nature ecology & evolution

Article

https://doi.org/10.1038/s41559-023-02047-3

# Mass production of unvouchered records fails to represent global biodiversity patterns

Received: 26 July 2022

Accepted: 26 March 2023

Published online: 01 May 2023

Check for updates

Barnabas H. Daru [1] & Jordan Rodriguez [2]

The ever-increasing human footprint even in very remote places on Earth has inspired efforts to document biodiversity vigorously in case organisms go extinct. However, the data commonly gathered come from either primary voucher specimens in a natural history collection or from direct field observations that are not traceable to tangible material in a museum or herbarium. Although both datasets are crucial for assessing how anthropogenic drivers affect biodiversity, they have widespread coverage gaps and biases that may render them inefficient in representing patterns of biodiversity. Using a large global dataset of around 1.9 billion occurrence records of terrestrial plants, butterflies, amphibians, birds, reptiles and mammals, we quantify coverage and biases of expected biodiversity patterns by voucher and observation records. We show that the mass production of observation records does not lead to higher coverage of expected biodiversity patterns but is disproportionately biased toward certain regions, clades, functional traits and time periods. Such coverage patterns are driven by the ease of accessibility to air and ground transportation, level of security and extent of human modification at each sampling site. Conversely, voucher records are vastly infrequent in occurrence data but in the few places where they are sampled, showed relative congruence with expected biodiversity patterns for all dimensions. The differences in coverage and bias by voucher and observation records have important implications on the utility of these records for research in ecology, evolution and conservation research.

The ongoing biodiversity crisis has inspired efforts to gather large amounts of biodiversity data in case organisms go extinct even before being discovered[1–4]. However, the biodiversity information commonly gathered is derived from either vouchers or direct field observations. A voucher is a specimen or sample preserved in a natural history collection that documents the existence of an organism at a particular time and space in a way that ensures scientific reproducibility ('voucher' records henceforth)[5]. The physical specimen can additionally serve as a tangible and verifiable source for new studies such as biotic interactions, disease lesions and imprints of physiological processes[6–10]. On the other hand, a field observation is a secondary voucher that captures supplemental information about an organism but is not traceable to tangible physical material in a museum or herbarium ('observation' records henceforth). The development of mobile applications and citizen-science programmes such as iNaturalist[11] or eBird[12] allows amateur observers to collect large volumes of observations and submit them electronically to centralized databases in a democratic way[13]. In some cases, structured surveys by professional scientists and researchers can be registered as observations[14,15]. Such observation-based data are generated inexpensively and are invaluable for exploring spatial and temporal patterns for many taxonomic groups[16,17]. The extent to which these disparate records differ from each

[1]Department of Biology, Stanford University, Stanford, CA, USA. [2]Department of Biology, Institute of Ecology and Evolution, University of Oregon, Eugene, OR, USA. e-mail: bdaru@stanford.edu

other and, consequently, represent different patterns of biodiversity remains less understood.

Biodiversity records are generally products of non-random sampling and thus prevalent with coverage gaps and biases that can render them of limited use in global biodiversity science as resulting coverage gaps and biases can potentially lead to spurious description and interpretation of biodiversity patterns[18–20]. Such coverage gaps and biases are human artifacts that can manifest (1) geographically in the disproportionate coverage of a species in some regions of its range relative to others[21]; (2) taxonomically in the tendency of some taxa or lineages to be more or less covered over others[22]; (3) temporally in the unbalanced collecting of specimens in some years or parts of the year[23,24]; and (4) functional traits in the disproportionate coverage of species on the basis of intrinsic life-history traits, including life cycle, size, growth form and rarity[25].

Identifying coverage of biodiversity patterns by biodiversity records is a priority[26–30] but a global assessment of how different this phenomenon is captured by voucher versus observation records has not been completed to date. In comparative temporal analyses, the increasing disconnection between voucher and observation records results from a change in collection practices in which the very nature of observation records allows them to be collected in larger quantities than voucher collections[31]. Related to this disconnection is the decline in public funding for museums and herbaria worldwide[32–42] (except for a few programmes like the US National Science Foundation and European Union's Horizon 2020 that still award collection grants[42]), which may also contribute to the decline in the number of trained professionals (taxonomists and systematists) in charge of collecting and maintaining voucher specimens[43]. When considered separately, there is some evidence that these disparate species records might be unveiling inaccurate patterns of expected biodiversity[30,44,45], which could hamper prospects of addressing questions in ecology, evolution and conservation[26]. One Norwegian study, for instance, demonstrated contrasting biodiversity patterns by voucher versus observation records along geographic, temporal and taxonomic axes but was focused on only plants[46]. The variability by voucher and observation records in describing expected biodiversity patterns for different taxa and across the globe remains poorly understood but important for prioritizing future data mobilization in the Anthropocene[47,48].

Here, we quantify coverage and biases of expected biodiversity patterns by voucher and observation records of major terrestrial species groups including vascular plants, butterflies, amphibians, birds, reptiles and mammals. We focused on these groups because they have been studied and explored for much longer than other taxonomic groups and consequently there is more occurrence information available on a global scale[49,50]. For each group, we also assess the factors limiting coverage and how they covary across taxa and grain size. We define coverage as the number of records required to inventory species in terms of their richness or abundance along taxonomic, geographic, temporal and functional trait dimensions[26,27]. However, coverage may be biased because certain species, regions, time periods or traits may be more or less covered over others. Despite the mass production of observation records, we found that their coverage of expected biodiversity patterns is incongruent with areas of high species richness. Such coverage patterns are driven by how accessible and secure a sampling site is and, in turn, reflect highly human-modified areas. Voucher records are vastly infrequent in occurrence data but, in the few places where they are sampled, showed relative congruence with expected biodiversity patterns for all dimensions. The differences in coverage by voucher and observation records have important implications for the utility of these records for research in ecology, evolution and conservation.

## Results and discussion

### Taxonomic coverage of lineages and grid cells

We assessed taxonomic coverage of lineages (families) and grid cells by voucher and observation records using a large global dataset, including ~1.9 billion occurrence records of terrestrial plants, butterflies, amphibians, birds, reptiles and mammals (Supplementary Table 1) across six variations of spatial grain (50, 100, 200, 400, 800 and 1,600 km). Taxonomic coverage was assessed as the ratio of documented species richness of a family or grid cell to expected species richness in the family or grid cell based on expert opinion (Extended Data Fig. 1). Biases in taxonomic coverage were assessed using phylogenetic signal test for lineages (Supplementary Table 2) and Moran's *I* spatial autocorrelation measure for grid cells (where Moran's $I = 1$ indicates biased geographic coverage and 0, even or random coverage).

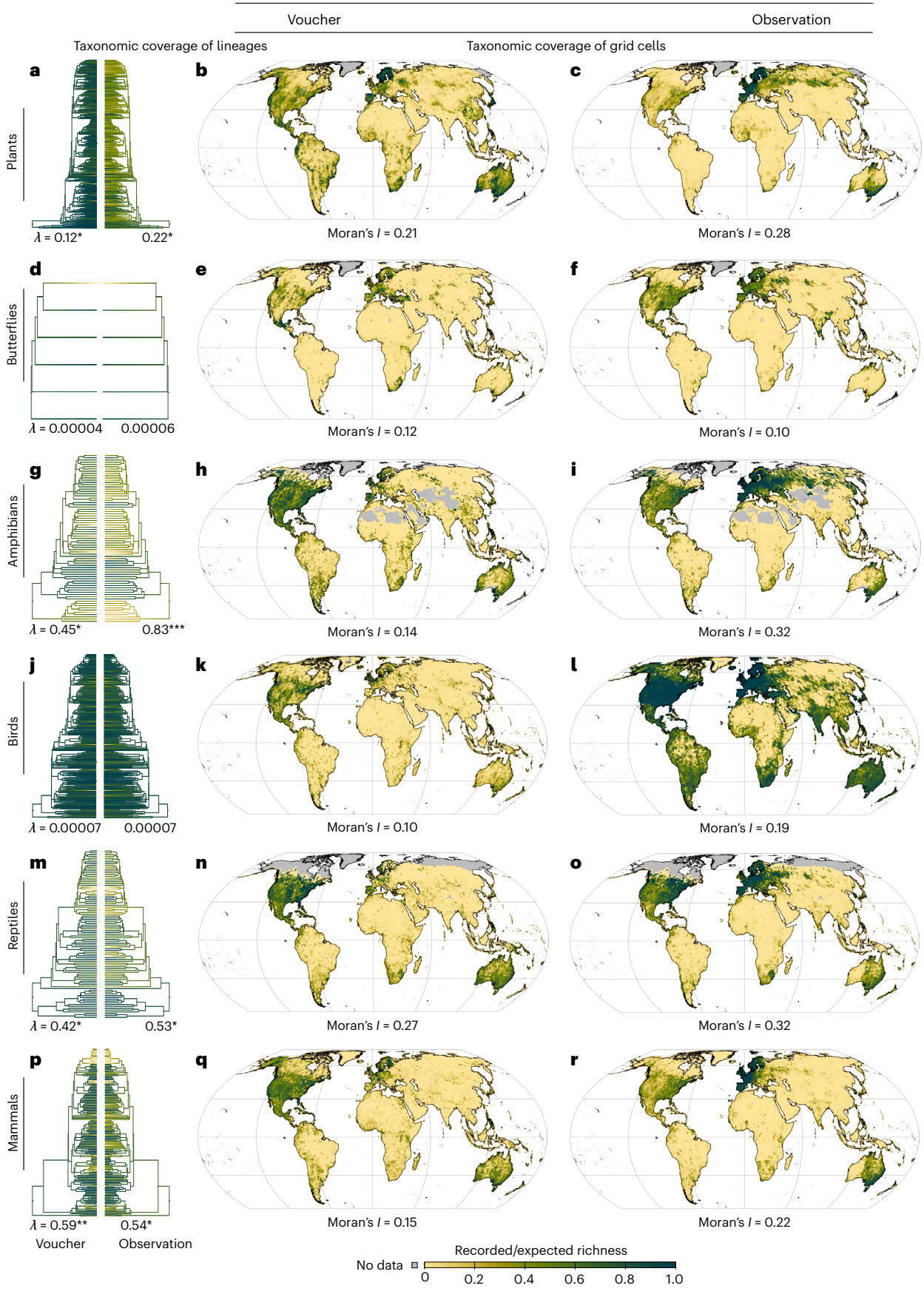
Both voucher and observation records of most taxonomic groups showed massive gaps in taxonomic coverage. When we weighted the expected species richness of a lineage (family) or grid cell by the actual documented richness of species within the lineage or grid cell, we found that the species richness of lineages derived from observation records tended to be phylogenetically biased and showed less congruence with expected richness for most taxonomic groups (Fig. 1). On the other hand, the taxonomic coverage of voucher records is more phylogenetically random and showed relative concordance to expected family richness across most taxonomic groups (Supplementary Table 2). Voucher records may be more relevant for many questions in taxonomy, systematics and conservation, where species identity is essential. For instance, the physical specimens can capture additional data, including nutrients, defensive compounds, herbivore damage, disease lesions and signatures of physiological processes, which are crucial for understanding the ecological and evolutionary responses of species in the Anthropocene but remain unrealized[9].

The taxonomic coverage of grid cells by occurrences of voucher records showed moderate to strong concordance to the expected species richness of most taxonomic groups ($r_s = 0.09–0.55$ for vouchers versus 0.011–0.92 for observations, all $P < 0.05$, from a modified *t*-test of spatial association; Supplementary Table 3), except birds in which observation records showed stronger associations to expected richness and these are consistent across grain sizes (Supplementary Table 3). Birds have been studied much longer because they are charismatic[51,52] and consequently bird sightings and observations are better reported in databases than other taxonomic groups[17]. Although well-sampled regions such as North America, western Europe and Australia tend to correspond with high levels of taxonomic coverage under observation records (Fig. 1), voucher records additionally capture expected richness in other regions known to harbour high concentrations of biodiversity

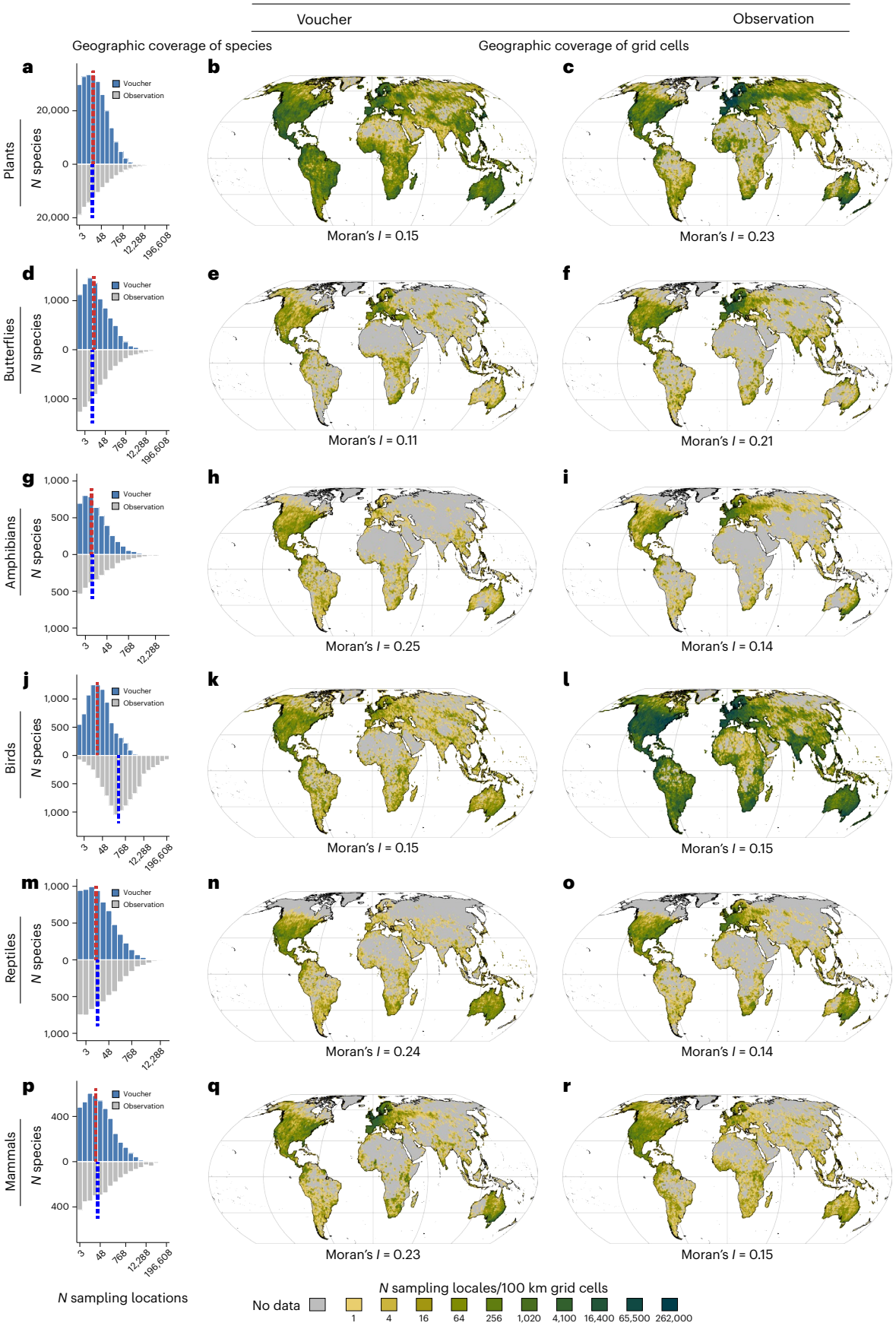
**Fig. 1 | The taxonomic coverage of lineages and grid cells by observation records are more biased and less congruent to expected richness patterns.** **a**–**r**, Taxonomic coverage across lineages and grid cells for: plants ($n = 240,377$ species, 423 families) (**a**–**c**), butterflies ($n = 9,809$ species, 6 families) (**d**–**f**), amphibians ($n = 4,862$ species, 71 families) (**g**–**i**), birds ($n = 9,380$ species, $n = 242$ families) (**j**–**l**), reptiles ($n = 7,259$ species, 88 families) (**m**–**o**) and mammals ($n = 4,508$ species, 141 families) (**p**–**r**). Taxonomic coverage was assessed as the ratio of documented species richness of a family or grid cell to expected species richness in the family or grid cell based on expert opinion. Coverage of voucher records are relatively phylogenetically random and showed relative concordance to expected richness across most taxonomic groups. Biases in taxonomic coverage were assessed using Pagel's $\lambda$ phylogenetic signal test for lineages and Moran's *I* spatial autocorrelation measure in the case of grid cells (Monte Carlo test, 999 randomizations) with values of 1 indicating clustered/biased taxonomic coverage and 0 corresponding to taxonomically even coverage of grid cells. The bamako colour palette is common to all panels, with dark green indicating high coverage and yellow indicating low coverage. Tests of phylogenetic signal in taxonomic coverage using other metrics (Blomberg's *K* and Abouheif's $C_{mean}$) are presented in Supplementary Table 2. Significance codes: $^{***}P < 0.001$, $^{**}P < 0.01$, $^{*}P < 0.05$. The maps are in the Wagner IV projection.

https://doi.org/10.1038/s41559-023-02047-3

such as South America, South Africa and Himalaya-Hengduan in Southeast Asia (Fig. 1). When we contrasted the relationship between taxonomic coverage versus dissimilarity (measured as spatial composition of beta diversity), we found that occurrence records of most taxonomic groups showed high dissimilarity in less frequently sampled regions of South America, Central Africa and Southeast Asia and low dissimilarity



https://doi.org/10.1038/s41559-023-02047-3



https://doi.org/10.1038/s41559-023-02047-3

**Fig. 2 | Patterns of geographic coverage of species and grid cells by voucher and observation records of individual taxa. a**–**r**, Geographic coverage across species and 100 × 100 km$^2$ grid cells for: plants ($n$ = 240,377 species) (**a**–**c**), butterflies ($n$ = 9,809 species) (**d**–**f**), amphibians ($n$ = 4,862 species) (**g**–**i**), birds ($n$ = 9,380 species) (**j**–**l**), reptiles ($n$ = 7,259 species) (**m**–**o**) and mammals ($n$ = 4,508 species) (**p**–**r**). Geographic coverage (number of unique collection locales for each species or grid cell) showed higher biases under observation records peaking in well-sampled regions of the world. Dashed vertical lines in **a**, **d**, **g**, **j**, **m** and **p**, indicate the median geographic coverage. Evenness or clustering of geographic coverage indicated by Moran's $I$ (Monte Carlo test, 999 randomizations) with values of 1 indicating clustered/biased coverage and 0 corresponding to geographically even coverage. The bamako colour palette is common to all panels, with dark green indicating high coverage and yellow indicating low coverage. Geographic coverages of grid cells at other spatial scales (50, 100, 200, 400, 800 and 1,600 km) are presented in Extended Data Fig. 4–9. The maps are in the Wagner IV projection.

in frequently sampled regions of Europe and North America (Extended Data Fig. 2). The negative correlation between dissimilarity and sampling effort is particularly strong for observation records of amphibians, birds and reptiles (Extended Data Fig. 3). Thus, biodiversity centres viewed from occurrence records alone may depict a signature of better data mobilization in frequently sampled Europe and North America and not true diversity patterns, corroborating previous observations of prevailing sampling biases in biodiversity-rich but infrequently sampled regions[26,27]. Such observations are invaluable for identifying priority sites for monitoring species distributions, demographic change and conservation through time[12,53]. Conversely, obstacles such as strict permitting regulations, ethical guidelines and challenges in specimen preparation and preservation, may impede the collection of vouchers[54]. As the global biodiversity crisis unfolds, the need for continued collecting of voucher specimens in a responsible way that adheres to best practices is necessary because of the irreplaceable contributions of vouchers to many fields beyond taxonomy[55].

## Geographic coverage of species and grid cells

We assessed geographic coverage by voucher and observation records as the number of unique collection sites of each species or grid cell. We found that the available records of plants and butterflies showed significant biases under observation records (median coverage of individual species: plants 12 and 10, butterflies 10 and 8, for voucher and observation records, respectively; from a two-sample $t$-test between voucher and observation records, all $P < 0.01$; Fig. 2), peaking in well-sampled regions especially North America, western Europe, Australia and South Africa (Fig. 2). In the case of tetrapods including amphibians, reptiles and mammals, we found geographic biases toward voucher records (Moran's $I$: 0.25, 0.24, 0.23, all $P = 0.01$, for amphibians, reptiles and mammals, respectively) clustering in eastern North America, western Europe and southeast Australia. Birds were an exception, showing similar biases for voucher and observation records (Moran's $I$ Monte Carlo test: 0.15, $P = 0.01$) but higher collection density by observation records (Fig. 2j–l). Coverage biases tended to increase at coarser grains for both voucher and observation records and across taxa (Extended Data Fig. 4–9), supporting previous observations of increasing disconnection between record types at different scales[31,46]. Most of the tropics including South America, Africa and Southeast Asia have essentially no records available. Many of these places are both biologically rich and heavily threatened[56,57] but lack mobilized records for modelling species distributions or setting conservation priorities. Although geographic coverage gaps and biases have been previously documented for biodiversity data[26,27,43,58–60], our study is, to our knowledge, the first global assessment to tease apart how different this phenomenon is for voucher and observation records.

## Temporal coverage of species and grid cells

We assessed temporal coverage as the time intervals spanned by voucher and observation records between all possible months between 1950 to 2021 to their respective closest months with available records for species and grid cells. Across all groups, voucher records showed significantly higher temporal coverage of species than observation records for most taxonomic groups with median time interval between records for plants (−4.74 and −8.63 years, $P < 0.01$), butterflies (−4.73 and −7.31 years, $P < 0.01$), amphibians (−8.06 and −8.50, $P = 0.076$) and mammals (−3.80 and −4.70, $P = 0.013$) (from a two-sample $t$-test for voucher and observation records; Fig. 3a,d,g,p). However, the high temporal coverage by voucher records is geographically biased toward regions of high collection density such as Western Europe, North America and Australia (Fig. 3), suggesting a long history of recording biodiversity rather than true diversity in these regions[61]. The time intervals spanned by observation records of most species groups especially in megadiverse but less frequently collected areas such as Africa, South America and Southeast Asia, are characterized by large temporal gaps (median interval across all groups: −11.73, −11.01 and −8.94 yr, respectively). While these patterns may be indicative of broader societal factors influencing biological collection such as regional conflicts, political instability or world wars, over the years, they also imply that observation records are missing for many years which has consequences for modelling demographic change through time.

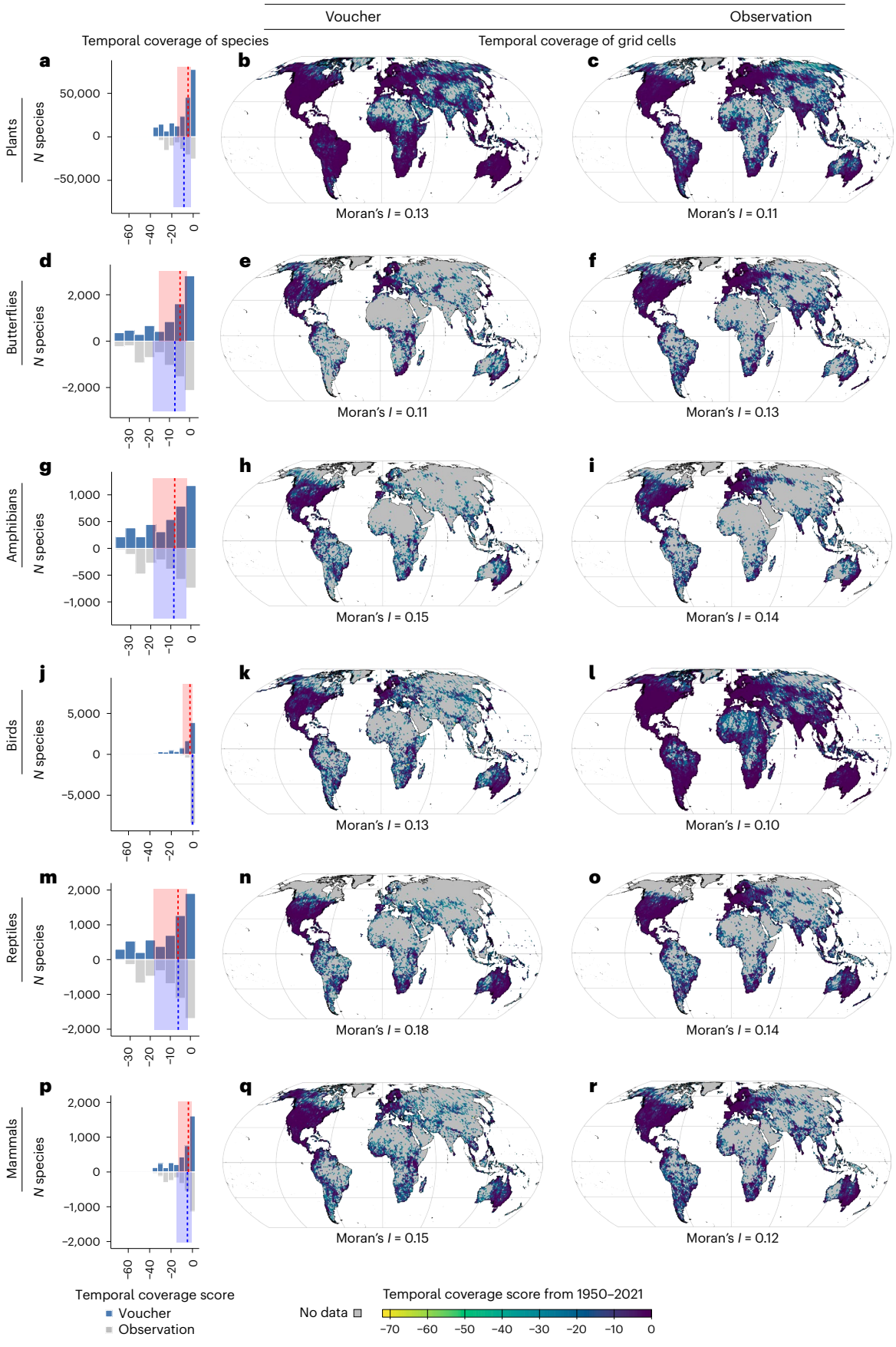
## Functional trait coverage

Species can be preferentially collected on the basis of traits innate to their ecology, morphology or life history. Across groups, we found that, while the number of records per species tended to be more frequent by observation records, the coverage of functional traits innate to species' ecology, morphology or life history are more evenly distributed under voucher records (Fig. 4). In the case of functional traits linked to the size of species such as snout–vent length, body mass or plant height, we found more frequency of relatively large-sized organisms by observation records compared to voucher records which tended to capture more variable functional spectra of organism sizes. For traits related to the age of organisms such as plant duration and longevity, we found that short to median-life span organisms or biennial species (in the case of plants) tend to be more frequently documented by observation records for most taxonomic groups. This is not true for amphibians, with small and short-lived species such as the Blanchard's cricket frog (*Acris blanchardi*) recorded more often by voucher records. Such findings could indicate that collectors of voucher records, who are often trained professionals, prefer collecting specimens with a suite of morphological features that are crucial for research in taxonomy and systematics as opposed to a select few traits that are favoured by amateur collectors

**Fig. 3 | Temporal coverage of species and grid cells by voucher and observation records. a**–**r**, Temporal coverage across species (left panel) and 100 × 100 km$^2$ grid cells (right panel) for: plants ($n$ = 240,377 species) (**a**–**c**), butterflies ($n$ = 9,809 species) (**d**–**f**), amphibians ($n$ = 4,565 species) (**g**–**i**), birds ($n$ = 9,358 species) (**j**–**l**), reptiles ($n$ = 6,889 species) (**m**–**o**) and mammals ($n$ = 4,364 species) (**p**–**r**). Temporal coverage of species and of grid cells was calculated as the negative mean minimum time interval between all possible months between 1950 and 2021 to their respective closest months with available records, for a species or grid cell, respectively. Less negative values indicate higher temporal coverage and large negative values if the time interval contains large temporal gaps without any records. Voucher records showed relatively higher temporal coverage of species but with large gaps in biodiversity-rich tropics. Dashed vertical lines and light shadings indicate the median and interquartile range (25/75%), respectively, of temporal coverage. Evenness or clustering of temporal coverage of grid cells indicated by Moran's $I$ with values of 1 indicating clustered/biased coverage and 0 corresponding to even coverage. The maps are in the Wagner IV projection.

https://doi.org/10.1038/s41559-023-02047-3

of observation records such as showy flowers or detectability. Butterfly host plants documented by observation records were found to be more frequent than those documented by voucher records. Specifically, 17.20% of butterflies documented by observation records and 15.76% of those documented by voucher records were associated with more than ten host plants (Fig. 4). For diet type, we found that bird species



https://doi.org/10.1038/s41559-023-02047-3

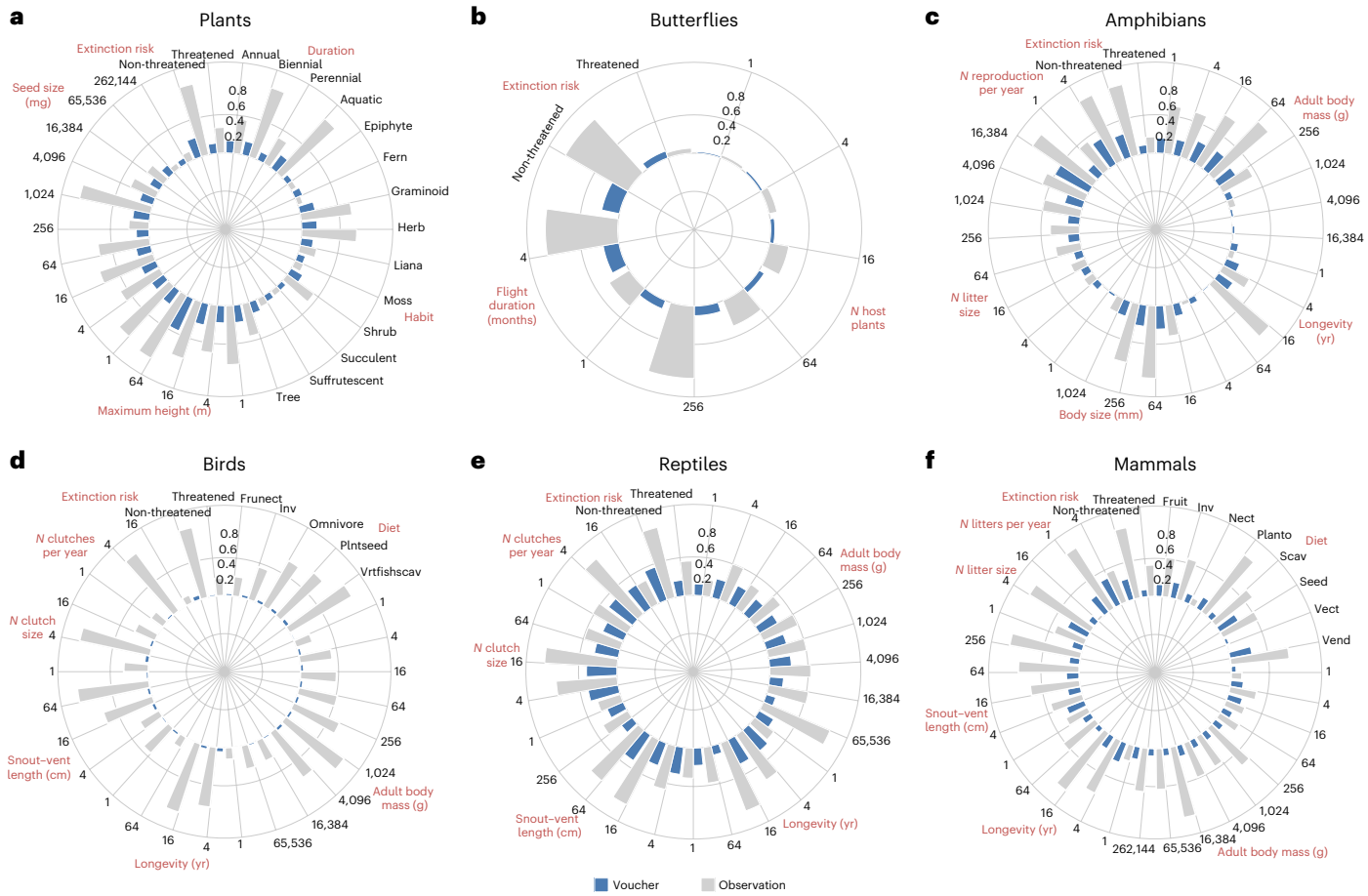



**Fig. 4 | Coverage of functional traits documented by voucher and observation records. a–f**, Coverage of functional traits in the number of specimens per species of plants (**a**), butterflies (**b**), amphibians (**c**), birds (**d**), reptiles (**e**) and mammals (**f**). For each category of functional traits, coverage of voucher and observation records were determined according to the number of specimens per species in each category and were arcsine-square-root-transformed. Despite higher collection frequencies of observation records, the coverage of functional traits are more evenly distributed under voucher records. Blue bars indicate voucher records, whereas grey bars indicate observation records. The dominant diets in birds are indicated as: frunect, fruits and nectar; inv, invertebrates; omnivor, diet consisting of plant and animal matter; plntseed, plant and seeds; vrtfishscv, vertebrates and fish and carrion. The dominant diets in mammals are indicated as: fruit, fruits; inv, invertebrates; nect, nectar; planto, other plant material; scav, diet consisting of carrion; seed, seeds; vect, diet consisting of ectotherms (reptiles and amphibians); vend, diet consisting of endotherms (mammals and birds).

whose diet comprised of plants and seeds or vertebrates, fish and carrion were recorded more often by observation records. Likewise, mammal species whose diet comprised of fruits, invertebrates, other plant material, carrion, seed and endotherms (birds and mammals) were recorded more often by observation records. Overall, the pattern of trait coverage for birds was more pronounced with the full spectrum of bird form and function (including adult body mass, longevity, size and clutch size) recorded more often by observation records. This confirms previous studies which suggest that practicalities such as government permit issues and conservation endangerment in which indiscriminate collecting may impact populations, may hinder the collecting of voucher specimens[62,63].

We also found that threatened species were represented by fewer collections on average than non-threatened species for both voucher and observation records. It is expected for rare species to be infrequent given their limited abundance[64] and there are justifiable restrictions on collecting rare or threatened species to avoid the further decline of wild populations[65]. However, this bias can potentially lead to erroneous extinction risk assessments and reduce opportunities for using information on historical populations and biogeography to guide species conservation and restoration.

## Association between coverage and socioeconomic conditions

Previous studies have indicated that the sampling of biodiversity data tended to occur near roadsides, in proximity to airports or in accessible places such as mountains[66–71] but it remains unclear how different the coverage of voucher and observation records are driven by these factors. We supplemented our analyses of taxonomic, geographic and temporal coverage of grid cells with spatial analyses of six socioeconomic variables using a spatial error model (Fig. 5; Extended Data Fig. 10).

For taxonomic coverage of grid cells, the representation of expected species richness by both voucher and observation records tended to be negatively driven by similar socioeconomic factors, such as areas of high human influence and easy access to airports, with more substantial effects for observation than voucher records across all taxonomic groups (Fig. 5a,d,g,j,m,p). For plants and amphibians in particular, the relationship of taxonomic coverage with the security of a region is strong but the effects for voucher and observation records are in opposite ways (correlation coefficients for plants (−0.49) versus (0.11) and amphibians (−0.69) versus (0), for vouchers versus observations, respectively, all $P < 0.05$; Fig. 5a,g). Although wars and regional conflicts have negative consequences on biodiversity[72], it is possible that occurrence records from regions impacted by war are the result of

surveys carried out by local scientists in partnership with international non-governmental organizations[72].

In the case of geographic coverage of grid cells, we showed that regions of high sampling density of both voucher and observation records generally correspond to low altitudes, high human footprint (inverse of wilderness), easy accessibility on ground and to airports and how secure a region is (Fig. 5b,e,h,k,n,q). However, the strength of the relationship with these variables is more pronounced for observation than for voucher records. Such a tendency to collect near accessible and appealing regions has been previously reported[21,67,73] but the differences between voucher and observation records probably reflect the proclivity of collectors. The collectors of observation records are probably a diverse team of amateurs who might focus their collecting activities around areas of increased human influence, such as hiking trails around community parks and neighbourhoods; whereas collectors of voucher records are often trained professionals who collect specimens as part of their formal jobs and tend to target their collecting in remote and wilderness areas. Additionally, urban heat effects from towns and cities, where many modern-day collecting activities are based, can alter the phenology of species nearby[73,74] or lead to erroneous estimation of species distribution models. Thus, the mass collection of observation records which tends to occur around accessible areas, secure conditions and in turn highly human-modified areas, suggests that available biodiversity records are not only unrepresentative of local or regional biodiversity but may not even reflect the general characteristics of the species.

Likewise, the drivers of temporal coverage by voucher and observation records tended to be similar, with regions of high temporal coverage corresponding to areas of high human influence, proximity to airports and how secure a region is (Fig. 5c,f,i,l,o,r). By contrast, while temporal coverage by voucher records is less affected by elevation and national research funding across all taxonomic groups, this is not true for observation records in which frequently sampled areas correspond to those of slightly higher elevations and regions with strong national research funding. These findings show that coverage of voucher records is relatively more even and reflective of expected biodiversity patterns than observations which tended to be clustered in a few regions that are easily accessible, secure and relatively influenced by human activities.

A voucher represents primary biodiversity data in taxonomy and systematics because it provides documentary evidence for species identification, re-examination and supporting material for conclusions reached in a study[5]. As new lines of investigation emerge, such as change through time[75] and emerging zoonoses[76], all of which require past and current specimens, scientists are seeing applications for collections beyond taxonomy and systematics. However, the rate at which vouchers are gathered in natural history museums and herbaria has slowed down, making it challenging to keep pace with modern scientific activity[77–89]. Instead, vouchers are rapidly being overwhelmed by the mass production of observation records[31]. Nevertheless, both voucher and observation data are complementary. Indeed, the biases we uncovered between voucher and observation records tended to be closer to each other than random expectations, suggesting similarities in underlying factors. Observation records increasingly capture a wide variety of derived information about the existence of an organism such as images, sound recordings, videos and behaviour beyond that of voucher specimens. Likewise, the foundation of research would be weak and future studies less supportive without vouchers[5]. Together, vouchers and the availability of different kinds of observation data will continue to be vital resources for assessing how anthropogenic drivers affect biodiversity in the past, present and future[47,90].

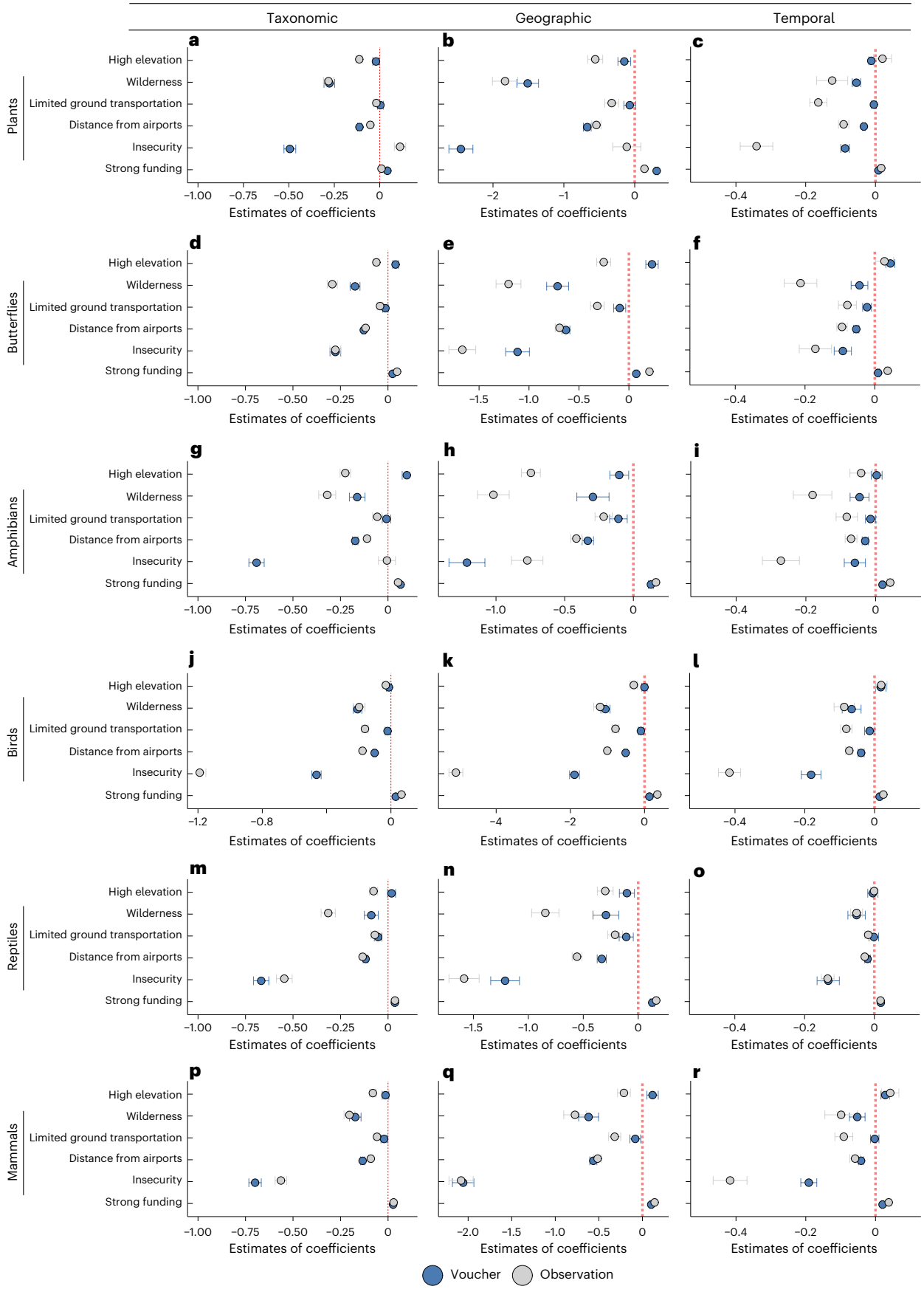
This study is not without limitations. First, not all voucher specimens have been digitized and mobilized online[91]. There is a lag time between collecting, processing and mobilization online. Second, although we recognize the potential circularity in using range maps as expected distributions of species, accounting for expected patterns of richness or abundance is difficult because our knowledge of such patterns often derives from (and may not necessarily be independent from) these biased collections. However, we argue that range maps are not derived from point records alone but integrated from other different types of data including expert knowledge of species' ecology and distributions, local inventories, atlas and literature. It is this integrative nature that inform the consideration of range maps as the most authoritative and only datasets available at a global scale of our knowledge of the species distributions and are used as baseline estimates of expected distributions of species (for example, refs. 92–94). Third, the species occurrences that we considered in this study are focused on charismatic groups with well-mobilized occurrence records and by no means exhaustive; we lack quantitative records for most species. For instance, beetles (Coleoptera) are the largest known order of insects of about 350,000 species and are ecologically important in nutrient cycling and relationships with other organisms[95] but not targeted for conservation planning[96]. Fourth, the available socioeconomic variables are aggregated averages for time periods that do not represent a one-to-one match to occurrence records. It would be interesting to explore whether analyses using variables spanning the same time as occurrences show similar patterns to those revealed here. However, the dataset to test this further is unavailable. Many impediments such as preservation practices, storage space and funding shortages for most museums and herbaria can bias occurrence records to have more frequency of observation records. In the case of organisms that are either too large to be collected and stored, such as large mammals, or too small to be seen and vouchered, such as ants and microbes, detailed ancillary data such as pictures and sounds should be documented.

## Dealing with coverage gaps and biases in future collecting

It is important to ameliorate the biases and coverage gaps for the different taxonomic groups and record types examined here to ensure that species occurrences remain vital for ecological and evolutionary research. The biodiversity-rich tropics including South America, West and Central Africa and Southeast Asia account for most coverage gaps which can be filled by targeting new surveys in these regions in ways that reduce the gaps and biases. This could be achieved through collaborative explorations and structured collecting between local researchers in these regions and those from well-funded museums in developed countries. Indeed, a few institutions from developed countries such as the Royal Botanic Gardens Kew or Missouri Botanical Garden already have long history of biodiversity exploration and partnerships with nations in sub-Saharan Africa and Madagascar, respectively. We encourage similar partnerships from other developed countries to support data mobilization efforts in countries limited by expertise or financial resources. We also found widespread temporal gaps across most species and sites.

**Fig. 5 | The estimates and 95% confidence intervals predicted by a spatial autoregressive error model of coverage (taxonomic, geographic and temporal) by voucher and observation records with socioeconomic predictors. a–r**, Estimated effects from spatial simultaneous autoregressive error models between taxonomic (left column), geographic (middle column) and temporal (right column) coverage of grid cells by voucher (blue) and observation (grey) records with each predictor for plants (**a**–**c**), butterflies (**d**–**f**), amphibians (**g**–**i**), birds (**j**–**l**), reptiles (**m**–**o**) and mammals (**p**–**r**). The centre point for the error bars is indicated with dots and represents the estimated coefficient for a one-unit change in the independent variable. The error bars show the 95% confidence intervals. The red dotted vertical line indicates the estimated effect at zero. These models indicate that socioeconomic variables have more substantial effects on the coverage of observation than voucher records. Higher scores for each variable correspond to increase in each descriptor. We present results for the spatial grain of 100 × 100 km$^2$. Estimated effects at other spatial grains (200 and 400 km) are presented in Supplementary Tables 4–6.

https://doi.org/10.1038/s41559-023-02047-3



To close such temporal gaps for voucher records, we encourage more systematic collecting at regular intervals by targeting clades and grid cells with the most out-of-date records and where species are known to occur but not yet recorded rather than in a select few places that are only accessible or appealing. For instance, targeting least-covered cells in eastern Saudi Arabia and Uzbekistan with −35.9 years away from when any plant voucher was collected or −72 years for eastern Mexico, southeast Egypt and southern India for butterfly vouchers since 1950, are highly desirable. On the other hand, future collecting of observation records, which are often guided by a mobile application, could combine educational feedback (regarding natural history, life cycle and endemism) and quizzes, to incentivize volunteer participation and steer users to explore less frequently sampled areas and species as validated by voucher collections[96–98]. Diminishing the gaps in trait coverage could include taking measurements directly from voucher specimens for species with missing trait information. If practicalities such as permitting issues or animal welfare[63,76] preclude the collecting of whole vouchers, particularly for large vertebrates, non-invasive collecting that combines high-resolution photographs, sounds, videos, molecular samples and other characteristics can suffice[99].

Overall, this study demonstrates that the nature and severity of coverage of expected biodiversity patterns can differ greatly between voucher and observation records. Such coverage gaps and biases in collection records place limitations on future studies[100] and can alter the interpretations in existing studies (for example, refs. 44,101,102). The differences in coverage and bias by voucher and observation records have important implications for the utility of these records for research in ecology, evolution and conservation research.

## Methods

### Data compilation

**Occurrence data.** We downloaded data of ~1.9 billion occurrence records from the Global Biodiversity Information Facility (GBIF) for six taxonomic groups for which occurrence records have been well-mobilized: plants (Plantae; $n$ = 374 million records), butterflies (Rhopalocera: Hedyloidea and Papilonoidea, including Heperiidae; $n$ = 72 million records), amphibians (8.1 million), birds (1.4 billion), reptiles (8 million) and mammals (29 million) (Supplementary Table 1). We distinguished the origin of each record on the basis of whether they came from material with a physical voucher specimen in museum or herbarium (referred to as 'voucher' record) or from observations that are not traceable to tangible physical material in a museum or herbarium ('observation' record). Our definition of voucher records therefore includes records labelled as 'preserved specimens', 'fossil specimens', 'living specimens' and 'material samples' whereas observations included only records labelled as 'direct observations', 'human observations', 'machine observations' and 'literature'. Occurrence records with unknown categories were removed from the analysis. The datasets were thoroughly cleaned to remove duplicates and records with erroneous localizations using the R package CoordinateCleaner v.2.0–20 (ref. 103). We then retained records with acceptable scientific names following currently accepted taxonomies of plants[104], butterflies[105], amphibians[106], birds[107,108], reptiles[109] and mammals[110].

Because occurrence records from GBIF often include both native and non-native distribution of a species[111], we restricted our analysis to species' native distributions, consistent with our goal of quantifying coverage of expected native biodiversity patterns by occurrence records. For tetrapods (amphibians, birds, reptiles and mammals), we restricted occurrences to their native ranges as determined by expert range maps available on the International Union of Conservation of Nature's (IUCN) Red List of Threatened Species spatial database. This was achieved by retaining records that fell within the native area of a species' expert range map[112,113] by setting the 'origin' field to 1, corresponding to the native range of the species following the IUCN map code designations. In the case of plants which have IUCN expert range maps for only 9% (33,573) of described plants[112], native distributions of plants were determined by overlaying species occurrence records against Kew Plants of the World Online database (POWO; http://www.plantsoftheworldonline.org/) and extracting records that fell within the boundaries of POWO. The POWO is a comprehensive database of native distribution maps for all plants of the world within biogeographic units defined by the World Geographical Scheme for Recording Plant Distributions[114]. In the case of butterflies, native occurrences were determined using range map overlays from a dataset of country-level species occurrences[105] which includes country-level species range maps from literature and publicly available occurrence records from GBIF[105].

**Expected species richness of families and grid cells.** To determine taxonomic coverage across lineages, we contrasted the number of species recorded in a family to the expected species richness in the family from literature. Information on expected species richness of plant families was derived from ref. 115, butterflies[105,116], amphibians[106], birds[107,108], reptiles[109] and mammals[110], which in turn, was used to quantify taxonomic coverage of families in a phylogenetic framework. The phylogeny of plants was derived from a dated phylogeny for seed plants of the world[117]; butterfly phylogeny was obtained from a Bayesian estimation of the age of butterflies[118], amphibians[119], birds[120], reptiles[121] and mammals[122]. For each taxonomic group, we sampled one species from each family to generate a family-level phylogeny for our analyses.

To understand patterns of taxonomic coverage in geographic space, we mapped expected species richness geographically across grid cells (Extended Data Fig. 1). In the case of plants, data on expected species richness across grid cells came from a co-kriging interpolation model of 1,032 regional floras worldwide[123], which we resampled across six spatial grain sizes of 50, 100, 200, 400, 800 and 1,600 km. For butterflies without any prior information on global distributions of species richness, pattern of expected richness was derived by extrapolating richness information from inventories, checklists, online regional databases and literature sources (Supplementary Note 1) and projecting across grid cells. We fitted a co-kriging interpolation model to estimate the probability of butterfly occurrence into unsampled areas on the basis of recorded richness across 543 geographic units (corresponding to centroids of national parks, reserves, gardens, survey plots, biogeographic regions, countries, states, provinces and counties) and four predictor variables including mean temperature, mean precipitation, elevation and potential evapotranspiration. The final output from the co-kriging model consisted of a modelled map in raster format at grid cell resolution of 0.5° equivalent to 50 km at the equator, which we resampled to six different grain sizes (50, 100, 200, 400, 800 and 1,600 km). In the case of tetrapods (amphibians, birds, reptiles and mammals), expected species richness was determined by overlaying expert-based range map of each species[112,113] with equal-area grid cells (Behrmann projection) across six different grain sizes (50, 100, 200, 400, 800 and 1,600 km) and counting the number of species in each grid cell.

### Data analysis

We assessed coverage by voucher and observation records of the six taxonomic groups (plants, butterflies, amphibians, birds, reptiles and mammals) along taxonomic, geographic, temporal and functional trait dimensions. All analyses described here were done using R v.4.2.2 (ref. 124) and packages phyloregion v.1.0.8 (ref. 125), terra v.1.7-3 (ref. 126), ape v.5.6-2 (ref. 127), spatialreg v.1.2-6 (ref. 128), gglot2 v.3.4.0 (ref. 129), adephylo v.1.1-13 (ref. 130) and phytools v.1.2-0 (ref. 131).

**Taxonomic coverage of species richness patterns.** Taxonomic coverage by voucher and observation records was assessed across lineages and geographically across grid cells as follows:

$$\text{Taxonomic coverage} = \frac{\sum S_{ij}}{\sum S} \quad (1)$$

where $S_{ij}$ is species richness within a lineage (family) or grid cell documented by a voucher $i$ or observation $j$ record and $S$ is the expected species richness within the lineage (family) or grid cell based on expert documentation from literature. First, we weighted the expected species richness of each lineage (family) or grid cell by the actual documented richness of species by voucher and observation records from GBIF (equation (1)). This means that coverages even for cells with expected species richness but no documented records were scored 0. Thus, a coverage score of 0 indicates no taxonomic coverage of expected richness and 1 indicates full coverage by voucher or observation records. For our purposes, coverage scores >1 were set to 1 for direct comparison as our goal was to analyse the level of completeness by occurrence records. Second, to estimate whether there was a phylogenetic signal in taxonomic coverage of lineages, we used three different and most used phylogenetic signal measures: Abouheif's $C_{mean}$ statistic[132], Blomberg's $K$[133] and Pagel's lambda ($\lambda$)[134]. Abouheif's $C_{mean}$ measures the autocorrelation coefficient of the relationship of cross-taxonomic trait variation on a phylogeny[135]; Blomberg's $K$ is the ratio of the variance among taxa divided by the contrasts variance[133]; and Pagel's $\lambda$ is the transformation of the phylogeny for the correlations between taxa, relative to the correlation expected under a Brownian motion model[134]. Statistical significance was assessed by calculating the standardized effective size of phylogenetic signal based on 1,000 randomizations of the coverage scores across the tips of the phylogeny. A strong phylogenetic signal, with values 1 and above (that is, close relatives share similar coverage of expected family richness) indicates biases in the taxonomic coverage of voucher and observation records; a value of 0 indicates no phylogenetic signal and taxonomic coverage can be considered statistically even or random. Abouheif's $C_{mean}$ was calculated using the R package adephylo[130] whereas both Blomberg's $K$ and Pagel's $\lambda$ were calculated using the R package phytools[131]. Third, we estimated biases in taxonomic coverage of grid cells using Moran's $I$ spatial autocorrelation measure (Monte Carlo test, 999 randomizations), with values of 1 indicating clustered/biased taxonomic coverage and 0 corresponds to taxonomically even coverage of grid cells.

To explore the differences between record types and how they relate to sampling effort, we used Simpson's β-diversity to measure the dissimilarity between pairs of grid cells within major biogeographical regions recognized by the World Geographical Scheme for Recording Plant Distributions[114]. We calculated β-diversity using the Simpson index, which considers differences in species composition, turnover and richness among different sites[136]. We then generated maps of dissimilarity between record types (Extended Data Fig. 2) and we used loess regression to analyse the correlation between sampling effort and dissimilarity (Extended Data Fig. 3).

**Geographic coverage of species and grid cells.** To assess the degree to which available records cover geographic regions across the globe, we assessed patterns of geographic coverage of species and grid cells as the number of unique collection locales for each species or grid cell, respectively, for each taxonomic group. Geographic coverage of grid cells was calculated by overlaying each species' unique point occurrence onto equal-area grid cells, returning a community matrix of abundance or absence of each occurrence in a grid cell. Spatial overlay of point occurrences was computed using the function points2comm in the R package phyloregion[125]. Geographic coverage of grid cells was assessed across six different grain sizes (50, 100, 200, 400, 800 and 1,600 km). We then tested the evenness or clustering of geographic coverage at each grain size using Moran's $I$ Monte Carlo test (999 randomizations) with Moran's $I$ of 1 indicating clustered/biased coverage and 0 corresponds to geographically even coverage.

**Temporal coverage.** To understand continuities in available records, we assessed temporal coverage of available records of each species or grid cell spanning 1950 and 2021. This period approximates the start of large-scale collection of occurrence records and thus provides reliable data for macroecological investigations[27,137]. We used the negative mean minimum time metric of ref. 27, defined as the negative mean minimum interval between a collection month and every other collection month for each species or grid cell between 1950 and 2021. We express temporal coverage as:

$$\text{Temporal coverage} = -\frac{1}{n} \times \sum_{i=1}^{n} \text{Min}T_i \quad (2)$$

where $\text{Min}T_i$ represents the minimum interval of a collection month $i$ to every other collection month $n$. The final outputs consisted of negative time intervals (rescaled in years), such that less negative values indicate higher temporal coverage and large negative values if not.

**Trait coverage.** We explored trait coverage—representation of expected biodiversity patterns attributable to intrinsic life-history characteristics, including life cycle, size and species conservation status—across the species groups. For plants, we included four trait categories known to capture the global spectrum of plant form and function[138]: maximum height, seed mass, growth duration and habit. Growth duration included three axes—annual, biennial and perennial; whereas habit included ten axes—aquatic, epiphyte, fern, graminoid, liana, moss, shrub, succulent, suffrutescent and tree. Most of the plant trait data were compiled from Missouri Botanical Garden's Tropicos database (https://tropicos.org, accessed March 2020). This dataset was supplemented by online regional databases (Supplementary Note 2). Each dataset was reviewed to synonymize terminologies for functional traits, for example, 'vines' versus 'lianas' for climbers or 'forbs' versus 'herbs' for herbaceous life forms. For species with multiple trait records, we took the maximum value for numeric variables and the mode for categorical traits. In the case of butterflies, we selected two traits with sufficient information: associations of butterflies with host plants and flight duration from LepTraits, a database of the world's butterfly traits[139]. For amphibians, we extracted five traits with sufficient information from AmphiBIO[140] including adult body mass, body size, longevity, reproductive output and maximum litter size. In the case of birds, reptiles and mammals, trait data were derived from Amniote[141] and Elton Traits 1.0 (ref. 142) life-history databases, repositories that include information on different functional traits. For these three taxonomic groups, we selected five trait categories with sufficient information that cover common spectra of vertebrate form and function[143]. For birds, we selected adult body mass, longevity, snout–vent length, clutch size (number of eggs), diet and number of clutches per year. We selected five traits for reptiles: adult body mass, longevity, snout–vent length, clutch size (number of eggs) and number of clutches per year. For mammals, we also selected six traits: adult body mass, longevity, snout–vent length, litter size (number of offspring), diet and number of litters per year.

We also quantified coverage on the basis of species' extinction risks under the assumption that extinction risk is phylogenetically non-random with species in some lineages at elevated risk of extinction[144,145]. For plants, because the IUCN Red List database contains extinction risk assessment for only 33,573 plant species, equating to only 9% of described plants (17 February 2022), we obtained extinction risk assessments from a machine-learning prediction of conservation status of over 150,000 land plant species[146]. These predicted conservation ratings were rescaled to two broader threat categories, threatened (extinction probability >0.01) and not threatened (extinction probability <0.1), following ref. 147. Data on extinction risk for butterflies, amphibians, birds, reptiles and mammals came from the conservation rankings of IUCN Red List database (www.iucnredlist.org, accessed 17 February 2022). These rankings were rescaled to the expected extinction probabilities over 100 years of each taxon following ref. 148 as follows: least concern (LC) = 0.001, near-threatened and

conservation-dependent (NT) = 0.01, vulnerable (VU) = 0.1, endangered (EN) = 0.67 and critically endangered (CR) = 0.999. These categories were also rescaled into non-threatened (NT + LC, that is, extinction probability <0.1) and threatened (CR + EN + VU, extinction probability >0.01). Species ranked as data deficient (DD), were excluded from our analysis (see also ref. 149).

Coverage of functional traits was estimated as a chi-squared test comparing the number of records per species documented by voucher or observation records for each type of functional trait. The number of records per species for each trait category was arcsine-square root-transformed before analysis.

**Effect of geographic and socioeconomic factors.** Six ecological and socioeconomic predictors (Supplementary Table 7) were compiled to assess drivers of coverage gaps by voucher and observation records. We grouped these predictors on the basis of appeal (elevation and wilderness), accessibility (travel time to cities and proximity to airports), security (secure conditions) and national research funding. These variables were chosen because of their limited collinearity (Extended Data Fig. 10), except for on-ground accessibility which strongly associates with wilderness (Pearson's $r = 0.87$) and with proximity to airports (Pearson's $r = 0.56$). The association of on-ground accessibility with wilderness and with proximity with airport reflects the rapid human-driven interconnectivity of regions through ground and air transportations leading to rapid erosion of wilderness[150]. We assumed that these six ecological and socioeconomic variables represent a broader spectrum of factors that can drive the collecting of biodiversity data[27] and thus used them in our analyses. Elevation, a measure of topography of an area, was obtained from WorldClim database[151] by calculating the difference between the minimum and maximum elevation values in each grid cell. Wilderness index (remoteness from modern human influence) was obtained from ref. 152. Information on ground accessibility, defined as the time to travel to major cities, was obtained from ref. 150. Accessibility of collecting sites to airports was estimated using a dataset on the locations of airports across the globe[153] and calculating the minimum distance of each grid cell centroid to the nearest airport. The security of an area was estimated using the Global Peace Index of countries[154]. In terms of how the financial resources of a country can influence collecting efforts, we obtained information on the per capita gross domestic expenditure on research and development[154–156]. Most of these variables (elevation, wilderness, on-ground accessibility and proximity to airports) were already aggregated as averages across years directly from their original sources. In the case of secure conditions (spanning 2008 to 2022) and national research funding (2015 to 2022), however, we took arithmetic averages to be consistent with the other variables. We extracted the values of each variable across grid cells and for six different grain sizes (50, 100, 200, 400, 800 and 1,600 km).

We analysed the effect of the six socioeconomic factors (elevation, wilderness, on-ground accessibility, proximity to airports, secure conditions and national research funding) on taxonomic, geographic and temporal coverage of grid cells for plants, butterflies, amphibians, birds, reptiles and mammals. Because our previous analysis of Moran's *I* statistic indicated positive spatial dependence among neighbouring grid cells for patterns of taxonomic, geographic and temporal coverage of grid cells, we used a spatial simultaneous autoregressive error model for the spatial linear regression analysis. Before analysis, the dependent variables were standardized using arcsine-square-root-transformed before log-transformation for taxonomic coverage and log-transformation for geographic coverage and temporal coverage. We also log-transformed all predictor variables before analysis. Our spatial error models included a neighbourhood structure and spatial weight matrix derived from the vector polygons of grid cells for each spatial grain. The neighbourhood structure across grid cells was created using the function poly2nb from the R package spdep[157], which was used to create the spatial weights on the basis of row standardization. We set the parameter, zero.policy to TRUE, in our spatial error models because some grid cells may not have neighbours. The fitted regression coefficients were reported.

### Reporting summary

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

## Data availability

The links to the species occurrence records downloaded from the GBIF are available at Zenodo (https://doi.org/10.5281/zenodo.6834577). The datasets, data tables, grid cell vector polygons and R codes are archived at Zenodo (https://doi.org/10.5281/zenodo.6834577).

## Code availability

All scripts, codes and data documentation necessary to repeat our analyses have been made available in the Zenodo database (https://doi.org/10.5281/zenodo.6834577) under the folder 'SCRIPTS'.

## References


1. Butchart, S. H. et al. Global biodiversity: indicators of recent declines. *Science* **328**, 1164–1168 (2010).
2. Tittensor, D. P. et al. A mid-term analysis of progress toward international biodiversity targets. *Science* **346**, 241–244 (2014).
3. Johnson, C. N. et al. Biodiversity losses and conservation responses in the Anthropocene. *Science* **356**, 270–275 (2017).
4. Díaz, S. et al. Pervasive human-driven decline of life on Earth points to the need for transformative change. *Science* **366**, eaax3100 (2019).
5. Kageyama, M. et al. in *Museum Studies: Perspectives and Innovations* (eds Williams, S. L. & Hawks, C. A.) 257–264 (Society for the Preservation of Natural History Collections, 2007).
6. Cook, J. A. et al. The Beringian Coevolution Project: holistic collections of mammals and associated parasites reveal novel perspectives on evolutionary and environmental change in the North. *Arct. Sci.* **3**, 585–617 (2016).
7. Jungblut, A. D. & Hawes, I. Using Captain Scott's Discovery specimens to unlock the past: has Antarctic cyanobacterial diversity changed over the last 100 years? *Proc. R. Soc. B* **284**, 20170833 (2017).
8. Daru, B. H., Bowman, E. A., Pfister, D. H. & Arnold, A. E. A novel proof-of-concept for capturing the diversity of endophytic fungi preserved in herbarium specimens. *Philos. Trans. R. Soc. B* **374**, 20170395 (2018).
9. Meineke, E. K., Davis, C. C. & Davies, T. J. The unrealized potential of herbaria for global change biology. *Ecol. Monogr.* **88**, 505–525 (2018).
10. Colella, J. P. et al. The open-specimen movement. *BioScience* **71**, 405–414 (2021).
11. Unger, S., Rollins, M., Tietz, A. & Dumais, H. iNaturalist as an engaging tool for identifying organisms in outdoor activities. *J. Biol. Educ.* **55**, 537–547 (2021).
12. Sullivan, B. L. et al. eBird: a citizen-based bird observation network in the biological sciences. *Biol. Conserv.* **142**, 2282–2292 (2009).
13. Dickinson, J. L. et al. The current state of citizen science as a tool for ecological research and public engagement. *Front. Ecol. Environ.* **10**, 291–297 (2012).
14. Miller-Rushing, A., Primack, R. & Bonney, R. The history of public participation in ecological research. *Front. Ecol. Environ.* **10**, 285–290 (2012).
15. Petersen, T. K., Speed, J. D. M., Grøtan, V. & Austrheim, G. Species data for understanding biodiversity dynamics: the what, where and when of species occurrence data collection. *Ecol. Solut. Evid.* **2**, e12048 (2021).

16. Dickinson, J. L., Zuckerberg, B. & Bonter, D. N. Citizen science as an ecological research tool: challenges and benefits. *Annu. Rev. Ecol. Evol. Syst.* **41**, 149–172 (2010).
17. Tulloch, A. I. T., Possingham, H. P., Joseph, L. N., Szabo, J. & Martin, T. G. Realising the full potential of citizen science monitoring programs. *Biol. Conserv.* **165**, 128–138 (2013).
18. Hortal, J. & Lobo, J. M. A synecological framework for systematic conservation planning. *Biodivers. Inform.* **3**, 16–45 (2006).
19. Lobo, J. M., Baselga, A., Hortal, J., Jimenez-Valverde, A. & Gomez, J. F. How does the knowledge about the spatial distribution of Iberian dung beetle species accumulate over time? *Divers. Distrib.* **13**, 772–780 (2007).
20. Sandel, B. et al. Estimating the missing species bias in plant trait measurements. *J. Veg. Sci.* **26**, 828–838 (2015).
21. Hijmans, R. J. et al. Assessing the geographic representation of genebank collections: the case of the Bolivian wild potatoes. *Conserv. Biol.* **14**, 1755–1765 (2000).
22. Hortal, J., Lobo, J. M. & Jimenez-Valverde, A. Limitations of biodiversity databases: case study on seed-plant diversity in Tenerife, Canary Islands. *Conserv. Biol.* **21**, 853–863 (2007).
23. Funk, V. A. & Morin, N. A survey of the herbaria of the southeast United States. *SIDA Contrib. Bot.* **18**, 35–52 (2000).
24. Norris, W. R., Lewis, D. Q., Widrlechner, M. P., Thompson, J. D. & Pope, R. O. Lessons from an inventory of the Ames, Iowa, flora (1859–2000). *J. Iowa Acad. Sci.* **108**, 34–63 (2001).
25. Schmidt-lebuhn, A. N., Knerr, N. J. & Kessler, M. Non-geographic collecting biases in herbarium specimens of Australian daisies (Asteraceae). *Biodivers. Conserv.* **22**, 905–919 (2013).
26. Meyer, C., Kreft, H., Guralnick, R. & Jetz, W. Global priorities for an effective information basis of biodiversity distributions. *Nat. Commun.* **6**, 8221 (2015).
27. Meyer, C., Weigelt, P. & Kreft, H. Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecol. Lett.* **19**, 992–1006 (2016).
28. Tingley, R., Meiri, S. & Chapple, D. G. Addressing knowledge gaps in reptile conservation. *Biol. Conserv.* **204**, 1–5 (2016).
29. Guedes, T. B. et al. Patterns, biases and prospects in the distribution and diversity of Neotropical snakes. *Glob. Ecol. Biogeogr.* **27**, 14–21 (2018).
30. Daru, B. H. et al. Widespread sampling biases in herbaria revealed from large-scale digitization. *New Phytol.* **217**, 939–955 (2018).
31. Troudet, J., Vignes-Lebbe, R., Grandcolas, P. & Legendre, F. The increasing disconnection of primary biodiversity data from specimens: how does it happen and how to handle it? *Syst. Biol.* **67**, 1110–1119 (2018).
32. Wheeler, Q. D. Insect diversity and cladistic constraints. *Ann. Entomol. Soc. Am.* **83**, 1031–1047 (1990).
33. Cotterill, F. P. D. Systematics, biological knowledge and environmental conservation. *Biodivers. Conserv.* **4**, 183–205 (1995).
34. Dalton, R. Natural history collections in crisis as funding is slashed. *Nature* **423**, 575 (2003).
35. Gropp, R. E. Are university natural science collections going extinct? *BioScience* **53**, 550 (2003).
36. Stokstad, E. Nebraska husks research to ease budget squeeze. *Science* **300**, 35 (2003).
37. Vollmar, A., Macklin, J. A. & Ford, L. Natural history specimen digitization: challenges and concerns. *Biodivers. Inform.* **7**, 93–112 (2010).
38. Andreone, F. et al. Italian natural history museums on the verge of collapse? *ZooKeys* **456**, 139–146 (2014).
39. Kemp, C. The endangered dead. *Nature* **518**, 292–294 (2015).
40. Paknia, O., Rajaei, Sh,H. & Koch, A. Lack of well-maintained natural history collections and taxonomists in megadiverse developing countries hampers global biodiversity exploration. *Org. Divers. Evol.* **15**, 619–629 (2015).
41. Nowogrodzki, A. Biological specimen troves threatened by funding pause. *Nature* **531**, 561 (2016).
42. Bakker, F. T. et al. The Global Museum: natural history collections and the future of evolutionary science and public education. *PeerJ* **8**, e8225 (2020).
43. Pyke, G. H. & Ehrlich, P. R. Biological collections and ecological/environmental research: a review, some observations and a look to the future. *Biol. Rev.* **85**, 247–266 (2010).
44. Maldonado, C. et al. Species diversity and distribution in the era of Big Data. *Glob. Ecol. Biogeogr.* **24**, 973–984 (2015).
45. Rudbeck, A. V. et al. The Darwinian shortfall in plants: phylogenetic knowledge is driven by range size. *Ecography* **2022**, e06142 (2022).
46. Speed, J. D. M. et al. Contrasting spatial, temporal and environmental patterns in observation and specimen based species occurrence data. *PLoS ONE* **13**, e0196417 (2018).
47. Meineke, E. K., Davies, T. J., Daru, B. H. & Davis, C. C. Biological collections for understanding biodiversity in the Anthropocene. *Philos. Trans. R. Soc. B* **374**, 20170386 (2018).
48. Pearson, K. D. & Mast, A. R. Mobilizing the community of biodiversity specimen collectors to effectively detect and document outliers in the Anthropocene. *Am. J. Bot.* **106**, 1052–1058 (2019).
49. *What is GBIF?* (GBIF, accessed 18 October 2022); https://www.gbif.org/what-is-gbif
50. *The IUCN Red List of Threatened Species. Version 2022-1* (IUCN, accessed 19 October 2022); https://www.iucnredlist.org
51. Fleishman, E. & Murphy, D. D. A realistic assessment of the indicator potential of butterflies and other charismatic taxonomic groups. *Conserv. Biol.* **23**, 1109–1116 (2009).
52. Troudet, J. et al. Taxonomic bias in biodiversity data and societal preferences. *Sci. Rep.* **7**, 9132 (2017).
53. Lehikoinen, A. et al. Declining population trends of European mountain birds. *Glob. Change Biol.* **25**, 577–588 (2019).
54. Allington-Jones, L. & Bailey, R. Treatments for lipid oxidation in taxidermy and impact on DNA recovery. *Stud. Conserv.* **66**, 463–476 (2021).
55. Rocha, L. A. et al. Specimen collection: an essential tool. *Science* **344**, 814–815 (2014).
56. Myers, N. et al. Biodiversity hotspots for conservation priorities. *Nature* **403**, 853–858 (2000).
57. Venter, O. et al. Targeting global protected area expansion for imperiled biodiversity. *PLoS Biol.* **12**, e1001891 (2014).
58. Loiselle, B. A. et al. Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes? *J. Biogeogr.* **35**, 105–116 (2008).
59. Newbold, T. Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. *Prog. Phys. Geogr.* **34**, 3–22 (2010).
60. Mair, L. & Ruete, A. Explaining spatial variation in the recording effort of citizen science data across multiple taxa. *PLoS ONE* **11**, e0147796 (2016).
61. Yang, W., Ma, K. & Kreft, H. Geographical sampling bias in a large distributional database and its effects on species richness–environment models. *J. Biogeogr.* **40**, 1415–1426 (2013).
62. Yates, T. L. The role of voucher specimens in mammal collections: characterisation and funding responsibilities. *Acta Zool. Fenn.* **170**, 81–82 (1985).
63. Donegan, T. M. New species and subspecies descriptions do not and should not always require a dead type specimen. *Zootaxa* **1761**, 37–48 (2008).
64. Palmer, M. W., Earls, P. G., Hoagland, B. W., White, P. S. & Wohlgemuth, T. Quantitative tools for perfecting species list. *Environmetrics* **13**, 121–137 (2002).

65. Robinson, J. G. in *Conservation of Exploited Species* (eds Reynolds, J. D. et al.) 485–498 (Cambridge Univ. Press, 2001).
66. Freitag, S., Hobson, C., Biggs, H. C. & Jaarsveld, A. S. Testing for potential survey bias: the effect of roads, urban areas and nature reserves on a southern African mammal data set. *Anim. Conserv.* **1**, 119–127 (1998).
67. Funk, V. A. & Richardson, K. Biological specimen data in biodiversity studies: use it or lose it. *Syst. Biol.* **51**, 303–316 (2002).
68. Soria-Auza, R. W. & Kessler, M. The influence of sampling intensity on the perception of the spatial distribution of tropical diversity and endemism: a case study of ferns from Bolivia. *Divers. Distrib.* **14**, 123–130 (2008).
69. Ballesteros-Mejia, L., Kitching, I. J., Jetz, W., Nagel, P. & Beck, J. Mapping the biodiversity of tropical insects: species richness and inventory completeness of African sphingid moths. *Glob. Ecol. Biogeogr.* **22**, 586–595 (2013).
70. Yang, W., Ma, K. & Kreft, H. Environmental and socio-economic factors shaping the geography of floristic collections in China. *Glob. Ecol. Biogeogr.* **23**, 1284–1292 (2014).
71. Zizka, A., Antonelli, A. & Silvestro, D. sampbias, a method for quantifying geographic sampling biases in species distribution data. *Ecography* **44**, 25–32 (2021).
72. Hanson, T. et al. Warfare in biodiversity hotspots. *Conserv. Biol.* **23**, 578–587 (2009).
73. Zipper, S. C. et al. Urban heat island impacts on plant phenology: intra-urban variability and response to land cover. *Environ. Res. Lett.* **11**, 054023 (2016).
74. Li, D., Stucky, B. J., Deck, J., Baiser, B. & Guralnick, R. P. The effect of urbanization on plant phenology depends on regional temperature. *Nat. Ecol. Evol.* **3**, 1661–1667 (2019).
75. Jeppsson, T., Lindhe, A., Gärdenfors, U. & Forslund, P. The use of historical collections to estimate population trends: a case study using Swedish longhorn beetles (Coleoptera: Cerambycidae). *Biol. Conserv.* **143**, 1940–1950 (2010).
76. Yates, T. L. et al. The ecology and evolutionary history of an emergent disease: hantavirus pulmonary syndrome. *Bioscience* **52**, 989–998 (2002).
77. O'Connell, A. F. Jr, Gilbert, A. T. & Hatfield, J. S. Contribution of natural history collection data to biodiversity assessment in national parks. *Conserv. Biol.* **18**, 1254–1261 (2004).
78. Prather, L. A., Fuentes, O. A., Mayfield, M. H. & Ferguson, C. J. The decline of plant collecting in the United States: a threat to the infrastructure of biodiversity studies. *Syst. Bot.* **29**, 15–28 (2004).
79. Winker, K. Natural history museums in a postbiodiversity era. *BioScience* **54**, 455–459 (2004).
80. Bortolus, A. Error cascades in the biological sciences: the unwanted consequences of using bad taxonomy in ecology. *Ambio* **37**, 114–118 (2008).
81. Joseph, L. Museum collections in ornithology: today's record of avian biodiversity for tomorrow's world. *Emu* https://doi.org/10.1071/MUv111n3_ED (2011).
82. Bradley, R. D., Bradley, L. C., Garner, H. J. & Baker, R. J. Assessing the value of natural history collections and addressing issues regarding long-term growth and care. *BioScience* **64**, 1150–1158 (2014).
83. Renner, S. S. & Rockinger, A. Is plant collecting in Germany coming to an end? *Willdenowia* **46**, 93–97 (2016).
84. Spear, D. M., Pauly, G. B. & Kaiser, K. Citizen science as a tool for augmenting museum collection data from urban areas. *Front. Ecol. Evol.* **5**, 86 (2017).
85. Dunnum, J. L., McLean, B. S. & Dowler, R. C. Mammal collections of the Western Hemisphere: a survey and directory of collections. *J. Mammal.* **99**, 1307–1322 (2018).
86. Malaney, J. & Cook, J. A perfect storm for mammalogy: declining sample availability in a period of rapid environmental degradation. *J. Mammal.* **99**, 773–788 (2018).
87. Ferguson, A. W. On the role of (and threat to) natural history museums in mammal conservation: an African small mammal perspective. *J. Vert. Biol.* **69**, 20028–1 (2020).
88. Salvador, R. & Cunha, C. Natural history collections and the future legacy of ecological research. *Oecologia* **192**, 641–646 (2020).
89. Fischer, E. E., Cobb, N. S., Kawahara, A. Y., Zaspel, J. M. & Cognato, A. I. Decline of amateur Lepidoptera collectors threatens the future of specimen-based research. *BioScience* **71**, 396–404 (2021).
90. Heberling, J. M., Miller, J. T., Noesgaard, D., Weingart, S. B. & Schigel, D. Data integration enables global biodiversity synthesis. *Proc. Natl Acad. Sci. USA* **118**, e2018093118 (2021).
91. Hedrick, B. P. et al. Digitization and the future of natural history collections. *BioScience* **70**, 243–251 (2020).
92. Sandel, B. et al. The influence of Late Quaternary climate-change velocity on species endemism. *Science* **334**, 660–664 (2011).
93. Holt, B. G. et al. An update of Wallace's zoogeographic regions of the world. *Science* **339**, 74–78 (2013).
94. Mainali, K., Hefley, T., Ries, L. & Fagan, W. F. Matching expert range maps with species distribution model predictions. *Conserv. Biol.* **34**, 1292–1304 (2020).
95. McKenna, D. D. & Farrell, B. D. in *The Timetree of Life* (eds Hedges, S. B. & Kumar, S.) 278–289 (Oxford Univ. Press, 2009).
96. Xue, Y., Davies, I., Fink, D., Wood, C. & Gomes, C. P. in *Principles and Practice of Constraint Programming* (ed. Rueher, M.) 707–719 (Springer, 2016).
97. Robinson, O. J., Ruiz-Gutierrez, V. & Fink, D. Correcting for bias in distribution modelling for rare species using citizen science data. *Divers. Distrib.* **24**, 460–472 (2018).
98. Callaghan, C. T., Rowley, J. J. L., Cornwell, W. K., Poore, A. G. B. & Major, R. E. Improving big citizen science data: moving beyond haphazard sampling. *PLoS Biol.* **17**, e3000357 (2019).
99. Clemann, N. et al. Value and impacts of collecting vertebrate voucher specimens, with guidelines for ethical collection. *Mem. Mus. Vic.* **72**, 141–151 (2014).
100. Syfert, M. M., Smith, M. J. & Coomes, D. A. The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. *PLoS ONE* **8**, e55158 (2013).
101. Soberón, J., Jiménez, R., Golubov, J. & Koleff, P. Assessing completeness of biodiversity databases at different spatial scales. *Ecography* **30**, 152–160 (2007).
102. Marcer, A. et al. Uncertainty matters: ascertaining where specimens in natural history collections come from and its implications for predicting species distributions. *Ecography* **2022**, e06025 (2022).
103. Zizka, A. et al. CoordinateCleaner: standardized cleaning of occurrence records from biological collection databases. *Methods Ecol. Evol.* **10**, 744–751 (2019).
104. *World Flora Online* (WHO, accessed 12 July 2022); http://www.worldfloraonline.org
105. Pinkert, S., Barve, V., Guralnick, R. & Jetz, W. Global geographical and latitudinal variation in butterfly species richness captured through a comprehensive country-level occurrence database. *Glob. Ecol. Biogeogr.* **31**, 830–839 (2022).
106. Frost, D. R. *Amphibian Species of the World: An Online Reference (Version 5.3)* (American Museum of Natural History, 2009); http://research.amnh.org/vz/herpetology/amphibia/index.php
107. del Hoyo, J. & Collar, N. J. *HBW and BirdLife International Illustrated Checklist of the Birds of the World: Non-passerines* Vol. 1 (Lynx Edicions, 2014).

108. del Hoyo, J. & Collar, N. J. *HBW and BirdLife International Illustrated Checklist of the Birds of the World: Passerines* Vol. 2 (Lynx Edicions, 2016).
109. Uetz, P., Freed, P., Aguilar, R. & Hošek, J. (eds) *The Reptile Database* (accessed January 6, 2020); http://reptile-database.org/
110. Wilson, D. E. & Reeder, D. M. *Mammal Species of the World: A Taxonomic and Geographic Reference* 3rd edn (John Hopkins Univ. Press, 2005).
111. Soberón, J. & Peterson, T. Biodiversity informatics: managing and applying primary biodiversity data. *Philos. Trans. R. Soc. B* **359**, 689–698 (2004).
112. *The IUCN Red List of Threatened Species. Version 6.2* (IUCN, accessed 28 February 2022); https://www.iucnredlist.org
113. *Bird Species Distribution Maps of the World. Version 2020.1* (BirdLife International, 2020); http://datazone.birdlife.org/species/requestdis
114. Brummitt, R. K. *World Geographical Scheme for Recording Plant Distributions* 2nd edn (TDWG, 2001); http://www.tdwg.org/standards/109
115. Harris, L. W. & Davies, T. J. A complete fossil-calibrated phylogeny of seed plant families as a tool for comparative analyses: testing the 'time for speciation' hypothesis. *PLoS ONE* **11**, e0162907 (2016).
116. Shields, O. World numbers of butterflies. *J. Lepid. Soc.* **43**, 178–183 (1989).
117. Smith, S. A. & Brown, J. W. Constructing a broadly inclusive seed plant phylogeny. *Am. J. Bot.* **105**, 302–314 (2018).
118. Chazot, N. et al. Priors and posteriors in Bayesian timing of divergence analyses: the age of butterflies revisited. *Syst. Biol.* **68**, 797–813 (2019).
119. Jetz, W. & Pyron, R. A. The interplay of past diversification and evolutionary isolation with present imperilment across the amphibian tree of life. *Nat. Ecol. Evol.* **2**, 850–858 (2018).
120. Jetz, W., Thomas, G. H., Joy, J. B., Hartmann, K. & Mooers, A. O. The global diversity of birds in space and time. *Nature* **491**, 444–448 (2012).
121. Tonini, J. F. R., Beard, K. H., Ferreira, R. B., Jetz, W. & Pyron, R. A. Fully-sampled phylogenies of squamates reveal evolutionary patterns in threat status. *Biol. Conserv.* **204**, 23–31 (2016).
122. Bininda-Emonds, O. R. et al. The delayed rise of present-day mammals. *Nature* **446**, 507–512 (2007).
123. Kreft, H. & Jetz, W. Global patterns and determinants of vascular plant diversity. *Proc. Natl Acad. Sci. USA* **104**, 5925–5930 (2007).
124. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2022).
125. Daru, B. H., Karunarathne, P. & Schliep, K. phyloregion: R package for biogeographic regionalization and macroecology. *Methods Ecol. Evol.* **11**, 1483–1491 (2020).
126. Hijmans, R. terra: Spatial data analysis. R package version 1.7-3 https://CRAN.R-project.org/package=terra (2023).
127. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
128. Bivand, R. S., Hauke, J. & Kossowski, T. Computing the Jacobian in Gaussian spatial autoregressive models: an illustrated comparison of available methods. *Geogr. Anal.* **45**, 150–179 (2013).
129. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag, 2016).
130. Jombart, T. & Dray, S. adephylo: exploratory analyses for the phylogenetic comparative method. *Bioinformatics* **26**, 1907–1909 (2008).
131. Revell, L. J. phytools: An R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**, 217–223 (2012).
132. Abouheif, E. A method for testing the assumption of phylogenetic independence in comparative data. *Evol. Ecol. Res.* **1**, 895–909 (1999).
133. Blomberg, S. P., Garland, T. & Ives, A. R. Testing for phylogenetic signal in comparative data: behavioural traits are more labile. *Evolution* **57**, 717–745 (2003).
134. Pagel, M. Inferring the historical patterns of biological evolution. *Nature* **401**, 877–884 (1999).
135. Pavoine, S., Ollier, S., Pontier, D. & Chessel, D. Testing for phylogenetic signal in phenotypic traits: new matrices of phylogenetic proximities. *Theor. Popul. Biol.* **73**, 79–91 (2008).
136. Koleff, P. et al. Measuring beta diversity for presence–absence data. *J. Anim. Ecol.* **72**, 367–382 (2003).
137. Kingsland, S. The importance of history and historical records for understanding the Anthropocene. *Bull. Ecol. Soc. Am.* **98**, 64–71 (2017).
138. Díaz, S. et al. The global spectrum of plant form and function. *Nature* **529**, 167–171 (2016).
139. Shirey, V. et al. LepTraits 1.0: a globally comprehensive dataset of butterfly traits. *Sci. Data* **9**, 382 (2022).
140. Oliveira, B. et al. AmphiBIO, a global database for amphibian ecological traits. *Sci. Data* **4**, 170123 (2017).
141. Myhrvold, N. P. et al. An amniote life-history database to perform comparative analyses with birds, mammals, and reptiles. *Ecology* **96**, 3109–3109 (2015).
142. Wilman, H. et al. EltonTraits 1.0: species-level foraging attributes of the world's birds and mammals. *Ecology* **95**, 2027–2027 (2014).
143. Carmona, C. P. et al. Erosion of global functional diversity across the tree of life. *Sci. Adv.* **7**, eabf2675.
144. Davies, T. J. The macroecology and macroevolution of plant species at risk. *New Phytol.* **222**, 708–713 (2019).
145. Purvis, A. et al. Nonrandom extinction and the loss of evolutionary history. *Science* **288**, 328–330 (2000).
146. Pelletier, T. A. et al. Predicting plant conservation priorities on a global scale. *Proc. Natl Acad. Sci. USA* **115**, 13027–13032 (2018).
147. Yessoufou, K., Daru, B. H. & Davies, T. J. Phylogenetic patterns of extinction risk in the Eastern Arc ecosystems, an African biodiversity hotspot. *PLoS ONE* **7**, e47082 (2012).
148. Redding, D. W. & Mooers, A. Ø. Incorporating evolutionary measures into conservation prioritization. *Conserv. Biol.* **20**, 1670–1678 (2006).
149. Bielby, J., Cunningham, A. A. & Purvis, A. Taxonomic selectivity in amphibians: ignorance, geography or biology? *Anim. Conserv.* **9**, 135–143 (2006).
150. Nelson, A. *Travel Time to Major Cities: A Global Map of Accessibility* (Global Environment Monitoring Unit, 2008).
151. Fick, S. E. & Hijmans, R. J. WorldClim 2: new 1 km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* **37**, 4302–4315 (2017).
152. Allan, J., Venter, O. & Watson, J. Temporally inter-comparable maps of terrestrial wilderness and the Last of the Wild. *Sci. Data* **4**, 170187 (2017).
153. Partow, A. *The Global Airport Database. Release Version 0.0.1* (Partow, 2003); http://www.partow.net/miscellaneous/airportdatabase/
154. *Global Peace Index 2022: Measuring Peace in a Complex World* (Institute for Economics & Peace, accessed 9 July 2022); http://visionofhumanity.org/resources
155. Palmer, L. Show me the money. *Nat. Clim. Change* **1**, 376–380 (2011).
156. *Science and Technology Report* (UNESCO Institute for Statistics, 2012); http://www.uis.unesco.org/ScienceTechnology/Pages/research-and-development-statistics.aspx

157. Bivand, R. R packages for analyzing spatial data: a comparative case study with areal data. *Geograph. Anal.* https://doi.org/10.1111/gean.12319 (2022).

## Acknowledgements

We thank Stanford University and Texas A&M University-Corpus Christi for logistic support. B.H.D. was supported by the US National Science Foundation (awards 2031928 and 2113424). We are grateful to G. Nakamura, L. Ford and S. Pons for comments on earlier drafts of the paper. In addition, we are grateful to Holger Kreft for kindly sharing his data on the expected distribution of plants, which was instrumental in our analysis.

## Author contributions

The study was conceived and designed by B.H.D. Analyses were carried out by B.H.D. The paper was written by B.H.D and revised by B.H.D. with help from J.R.

## Competing interests

The authors declare no competing interests.

## Additional information

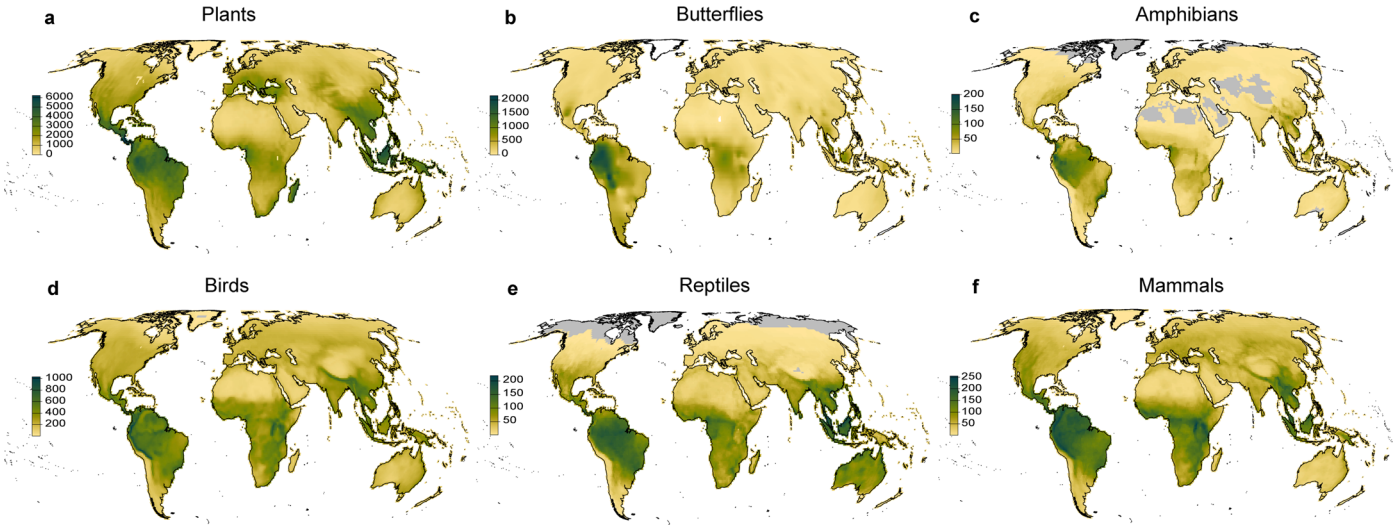
**Extended data** is available for this paper at https://doi.org/10.1038/s41559-023-02047-3.

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41559-023-02047-3.

**Correspondence and requests for materials** should be addressed to Barnabas H. Daru.

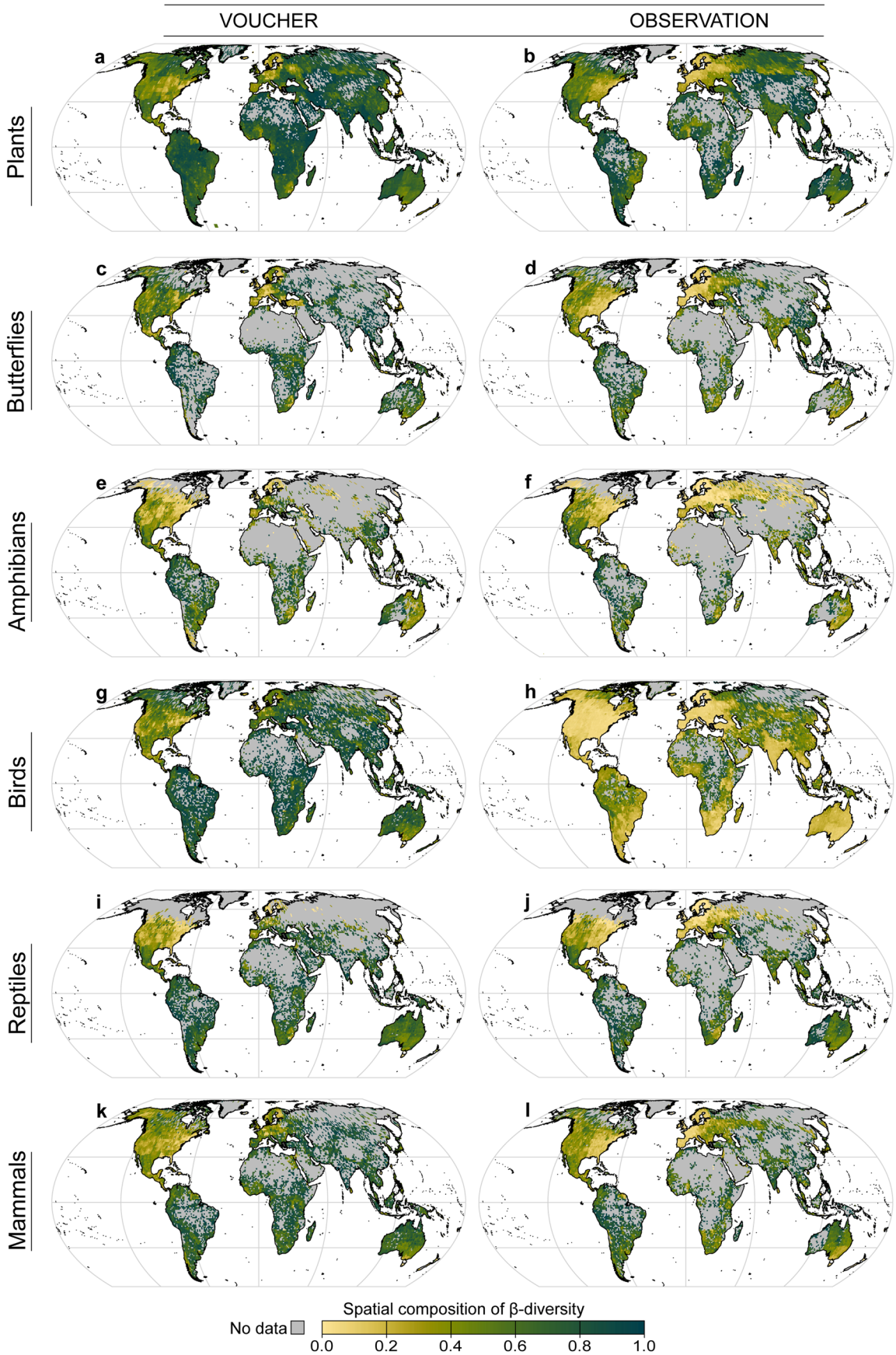
**Peer review information** *Nature Ecology & Evolution* thanks James Speed and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

https://doi.org/10.1038/s41559-023-02047-3




**Extended Data Fig. 1 | Patterns of expected species richness of terrestrial taxa.** The expected species richness of (**a**) Plants was derived from a co-kriging interpolation model of 1,032 regional floras worldwide, and (**b**) Butterflies, derived from a co-kriging interpolation of 543 geographic units covering the known inventory of butterflies, whereas the expected species richness of (**c**) Amphibians, (**d**) Birds, (**e**) Reptiles, and (**f**) Mammals, were generated by overlaying expert-based extent-of-occurrence range map of each species with equal-area grid cells of 100 km × 100 km. The bamako colour palette is common to all panels, with dark green indicating high coverage and yellow indicating low coverage. The maps are in the Wagner IV projection.

https://doi.org/10.1038/s41559-023-02047-3

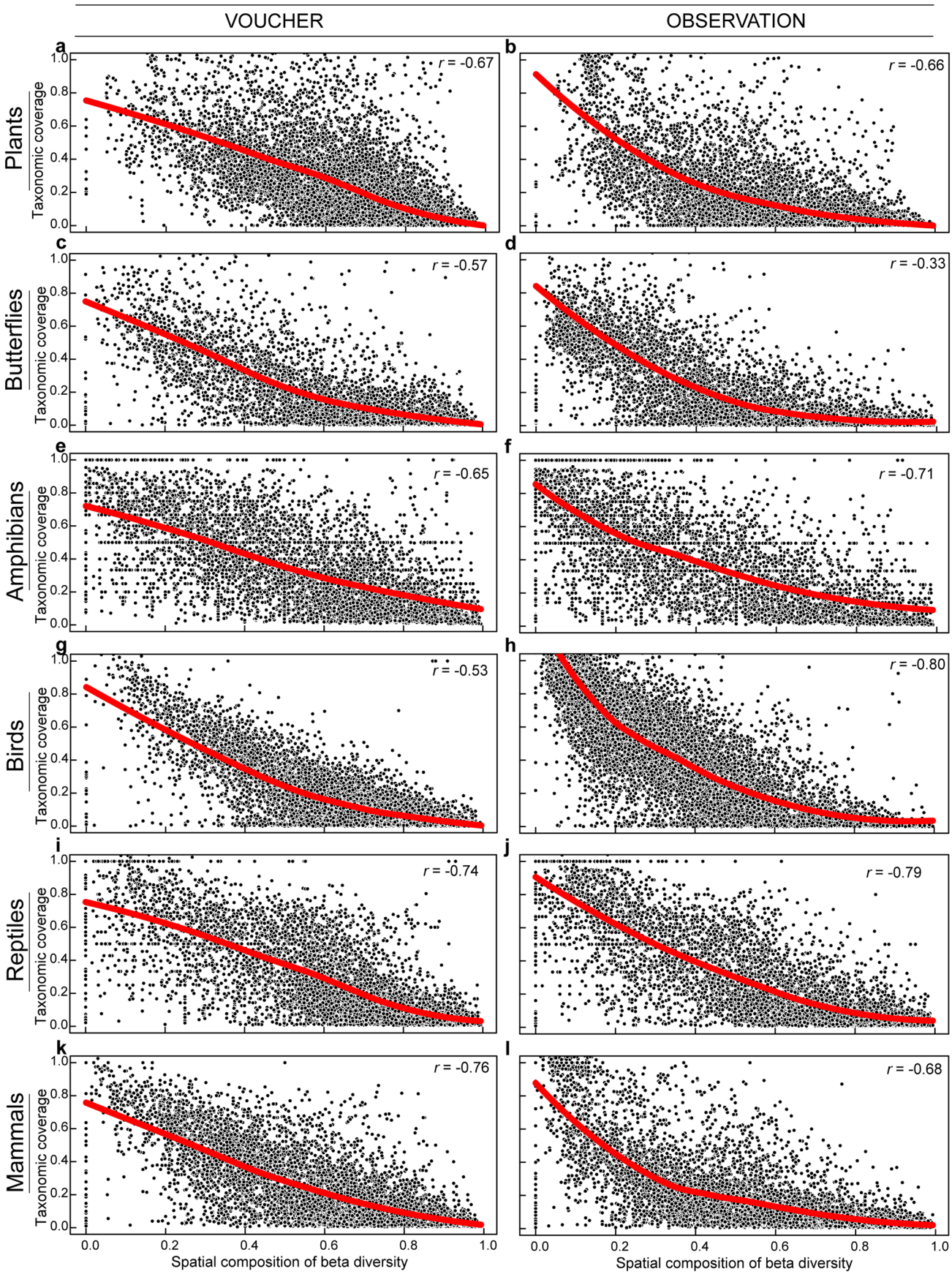



**Extended Data Fig. 2 | See next page for caption.**

https://doi.org/10.1038/s41559-023-02047-3

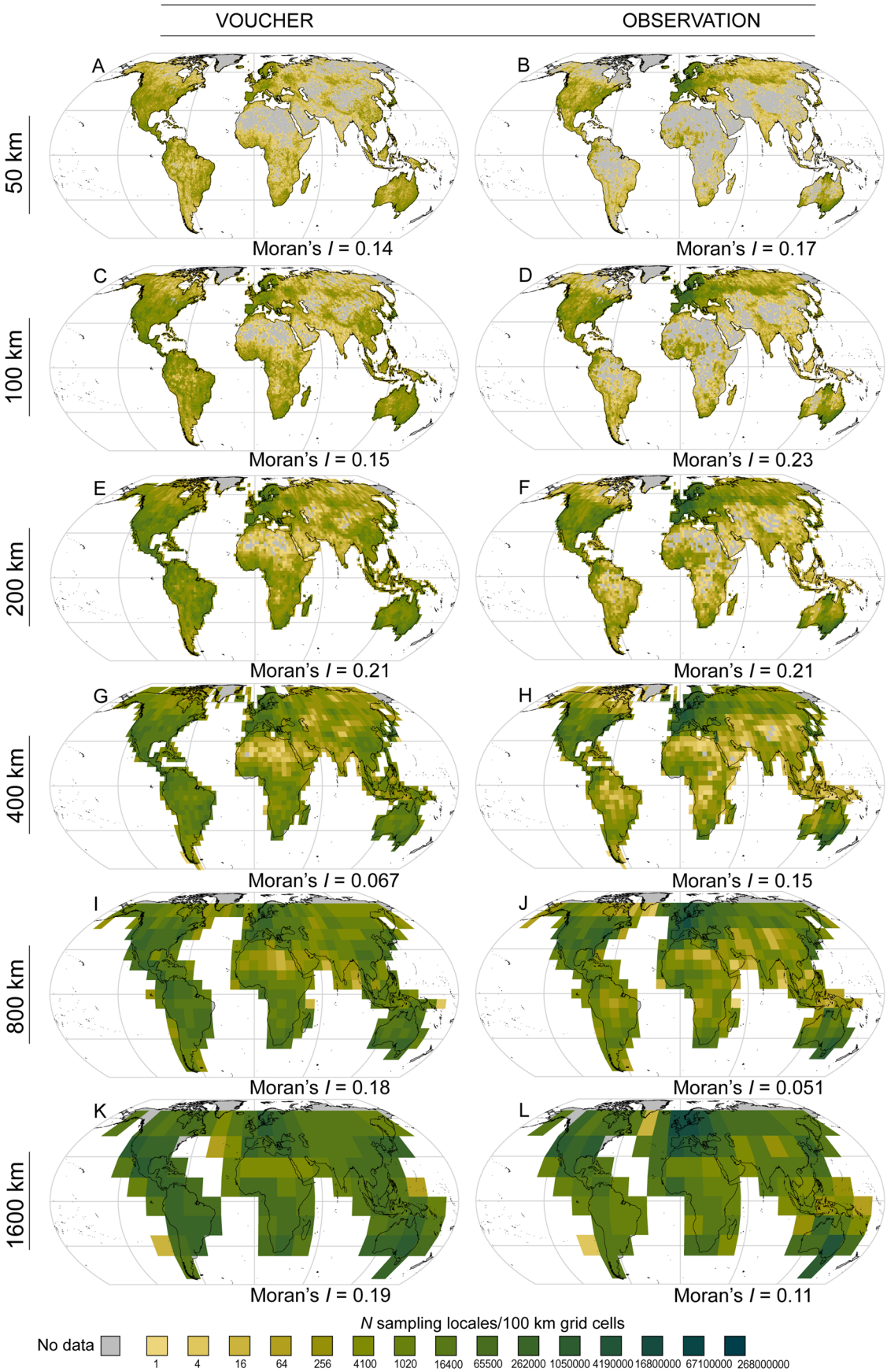
**Extended Data Fig. 2 | Spatial composition of β-diversity across grid cells by voucher and observation records.** Maps of dissimilarity between record types for: (**a, b**) Plants (n = 240,377 species), (**c**, **d**) Butterflies (n = 9809 species), (**e**, **f**) Amphibians (n = 4862 species), (**g**, **h**) Birds (n = 9380 species), (**i**, **j**) Reptiles (n = 7259 species), and (**k**, **l**) Mammals (n = 4508 species). Dissimilarity was assessed by generating pairwise distance matrices of Simpson's β-diversity between all pairs of grid cells within major biogeographically defined areas recognized by the Biodiversity Information Standards (also known as the Taxonomic Databases Working Group (TDWG)). Values of β vary between 0 (species composition is identical between grid cells) and 1 (high dissimilarity, no shared taxa). Both voucher and observation records of most taxonomic groups showed high dissimilarity in less frequently sampled regions of South America, Africa, and Southeast Asia, and decline in frequently sampled Europe and North America.

https://doi.org/10.1038/s41559-023-02047-3




**Extended Data Fig. 3 | Relationship between sampling effort (measured as taxonomic coverage) versus dissimilarity (measured as spatial composition of beta diversity) by voucher and observation records.** Indicated are the relationships between sampling effort and dissimilarity of record types for (**a**, **b**) plants, (**c**, **d**) butterflies, (**e**, **f**) amphibians, (**g**, **h**) birds, (**i**, **j**) reptiles, and (**k**, **l**) mammals. Trend line (in red) computed by evaluating the loess smooth at equally spaced points covering the range of dissimilarity values for each sampling effort.

https://doi.org/10.1038/s41559-023-02047-3

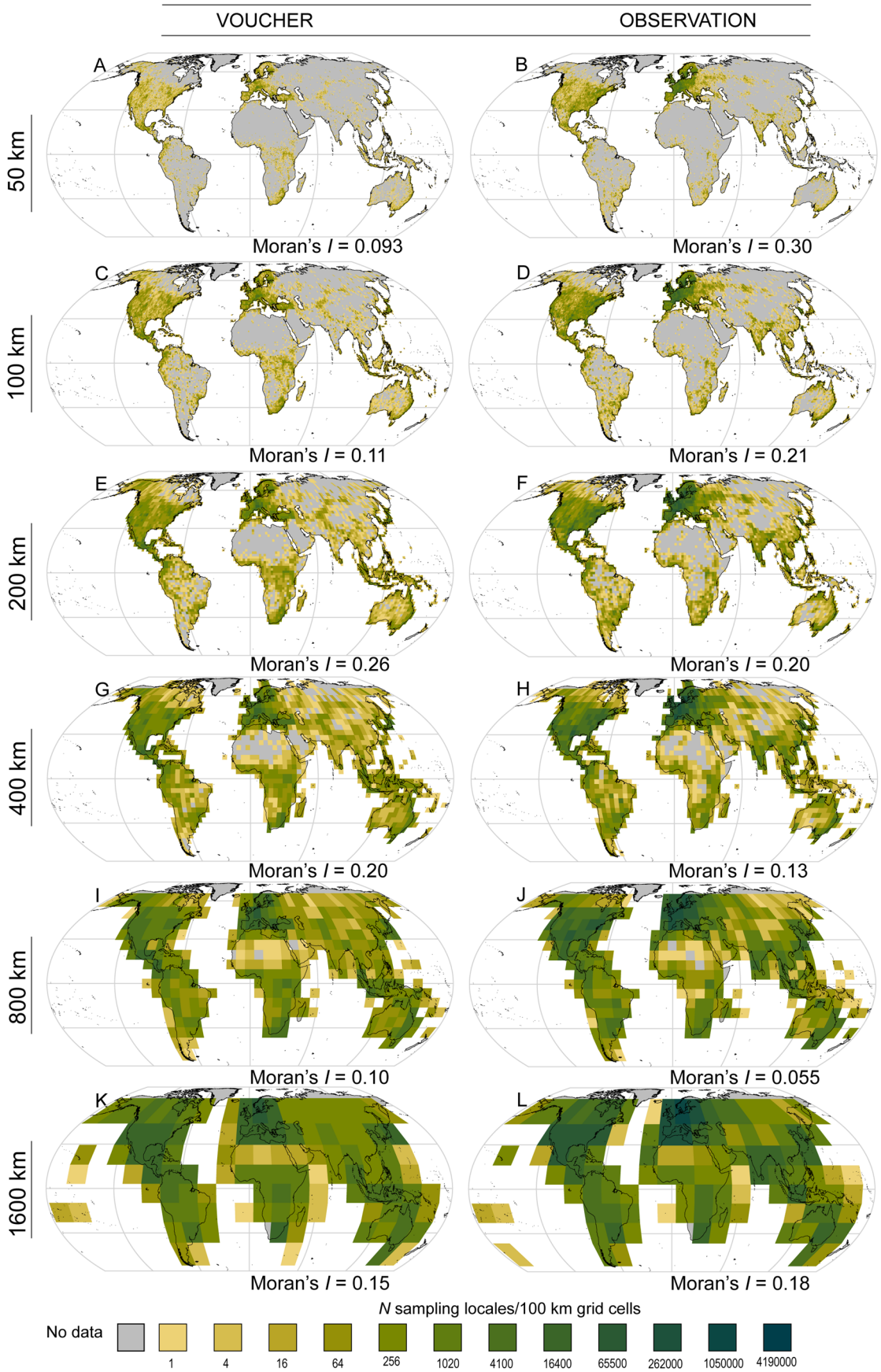



**Extended Data Fig. 4 | See next page for caption.**

https://doi.org/10.1038/s41559-023-02047-3

**Extended Data Fig. 4 | Patterns of geographic coverage of grid cells by voucher and observation records of plants across spatial grain (50 × 50, 100 × 100, 200 × 200, 400 × 400, 800 × 800 and 1600 km × 1600 km).** Geographic coverage of grid cells was calculated as number of unique collection locales for each grid cell. Evenness or clustering of geographic coverage indicated by Moran's *I* (Monte Carlo test, 999 randomizations) with values of 1 indicating clustered/biased coverage and 0 corresponding to geographically even coverage. The bamako colour palette is common to all panels, with darkgreen indicating high coverage and yellow indicating low coverage. The maps are in the Wagner IV projection.

https://doi.org/10.1038/s41559-023-02047-3

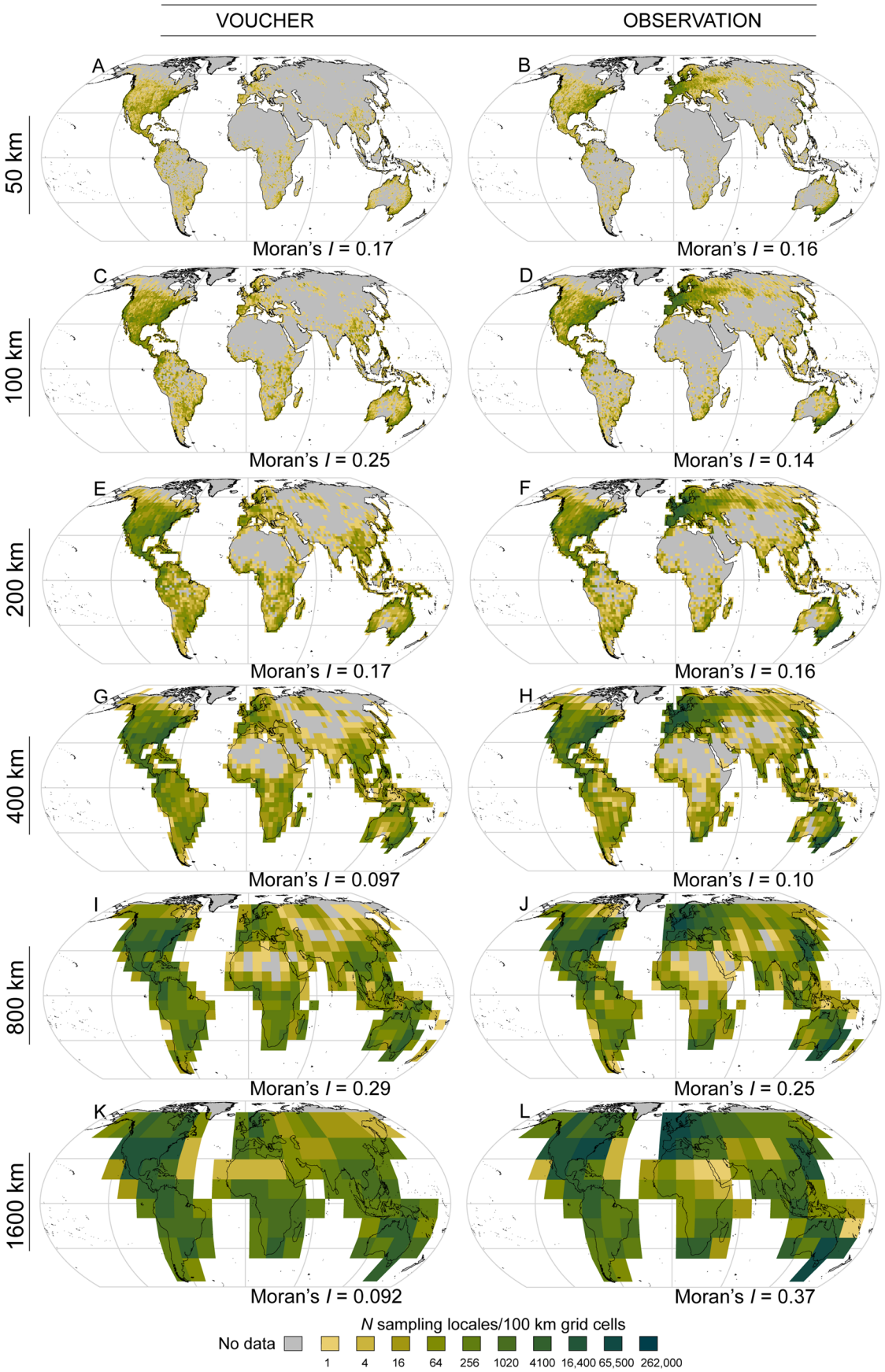



**Extended Data Fig. 5 | See next page for caption.**

https://doi.org/10.1038/s41559-023-02047-3

**Extended Data Fig. 5 | Patterns of geographic coverage of grid cells by voucher and observation records of butterflies across spatial grain (50 × 50, 100 × 100, 200 × 200, 400 × 400, 800 × 800 and 1600 km × 1600 km).** Geographic coverage of grid cells was calculated as number of unique collection locales for each grid cell. Evenness or clustering of geographic coverage indicated by Moran's *I* (Monte Carlo test, 999 randomizations) with values of 1 indicating clustered/biased coverage and 0 corresponding to geographically even coverage. The bamako colour palette is common to all panels, with darkgreen indicating high coverage and yellow indicating low coverage. The maps are in the Wagner IV projection.

https://doi.org/10.1038/s41559-023-02047-3

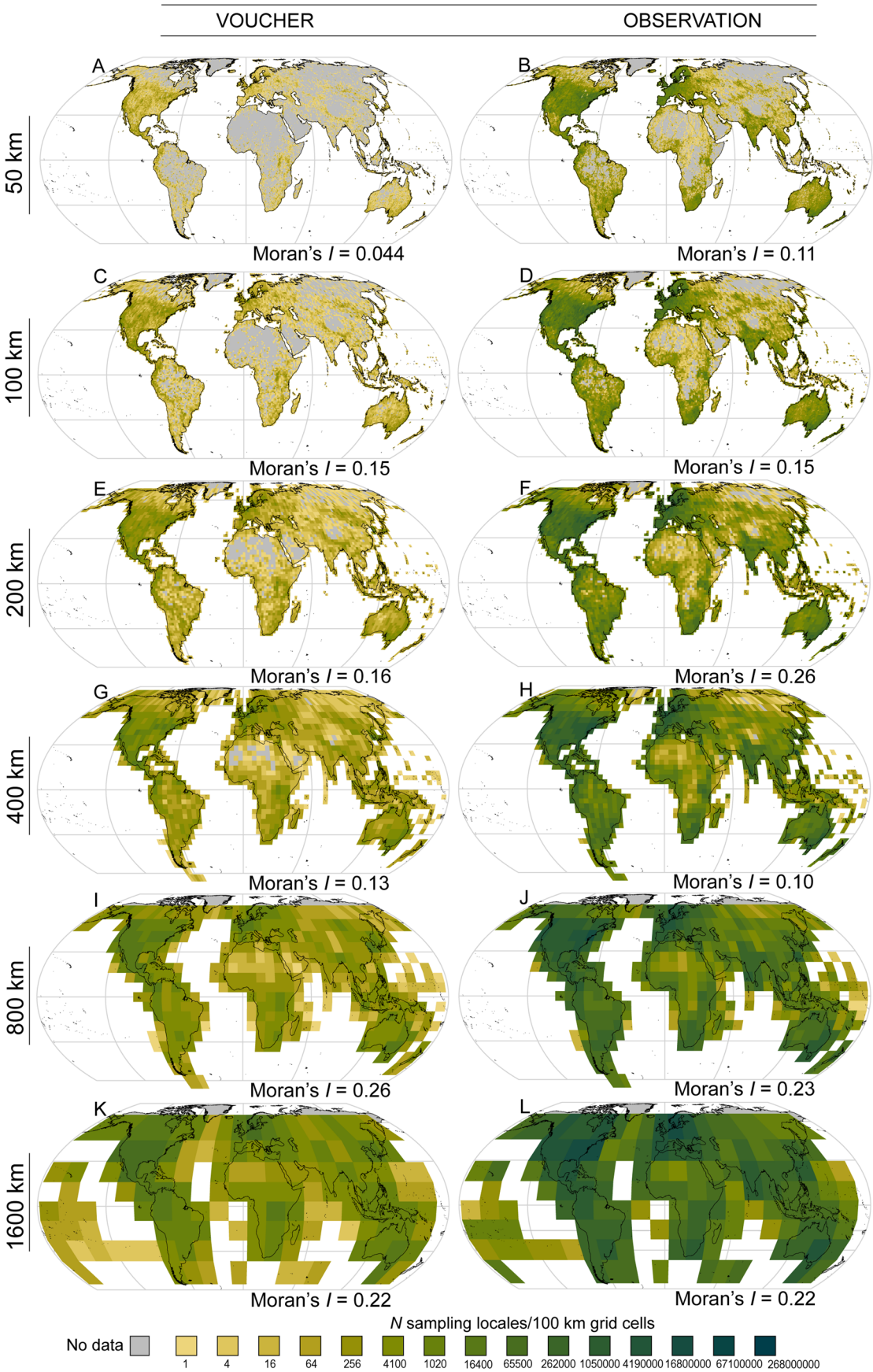



**Extended Data Fig. 6 | See next page for caption.**

https://doi.org/10.1038/s41559-023-02047-3

**Extended Data Fig. 6 | Patterns of geographic coverage of grid cells by voucher and observation records of amphibians across spatial grain (50 × 50, 100 × 100, 200 × 200, 400 × 400, 800 × 800 and 1600 km × 1600 km).** Geographic coverage of grid cells was calculated as number of unique collection locales for each grid cell. Evenness or clustering of geographic coverage indicated by Moran's *I* (Monte Carlo test, 999 randomizations) with values of 1 indicating clustered/biased coverage and 0 corresponding to geographically even coverage. The bamako colour palette is common to all panels, with darkgreen indicating high coverage and yellow indicating low coverage. The maps are in the Wagner IV projection.

https://doi.org/10.1038/s41559-023-02047-3

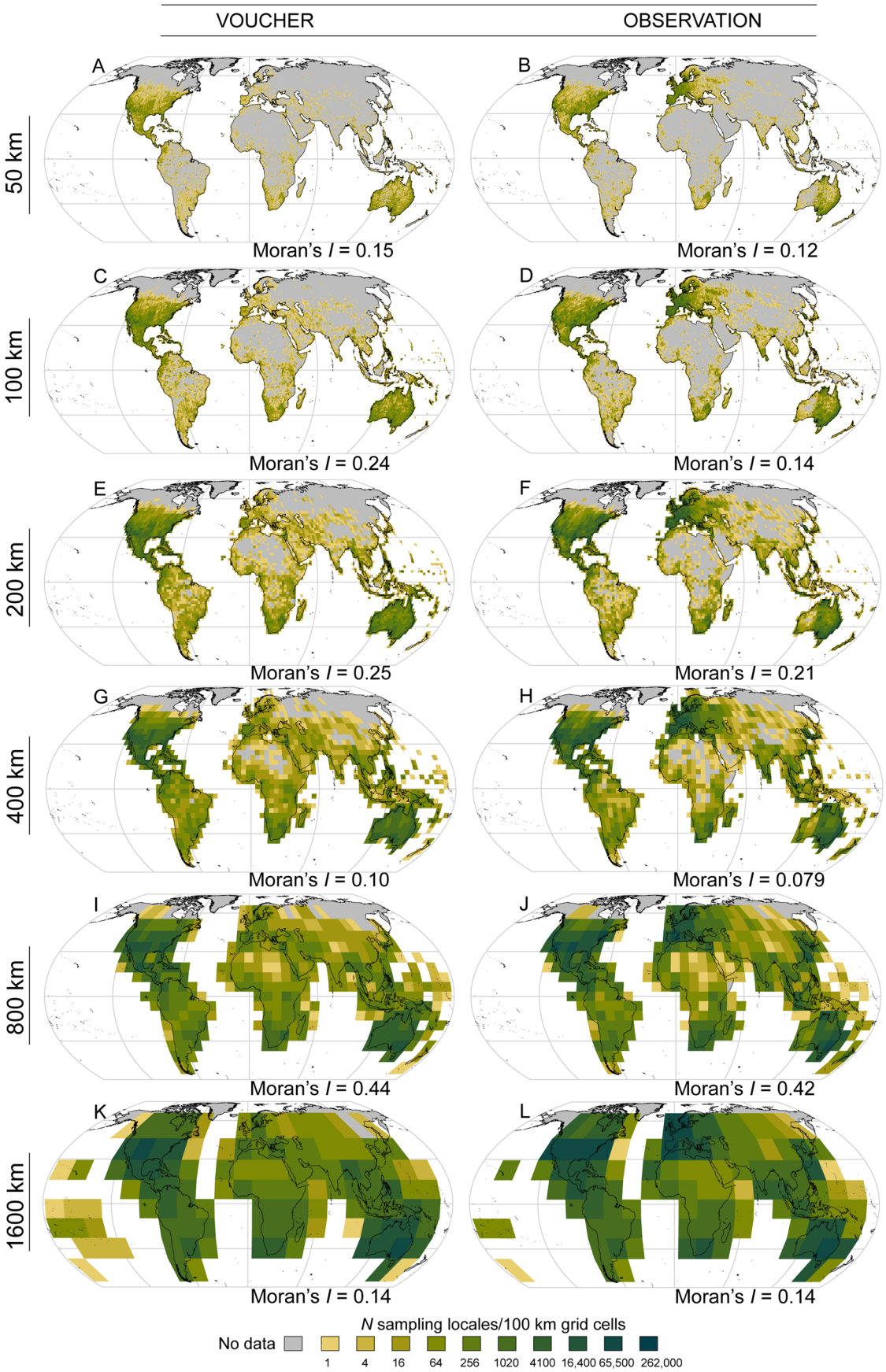



**Extended Data Fig. 7 | See next page for caption.**

https://doi.org/10.1038/s41559-023-02047-3

**Extended Data Fig. 7 | Patterns of geographic coverage of grid cells by voucher and observation records of birds across spatial grain (50 × 50, 100 × 100, 200 × 200, 400 × 400, 800 × 800 and 1600 km × 1600 km).** Geographic coverage of grid cells was calculated as number of unique collection locales for each grid cell. Evenness or clustering of geographic coverage indicated by Moran's *I* (Monte Carlo test, 999 randomizations) with values of 1 indicating clustered/biased coverage and 0 corresponding to geographically even coverage. The bamako colour palette is common to all panels, with darkgreen indicating high coverage and yellow indicating low coverage. The maps are in the Wagner IV projection.

https://doi.org/10.1038/s41559-023-02047-3

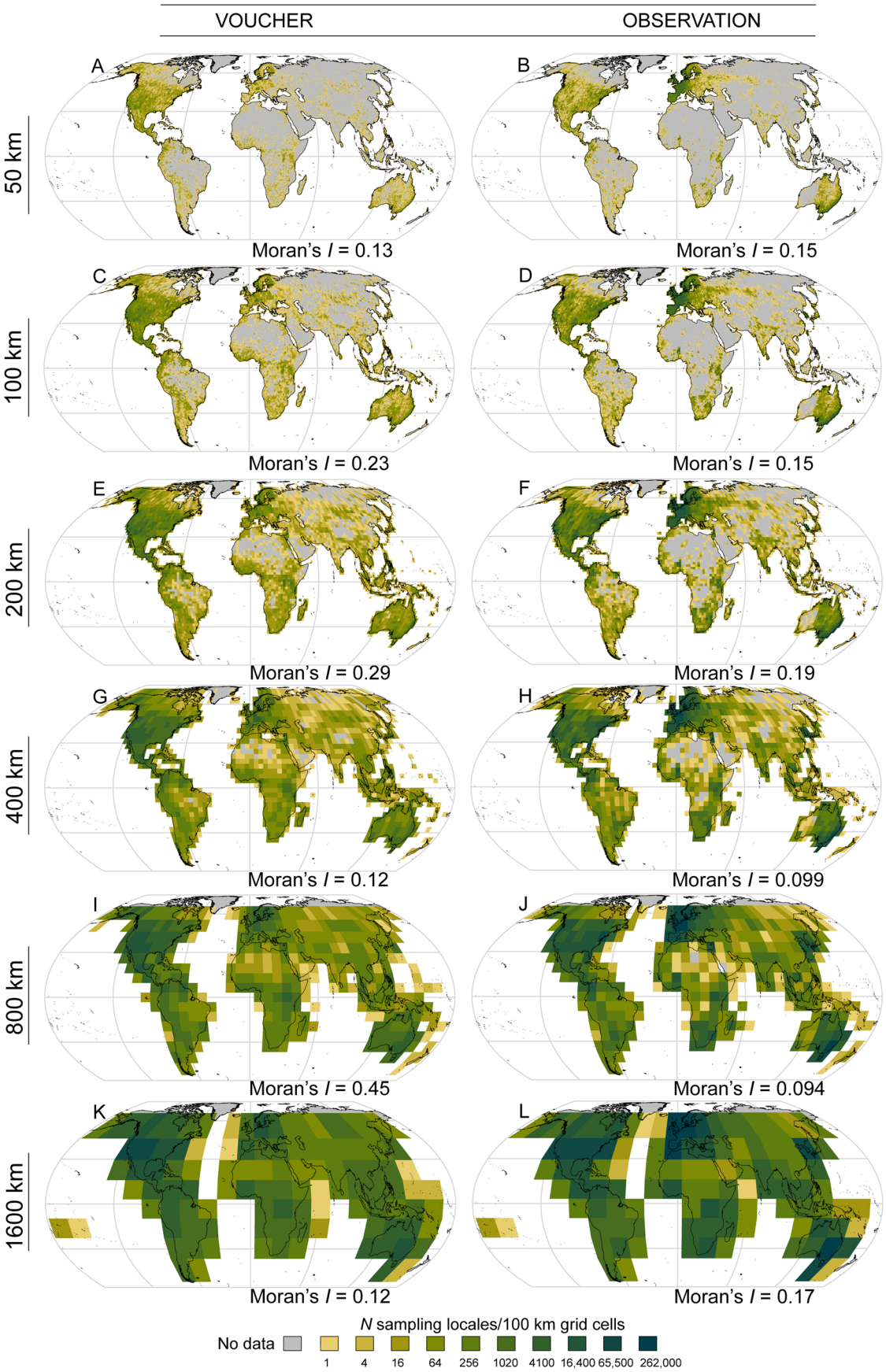



**Extended Data Fig. 8 | See next page for caption.**

https://doi.org/10.1038/s41559-023-02047-3

**Extended Data Fig. 8 | Patterns of geographic coverage of grid cells by voucher and observation records of reptiles across spatial grain (50 × 50, 100 × 100, 200 × 200, 400 × 400, 800 × 800 and 1600 km × 1600 km).** Geographic coverage of grid cells was calculated as number of unique collection locales for each grid cell. Evenness or clustering of geographic coverage indicated by Moran's *I* (Monte Carlo test, 999 randomizations) with values of 1 indicating clustered/biased coverage and 0 corresponding to geographically even coverage. The bamako colour palette is common to all panels, with darkgreen indicating high coverage and yellow indicating low coverage. The maps are in the Wagner IV projection.

https://doi.org/10.1038/s41559-023-02047-3

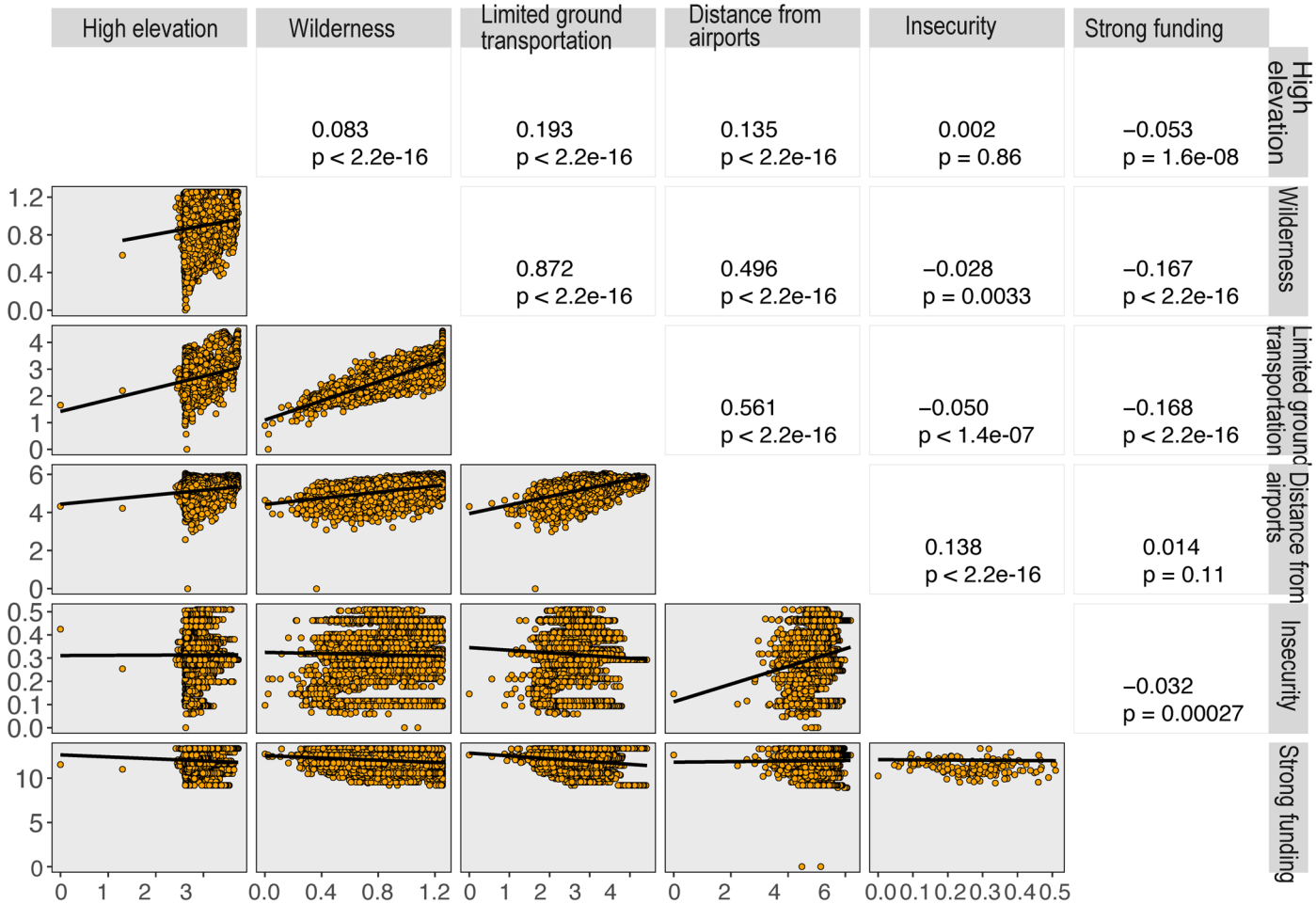



**Extended Data Fig. 9 | See next page for caption.**

https://doi.org/10.1038/s41559-023-02047-3

**Extended Data Fig. 9 | Patterns of geographic coverage of grid cells by voucher and observation records of mammals across spatial grain (50 × 50, 100 × 100, 200 × 200, 400 × 400, 800 × 800 and 1600 km × 1600 km).** Geographic coverage of grid cells was calculated as number of unique collection locales for each grid cell. Evenness or clustering of geographic coverage indicated by Moran's *I* (Monte Carlo test, 999 randomizations) with values of 1 indicating clustered/biased coverage and 0 corresponding to geographically even coverage. The bamako colour palette is common to all panels, with darkgreen indicating high coverage and yellow indicating low coverage. The maps are in the Wagner IV projection.

https://doi.org/10.1038/s41559-023-02047-3




**Extended Data Fig. 10 | Pairwise relationships between 6 socioeconomic and ecological variables.** Correlations based on pairwise Spearman-rank correlations between the variables at spatial grain of 100 km. All variables were log-transformed before analysis. The statistical test used was two-sided. Exact p values are indicated below correlation coefficients.

nature portfolio

Corresponding author(s): BARNABAS H. DARU

Last updated by author(s): Mar 22, 2023

# Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

| n/a | Confirmed | |
|---|---|---|
| ☐ | ☒ | The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☒ | ☐ | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ | The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☐ | ☒ | A description of all covariates tested |
| ☐ | ☒ | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ | For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☒ | ☐ | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☐ | ☒ | Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |

*Our web collection on statistics for biologists contains articles on many of the points above.*

## Software and code

Policy information about availability of computer code

| | |
|---|---|
| Data collection | The open source software used to collect the is the R computing language version 4.2.1 (2022-06-23) -- "Funny-Looking Kid". The expected species richness for butterflies was newly generated in this study. This was achieved by extrapolating richness information from inventories, checklists, online regional databases, and literature sources and projecting across grid cells. We then fitted a co-kriging interpolation model to estimate the probability of butterfly occurrence into unsampled areas based on recorded richness across 543 geographic units, and four predictor variables. The final output from the co-kriging model consisted of a modeled map in raster format at grid cell resolution of 0.5 degree equivalent to 50 km at the equator, which we resampled to six different grain sizes (50, 100, 200, 400, 800, and 1600 km).<br>The R code and data documentation necessary to repeat our analyses have been made available in the Zenodo database [https://doi.org/10.5281/zenodo.6834577] under the folder "SCRIPTS". |
| Data analysis | The following open source software and R scripts were used to analyze the data: R v.4.2.2 and packages phyloregion v.1.0.8, terra v.1.7-3, ape 5.6-2, spatialreg v.1.2-6, gglot2 v.3.4.0, adephylo v.1.1-13, phytools v.1.2-0, scico 1.3.0, and ggtree 3.4.1. Custom R scripts were developed to analyze the spatial distribution data and are permanently available at the Zenodo database [https://doi.org/10.5281/zenodo.6834577]. |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The links to the species occurrence records downloaded from the Global Biodiversity Information Facility are available at Zenodo (https://doi.org/10.5281/zenodo.6834577) and provided in Supplementary Table 1. The datasets, data tables, grid cell vector polygons, and R codes are archived at Zenodo (https://doi.org/10.5281/zenodo.6834577).

## Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

| | |
|---|---|
| Reporting on sex and gender | Not applicable in this study |
| Population characteristics | Not applicable in this study |
| Recruitment | Not applicable in this study |
| Ethics oversight | Not applicable in this study |

Note that full information on the approval of the study protocol must also be provided in the manuscript.

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

# Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Study description | This study quantifies coverage and biases of expected biodiversity patterns by voucher and observation records of major terrestrial species groups including vascular plants, butterflies, amphibians, birds, reptiles, and mammals. Data were compiled primarily from GBIF, functional traits for each taxonomic group compiled from various sources. The data were analyzed taxonomically, geographically, temporally, and along functional trait axes, to contrast coverage by voucher and observation records along these dimensions. |
| Research sample | For this study, a species occurrence record was considered as a research sample. To this end, we downloaded data of c. 1.9 billion occurrence records from the global biodiversity information facility (GBIF) for the six taxonomic groups: plants (Plantae; n = 374 M records), butterflies (Rhopalocera: Hedyloidea and Papilonoidea including Heperiidae; n = 72 M records), amphibians (8.1 M), birds (1.4 B), reptiles (8 M), and mammals (29 M). |
| Sampling strategy | We distinguished the origin of each record based on whether they came from material with a physical voucher specimen in museum or herbarium or from observations that are not traceable to tangible physical material in a museum or herbarium. Occurrence records with unknown categories were removed from the analysis. The datasets were thoroughly cleaned to remove duplicates and records with erroneous localizations using the R package CoordinateCleaner v.2.0-20. |
| Data collection | Data on species occurrences were collected primarily from GBIF by Barnabas H. Daru. All analyses were carried out by Barnabas H. Daru. |
| Timing and spatial scale | Data collection from GBIF is in contemporary times and started from March 20, 2022 to June 3, 2022 (Table S1). The spatial scope of the analysis is global, covering all terrestrial areas. |
| Data exclusions | No data were intentionally excluded from the analyses. However, occurrence records with unknown categories were removed from the analysis. The datasets were thoroughly cleaned to remove duplicates and retain records with valid geographic localities and acceptable scientific names following currently accepted taxonomies. Species categorized by the IUCN as 'Data deficient' were also excluded from our analysis of extinction risk. |

| | |
|---|---|
| Reproducibility | All scripts, codes, and data documentation necessary to repeat and reproduce our analyses have been made available in the Zenodo database [https://doi.org/10.5281/zenodo.6834577] under the folder "SCRIPTS". |
| Randomization | For our analysis of estimating biases in taxonomic coverage of lineages, we used three common indices of phylogenetic signal. Statistical significance was assessed by calculating the standardized effective size of phylogenetic signal based on 1000 randomizations of the trait values across the tips of the phylogeny. Biases in taxonomic coverage of grid cells were analyzed using Moran's I spatial autocorrelation measure (with Monte Carlo test, 999 randomizations). |
| Blinding | Blinding was not relevant in this study. |

Did the study involve field work? ☐ Yes ☒ No

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ Antibodies |
| ☒ | ☐ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology and archaeology |
| ☒ | ☐ Animals and other organisms |
| ☒ | ☐ Clinical data |
| ☒ | ☐ Dual use research of concern |

## Methods

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ ChIP-seq |
| ☒ | ☐ Flow cytometry |
| ☒ | ☐ MRI-based neuroimaging |