ECOGRAPHY

Research

Sampling biases shape our view of the natural world

Alice C. Hughes, Michael C. Orr, Keping Ma, Mark J. Costello, John Waller, Pieter Provoost, Qinmin Yang, Chaodong Zhu and Huijie Qiao

A. C. Hughes (https://orcid.org/0000-0002-4220-1033) ✉ (ach_conservation2@hotmail.com), Center for Integrative Conservation, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Menglun, China. – M. C. Orr (https://orcid.org/0000-0002-9096-3008), C. Zhu and H. Qiao (https://orcid.org/0000-0002-5345-6234) ✉ (huijieqiao@gmail.com), Key Laboratory of Zoological Systematics, Inst. of Zoology, Chinese Academy of Sciences, Beijing, China. – K. Ma (https://orcid.org/0000-0001-9112-5340), Inst. of Botany, Chinese Academy of Sciences, Beijing, China. – M. J. Costello, School of Environment, Univ. of Auckland, Auckland, New Zealand. – J. Waller, GBIF, Global Biodiversity Information Facility, Copenhagen, Denmark. – P. Provoost, OBIS, Intergovernmental Oceanographic Commission (IOC) of UNESCO, Paris, France. – Q. Yang, State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang Univ., China.

Ecography
44: 1259–1269, 2021
doi: 10.1111/ecog.05926

Subject Editor: Jorge Soberon
Editor-in-Chief: Miguel Araújo
Accepted 19 May 2021

Spatial patterns of biodiversity are inextricably linked to their collection methods, yet no synthesis of bias patterns or their consequences exists. As such, views of organismal distribution and the ecosystems they make up may be incorrect, undermining countless ecological and evolutionary studies. Using 742 million records of 374 900 species, we explore the global patterns and impacts of biases related to taxonomy, accessibility, ecotype and data type across terrestrial and marine systems. Pervasive sampling and observation biases exist across animals, with only 6.74% of the globe sampled, and disproportionately poor tropical sampling. High elevations and deep seas are particularly unknown. Over 50% of records in most groups account for under 2% of species and citizen-science only exacerbates biases. Additional data will be needed to overcome many of these biases, but we must increasingly value data publication to bridge this gap and better represent species' distributions from more distant and inaccessible areas, and provide the necessary basis for conservation and management.

Keywords: biodiversity, data, distributions, global, macroecology, species richness

Introduction

Human knowledge of biodiversity is based on observations or specimens of different species that are used to determine their distributions. Given accessibility biases, models are often necessary to improve the limited resolution at which we can map life (Gaston 2000, Jetz et al. 2012). However, if occurrence records are biased in their collection, models may be unrealistic (Beck et al. 2014, Costello et al. 2015a, Qiao et al. 2015). Similarly, in correlative studies, trends may actually be reversed when accounting for sampling effort versus not (Hughes 2017). The ability to understand and protect life on Earth is, in turn, limited by present knowledge of the biases underlying the data, as these biases frame resulting perspectives and influence all analytical outcomes.

We must understand the spatial structure of biases to accurately reconstruct large-scale patterns, but prior studies have focused on specific regions, systems or taxa

NORDIC SOCIETY OIKOS

www.ecography.org

(Yesson et al. 2007, Mora et al. 2008, Daru et al. 2018). Whilst former analyses have explored some of these analyses, including geographic, political and accessibility biases (Kadmon et al. 2004, Boakes et al. 2010, Amano and Sutherland 2013, Meyer et al. 2016, Moudrý and Devillers 2020) most of these focus on limited regions or taxa (the largest analyzing terrestrial vertebrates only). Although countless algorithms have been developed for estimating biodiversity patterns (Soberón and Llorente 1993, Colwell and Coddington 1994, Gotelli and Colwell 2001, Mora et al. 2008, Chao et al. 2014, Colwell and Elsensohn 2014, Hsieh et al. 2016), the value and applicability of statistical method will always be determined by available data. One might assume that with millions of newly available records we would have improved our spatial knowledge and representation of life on the planet, yet progress has not been globally assessed, and biases have yet to be compared between the terrestrial and marine realms (Webb et al. 2010). Further, the proximal drivers of biases remain unexplored at the global scale. Consequently, it is tremendously difficult to map any single group at the global scale, much less all of life, and even then many analytical methods may fall short (García-Roselló et al. 2015, Orr et al. 2021a), except for the best-known groups such as birds (Rahbek et al. 2007, Jetz et al. 2012). Here, we explore the bias dynamics of some of the 'best-sampled' animals across terrestrial and marine systems to determine how accessibility has shaped our view of the natural world.

Whilst former analyses have highlighted that biases within globally available datasets mean that commonly used indices for biodiversity mapping do not reliably recreate richness patterns (and risk conflating richness with sampling intensity, and turnover with spatial gaps; Engemann et al. 2015), other indices which can reconstruct diversity patterns despite these biases are less studied. This analysis represents the first comprehensive global analysis representing both marine and terrestrial data, analyzing their spatial and taxonomic coverage, the biases encountered and the drivers of these biases. Prior analyses have unpacked various parts of these trends, but a holistic and standardized view across regions and taxa was lacking; thus, we selected taxa with both marine and terrestrial distributions (or saltwater and freshwater) so relative coverage could be compared in both contexts within taxa. Furthermore, whilst approaches have been developed to 'clean data' (Zizka et al. 2019, 2021, Jin and Yang 2020) these do not provide mechanisms to use existing data better to assess global diversity patterns or their completeness. Here, we also discuss the limits of effective use of existing data, and how gaps might most effectively be targeted.

Methods

Detailed methods are given in the Supporting information, so here we provide a briefer summary. A representative sample of higher taxa of vertebrates and invertebrates was selected for based on their distribution within both terrestrial and ocean systems. All four major terrestrial vertebrate groups were selected (Aves, Amphibia, Mammalia and Reptilia) along with a selection of groups found in both realms (Actinopterygii, Annelida, Arachnida, Cnidaria, Elasmobranchii, Gastropoda and Malacostraca). In total, 742 161 633 records were analyzed, including 38 313 609 of 57 252 510 potential OBIS records for marine systems (80% of animal records, 67% of all records) and 703 848 024 of 1.23 billion potential GBIF records. Records were filtered for synonyms, then distributions analyzed to assess the percentage of records in relation to roads, cities, shipping routes and coastlines (percentage of records at 0–1, 1–2.5, 2.5–5 and over 5 km) in R. We also assayed the levels of spatial coverage at different elevations (and elevation zones, including above and immediately below the treeline), biomes and within protected areas and KBAs.

Results

Global bias patterns

Terrestrial and marine systems are mostly unsampled (based on all databased records). At a 5 km resolution, < 7% of the Earth's surface was sampled, only 5% of the ocean and 11% of land (Fig. 1). A 10 km grid inflates sampling coverage up to three times for most groups (Supporting information). If birds are removed (87% of all GBIF and OBIS records), coverage drops to 4% for oceans and 7% for land (5 km grid; Table 1). Decreasing resolution inflates perceived coverage (Supporting information); for example, for ocean samples of Actinopterygii coverage changes from 0.21% at a 0.01° resolution to 47% at 1°, whilst on land coverage shifts from 0.24% to 46%. Likewise, mammals shift from 0.22% in ocean and 0.45% of land at 0.01° to 52 and 51% at 1°.

Taxonomic biases pervade; for OBIS, just 155/31 859 genera account for 50% of records, whereas in GBIF, 100 bird species account for 56% of the records (0.027% of species in this analysis) and 38.4% of total animal records. Further, 2% of GBIF animal records come from just *Anas platyrhynchos* (mallards) and *Sturnus vulgaris* (starlings); 11.4% of animal records come from ten bird species (Supporting information).

GBIF data fail to represent diversity across groups. For all taxa examined, 10% of all records covered < 0.1% of species and 25% of records covered < 0.5% of species. Surprisingly, the top 50% of records for each taxon represented < 4% of species with the exception of Cnidaria (7%). Birds are hugely overrepresented; the 85 most-sampled bird species each have more records individually than all reptiles. Despite similar numbers of species, reptiles have 0.7% of the number of bird records (Supporting information). For many taxa, a significant proportion of species is represented by a single record, while many may have no records (in OBIS, from 4% in mammals to 50% in Arachnida – Supporting information). This dominance is even more apparent when the percentage of observations from birds is explored at different resolutions (Supporting information): at high resolutions all terrestrial areas are dominated by birds, and at coarser resolutions,

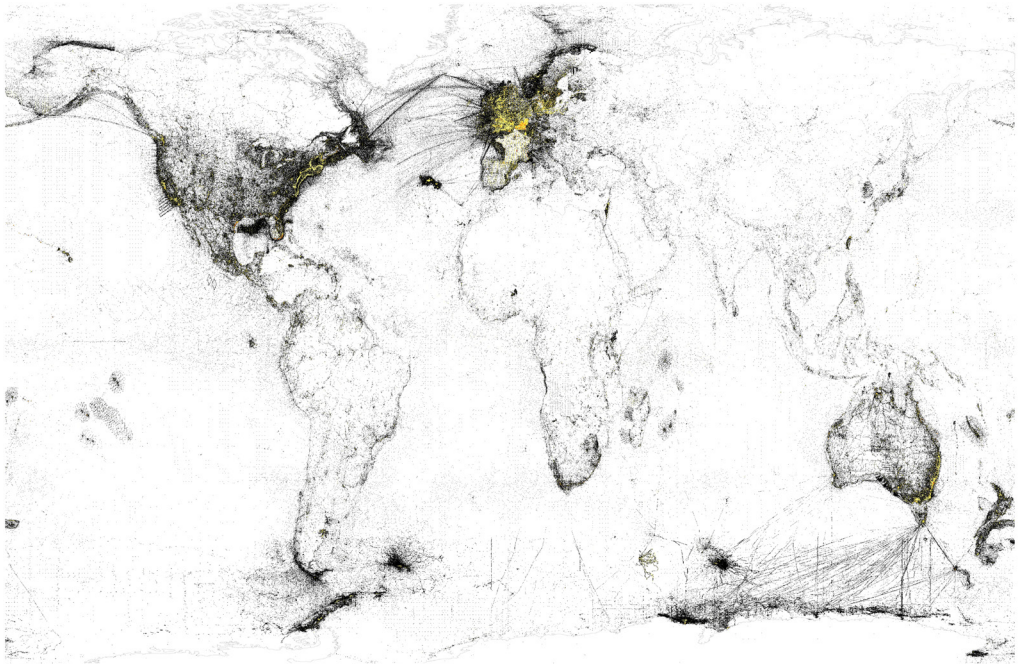

Figure 1. Areas with high numbers of records in GBIF and OBIS databases. Black 1–50 records, Yellow-red > 50 records at a 5 km resolution.

whilst developed countries achieve a better representation of other taxa, developing countries are still almost exclusively birds. Furthermore, even from the subset of areas with data, over 80% of samples from most taxa come from under 10% of sampled areas, though sampling is more concentrated into a smaller percentage of sampled areas in terrestrial than oceanic systems (Supporting information). Data are clustered in very small regions; for example, at 0.01° (1 km) resolution 1% of the sampled area includes on averaged 57% of sampled ocean species and 51.23% of terrestrial species. However, it should be noted that this concentration of data varies and can be as high as 77.2% of ocean species and 79.9% of terrestrial species in just 1% of sampled area (in the case of birds), compared to less well sampled groups (invertebrates are often under 40% of species, with annelids and arachnids showing low values in oceanic and terrestrial systems).

Globally, huge spatial biases exist (Supporting information); 79% of GBIF data comes from ten countries, 37% from USA. When terrestrial political areas < 100 km^2 are excluded, sampling coverage and GDP per capita are strongly related ($y = 0.2967x + 2.8446$, $R^2 = 0.2511$), with higher-GDP countries better covered. The GDP-per-capita of territories also influences sampling coverage, with developing countries less known despite many more species in tropical areas (via Exclusive Economic Zones; Supporting information).

Country centroid and gridding of points were less impactful. Gridding was largely from genuinely-gridded plant surveys (572 datasets accounting for 8.1% of all records, with the largest at 20 999 334 records), and only 0.01% of all non-plant records were located in country centroids, 0.02% at country or province.

Table 1. Percentage global coverage for terrestrial (including freshwater) and ocean areas at a 5 km resolution per-taxon. Coverage of areas 500 m under the treeline (TL500), above the treeline (Nival) and top global quartile (quart) and deep-sea (DS) sampling coverage.

	Ocean	Terrestrial	TL500	Nival	Quart	DS 1000–1800	DS > 1800
Actinopterygii	1.58	2.4	0.44	1.69	0.32		
Amphibia	0	2.02	0.48	1.39	0.39	0	0
Annelida	0.36	0.42	0.37	1.48	0.12	0.1	0.01
Arachnida	0.01	1.17	1.88	4.49	0.68	0	0
Aves	1.93	8.04	1.05	2.74	0.23	0.6	0.02
Cnidaria	0.54	0.21	0.03	0.05	0	0.42	0.01
Elasmobranchii	0.61	0.07	0.42	0.25	0.02	0.49	0.02
Gastropoda	0.61	1.41	3.87	8.96	1.24		
Malacostraca	1.48	0.8	0.12	0.47	0.02	0.1	0.01
Mammalia	1.2	3.35	0.42	1.35	0.17	0	0
Reptilia	0.18	2.59	0.19	0.67	0.03	0.4	0.02

Trends in accessibility

High mountains and deep seas

Sampling is limited by elevation/depth and ecosystem (Supporting information), with most roads and therefore sampling at lower elevations (Supporting information). When high elevations are examined, coverage for all groups is low, at < 1% coverage for most in the area abutting treeline (nival zone) and for most groups, except birds, above the treeline (Table 1, Supporting information); the top quartile of global elevations also has < 1% coverage for all groups except birds (1.24%), typically with just 1–5% of each taxon's records there.

Coverage is notably lower in marine systems overall, though deep-sea areas are particularly unsampled, with < 0.6% coverage for all groups between 1000 and 1800 m, and < 0.02% at depths below 1800 m for all groups. However, between 1000 and 1800 m contains 5.1% of records for specific groups (Cnidaria, Table 1).

Accessibility

Spatial biases are high across taxa, ranging from 41% (Actinopterygii) to 65% (Elasmobranchia) of non-marine records within 1 km of roads, with a further 40% within 2.5 km of roads (Fig. 2). At least of 80% of records were within 2.5 km of roads for each taxon independently. If genus averages are examined, the average percentage of localities > 5 km from roads increases across groups (Supporting information), as large numbers of rarely-recorded genera are found farther from roads. However, when examined separately, the proportion of some extinct genera (e.g. reptiles) away from roads increases, indicating that targeted sampling can overcome accessibility biases.

In oceans, coastal records (within 5 km) make up 30–50% of records for most taxa, but exceptions in the best-studied groups bias relative sampling levels and, thus, the mean patterns when points are aggregated. For example, in marine mammals and elasmobranchs, 4% of genera distributed near coastlines comprise 66% of records. Thus, these few, well-studied groups change overall patterns if considering only sample numbers (Supporting information).

The busiest shipping routes in the ocean only cover 2% of ocean area, but contain 18% of records and 41% of species. These include millions of records from the century-old Continuous Plankton Recorder surveys, where sampling nets are towed behind commercial ships (Reid et al. 2003). Other shipping routes contain 50% of ocean records, whilst covering 32% of the ocean and the open ocean has 32% of records, despite covering > 65% of the ocean (Supporting information).

In terrestrial systems, sampling increases > 5 km from coasts, except on islands (Supporting information). Sampling is also closely-associated with cities, with 22% (mammal) to 47% (arachnid) of records found within 1 km of cities (Supporting information). The average number of records per genus near cities increases for certain groups (birds, Supporting information), indicating that some genera are seen almost entirely near cities, especially in arid countries with limited agriculture (Supporting information).

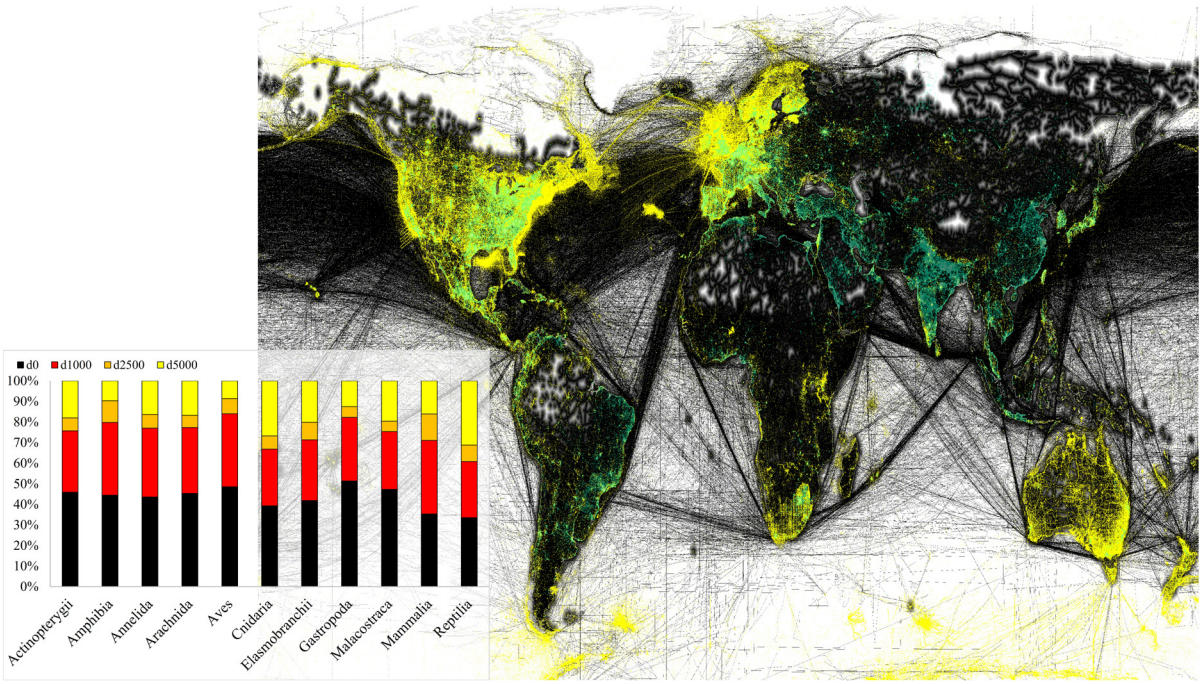

Figure 2. Map of global OBIS and GBIF data for selected taxa (yellow) with roads and shipping routes (black) and cities (green). Species data fall almost exclusively on these access routes, (barplot) with most distribution data within 2.5 km of roads on land, or either on the coast or a shipping route on the ocean. Regional spatial biases are also clear.

Observation type

How and by whom data were collected strongly impacts biases where citizen observations are popular. Thus, countries with more human observations and fewer specimen records have a higher percentage of records within 2.5 km of roads for birds (Supporting information, other groups showing little consistent impact). Consequently, the number of records relative to the richness within each group varies dramatically, with high numbers of easily-observed, common species such as ducks and rabbits (Supporting information). Observation type also relates to GDP, with most higher-GDP-c countries holding larger proportions of human-observed records, especially birds (birds: $y = 429.82x - 13\,634$, $R^2 = 0.1165$; overall: $y = 431.64x - 8611.4$, $R^2 = 0.211$, $p < 0.001$). Additionally, annual tourist records suggest that countries with low GDPs and high proportions of human-observation data relate to greater tourism (Supporting information), whereas smaller proportions of human-observed records relate to low GDP and tourism, such as in central Africa and most Pacific islands.

Biomes and realms

Terrestrial biomes are unevenly sampled, and 22% of the area within 2.5 km of roads falls in temperate broadleaf forests, containing over > 50% of mammal records and almost 50% of bird and amphibian records, despite representing 9% of land (Fig. 1). This biome has a mean of 200 (reptiles) to 57 074 (birds) records per genus recorded, in contrast with mangroves at 14 and 360 records per genus recorded, respectively. Biome sampling biases link to inaccessibility, with some of the most-diverse biomes (i.e. tropical) undersampled versus temperate biomes. Natural grasslands are even worse sampled, Montane grassland has one mammal record approximately every 32 km^2, whereas moist tropical forest has 1/15 km^2 and temperate forest 1/1 km^2. If sampling is plotted as a cartogram (Fig. 3A), a bimodal latitudinal gradient in sampling results, showing how under-represented tropical biomes are globally (Fig. 3B). Marine realms show similar biases, with just two of 30 marine realms comprising 47% of records, yet only 10% of genera, while the two richest realms (17% of genera) include 9% of records (Supporting information).

Protection and KBAs

Key biodiversity areas have 27% area sampling coverage for land and 18% for oceans. Marine protected areas (MPAs) have lower sampling coverage (10%) than terrestrial protected areas (16%). Some large MPAs in the open ocean are consequently undersampled, ranging from 0.02% area coverage in Arachnida (largely mites) to 5% in birds (Supporting information). Unsurprisingly, birds have the greatest terrestrial coverage (16%), nearly double that of the next-highest group (mammals, 8%), quadrupling the best-sampled invertebrates (Gastropoda at 3%), reflecting both the numerosity of bird records and the emphasis on birds in KBA designation. Marine KBAs have lower coverage at 7% for birds and also lower coverage for all other groups (Supporting information). Within protected areas, 7% fewer records were located within 2.5 km of roads than outside protected areas on average per genus.

Discussion

Overview

The digital revolution has transformed the sciences. Ecologists, once limited to single-site studies, are now challenged by sheer data volume. However, these data represent a tiny proportion of the planet, and are unrepresentative across space and the tree of life. Though surveys have been conducted globally, a lack of institutional support, recognition and capacity exacerbates existing trends on data availability, leading to the biased global databases, providing a coverage of under 7% of the world's surface at even a moderate resolution (5 km) and under 1% for most taxa at higher resolutions. Whilst some of these patterns have been explored previously, even some of the most comprehensive (Meyer et al. 2015) only includes 21% of the data and 5.6% of species in our study, and significant effort has been made to improve the quality of GBIF data and correct bias in the intervening years (Moudrý and Devillers 2020) based upon recommendations to improve accuracy and usability of the data (Anderson et al. 2015). Sampling is universally poor at < 11% for terrestrial and ~5% for marine areas, barely touching deep sea or high elevations. Regional biases are well-known (Supporting information; Martin et al. 2012); the US alone represents 44% of available terrestrial vertebrate records. The top ten countries have 82% of records; yet, this is limited to Europe, USA, Australia and South Africa, leaving 18% to the remaining 240 (96%) of countries. In addition, whilst coarse (1-degree) cells show that over 50% of the planet is sampled, this reduces to under 1% when 1 km cells are used as a basis, and to understand community dynamics (especially in heterogeneous landscapes) high resolutions are necessary to avoid conflating richness with turnover.

Although birds are better sampled than other vertebrates, invertebrates pale in comparison to vertebrates despite comprising the vast majority of all named species (Costello et al. 2013a), showing that both regional and taxon-specific efforts are necessary to improve our view of the natural world. Whilst these taxonomic biases have been documented previously (Troudet et al. 2017), increasing data availability has actually exacerbated this disparity. However, even these basic coverage statistics can be misleading, as changing grain-size dramatically alters area coverage estimates two- or three-fold. Many macroecological studies (Tittensor et al. 2010) make BAD (best-available-data) arguments and use coarse resolutions to explore ecological patterns, yet unaccounted-for topographic and climatic heterogeneity limit the meaningfulness of such analyses.

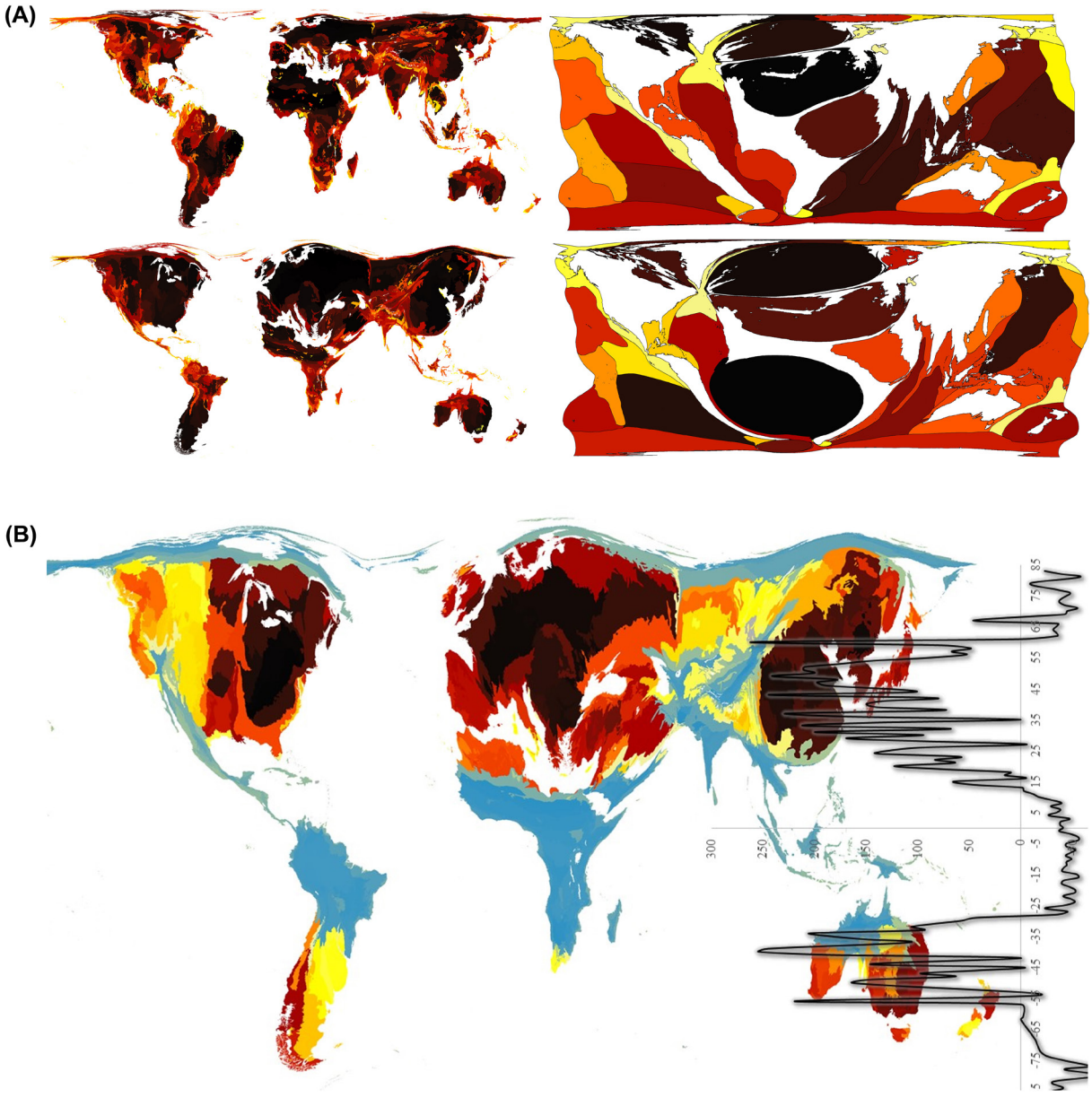

Figure 3. (A) Area cartograms for sampling and species richness per-biome using GBIF (left) and OBIS (right) data. Areas are resized based on relative species numbers (top) and sampling density (bottom). Diversity varies from pale yellow (low diversity) through red-black for increasing diversity. (B) Change in size relative to sampling density-terrestrial areas oversampled more than average relative to size are coloured orange-red-black, those undersampled green-blue. Graph shows relative size change based on increases or decreases relative to average sampling latitudinally per unit area.

Though publicly-available distribution data provide useful temporal (seasonal change) data for birds, coarse resolution data cannot be used for any form of regional assessment for most areas (La Sorte and Somveille 2019). In groups like mammals, this may be not only due to the lack of tropical data, but the need for specialist skills for key, understudied groups like bats, which often have the highest mammalian diversity in tropical regions. For other vertebrates and virtually all invertebrates, analysis on biodiversity would prove challenging. Taxonomic coverage is also a challenge, with 50% of records limited to < 2% of species within any given taxon for most groups.

Drivers of bias

The majority of records for all non-marine groups fell within 2.5 km of a road (averaging 47–56% within 1 km, and a further 36% between 1 and 2.5 km), and these biases become particularly pronounced where citizen observations dominate (Fig. 2, Supporting information). For birds, citizen science exacerbates biases while scientific surveys reduce them in terms of biome as well as road proximity bias. Further, the percentage of records made through human observation shows a strong relationship with GDP/capita, exacerbating

global and regional biases as countries with lower GDP may have lower research capacity and access to resources. Consequently, high sampling coverage is almost exclusive to developed countries (Fig. 1), this is in part responsible for the bimodal gradient in global terrestrial sampling (Fig. 3), which has also been documented in marine environments (Menegotto and Rangel 2018).

Roads provide access to land, rivers and lakes, yet for the ocean a much smaller area is easily accessible and citizen science is thereby limited. Regardless, disproportionate sampling falls within the immediate vicinity of roads, coasts and shipping routes, restricting sampling to a subset of systems and species. This limits our knowledge of species distributions to disturbed, sub-optimal habitats where common, human-associated species thrive. These differences fall in no small part to the inputs of citizen scientists, exacerbating biases and vastly increasing the representation of certain groups of 'supertramps' such as ducks and rabbits (Supporting information).

We do not seek to dismiss the use of citizen science data for understanding biological patterns (Amano et al. 2016), and although biases that can exist in such data have been acknowledged (Isaac and Pocock 2015), its impact on global diversity mapping is less well described. Our results highlight the need for comparable data from less-accessible areas where citizen science approaches are insufficient; this is possible through museum specimen digitization and data sharing, but such efforts are limited by poor funding and academic models that fail to recognize or reward generating and sharing distributional data (Costello 2009). Data for the most popular taxa in both terrestrial and marine realms (i.e. birds, cetaceans, elasmobranchs, primates) show the lowest proximity bias to roads, seaways and coasts, demonstrating that popularity (and funding) enables more-representative data.

The consequences of biased data

The consequences of lacking data from more-diverse, intact ecosystems and their species are manifold (Fahrig and Rytwinski 2009). This shifts the understanding of species requirements to the most-disturbed, often least-optimal areas, impeding knowledge of their optimal habitats and tolerances. Global maps now show the extent of anthropogenic impacts for a range of drivers (Bowler et al. 2020), but the information on what species and habitats are being impacted is highly-biased. Species distribution models failing to include additional records from intact habitat cannot generate accurate species ranges or biodiversity patterns, even when compensating with sophisticated statistical techniques (Graham et al. 2004, Qiao et al. 2017). Consequently, biodiversity in intact areas may be underestimated, undervaluing their conservation status. Thus, though studies have frequently found that carefully-used data can be highly informative for understanding patterns (García-Roselló et al. 2015), additional data are needed to address current biases in representation of various environmental facets even at smaller, regional scales for poorly-known areas, as resolution is critically important in deciding analysis outcomes (Peterson and Watson 1998, Stockwell and Peterson 2003, Rahbek 2005, Hurlbert and Jetz 2007, Lira-Noriega et al. 2007). Though statistical and model-based approaches are often applied to attempt to correct for data shortfalls and biases, even the best of interpolations and extrapolations need to be verified and calibrated by observations. Science advances based on records of the real world and our analyses illustrate that circumspection is needed about any generalities in spatial and temporal trends in biodiversity for most of the world.

Pervasive biases inhibit the ability to predict and prevent biodiversity loss to global changes, and even 'key biodiversity areas' lack data for most taxa for comparable analysis within or between regions. Put simply, if one does not know the true past or present state of habitat, they cannot reasonably know its future, regardless of the methods used. Montane areas are regarded to be at great ecological risk from climate change, but our analyses show that they are some of the worst-sampled. High-elevation areas are sampled as poorly as the deep sea, making it impossible to map most species sensibly. These biases may also hinder assessments of fragmentation or edge effects by obscuring the negative impacts of disturbance via over-inflated richness in such areas. In the oceans, where the impact of sound and vibration are increasingly well-known (Nagelkerken et al. 2019), data are similarly almost entirely from the most-disturbed areas, with the greatest ship-traffic (coasts and seaways). This dramatically limits an understanding of the wider impacts of disturbance, because the animals in these systems have been exposed to such disturbances for decades, with more sensitive species already extirpated. These habitats are no longer what they were even when early expeditions took place, making what few data are available, in many cases, unrepresentative.

Unfortunately, the resources presently available are not yet fit for the purpose of understanding and protecting global biodiversity, although many researchers attempt to do this. In common practice, current views of the natural world are limited either to 'expert opinion' IUCN maps, where inherent knowledge gaps have huge consequences of data use, including persistent administrative-area biases (Hughes et al. 2021), or a coverage completeness for < 6.74% of the planet, based on terrestrial, temperate lowlands near roads and cities in developed countries and their coastal regions. Neither data set is currently sufficient for truly global analyses, and efforts must be made to better integrate these data, such that they may help alleviate the weaknesses and biases of each other. A major failing of expert-opinion maps is that source data are rarely available so neither the spatial nor temporal evidence behind the map is known, preventing uncertainty analysis. These maps may be improved by modelling the geographic ranges of species based on the relationship of field observations to environmental variables, such as developed by AquaMaps for 25 000 aquatic species (Kaschner et al. 2019). While point samples suffer from errors of omission, as demonstrated here, species range maps have errors of commission. Species are unlikely to be present at every location within their geographic range due to local habitat suitability,

fluctuating abundance and variable detectability and only recent field records can detect changes in species' abundances. Furthermore, whilst approaches such as species distribution models are a popular way to overcome such biases and reconstruct species ranges, less than 1% of the planet has records available at the 1 km resolution that is most popular for such analysis.

Overcoming current biased data

Our biased worldview cannot be rectified by further data that build upon these access-driven biases, through citizen science or other means, requiring involvement of the global scientific community, and further efforts to mobilise existing inaccessible data, such as GBIF's BIFA and BID initiatives. Thus, simply sampling more may as much perpetuate biases as address them. A strategic approach to share data and fill gaps is needed. We do not suggest that everywhere and every taxon needs regular sampling, because some assumptions allow filling of gaps in species distributions. Rather, a stratified sampling may be more representative and cost-efficient than the present idiosyncratic approach, as suggested for oceans based on environmental heterogeneity (Costello et al. 2018). However, an important first step will be to know what data exist in useable form, as if we judge data biases on only public data then we may prioritize the wrong regions (Orr et al. 2021a).

Part of the reluctance to share data in ecology comes from this discipline having evolved on a local level, before access to data became a fundamental necessity to understand and manage global diversity (Costello 2009, Stork 2018). Unlike molecular biology, where resources are generally archived on the singular repository GenBank, data are more diffuse in other biological fields. The late recognition of the need for such standardization has led to the fragmentation of knowledge and data in many forms in the literature, and hundreds of online, unlinked databases, precluding easy analysis (Poisot et al. 2019). Established regional databases may also prohibit access by international researchers, such as the Malaysian Mybis database, preventing even the analysis of ranges for endemic or small-ranged species with much of their range extending between Malaysia and neighboring countries. Similar lacking data for most taxa in China, Russia, India and others may also prove problematic. In general, Africa and Asia will likely require the most effort to mobilize sufficient data for reliable biodiversity mapping and management, which have been major focuses of GBIF data mobilization efforts. Understanding global biodiversity patterns will require not simply the generation of new data, but the liberation and improvement of existing data which may be online on platforms like Dryad or the literature, in museum collections, on computer hard-drives, or is available in partially complete formats. Many data exist which have essentially 'leaked' out of the mobilization pipeline, such as at the stage of georeferencing, and this is a major challenge to leveraging databased specimen records (Soberón and Peterson 2004, Peterson et al. 2018). Biases in sampling (Supporting information) do not represent all data, instead representing the combination of genuine gaps and hidden data, as huge, inaccessible collections exist globally and the lack of access precludes analysis or complete knowledge of which gaps most need filling.

These systemic data gaps can be overcome through several means. First, strategic inventorying and digitization can produce less-biased information (Meyer et al. 2016). For example, standardized surveys such as the Continuous Plankton Recorder and Reef Life Survey, have enabled better sampling of plankton and reef fauna across ocean realms although biases remain (Costello et al. 2017). Second, existing data can be augmented with additional metadata to enable bias accounting. GBIF and OBIS recently developed the 'EventCore' to further standardize data collection events, enabling inclusion of sample data as well species records. However, an overarching Project-Core framework (De Pooter et al. 2017) could enable associated metadata such as the collector effort (hours) for a project, its mapped geographic scope, and sampling methods, which could then be compared across projects to methodologically and spatially control for collection bias.

Finally, increased funding, institutional and data sharing requirements within grants, and career recognition of data generation could all greatly enhance data availability for other taxa from more-diverse and less-accessible areas and facilitate the sharing of the data needed to understand global biodiversity patterns (Costello et al. 2013b, 2015b, La Sorte and Somveille 2019). Though more difficult than making BAD arguments and simply using what data are available, these steps will be necessary for scientists to realistically predict and prevent biodiversity loss.

Synthesis

The current global view of biodiversity is constrained to what can be seen from easily-accessible areas (roads, coast, etc.). The most-diverse ecosystems and specialist species are underrepresented, preventing management and conservation of diversity or prediction of how ecosystems will be impacted by global changes. Before we construct a global, well-organized and unbiased database, we must recognize the uncertainty of the conclusions that we make via online data sources (GBIF/OBIS). We also better acknowledge the limitations of knowledge, and ensure that the assumptions most analyses include are met to ensure meaningful analysis. Remedying these biases is not possible through modelling alone without sacrificing the least-known areas and systems. Whilst data sharing is now often mandated by journals, the platform is often unspecified, leading to a fragmentation of species data, from supplements and appendices, as well as online repositories such as Zenodo, Figshare and Dryad (Guralnick and Hill 2009, Rüegg et al. 2014), making comprehensive assessment impossible. For example, most public bee data are spread across five major online repositories, but this does not include many single-institution databases and private datasets yet to

be shared (Orr et al. 2021a). A standardized, singular platform for data would be ideal (distributional, molecular, morphological, taxonomic, etc.; Orr et al. 2020, 2021b), but it may be that the most we can hope for is a better interlinking of extant databases, which itself may already prove quite challenging. Such efforts could be explicitly linked to National Biodiversity and Action Plans (NBSAPs) that countries create within the framework of the Convention of Biological Diversity. Such targets for the post-2020 biodiversity framework are in discussion and would greatly facilitate these efforts, as in molecular biology (GenBank), enabling scientists to better mobilize and be recognized for their contributions to protecting global biodiversity, and better analyze and target the various data-gaps to provide a more complete understanding of the natural world.

Acknowledgements – We thank Townsend Peterson for useful comments.
Funding – ACH was supported by the Chinese National Natural Science Foundation (NSFC) no. U1602265. HJQ was supported by NSFC31772432. MCO was supported by The National Science Fund for Distinguished Young Scholars (31625024), and partially by the NSFC International Young Scholars Program (31850410464) and the Chinese Academy of Sciences (CAS) President's International Fellowship Initiative (2018PB0003, 2020PB0142). ACH also acknowledges by the Strategic Priority Research Program of CAS (XDA20050202), the High-End Foreign Experts Program of Yunnan. Province (Y9YN021B01), and the CAS 135 program (2017XTBG-T03).

Author contributions

Alice Hughes and **Michael Orr** are co-first, and equally contributing, authors**. Alice Hughes**: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Resources (equal); Software (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review and editing (equal). **Michael Orr**: Conceptualization (equal); Data curation (equal); Funding acquisition (equal); Investigation (equal); Project administration (equal); Resources (equal); Writing – original draft (equal); Writing – review and editing (equal). **Keping Ma**: Writing – review and editing (equal). **Mark Costello**: Conceptualization (equal); Validation (equal); Writing – review and editing (equal). **John Waller**: Methodology (equal); Validation (equal); Writing – review and editing (equal). **Pieter Provoost**: Methodology (equal); Validation (equal); Writing – review and editing (equal). **Qinmin Yang**: Resources (equal); Writing – review and editing (equal). **Chaodong Zhu**: Writing – review and editing (equal). **Huijie Qiao**: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Resources (equal); Validation (equal); Visualization (equal); Writing – review and editing (equal).

Data availability statement

All code files are available on the GitHub of HQ <https://github.com/qiaohj/gbif_obis>, data layers are available from Dryad <https://doi.org/10.5061/dryad.zw3r2287z>, and original data are open access via GBIF and OBIS via download links given in the Supplementary information.

References

Amano, T. and Sutherland, W. J. 2013. Four barriers to the global understanding of biodiversity conservation: wealth, language, geographical location and security. – Proc. R. Soc. B 280: 20122649.
Amano, T. et al. 2016. Spatial gaps in global biodiversity information and the role of citizen science. – Bioscience 66: 393–400.
Anderson, R. et al. 2015. Final report of the Task Group on GBIF Data Fitness for use in distribution modelling. – Global Biodiversity Information Facility.
Beck, J. et al. 2014. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. – Ecol. Info. 19: 10–15.
Boakes, E. H. et al. 2010. Distorted views of biodiversity: spatial and temporal bias in species occurrence data. – PLoS Biol. 8: e1000385.
Bowler, D. E. et al. 2020. Mapping human pressures on biodiversity across the planet uncovers anthropogenic threat complexes. – People Nat. 2: 380–394.
Chao, A. et al. 2014. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. – Ecol. Monogr. 84: 45–67.
Colwell, R. K. and Coddington, J. A. 1994. Estimating terrestrial biodiversity through extrapolation. – Phil. Trans. R. Soc. B 345: 101–118.
Colwell, R. K. and Elsensohn, J. E. 2014. EstimateS turns 20: statistical estimation of species richness and shared species from samples, with non-parametric extrapolation. – Ecography 37: 609–613.
Costello, M. J. 2009. Motivating online publication of data. – BioScience 59: 418–427.
Costello, M. J. et al. 2013a. Can we name Earth's species before they go extinct? – Science 339: 413–416.
Costello, M. J. et al. 2013b. Biodiversity data should be published, cited and peer reviewed. – Trends Ecol. Evol. 28: 454–461.
Costello, M. J. et al. 2015a. Factors influencing when species are first named and estimating global species richness. – Global Ecol. Conserv. 4: 243–254.
Costello, M. J. et al. 2015b. Conservation of biodiversity through taxonomy, data publication and collaborative infrastructures. – Conserv. Biol. 29: 1094–1099.
Costello, M. J. et al. 2017. Methods for the study of marine biodiversity. – In: The GEO handbook on biodiversity observation networks. Springer, pp. 129–163.
Costello, M. J. et al. 2018. Stratifying ocean sampling globally and with depth to account for environmental variability. – Sci. Rep. 8: 11259.
Daru, B. H. et al. 2018. Widespread sampling biases in herbaria revealed from large-scale digitization. – New Phytol. 217: 939–955.

De Pooter, D. et al. 2017. Toward a new data standard for combined marine biological and environmental datasets-expanding OBIS beyond species occurrences. – Biodivers. Data J. 5: e10989.
Engemann, K. et al. 2015. Limited sampling hampers 'big data' estimation of species richness in a tropical biodiversity hotspot. – Ecol. Evol. 5: 807–820.
Fahrig, L. and Rytwinski, T. 2009. Effects of roads on animal abundance: an empirical review and synthesis. – Ecol. Soc. 14(1): 21.
García-Roselló, E. et al. 2015. Can we derive macroecological patterns from primary Global Biodiversity Information Facility data? – Global Ecol. Biogeogr. 24: 335–347.
Gaston, K. J. 2000. Global patterns in biodiversity. – Nature 405(6783): 220–227.
Gotelli, N. J. and Colwell, R. K. 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. – Ecol. Lett. 4: 379–391.
Graham, C. H. et al. 2004. New developments in museum-based informatics and applications in biodiversity analysis. – Trends Ecol. Evol. 19: 497–503.
Guralnick, R. and Hill, A. 2009. Biodiversity informatics: automated approaches for documenting global biodiversity patterns and processes. – Bioinformatics 25: 421–428.
Hsieh, T. C. et al. 2016. iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). – Methods Ecol. Evol. 7: 1451–1456.
Hughes, A. C. 2017. Mapping priorities for conservation in Southeast Asia. – Biol. Conserv. 209: 395–405.
Hughes, A. C. et al. 2021. Effectively and accurately mapping global biodiversity patterns for different regions and taxa. – Global Ecol. Biogeogr. doi: 10.1111/geb.13304
Hurlbert, A. H. and Jetz, W. 2007. Species richness, hotspots and the scale dependence of range maps in ecology and conservation. – Proc. Natl Acad. Sci. USA 104: 13384–13389.
Isaac, N. J. B. and Pocock, M. J. O. 2015. Bias and information in biological records. – Biol. J. Linn. Soc. 115: 522–531.
Jetz, W. et al. 2012. The global diversity of birds in space and time. – Nature 491: 444–448.
Jin, J. and Yang, J. 2020. BDcleaner: a workflow for cleaning taxonomic and geographic errors in occurrence data archived in biodiversity databases. – Global Ecol. Conserv. 21: e00852.
Kadmon, R. et al. 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. – Ecol. Appl. 14: 401–413.
Kaschner, K. et al. 2019. AquaMaps: Predicted range maps for aquatic species. – Electronic publication, <www.aquamaps.org>, ver. 09/2019.
La Sorte, F. A. and Somveille, M. 2020. Survey completeness of a global citizen-science database of bird occurrence. – Ecography 43: 34–43.
Lira-Noriega, A. et al. 2007. Scale dependency of diversity components estimated from primary biodiversity data and distribution maps. – Divers. Distrib. 13: 185–195.
Martin, L. J. et al. 2012. Mapping where ecologists work: biases in the global distribution of terrestrial ecological observations. – Front. Ecol. Environ. 10: 195–201.
Menegotto, A. and Rangel, T. F. 2018. Mapping knowledge gaps in marine diversity reveals a latitudinal gradient of missing species richness. – Nat. Commun. 9: 4713.
Meyer, C. et al. 2015. Global priorities for an effective information basis of biodiversity distributions. – Nat. Commun. 6: 8221.
Meyer, C. et al. 2016. Multidimensional biases, gaps and uncertainties in global plant occurrence information. – Ecol. Lett. 19: 992–1006.
Mora, C. et al. 2008. The completeness of taxonomic inventories for describing the global diversity and distribution of marine fishes. – Proc. R. Soc. B 275: 149–155.
Moudrý, V. and Devillers, R. 2020. Quality and usability challenges of global marine biodiversity databases: an example for marine mammal data. – Ecol. Info. 56: 101051.
Nagelkerken, I. et al. 2019. Consequences of anthropogenic changes in the sensory landscape of marine animals. Oceanography and marine biology. – CRC Press, Taylor & Francis Group, Boca Raton.
Orr, M. C. et al. 2020. Three questions: how can taxonomists survive and thrive worldwide? – Megataxa 1: 19–27.
Orr, M. C. et al. 2021a. Global patterns and drivers of bee distribution. – Curr. Biol. 31: 451–458.
Orr, M. C. et al. 2021b. Taxonomy must engage with new technologies and evolve to face future challenges. – Nat. Ecol. Evol. 5: 3–4.
Peterson, A. T. et al. 2018. Data leakage and loss in biodiversity informatics. – Biodivers. Data J. 6: e26826.
Peterson, A. T. and Watson, D. M. 1998. Problems with areal definitions of endemism: the effects of spatial scaling. – Divers. Distrib. 4: 189–194.
Poisot, T. et al. 2019. Ecological data should not be so hard to find and reuse. – Trends Ecol. Evol. 34: 494–496.
Qiao, H. et al. 2015. No silver bullets in correlative ecological niche modelling: insights from testing among many potential algorithms for niche estimation. – Methods Ecol. Evol. 6: 1126–1136.
Qiao, H. et al. 2017. Using data from related species to overcome spatial sampling bias and associated limitations in ecological niche modelling. – Methods Ecol. Evol. 8: 1804–1812.
Rahbek, C. 2005. The role of spatial scale and the perception of large-scale species-richness patterns. – Ecol. Lett. 8: 224–239.
Rahbek, C. et al. 2007. Predicting continental-scale patterns of bird species richness with spatially explicit models. – Proc. R. Soc. B 274: 165–174.
Reid, P. C. et al. 2003. The Continuous Plankton Recorder: concepts and history, from Plankton Indicator to undulating recorders. – Progr. Oceanogr. 58: 117–173.
Rüegg, J. et al. 2014. Completing the data life cycle: using information management in macrosystems ecology research. – Front. Ecol. Environ. 12: 24–30.
Soberón, J. and Llorente, J. 1993. The use of species accumulation functions for the prediction of species richness. – Conserv. Biol. 7: 480–488.
Soberón, J. and Peterson, T. 2004. Biodiversity informatics: managing and applying primary biodiversity data. – Phil. Trans. R. Soc. B 359: 689–698.
Stockwell, D. and Peterson, A. T. 2003. Comparison of resolution of methods used in mapping biodiversity patterns from point-occurrence data. – Ecol. Indic. 3: 213–221.
Stork, N. E. 2018. How many species of insects and other terrestrial arthropods are there on Earth? – Annu. Rev. Entomol. 63: 31–45.

Tittensor, D. P. et al. 2010. Global patterns and predictors of marine biodiversity across taxa. – Nature 466: 1098–1101.
Troudet, J. et al. 2017. Taxonomic bias in biodiversity data and societal preferences. – Sci. Rep. 7: 9132.
Webb, T. J. et al. 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: e10223.
Yesson, C. et al. 2007. How global is the global biodiversity information facility? – PloS One 2(11): e1124.
Zizka, A. et al. 2019. CoordinateCleaner: standardized cleaning of occurrence records from biological collection databases. – Methods Ecol. Evol. 10: 744–751.
Zizka, A. et al. 2021. sampbias, a method for quantifying geographic sampling biases in species distribution data. – Ecography 44: 25–32.