps for each region with the categorisation set to "minimum" to ensure the minimum distance was selected in areas adjacent to several countries. All datasets used are listed in Table S1.

In addition we calculated the percentage of the area covered (based on a 5km resolution of GBIF data with at least one record per 5km cell) at different distances from the roads, including 1km, 5km, 10 km, 50km, 100km, 250km, 500km, and >750km. This was calculated for each region for both marine and terrestrial areas, as well as overall in marine systems, and then the relationship between distances and coverage analysed using regressions.

The GBIF 2025 record density dataset was then converted into points of each geographic region and the extract values to points tool used to extract the distance to the road. The same procedure was then applied to extract out elevation, Human modification index and conversion pressure. For coastlines the "distance to coast" was calculated using a world country border map. The EEZs for each region were then dissolved based on the geographic region and used to clip the GBIF points for areas within each regional EEZ as well as the HighSeas. These points were then used to assess the relationship between the density of points and the distance to the coast. An exploratory regression was then applied to explore the relationship with sampling intensity and all of the factors, and the best model as well as the significance of all factors calculated.

Supplementary Methods 2:

Species population trends monitoring datasets

BioTIME, PREDICTS, FishRivTime and the Living Planet index (LPI) are all composite datasets aiming to measure species abundance changes over time or due to land or water use changes, which include both direct distribution data (OBIS and GBIF), published studies, and other sources of data. Thus whilst this data does overlap with the above sources of data, only data that satisfies certain criteria are included to attempt to assess changes in populations, diversity, or community structure over space or time.

A. BioTIME

For BioTIME data, spreadsheets were obtained of the number of species and occurrence records per country for the same taxonomic groups as used in the GBIF analysis, as well as the area with data (using a 5km resolution) for EEZs, Longhurst regions and ecoregions. Summaries of the number of species, and samples, as well as the percentage of each zone (Longhurst, EEZ, ecoregions) by calculating the total area of each of these and then cross-referencing the area covered by BioTIME data to provide the percentage covered. These spreadsheets were then connected to shapefiles of each region using "joins and connects" then mapped out to show patterns of coverage. Further complementary material was drawn from the recent paper (Dornelas et al., 2025). Complementary metrics were also recorded from RivFishTime using the same approach.

B. Living Planet Index and other monitoring databases

Different approaches were applied due to different availability of data and different modes of reporting and recording for each database. The Living Planet database makes the data available when a new report is released. Using the latest Living Planet data (WWF 2025) we downloaded all data, then calculated the number of species and populations monitored for each country. Data was then attached to shapefiles of each region using "joins and relates" in ArcMap to visualise patterns of survey effort. Additionally the number of times each population was observed was calculated, and the average timing between surveys calculated for different taxa to assess the effectiveness of data for monitoring. PREDICTS was summarised in the same way (numbers of samples and species per biome and country), whilst the monitoring data (Moussey et al., 2022) was analysed to show the number of monitoring programs per country for each taxa separately in marine, freshwater and coastal ecosystems.

Supplementary Results

Supplementary Results S1.

Taxonomic biases and patterns in GBIF data

Understanding the distribution of records can also facilitate understanding of how representative data is. For mammals the United States, followed by Australia in leading records until 2022 when France took over, representing 18% of all records by 2025 (whereas the US represented 12% of records). Whilst Brazil has records on the largest number of amphibian species since surpassing Colombia in 2015, the United States has had the largest number of total occurrence records for amphibians since 2010 (currently 22%), followed by Australia (16%). Birds show similar patterns with the US representing 47% of total occurrence records, and at least 39 countries with incomplete records, largely in Africa, Central Asia and various Pacific islands. The same pattern is true of ray-finned fishes (Actinopterygii) with 29% of all occurrence records in the United States, followed by France and Canada. Invertebrate taxa, however, show quite different patterns, for example Indonesia has records on the largest number of dragonfly and damselfly (Odonata) species, whilst the Netherlands, France and the UK have the largest number of Odonata occurrence records. The UK also leads for insects

overall with 28% of all records. However, for molluscs and arachnids the United States leads, with 20% and 16% of records respectively. For plants, European countries have the best coverage of data, with the UK leading for mosses (21% of records), followed by Sweden, Spain leading for Gymnosperms (13%) followed by the UK, and France leading for Angiosperms (18%). In all these instances, consistent increases are present globally, but in all instances the US has the greatest number of recorded species (though for Angiosperms Brazil did until recently). For Ferns however, whilst France leads for records (12%) China and Indonesia have the greatest number of recorded species. For Fungi the UK also led for both Sac Fungi (17% records) followed by Sweden, and Sweden led for Basidiomycota with 16% of records. Data is provided in Data S1).

Supplementary Results S2

Geospatial biases in GBIF data

In most regions distance to road had a significant relationship with sampling density 100% of the time (in Africa, Latin America and Oceania it was only 75%). However, what that relationship looked like varied. In many regions there was an equal split between positive and negative relationships, but for Canada and Alaska, the US, Northern Asia, Latin America 75% of instances were negative, and in Southeast and Southern Asia there was a significant negative relationship for 100% of instances. Only the Middle East always showed a positive relationship with sample density and road distance (likely associated with off-road and desert driving). Elevation was also generally negatively associated with sampling intensity, however, this was only significant for 100% of cases in Europe, Africa, Central Asia, the Middle East, the US and Northern Asia. Furthermore, of these, Africa and Latin America have a positive relationship, as does South Asia (though this is not always significant). Human modification was also a significant influence in most cases (only 75% in Africa, and not significant in Latin America). However, in all other incidences it has a 100% significant positive relationship, showing that the most modified areas are the best sampled. Conversion pressure (Oakleaf et al., 2024) shows significant relationships for most regions, however the outcomes are more varied, predominantly negative in Europe, the US, North Asia, and Southeast Asia; and predominantly positive in South Asia, and Africa. However only looking at occurrences per sampled cell will inevitably not reflect unsampled cells well, and area covered also needs to be considered.

For most regions the percentage of the area covered rapidly declined with distance from the road; though this was not universally the case, and it should be noted that as only major roads were included this necessarily omits smaller roads (as well as tracks) where observations may frequently take place. Significant negative relationships (based on an logarithmic regression) occurred in the case of Central Asia (R2: -0.932; Oceania -0.9677; SEA -0.82; SOA -0.7; LAM -0.971; EU-0.972; Can -0.978).

In terms of area from the coast there is both a strong overall negative trend (-0.909), and strong trends in all regions except Antarctica (US: -0.9955; EU -0.958; Af -0.92; L America -0.956; Middle East -0.877; N Asia -0.86; South Asia -0.83; Southeast Asia -0.79; Oceania -0.94). This

highlights how well coastlines are studied relative to all other regions, with an exponential decrease in the coverage of samples.

Supplementary Results S3. Biases in monitoring databases

BioTIME

BioTIME includes 56,400 taxa based 1,989,233 records extracted from 1,603,067 sample events, from 553,253 sampling locations, taken from 708 studies, all of which have a minimum of two sampling events taken at least two years apart (Dornelas et al., 2025). Taxonomic coverage has improved considerably between versions, most notably now including over 12% of Chordate species and 11% of described Annelid species, though remaining lower for other taxa (Data S3). For terrestrial regions, again the United States dominated the number of samples for most taxa, though for some (i.e Reptiles- Australia (27.8% species), Cyclostomata – France (34.6%), Fungi- UK (61.4%), Echinodermata- Canada (just more than the US at 39.7% species)) other regions had a larger share; these biases showing similarities to those shown within GBIF. For species, however, the US only dominated for Molluscs (39.7% species), Annelids (48.7% species) and Echinoderms (28.7% species). For reptiles, Australia again dominated 25.4% of species. Brazil had 41.1% of Amphibian species data, despite only having 2.1% of records, Germany had 50.5% of fungi species despite only hosting 13.4% of records, but for other taxa no single region held more than 25% of species or records. In total 51 terrestrial ecoregions had data within BioTIME, with only seven (largely small island) regions showing over 60% data coverage, with African and South American ecoregions having particularly low coverage, and few Asian regions with any data.

Spatial biases on land mirror those of GBIF, though taxonomic biases are less pronounced, with a better relative reflection of invertebrates. In Ocean regions sampling patterns also mirror those of GBIF with the best coverage on the Eastern US, around Europe, and to the South of Australia. Atlantic ocean coasts and continental shelf had much stronger data coverage than in the Indian or South Pacific oceans, and particularly deep sea (slope, abyssal and hadal) were poorly represented. Southern Ocean data is very sparse, with the entire West Antarctic (Weddell, Scotia, Bellingshausen, Amundsen and Ross seas) represented by a single location, except for plankton sampling. Arctic coverage is extremely patchy, even around the same island, such as Greenland. Surprisingly, some well known marine biodiversity hotspots, such as SE Asia and moderate richness of East Africa and the west coast of South America are poorly represented. Amongst the marine realms, nowhere had over 50% of their area covered, with the best coverage (46%) in the NW Atlantic Shelves, followed by New Zealand Coastal province at 13%. For EEZs, five had an over 50% data coverage, though three of these were around the Eastern US and Canada, with the other two around Belgium and Israel, conversely all tropical, and most oceanic island EEZs had little or no data.

RivFishTIME

As BioTIME has little inclusion of freshwater systems, RivFishTIME was created to monitor freshwater fish. The RivFishTIME database has 11,072 locations from 402 basins in 19 countries, and 944 fish species (Comte et al., 2021). The times between samples, methods, and

completeness of RivFishTIME is high enough to infer changes over time, however almost all analysis is from high income economies, with only one South American, 2 African countries within tropical ecosystems. Further time series have been conducted for over 6200 taxa in Europe have shown population changes over time (Pilotto et al., 2020), but is limited to Europe.

Living Planet Index

The LPI database shows that 20% of counts include only two population counts, whereas larger and more representative censuses were limited to few taxa (particularly birds and fish) (Data S4). Many counts also indicate activity rather than abundance, and may not be comparable due to short sampling periods. In total 2031 of the species in the LPR are Actinopterygii fish (39%) mainly due to fishing for food, and 24.3% of samples are from Canada alone (despite the highest diversity the highest diversity being in Brazil (15.42%), 1625 are birds (31.4%, but 52.3% are from Australia and Canada, representing only 10.7% of species), and 796 (15.4%) are mammals, with 329 amphibians (6%) monitored and all other groups populations measured even less. At a country level, Australia and Canada frequently have the highest number of samples, whilst by region Latin America had the greatest number (1758 species monitored), followed by North America (1367), whereas all other regions had fewer. Notably species counted the most often were almost all migratory wading birds, for example the Charadriiformes had 9273 populations of 211 species monitored (with 1083 populations of *Calidris ruficollis* alone), this was followed by Perciformes (3200 populations of 707 species, due to monitoring for food). For mammals artiodactyls (1280 populations, 125 species) and carnivores (1009 populations, 122 species) were the best monitored. The biases here replicate those of other datasets, though a larger proportion of fish are monitored than in other databases.

PREDICTS

PREDICTS assesses change on a gradient of disturbance. Like other datasets PREDICTS data demonstrates considerable biases (especially geographic), but has a greater inclusion of invertebrates than many other databases (Data S5). The greatest number of species was for arthropods (41724 including duplicates between countries) followed by plants at 28798, which also had the greatest number of samples, though plants had only marginally more samples than birds (11354 vs 11331) despite considerably more plant species (8212). Like other datasets PREDICTS is also dominated by high income economies, for example the UK has the most annelid samples and species (59.4% and 50.2%), (followed by New Zealand - 23% and 34%), and the highest number of mollusc samples (32.3%) despite low diversity (3.4%). Likewise, Japan has the highest number of Amphibian samples (41.8%) despite very low diversity (0.6%), whilst Madagascar had only 0.5% of amphibian samples, but these included 15.7% of species. Reptiles had relatively more balance in sampling with Australia having 35.3% of samples to 29% of species and Mexico having 23.9% of samples and 18.4% of species. For mammals the greatest number of samples also came from Australia (20%), but only 8.8% of species, whereas Brazil had 5.7% of samples, but 17% of species. For Arthropods, Brazil has 20.5% of samples and 11.47% of species, whereas Asia in particular shows smaller proportions of samples and high proportions of species (Data S3). For Fungi Italy had the most samples and species (29.8%, 30%) with the UK having the second most samples (21.9%) but few species, whereas Australia has 5% of samples, but has 24.7% of species. For biomes,

Temperate Broadleaf & Mixed Forests was by far the best sampled for Annelids, Arthropods, Molluscs, Amphibians, Plants and fungi (though for fungi Mediterranean Forests were almost as well sampled), but for most of these Tropical & Subtropical Moist Broadleaf Forests had more species, and many biomes were unsampled for the majority of taxa.

Species monitoring programs

The Monitoring programs dataset assembled national monitoring programs across taxa for terrestrial, freshwater, marine and coastal regions, and have demonstrably different patterns to other databases (Moussy et al., 2022; Data S6). Terrestrial monitoring was the most common with 820 monitoring programs from 94 countries. China had the most monitoring programs (62 programs) followed by France (53), primarily driven by birds (388 overall, 34 in China, 25 in France), followed by mammals (264 overall, with 26 in China and 22 in South Africa). This was followed by Plants (105 overall, 11 in South Africa) and reptiles (70 overall). Regional patterns were similar, with 402 overall, led by Europe at 323, programs 131 focused on birds, followed by Africa with 48 on birds and mammals. The next most monitored was Freshwater systems, which included 66 countries and 328 programs, with China leading at 39 programs (though 35 of these are birds). This was followed by mammals (51 programs) and fish (45 programs) with few countries showing large numbers of programs. At a regional level, birds also led with 178 programs, whilst geographically Europe led with 151 programs (65 on birds). Coastal programs also focused almost entirely on birds (with most in Europe), whilst mammals and birds were the main focus of marine programs (principally European mammals; 5 of the 10 European programs).

Data type	**Source/citation**	**Link**
Longhurst regions	MarineRegions (2025) Longhurst Provinces Ecological geography of the Sea (Longhurst, 1998) https://www.marineregions.org/gazetteer.php?p=details&id=22538	*https://hub.arcgis.com/datasets/schools-BE::longhurst-biogeographical-provinces/explore*
Exclusive Economic Zones	Flanders Marine Institute (2024). The intersect of the Exclusive Economic Zones and IHO sea areas, version 5. Available online at	https://www.marineregions.org

	https://www.marineregions.org/. https://doi.org/10.14284/699	
Ecoregions	Dinerstein, E., Olson, D., Joshi, A., Vynne, C., Burgess, N. D., Wikramanayake, E., ... & Saleem, M. (2017). An ecoregion-based approach to protecting half the terrestrial realm. BioScience, 67(6), 534-545.	*https://data-gis.unep-wcmc.org/portal/home/item.html?id=0127920779a64e3f98925f2d3da3b847*
Administrative areas	World Bank Official Boundaries	*https://datacatalog.worldbank.org/search/dataset/0038272/world-bank-official-boundaries*
UN regions	Unicef Regional Classifications	*https://data.unicef.org/regionalclassifications/*
Coral reefs	UNEP-WCMC (2022) Global Distribution of Coral Reefs https://data-gis.unep-wcmc.org/portal/home/item.html?id=0613604367334836863f5c0c10e452bf	*https://data-gis.unep-wcmc.org*
Road data	Geofabrik	*https://download.geofabrik.de/north-america.html*
Road data	European Environment Agency	*https://www.eea.europa.eu/data-and-maps/data/eea-reference-grids-2/gis-files/germany-shapefile*
Road data	Humanitarian data	*https://data.humdata.org/dataset/?q=Netherlands+roads&sort=score+desc%2C+last_modified+desc&ext_page_size=25*

Road data	Ordnance Survey UK	*https://osdatahub.os.uk/downloads/open/OpenRoads*
Global Human Modification Index	Theobald, D. M., Oakleaf, J. R., Moncrieff, G., Voigt, M., Kiesecker, J., & Kennedy, C. M. (2025). Global extent and change in human modification of terrestrial ecosystems from 1990 to 2022. Scientific Data, 12(1), 606.	*https://figshare.com/articles/dataset/Global_Human_Modification/7283087*
Global Human Modification Index	Theobald, D.M., Oakleaf, J.R., Moncrieff, G., Voigt, M., Kiesecker, J., and Kennedy, C.M. <in review>. Global extent and change in human modification of terrestrial ecosystems from 1990 to 2022. Scientific Data.	*https://zenodo.org/records/16907328*
Digital Elevation Model	Ince, E. S., Abrykosov, O., & Förste, C. (2024). GDEMM2024: Global Digital Elevation Merged Model 2024 for surface, bedrock, ice thickness, and land-type masks. Scientific Data, 11(1), 1087.	*https://datapub.gfz-potsdam.de/download/10.5880.GFZ.1.2.2024.002-Veebui/GDEMM2024_SUR.30s.tif*
Conversion pressure	Oakleaf, J., Kennedy, C., Wolff, N. H., Terasaki Hart, D. E., Ellis, P., Theobald, D. M., ... & Kiesecker, J. (2024). Mapping global land conversion pressure to support conservation planning. Scientific Data, 11(1), 830.	https://www.nature.com/articles/s41597-024-03639-9

Table S1. Data types used for analysis