*Diversity and Distributions*

WILEY

Diversity and Distributions

DATA PAPER OPEN ACCESS

# A Global, Taxon-Stratified, High-Resolution Sampling-Effort Dataset From GBIF for Bias-Aware Ecological Modelling

Ahmed El-Gabbas

Department of Community Ecology, Helmholtz-Centre for Environmental Research – UFZ, Halle (Saale), Germany

**Correspondence:** Ahmed El-Gabbas (elgabbas@outlook.com)

**Received:** 19 November 2025 | **Revised:** 25 April 2026 | **Accepted:** 5 May 2026

**Editor:** Cesar Capinha

**Keywords:** biodiversity informatics | data gaps | GBIF | sampling bias | sampling efforts | species distribution modelling

## ABSTRACT

**Introduction and Aim:** Spatiotemporal and taxonomic sampling bias in biodiversity occurrence data poses critical challenges for robust ecological inference, species distribution models (SDMs), and conservation planning. Despite the exponential growth in global biodiversity records over recent decades, these biases persist. This study converts raw occurrence records from the Global Biodiversity Information Facility (GBIF) into global, publicly available, taxon-stratified, and temporally resolved sampling-effort rasters using a reproducible workflow, providing transparent and standardised measures of observation count and species richness to support bias-aware ecological analyses.

**Main Variables Included:** Two complementary raster variables: observation count and species richness, each provided across major taxonomic groups and their descendant levels (e.g., classes, orders, families).

**Time Coverage:** Annual and cumulative rasters span 1980–2025.

**Spatial Coverage:** Global; four spatial resolutions (~1, 5, 10, and 20 km).

**Taxa:** Nine major taxonomic groups: Amphibia, Arachnida, Aves (birds), Fungi, Insecta, Mammalia, Mollusca, Reptilia, and Tracheophyta (vascular plants), with descendant-level outputs.

**Applications:** Based on ~3 billion records for > 730,000 species, this study provides annual and cumulative global rasters quantifying observation count and species richness at four resolutions, stratified by nine taxonomic groups and their descendants. At 1 km resolution, 95% of records occupy merely 0.33% of Earth's surface (0.93% of land), whilst the remaining data extend across only 1.77% (3.88% of land), leaving approximately 98% (95% of land) unsampled. This extreme concentration persists across all taxonomic groups, underscoring the need for taxon-specific bias correction. Annual data enable exploration of long-term trends in data mobilisation and sampling effort. These rasters enable bias correction in presence-only SDMs, including MaxEnt bias files, target-group backgrounds, and model-based approaches. Beyond SDMs, they can inform macroecological synthesis, biodiversity monitoring, and systematic conservation planning by identifying spatial and temporal knowledge gaps. All data and code are openly available under FAIR principles, promoting transparent and reproducible biodiversity science.

## 1 | Introduction

Recent decades have witnessed a rapid growth in the mobilisation of biodiversity data, driven by the digitisation of collections and the growth of citizen-science platforms (Heberling et al. 2021; Hughes et al. 2021; Wüest et al. 2019). Global biodiversity repositories, such as the Global Biodiversity Information Facility (GBIF; https://www.gbif.org/), aggregate billions of georeferenced species occurrence records, providing centralised access to heterogeneous observations from museums, herbaria,

personal collections, monitoring schemes, and citizen scientists (Beck et al. 2014; Hobern et al. 2019). As of November 2025, GBIF held ~3.56 billion occurrence records. This enormous, rapidly growing resource underpins analyses in macroecology, biogeography, and conservation science by enabling large-scale inferences about species distributions, richness patterns, and environmental associations (Steinke et al. 2025). Among other uses, these datasets power species distribution models (SDMs), which are applied to estimate current species ranges, project climate-driven range shifts, and evaluate invasion or conservation risk (Araujo et al. 2019; Elith and Leathwick 2009; Jiménez-Valverde et al. 2011; Phillips et al. 2009). These data also enable macroecological analyses, inform extinction-risk assessments, and support systematic conservation planning and prioritisation (El-Gabbas et al. 2020; Guisan et al. 2013; Syfert et al. 2013).

However, opportunistic presence-only observations are frequently collected without standardised protocols, yielding biased samples of species' distributions (Beck et al. 2014; Isaac and Pocock 2015; Phillips et al. 2009). Sampling effort overlaps strongly with human accessibility, including proximity to roads, urban centres, and research institutions (Meyer et al. 2015; Warton et al. 2013). Determinants of spatial sampling effort and record availability include distribution of protected areas, local research capacity, historical collecting practices, colonial legacies, data-sharing policies, and socioeconomic factors (Amano and Sutherland 2013; El-Gabbas and Dormann 2017; Faxon and Chapman 2025; Kadmon et al. 2004; Meyer et al. 2015). Importantly, the structure and drivers of sampling bias differ substantially among taxonomic groups (Rocchini et al. 2023). For example, birds are frequently documented through large-scale citizen-science networks that generate dense spatial clustering around observation hotspots and accessible areas (Callaghan et al. 2021; La Sorte et al. 2024). In contrast, mammals are often recorded using more heterogeneous and method-dependent approaches, including camera trapping, roadkill monitoring, targeted surveys, and opportunistic observations (Green et al. 2020; Périquet et al. 2018). These differences in observer communities, detection methods, and logistical constraints generate taxon-specific spatial and temporal bias patterns that must be considered when interpreting occurrence-based analyses. Such preferential sampling produces systematic spatial bias, with record clusters concentrated in regions with vigorous research activity and infrastructure, such as Western Europe, North America, and Australia (Christie et al. 2021; Fithian et al. 2015; Hughes et al. 2021; Meyer et al. 2015). In contrast, persistent gaps remain across many tropical and remote regions (Hughes et al. 2021; Meyer et al. 2015). These spatial imbalances are compounded by taxonomic disparities in sampling effort. Well-studied groups such as birds and butterflies are comparatively well documented, whereas fungi and many invertebrate taxa remain poorly sampled (Meyer et al. 2015; Troudet et al. 2017).

Spatial sampling bias often leads to environmental bias when heavily sampled areas over-represent particular climatic or habitat conditions, further compromising ecological inference (Aiello-Lammens et al. 2015; Guillera-Arroita et al. 2015; Hirzel and Guisan 2002; Phillips et al. 2009; Pili et al. 2025; Ranc et al. 2017). SDMs built with uncorrected occurrence data may fit observer effort rather than the true environmental suitability, biasing the resulting projections (Phillips et al. 2009; Pili et al. 2025; Randin et al. 2006). Further, sampling effort often correlates with ecologically relevant climatic and land-cover variables, making it challenging to disentangle observation processes from ecological signals using presence-only data alone (Fithian et al. 2015; Kadmon et al. 2004). If unaddressed, these biases propagate through modelling workflows, degrading model accuracy and transferability (Hughes et al. 2021; Phillips et al. 2009; Pili et al. 2025) and artificially inflate habitat suitability estimates, range extents, or species richness in heavily sampled regions. In data-poor areas, models often underestimate suitability, resulting in omission errors that may overlook important conservation areas and distort climate change projections (e.g., overestimating range contractions) (Fithian et al. 2015; Kramer-Schadt et al. 2013). Such omission errors may also compromise risk assessments, for example, by obscuring expansion fronts or areas of elevated concern for invasive alien species (Broennimann et al. 2021). Ultimately, these distortions can lead to suboptimal conservation prioritisation by systematically overlooking undersampled but biodiverse regions and thus requiring more area or resources to achieve equivalent species conservation (El-Gabbas et al. 2020; Grand et al. 2007; Kramer-Schadt et al. 2013; Watson et al. 2011).

These challenges have prompted methodological advances to address sampling bias in SDMs (Baker et al. 2024). Key strategies include spatial and environmental filtering (Aiello-Lammens et al. 2015; Pili et al. 2025; Varela et al. 2014), target-group background sampling (Phillips et al. 2009), and model-based bias correction (Fithian et al. 2015; Warton et al. 2013). The latter two methods rely on quantitative, spatially explicit proxies of sampling effort, typically implemented as gridded surfaces that summarise observation intensity, spatial distribution, and taxonomic coverage through time (El-Gabbas and Dormann 2017; Fithian et al. 2015; Phillips et al. 2009; Warton et al. 2013). The effectiveness of bias correction using these methods depends critically on the spatial resolution, temporal alignment, and taxonomic specificity of these proxies. Bias-aware modelling approaches are motivated by the recognition that biodiversity occurrence data are shaped as much by observation processes as by ecological patterns. They assume that species belonging to the same taxonomic or ecological group experience similar sampling mechanisms because field surveys, museum collections, and citizen-science activities typically record multiple co-occurring species during efforts targeted at particular taxa, regions, or research objectives (Phillips et al. 2009). Under this framework, aggregated occurrence data from related taxa provide information on the spatial, temporal, and environmental structure of sampling effort experienced by individual species. This assumption reflects established ecological sampling practice and is consistent with recent syntheses highlighting that uncertainty in biodiversity assessments often arises from uneven, taxon-specific observation processes rather than true ecological change (Johnson et al. 2024). Explicitly accounting for these shared sampling processes enables bias-aware models to better separate observation intensity from biological signal, thereby improving the interpretability and robustness of SDMs and biodiversity indicators.

Many existing proxies of sampling effort suffer from three interrelated limitations that constrain their usefulness for

bias-aware ecological modelling. First, spatial resolution is often too coarse to support analyses at multiple scales. Available proxies range from coarse-resolution global indicators (e.g., Davis et al. 2023; Hortal et al. 2015) to fine-scale record-count maps generated ad hoc from occurrence databases (e.g., user-derived grid-cell aggregation of GBIF downloads; El-Gabbas and Dormann 2017). However, these outputs are typically produced for specific studies and are not available as consistent, standardised multi-resolution global products that enable systematic cross-scale sensitivity analyses. Second, taxonomic specificity is frequently lacking. Many global indicators pool records across taxa (Davis et al. 2023), thereby masking strong heterogeneity in sampling effort within taxonomic groups (Troudet et al. 2017). As a consequence, well-studied or charismatic taxa dominate these proxies, while cryptic or less conspicuous groups remain underrepresented (Meyer et al. 2015; Seddon et al. 2005; Troudet et al. 2017), rendering all-taxa effort surfaces potentially misleading when applied to specialist taxa. Third, temporal coverage is commonly overlooked. Cumulative effort layers conflate historical collection biases with more recent contributions from citizen-science platforms, obscuring temporal changes in sampling effort and limiting their suitability for analyses focused on recent decades, temporal trends, or dynamic species responses. Together, these limitations hinder the accurate characterisation and correction of sampling bias in biodiversity analyses (Hughes et al. 2021). In addition, sampling effort proxies based solely on observation counts conflate sampling intensity with taxonomic coverage. High record numbers may result either from intensive sampling of a few taxa or from broader coverage across many taxa, and these two situations have very different implications for bias correction and conservation prioritisation. Without spatially explicit, taxon-stratified, and temporally resolved effort layers, downstream applications, such as SDMs, trend assessments, and conservation prioritisation, risk propagating biased inferences and misallocating research and management effort.

To address these gaps, this study provides a global collection of high-resolution, taxon-stratified sampling-effort rasters designed for bias-aware ecological modelling and biodiversity assessment. These raster products quantify spatial and temporal variation in observation effort using two complementary metrics: observation count and survey species richness (unique species counts), and are available for nine major taxonomic groups and their immediate descendants (e.g., orders or classes, where applicable), revealing within-group heterogeneity that is masked by all-taxa aggregates. Annual layers spanning 1980–2025, together with cumulative surfaces, enable temporal alignment with ecological analyses and the assessment of data mobilisation trajectories. Multi-resolution outputs at approximately 1, 5, 10, and 20 km support applications ranging from local to global scales, enabling direct alignment with widely used open-access environmental variables commonly employed as predictors in macroecological analyses. These outputs also facilitate sensitivity analyses (e.g., evaluating how model outcomes change when using sampling-effort layers at different spatial resolutions or taxonomic levels). The sampling-effort rasters are generated through an accompanying open-source R workflow that is provided to ensure transparency, reproducibility, and flexibility, allowing users to reproduce, update, or tailor the products to specific spatial, temporal, or taxonomic requirements. All final rasters and programmatic access functions are publicly archived and openly accessible.

## 2 | Methods

The workflow used to create the sampling-effort rasters comprises a modular, documented R pipeline (R Core Team 2025) (available at https://github.com/elgabbas/global_sampling_efforts/), prioritising reproducibility, scalability, and robustness. All steps, from GBIF queries to final outputs, are programmatically defined. Downloaded metadata and intermediate products are persisted for replication and auditing purposes. Spatial tiling, chunk-wise streaming of large files, and parallelisation enable the processing of hundreds of millions of records. Download validation, file integrity checks, and optional purging of failed downloads ensure reliable execution. Dependence on established cross-platform R packages and on standard geospatial formats (GeoTIFF) maximises accessibility and compatibility with standard analytical tools (Kass et al. 2024). The core orchestration function, *process_efforts*, coordinates the workflow and serves as the single entry point for reproducing the results of this study. The orchestration is supported by specialised routines: *check_requests* (request validation), *read_chunk_data* (data cleaning), *process_chunk* (tile-wise rasterisation), and *merge_tiles_annual*/*merge_tiles_total* (global mosaicking).

The full global workflow processes billions of GBIF occurrence records across multiple taxonomic ranks and spatial resolutions. At global extent and high taxonomic ranks, computation requires high-performance computing (HPC) infrastructure, including large-memory nodes, parallel processing capabilities, and substantial intermediate storage. Reproducing the full pipeline, therefore, presumes advanced R programming experience, familiarity with distributed or parallel workflows, and access to institutional HPC resources. The pre-calculated sampling effort raster products provided with this study are intended to ensure accessibility for users without such infrastructure, while maintaining full methodological transparency.

### 2.1 | Taxonomic Scope

The workflow was executed for nine focal taxonomic groups: Amphibia, Arachnida, Aves (birds), Fungi, Insecta, Mammalia, Mollusca, Reptilia, and Tracheophyta (vascular plants). For each group, the GBIF taxonomic backbone key was resolved via the *rgbif* package (Chamberlain et al. 2023), which queries GBIF's taxonomic API to return the accepted 'usageKey' and hierarchy. Immediate taxonomic descendants at the next standard hierarchical rank (orders within classes, or classes within phyla, etc.) were retrieved while excluding synonyms and extinct taxa. This automated enumeration accommodates taxonomic updates and ensures consistency with the current GBIF backbone taxonomy. For Reptilia, owing to paraphyly and systematic revisions, descendants were defined as four classes (Sphenodontia, Crocodylia, Squamata, and Testudines) with hard-coded 'usageKey' values. Refer to Table 1 for the complete list of groups and their descendants.

**TABLE 1** | Summary of GBIF occurrence records and cleaned datasets.

| Group | # Observations (raw/cleaned) * 1000 | # Species | Descendants |
|---|---|---|---|
| Class Amphibia | 8530.6/7520.7 | 6125 | *3 orders*: Anura (6107.3/5366); Caudata (1407/637); Gymnophiona (6.4/122) |
| Class Arachnida | 9122.4/8119.7 | 29,736 | *16 orders*: Amblypygi (5.2/182); Araneae (5788.5/18,796); Holothyrida (0.06/5); Ixodida (439.4/309); Mesostigmata (491/1196); Opilioacarida (0.5/16); Opiliones (320.9/1815); Palpigradi (0.2/15); Pseudoscorpiones (48.2/547); Ricinulei (0.617/34); Sarcoptiformes (307.1/2225); Schizomida (3.05/115); Scorpiones (126.2/1400); Solifugae (8.2/259); Trombidiformes (578.488/2789); Uropygi (2.1/33) |
| Class Aves | 2,249,518.9/2,184,792 | 11,177 | *42 orders*: Accipitriformes (107,781.8/270); Anseriformes (180,244.3/223); Apodiformes (34,423.9/488); Apterygiformes (11.3/5); Bucerotiformes (2894.7/74); Caprimulgiformes (2897.5/110); Cariamiformes (68.7/2); Casuariiformes (131.8/4); Charadriiformes (184,412.7/397); Ciconiiformes (3847.9/21); Coliiformes (711.2/6); Columbiformes (86,196.9/339); Coraciiformes (20,235.3/185); Cuculiformes (11,909.8/153); Eurypygiformes (48.4/2); Falconiformes (20,776/64); Galliformes (17,548.7/302); Gaviiformes (5583.6/6); Gruiformes (30,308.4/172); Leptosomiformes (5.8/1); Mesitornithiformes (3.2/3); Musophagiformes (525.4/23); Nyctibiiformes (154.9/7); Opisthocomiformes (61.1/1); Otidiformes (449.5/27); Passeriformes (1,210,209/6711); Pelecaniformes (81,465.7/113); Phaethontiformes (142.7/3); Phoenicopteriformes (796.2/6); Piciformes (95,366.4/467); Podicipediformes (17,494.4/22); Procellariiformes (5393.4/147); Psittaciformes (18,077.6/387); Pteroclidiformes (215.2/16); Rheiformes (62.3/2); Sphenisciformes (1280.7/19); Steatornithiformes (12.2/1); Strigiformes (10,941/246); Struthioniformes (129.7/2); Suliformes (29,658.5/60); Tinamiformes (817.1/47); Trogoniformes (1496.8/43) |
| Kingdom Fungi | 37,733.6/33,896 | 62,494 | *11 phyla*: Ascomycota (14,745.6/36,577); Basidiomycota (17,397.8/24,803); Blastocladiomycota (15.4/37); Chytridiomycota (360.9/197); Entomophthoromycota (14/110); Glomeromycota (369.5/245); Mucoromycota (903.0/466); Neocallimastigomycota (2.3/11); Sanchytriomycota (1.3/2); Zoopagomycota (62/42); Zygomycota (24.4/4) |
| Class Insecta | 245,117.3/221,258.4 | 290,768 | *31 orders*: Archaeognatha (13.9/53); Blattodea (283.9/1757); Cnemidolestodea (0.9/0); Coleoptera (22,152.4/75,782); Dermaptera (119.2/521); Diptera (22,857.8/40,651); Embioptera (2.2/32); Ephemeroptera (1587.8/1573); Grylloblattodea (0.1/15); Hemiptera (7783.8/24,681); Hymenoptera (18,672.5/43,711); Lepidoptera (126,241.9/74,410); Mantodea (294.9/1004); Mantophasmatodea (0.05/9); Mecoptera (83.7/304); Megaloptera (117.9/236); Neuroptera (340.3/1723); Odonata (11,356.4/4568); Orthoptera (5181.5/8430); Palaeodictyoptera (0.2/0); Phasmida (59.7/992); Plecoptera (871.8/1799); Protorthoptera (0.001/0); Psocodea (280.4/1116); Raphidioptera (15.1/124); Siphonaptera (66.4/554); Strepsiptera (4.6/62); Thysanoptera (114.7/802); Trichoptera (2724/5776); Zoraptera (0.2/6); Zygentoma (30.3/77) |

(Continues)

**TABLE 1** | (Continued)

| Group | # Observations (raw/cleaned) * 1000 | # Species | Descendants |
|---|---|---|---|
| Class Mammalia | 37,911/34,131 | 5224 | *29 orders*: Afrosoricida (4.6/50); Artiodactyla (6093.7/246); Carnivora (6670.4/289); Cetacea (1845.2/96); Chiroptera (7084/1240); Cingulata (48.9/22); Dasyuromorphia (207.5/70); Dermoptera (1.7/3); Didelphimorphia (89.1/101); Diprotodontia (2286.3/119); Erinaceomorpha (714.9/25); Hyracoidea (12.6/5); Lagomorpha (1632.4/101); Macroscelidea (1/18); Microbiotheria (1.8/2); Monotremata (99.9/4); Notoryctemorphia (0.3/1); Paucituberculata (0.5/8); Peramelemorphia (241.7/18); Perissodactyla (126.5/30); Pholidota (0.4/9); Pilosa (33.5/17); Primates (188.4/398); Proboscidea (9.4/7); Rodentia (6142.5/1947); Scandentia (2.5/18); Sirenia (49.9/6); Soricomorpha (540.7/373); Tubulidentata (0.9/1) |
| Phylum Mollusca | 13,095.6/12,121 | 45,995 | *10 classes*: Bivalvia (3996.4/6921); Caudofoveata (61.4/51); Cephalopoda (532/668); Cricoconarida (0.01/0); Gastropoda (7257/37,166); Monoplacophora (0.08/16); Polyplacophora (189.4/693); Rostroconchia (0.033/0); Scaphopoda (77.5/415); Solenogastres (7.2/65) |
| Reptilia | 8221/7002.3 | 9267 | *4 classes*: Crocodylia (178.2/28); Sphenodontia (0.02/1); Squamata (5747.511/8913); Testudines (1076.6/325) |
| Phylum Tracheophyta (vascular plants) | 353,837.5/332,403.4 | 269,600 | *8 classes*: Cycadopsida (22.3/318); Ginkgoopsida (53.6/1); Gnetopsida (43.6/99); Liliopsida (69,375/56,038); Lycopodiopsida (733.8/1125); Magnoliopsida (247,681.1/201,333); Pinopsida (5001.3/710); Polypodiopsida (9492.3/9976) |
| **Total** | **2,963,088/2,841,244** | **730,386** | |

*Note:* The "Group" column lists the nine major taxonomic groups analysed in this study. The "# observations" column shows the number of raw (original) and cleaned GBIF occurrence records (in thousands); raw = original aggregated records, cleaned = records after quality-control filters. The "# species" column shows the number of species retained after cleaning. The "Descendants" columns report the number of direct taxonomic descendants, followed by each descendant's name and its cleaned observations (in thousands) and species count. All cleaned counts reflect filtering, taxonomic standardisation, and other data-cleaning procedures described in the Methods section.

The nine focal groups were selected based on a combination of global data availability within GBIF, taxonomic consistency in the GBIF backbone, and their relevance to common applications in macroecology, SDMs, and biodiversity assessment. The framework itself is not restricted to these groups; in principle, additional taxa or alternative hierarchical levels can be processed using the same pipeline, subject to sufficient data availability and appropriate computational resources. The present selection, therefore, reflects a pragmatic balance between taxonomic coverage, global coverage, and reproducibility, rather than a conceptual limitation of the workflow. Although the majority of occurrence records within these groups are terrestrial, the datasets also include aquatic and marine-associated taxa (e.g., cetaceans, seabirds, and aquatic insects), reflecting the taxonomic composition of GBIF records.

## 2.2 | Spatial Tiling and Multi-Resolution Reference Grids

To manage enormous global-scale data volumes and enable parallel processing, the Earth's surface was partitioned into uniform 7.5° × 7.5° tiles (1152 tiles globally). This tile size balances manageable download sizes with computational efficiency. Four global reference rasters were constructed at resolutions approximating ~1 km (30 arc-seconds; 21,600 rows × 43,200 columns), ~5 km (2.5 arc-minutes; 4320 × 8640), ~10 km (5 arc-minutes; 2160 × 4320), and ~20 km (10 arc-minutes; 1080 × 2160). These resolutions approximate metric distances near the equator (actual cell dimensions vary with latitude). Because the output rasters are provided on a global unprojected WGS84 latitude-longitude grid (EPSG:4326), grid-cell surface area decreases systematically with latitude. Occurrence records were assigned directly to latitude-longitude grid cells based solely on their reported coordinates, and no area normalisation or density standardisation was applied. Consequently, the resulting layers represent raw counts of records or species per grid cell, not area-standardised densities (e.g., per $km^2$). These products are therefore intended to characterise relative sampling effort and spatial bias structure, and to serve as covariates in modelling frameworks, rather than to enable direct comparison of absolute sampling intensity across latitudes. Users requiring area-normalised metrics should compute per-cell geodesic areas and derive densities accordingly, or perform aggregation within an equal-area

spatial framework appropriate to their study design. For each 7.5° tile and each resolution, the corresponding reference raster was cropped to the tile bounding box and saved as a GeoTIFF mask. These pre-generated masks acted as spatial templates for aggregating occurrence records, avoiding redundant spatial operations and ensuring consistency across tiles.

### 2.3 | GBIF Data Retrieval and Quality Filtering

For each taxonomic group × tile combination, the workflow submitted a GBIF occurrence download request via *rgbif*. Requests restricted records to those meeting quality and temporal criteria: only georeferenced presence records were requested, excluding records with known geospatial issues, fossils, and captive/living (e.g., zoo or botanical garden) records lacking wild-occurrence information. Requests were limited to records from 1980 onwards to emphasise contemporary biodiversity patterns; the download format was set to "SIMPLE_CSV" to optimise size and processing. The year 1980 was selected as a starting point because global digitisation of occurrence records increased markedly thereafter, reducing artefacts associated with sparse early records. The workflow enforces a maximum of three concurrent requests (to comply with GBIF service-level agreements), tracks status, saves request metadata, and performs downstream validation. Each request's metadata includes a unique download/citation DOI, creation timestamp, and related details. DOIs for data requests containing species records are listed in Appendix S1.

Downloaded occurrence files arrived as ZIP archives (each containing a single CSV) varying from kilobytes to multiple gigabytes (e.g., birds in Western Europe). To avoid exhausting memory, a chunk-wise streaming strategy was implemented. System-level commands (*unzip*, *cut*, *sed*, and *split*) were used to extract and decompress the CSV in a streaming pipeline (piped rather than fully written to disk). Each tile's data was streamed and split into 100,000-line chunks (a compromise between memory footprint and I/O overhead), retaining only required columns and writing intermediate files to temporary storage. This design minimised peak memory usage and enabled parallelised cleaning; chunks were read with all columns as character strings initially to avoid parsing errors.

A rigorous filter sequence was applied. Records with common institutional default uncertainty values (301, 3036, 999, 9999 m) were excluded. The decimal precision of coordinates was computed, and records with ≤1 decimal place in either coordinate were discarded, as these typically indicate coarse georeferencing. Records with reported coordinate uncertainty greater than 10 km (as provided by GBIF) were excluded to reduce spatial imprecision. The 10 km threshold was selected as a pragmatic compromise corresponding approximately to the midpoint of the four spatial resolutions considered, ensuring that retained records did not exceed the coarsest grid cell size. Records with missing uncertainty values were retained to avoid disproportionate exclusion of historical museum specimens. Records with missing/invalid coordinates or equal latitude and longitude were removed. The 'CoordinateCleaner' R package (Zizka et al. 2019) was applied to flag further spatial anomalies, including proximity to known geographic features and patterns indicative of data entry errors. Records near country or province centroids, capital cities, biodiversity institutions, GBIF headquarters, and equal latitude-longitude pairs were excluded. Cleaned records were aggregated by descendant and saved as compressed data frames. This chunk-wise workflow ensured the production of high-quality, spatially explicit records while maintaining computational feasibility.

### 2.4 | Rasterisation and Global Mosaicking

For each tile and for each descendant taxon with cleaned records, spatial aggregation onto tile-specific masks was performed at four spatial resolutions. Occurrence records were assigned to grid cells solely based on their reported geographic coordinates, without spatial propagation or buffering. This means that for all remaining records, the reported coordinates were used directly regardless of the target spatial resolution. The *process_chunk()* function computed two metrics: observation count (record count per grid cell) and (survey) species richness (unique GBIF 'speciesKey' counts per grid cell, excluding non-species-level identifications). Species richness was included to distinguish grid cells with high observation count, driven by repeated sampling of a few taxa, from cells with broader taxonomic coverage. For each metric, cumulative and annual rasters (1980 to present; processed October 2025) were generated and saved as compressed GeoTIFFs.

Per-tile rasters were mosaicked into seamless global rasters for annual and cumulative outputs (overall total), for each metric and resolution. Cumulative rasters suit analyses spanning full periods (i.e., covering a temporal range of 1980–2025), aggregating all data for comprehensive bias proxies. Mosaicking utilised the *merge_tiles_annual()* and *merge_tiles_total()* functions and was parallelised across descendant–resolution–metric combinations. The workflow employed Virtual Rasters (VRTs) to reference per-tile GeoTIFFs without loading full rasters into memory. Cells with no retained occurrence records were assigned a value of zero, indicating the absence of GBIF-accessible records for the respective taxonomic group within that grid cell. Importantly, these values represent sampling outcomes rather than ecological absence or species non-occurrence. Global mosaics were written using high (ZSTD) compression to save disk space. The mosaicking workflow first produced global rasters at the descendant level. Descendant-level mosaics for each metric-resolution-year combination were then summed to obtain group-level mosaics for each focal taxon. The final outputs consisted of GeoTIFFs at both the descendant and group levels for each metric, resolution, and for both annual and cumulative time slices.

For each taxonomic group, GeoTIFF files were organised into eight resolution- and metric-specific subdirectories: "*res_<resolution>_<metric>*", where *resolution* is one of 1, 5, 10, and 20; and *metric* is either "*n_obs*" or "*n_sp*", for observation count and species richness, respectively. In each of these directories, file names follow the naming convention: "*<metric>_<group>_<time>_res_<resolution>.tif*", where *group* is either the name of the main taxonomic group or its descendant (such as Insecta or Coleoptera; see Table 1), and *time* is either the year (1981–2025) or "total" for cumulative outputs.

### 2.5 | Data Availability and Programmatic Access

Final raster products for all nine focal groups are archived on the Open Science Framework (OSF; http://osf.io/) and Zenodo (https://zenodo.org/) under the Creative Commons 4.0 licence, ensuring maximum reusability. Each group's archive contains all annual and cumulative rasters for both metrics at all four resolutions. Programmatic access is provided via the *ecokit* R package (El-Gabbas 2025), available on GitHub (https://github.com/elgabbas/ecokit), which exposes a *get_sampling_effort()* function that enables users to download and cache the rasters on demand for seamless integration into their workflows. For example, the following R code snippet demonstrates how to access a specific raster programmatically (the 2020 "Coleoptera" species count raster at 5 km resolution). Further usage examples illustrating how to retrieve and integrate the archived raster products are documented in the README of the companion GitHub repository (https://github.com/elgabbas/global_sampling_efforts/).

```
library(ecokit)
# Download and load the 2020 Coleoptera species count raster at ~5 km resolution
effort_raster <- ecokit::get_sampling_effort(
    group = "insecta", descendants = "coleoptera", metric = "n_sp",
    years = 2020, resolution = 5)
```

## 3 | Results

The workflow processed almost three billion raw GBIF records. After rigorous cleaning, ~2.84 billion records (80% of the total raw data) representing over 730,000 species remained (see Table 1 for a summary and Appendix S1 for the complete list of citations of the processed data). These comprise ~41% of all scientific names with observations available on GBIF as of November 2025 (3.56 billion observations; 1.75 million taxa[1]). Applied to nine major taxonomic groups (Table 1), the workflow produced annual (1980–2025) and cumulative sampling-effort rasters for observation count and species richness. For each group, global mosaics were generated at four spatial resolutions (~1, 5, 10, and 20 km) for both group-level and descendant-level taxa (e.g., orders within Insecta). All rasters are stored as compressed GeoTIFFs (EPSG:4326) with explicit zero values for unsampled cells and are archived with programmatic access via the *ecokit* R package (El-Gabbas 2025; see Methods).

Observation count is concentrated in temperate regions, including Western Europe, mid-North America (mainly the USA), southeastern Australia, South Africa, and southwest India, with smaller hotspots in coastal South America (Figure 1). Recorded species richness peaks in Western Europe and, to a lesser extent, the United States. Conversely, tropical regions (the Amazon and Congo basins, New Guinea, and much of Asia), the Middle East, and remote polar/high-altitude areas (the Arctic and Tibetan Plateau) remain sparsely sampled (Figure 1). High observation count does not necessarily coincide with high taxonomic coverage (i.e., the number and diversity of taxa represented), underscoring the value of both metrics for bias correction (Figures 1, 2 and S1). Patterns are qualitatively similar across spatial resolutions but differ in intensity (Figures S2 and S3): finer resolutions capture strong clustering around urban centres, whilst coarser grids smooth hotspots and increase apparent coverage in sparse regions. At coarser resolutions, country borders become visible, reflecting national variation in data sharing to GBIF. This scale dependence highlights the importance of aligning the spatial resolution of effort layers with that of environmental predictors in downstream analyses.

To quantify the spatial concentration of sampling effort, grid cells were ranked by cumulative record count and the proportion of surface area containing the top 95% and bottom 5% of all records at each spatial resolution and taxonomic group was calculated. Given the predominantly terrestrial nature of the data, land-area–restricted statistics provide a more informative representation for most applications. Across all taxa, 95% of cumulative GBIF records occupy only 0.33% of the Earth's surface at 1 km resolution, corresponding to 0.93% of the terrestrial land area. In contrast, the lowest 5% of records cover only 1.77% of the globe (3.88% of land). Consequently, approximately 98% of the Earth's surface (95% of land) remains unsampled (Table 2; Figures 3 and 4; Appendix S2). The extreme concentration of observations is consistent across all nine major taxa (Table 2; Figures 3 and 4; Appendix S2). With coarser resolution, apparent spatial coverage increases slightly (at 20 km resolution, 95% of records fall within 2% of Earth's surface; 5.6% of land), but even then, large parts of the planet remain devoid of biodiversity observations.

Although taxonomic groups show variation in spatial distribution, overall patterns are broadly consistent, reflecting a common global geography of sampling effort (Figures 2 and S1; Appendix S2). However, pronounced discrepancies exist between groups and between observation count and species richness (Figure 2). Birds account for more than 2.18 billion records (~77% of the dataset), producing spatial patterns that closely resemble full-data maps (Figure 1 and Appendix S2). Passeriformes dominate bird data, with ~1.21 billion records (~42% of total records) for only 6711 species (0.9% of total species; Figure 2 and Table 1). In contrast, highly diverse groups such as insects and vascular plants contribute only 221.26 and 332.4 million records, respectively, despite each exceeding 260,000 species (Table 1 and Figure 2). Excluding Passeriformes does not alter overall observation count patterns (Figures 1 and S4), indicating that sampling inequality reflects structural biases in observer accessibility, research focus, and data mobilisation shared across taxa rather than a single dominant lineage.

Annual trends (complete years, 1980–2024) show steady increases during the 1990s and early 2000s, followed by a sharp acceleration after the mid-2000s (Figures 5 and S5). Growth pace and magnitude vary among taxa. Bird records increased exponentially after 2000 (driven by Passeriformes), reaching ~25 million/year by 2010, ~175 million by 2020, and peaking at > 270 million in 2024. Passeriformes alone contributed ~25 million annual records from 2012, 50 million by 2016, and more than 100 million after 2020, peaking at around 151.82 million in 2024. The number of Passeriformes records (6711 species) in 2024 alone (151.82 million) represents > 68% of the total number of observations of all insects over almost 45 years (221.25 million records; 290,768 species). Vascular plants showed steady early increases, marked expansion during the 2000s, and a sharp

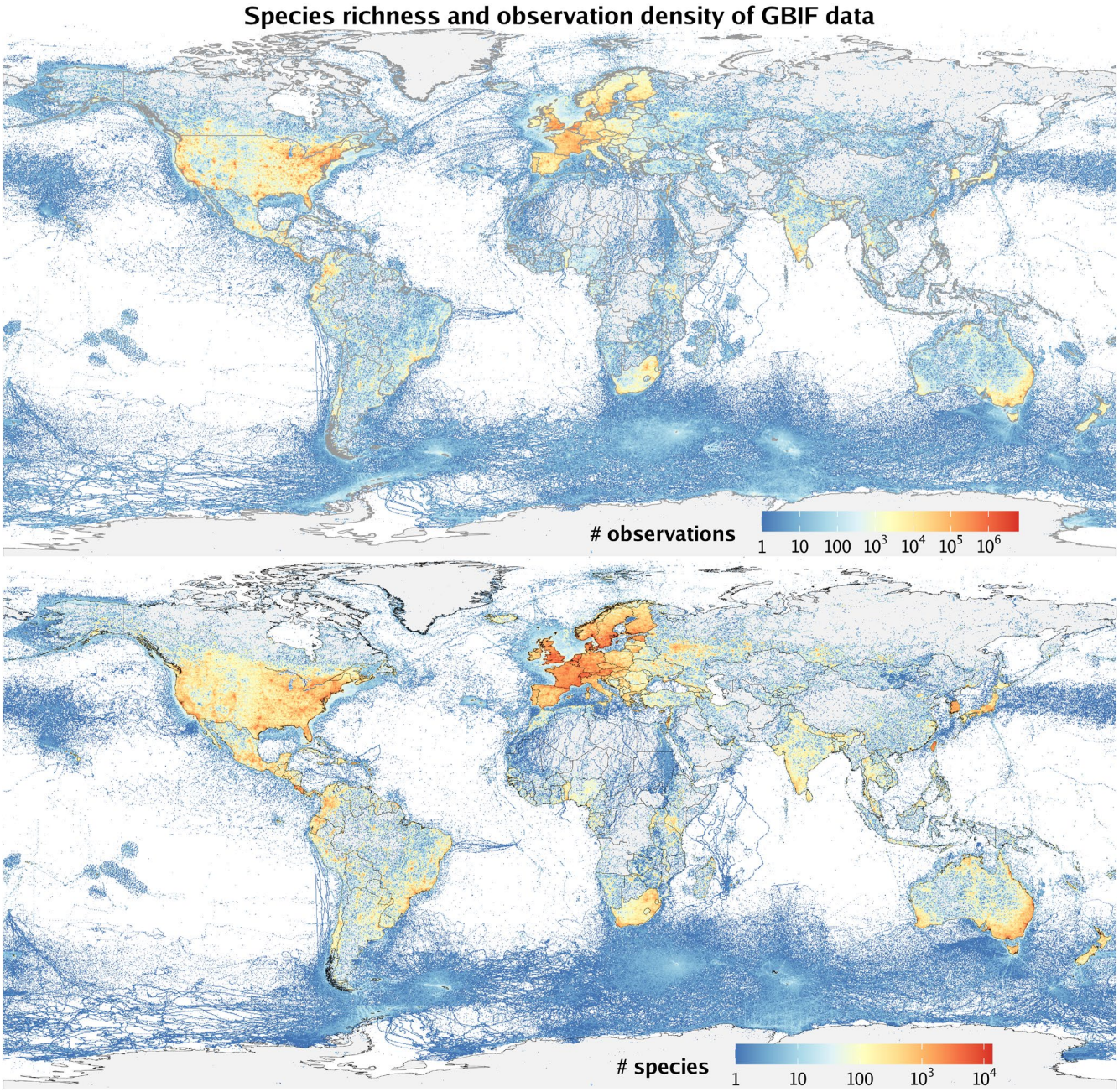


**FIGURE 1** | Global patterns of observation count (top) and species richness (bottom panel) across all cleaned GBIF data at ~20 km resolution (both at $\log_{10}$ scale). Warmer colours indicate higher values. Observation count is concentrated in temperate regions (Western Europe, North America, Australia), while extensive tropical areas remain sparsely sampled. Grey areas indicate no records. See Appendix S2 for taxon-specific maps.

recent surge. Among insects, Lepidoptera rose steeply from the 1990s, peaking in 2019 before declining slightly. Reptilian Squamata maintained high recording rates since 2010. Species richness trajectories generally mirror record trends but show steeper increases for insects and plants, indicating broader taxonomic mobilisation. Fungi and molluscs accumulated data more slowly, consistent with persistent identification challenges.

## 4 | Discussion

Sampling bias presents significant challenges for ecological modelling with opportunistic, presence-only records (Hughes et al. 2021; Merow et al. 2016; Phillips et al. 2009). This study addresses these challenges by converting raw GBIF records into taxon-stratified, temporally explicit, multi-resolution rasters directly usable for ecological analyses. Key strengths include scalability (processing billions of records into compact products covering the period 1980–2025), taxonomic stratification that exposes within-group heterogeneity, dual metrics that separate recording intensity from taxonomic coverage, and multi-resolution outputs for cross-scale sensitivity analyses. Existing global proxies of sampling effort are typically generated ad hoc for individual studies or restricted to narrow taxonomic scopes. Where global effort maps exist (e.g., Davis et al. 2023), they are static, low-resolution, temporally limited, and often


**FIGURE 2** | Taxonomic disparities in GBIF sampling effort. Relationship between number of species (thousands; *x*-axis) and total observations (millions; *y*-axis) across taxonomic groups. Each major taxonomic group is colour-coded with plus symbols indicating group-level totals and filled circles showing direct descendant taxa (not constrained to a uniform taxonomic rank across groups). The inset panel provides a magnified view of the boxed region for lower-observation taxa (bottom-left of the figure). Birds dominate observations (~77% of records) with Passeriformes alone contributing ~1.2 billion records for only 6711 species. In contrast, insects and vascular plants each exceed 260,000 species, but contribute only ~221 million and 333 million observations, respectively. See Table 1 for detailed counts.

non-reproducible. In contrast, the approach developed here provides a unified, fully scripted framework that generates consistent effort layers across taxa, time periods, and spatial scales. Public archival and programmatic access enhances reuse and reproducibility.

The global, taxon-stratified sampling-effort raster dataset generated in this study reveals pronounced spatial and temporal inequities that surpass earlier global assessments, consistent with known taxonomic and spatiotemporal biases in conservation research and biodiversity data (Di Marco et al. 2017; Mair et al. 2018). Previous studies have documented clustering of biodiversity records in temperate, economically developed regions (e.g., Hughes et al. 2021; Meyer et al. 2016). The present analysis corroborates these patterns while providing a more explicit and reproducible quantification across spatial resolutions and taxonomic groups. Sampling effort is extremely spatially concentrated. At 1 km resolution, less than 1% of Earth's surface (0.33% including oceans; 0.93% of land) holds 95% of all GBIF records, whilst the bottom 5% of records occupy < 2% of the planet (1.77% including oceans; 3.88% of land). Even at 20 km resolution, the spatial concentration of records remains extremely high. 95% of records are located within only ~2% of the global surface area (equivalent to 5.6% of land). Consequently, the majority of grid cells remain unsampled (67% when including oceans and 59.6% of land cells), with unsampled areas predominantly occurring in tropical, arid, and politically unstable regions (Table 2; Figures 3 and 4). This spatial inequality is evident across all major taxonomic groups, though its severity varies (Figure 4). Coverage for vascular plants and mammals is marginally better than for birds, and reptiles show slightly broader coverage at coarser (10–20 km) resolutions. Such heterogeneity underscores that sampling bias is taxon-specific and cannot be adequately

**TABLE 2** | Spatial concentration and coverage gaps in global GBIF biodiversity data.

| | Cumulative top 95% | | | | Cumulative lowest 5% | | | | % Uncovered | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Group** | **~1 km** | **~5 km** | **~10 km** | **~20 km** | **~1 km** | **~5 km** | **~10 km** | **~20 km** | **~1 km** | **~5 km** | **~10 km** | **~20 km** |
| All | 0.33 | 1.05 | 1.48 | 2.01 | 1.77 | 10.08 | 18.80 | 30.95 | 97.91 | 88.87 | 79.72 | 67.05 |
| | 0.93 | 2.98 | 4.20 | 5.60 | 3.88 | 15.45 | 24.38 | 34.83 | 95.19 | 81.58 | 71.42 | 59.56 |
| Amphibia | 0.07 | 0.48 | 0.81 | 1.20 | 0.04 | 0.60 | 1.41 | 2.90 | 99.89 | 98.92 | 97.79 | 95.89 |
| | 0.22 | 1.37 | 2.31 | 3.41 | 0.12 | 1.72 | 4.02 | 8.15 | 99.67 | 96.90 | 93.67 | 88.45 |
| Arachnida | 0.06 | 0.37 | 0.67 | 1.09 | 0.04 | 0.62 | 1.47 | 3.07 | 99.90 | 99.01 | 97.86 | 95.83 |
| | 0.17 | 1.07 | 1.91 | 3.09 | 0.13 | 1.77 | 4.18 | 8.54 | 99.70 | 97.16 | 93.90 | 88.37 |
| Aves | 0.18 | 0.74 | 1.15 | 1.65 | 1.02 | 6.29 | 12.24 | 21.78 | 98.81 | 92.97 | 86.61 | 76.56 |
| | 0.50 | 2.11 | 3.26 | 4.62 | 2.36 | 11.27 | 18.50 | 27.65 | 97.14 | 86.62 | 78.24 | 67.73 |
| Fungi | 0.06 | 0.32 | 0.53 | 0.82 | 0.11 | 0.94 | 1.91 | 3.65 | 99.82 | 98.74 | 97.56 | 95.53 |
| | 0.18 | 0.91 | 1.51 | 2.28 | 0.33 | 2.68 | 5.37 | 9.93 | 99.49 | 96.41 | 93.12 | 87.78 |
| Insecta | 0.10 | 0.44 | 0.70 | 1.02 | 0.35 | 2.12 | 4.04 | 7.22 | 99.55 | 97.44 | 95.26 | 91.77 |
| | 0.28 | 1.27 | 2.00 | 2.87 | 1.00 | 5.92 | 10.97 | 18.57 | 98.72 | 92.81 | 87.04 | 78.56 |
| Mammalia | 0.28 | 1.16 | 1.69 | 2.34 | 0.18 | 2.96 | 6.76 | 12.67 | 99.54 | 95.88 | 91.55 | 84.99 |
| | 0.64 | 2.53 | 3.70 | 5.09 | 0.34 | 2.76 | 6.03 | 11.62 | 99.24 | 94.85 | 90.34 | 83.30 |
| Mollusca | 0.06 | 0.42 | 0.78 | 1.28 | 0.06 | 0.73 | 1.83 | 4.10 | 99.89 | 98.85 | 97.39 | 94.62 |
| | 0.13 | 0.90 | 1.64 | 2.59 | 0.13 | 1.42 | 3.23 | 6.60 | 99.75 | 97.68 | 95.13 | 90.81 |
| Reptilia | 0.13 | 0.84 | 1.43 | 2.17 | 0.04 | 0.80 | 2.15 | 4.78 | 99.83 | 98.36 | 96.41 | 93.05 |
| | 0.34 | 2.24 | 3.83 | 5.75 | 0.10 | 1.69 | 4.21 | 8.60 | 99.57 | 96.09 | 91.98 | 85.65 |
| Tracheophyta | 0.31 | 0.97 | 1.36 | 1.81 | 0.63 | 3.37 | 5.95 | 9.61 | 99.06 | 95.66 | 92.69 | 88.58 |
| | 0.89 | 2.78 | 3.89 | 5.09 | 1.82 | 9.53 | 16.51 | 25.72 | 97.28 | 87.68 | 79.60 | 69.19 |

*Note:* Percentages of the Earth's surface represented by the cumulative top 95% and bottom 5% of cleaned GBIF occurrence records, together with the proportion of grid cells lacking any record ("Uncovered"), are shown for four spatial resolutions (~1, 5, 10, and 20 km). For each taxonomic group and for the overall dataset, the first row reports values relative to the Earth's surface (land + marine). The second row (shaded in grey) reports values restricted to terrestrial cells only. For predominantly terrestrial taxa, land-restricted statistics provide a more informative reference for interpreting spatial sampling concentration.

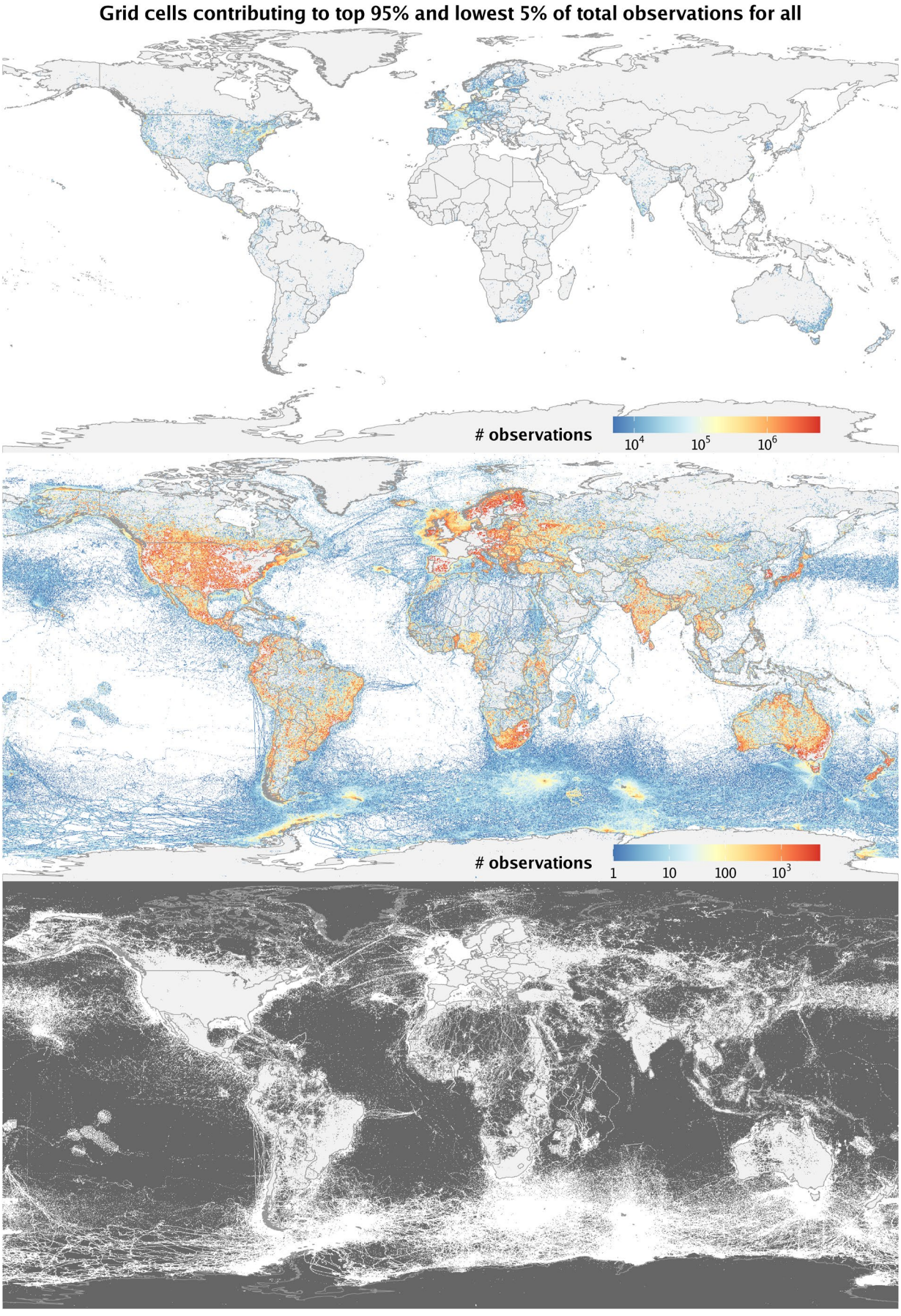


**FIGURE 3** | Legend on next page.

**FIGURE 3** | Spatial concentration and gaps in global sampling effort at 20 km resolution. Top panel: grid cells containing the cumulative top 95% of all observations ($\log_{10}$ scale); warmer colours indicate higher observation count. Middle panel: grid cells containing the cumulative bottom 5% of observations ($\log_{10}$ scale), revealing marginal coverage in data-sparse regions. Bottom panel: unsampled areas (dark grey) with zero observations; white areas indicate sampled grid cells. These three panels together show that, at the coarsest resolution considered (20 km), the majority of Earth's surface lacks any cleaned GBIF records. See Figure 4 and Table 2 for values at different resolutions for all taxonomic groups. See Appendix S2 for taxon-specific maps.

represented by a single all-taxa proxy. The findings corroborate earlier global syntheses (Hughes et al. 2021) that estimated < 7% coverage of GBIF and the Ocean Biogeographic Information System (OBIS; https://obis.org/) at 5 km resolution, but extend these by providing a continuous, reproducible quantification across scales. The consequences of this strong spatial skew are substantial. Analyses relying on raw occurrence data without sampling bias correction will overrepresent heavily sampled regions, artificially inflating apparent richness and environmental niche, while underestimating biodiversity in data-poor areas. Beyond technical sampling bias correction, these maps supply an empirical basis for equitable global biodiversity monitoring, shifting focus from where data are abundant to where biodiversity knowledge is most deficient (Johnson et al. 2024).

Taxonomic disparities are equally pronounced. Birds alone contribute > 2.18 billion records (~77% of all observations; representing only 11,177 species), whereas insects and vascular plants, each comprising > 260,000 species, contribute only a few hundred million records each (Table 1 and Figure 2). GBIF data show a global average of ~4300 records per bird species per year, compared to only 17 and 27 per insect and plant species, respectively; a taxonomic imbalance consistently documented in previous studies (Gaiji et al. 2013; Meyer et al. 2016; Troudet et al. 2017). Recent increases in biodiversity data availability have been shown to disproportionately benefit already heavily sampled taxa, thereby reinforcing existing disparities, particularly for birds (Hughes et al. 2021; Troudet et al. 2017). Pooling records across taxa, even within the same high-level taxonomic group, overlooks heterogeneity in sampling effort and masks group-specific gaps. Broad groupings (e.g., Mammalia or Insecta) may combine species that are surveyed using fundamentally different observation methods and sampling infrastructures. For example, mammals may include species recorded through acoustic monitoring (e.g., bats), camera trapping, hunting records, roadkill surveys, or opportunistic sightings. Such aggregated data should therefore be interpreted as a coarse proxy for overall sampling effort rather than a direct representation of survey intensity for individual species. Where finer methodological alignment is required, users are encouraged to rely on lower taxonomic levels (e.g., analysing descendant-level groups such as orders or families rather than aggregating at broad class-level categories) or to generate custom bias layers tailored to their focal taxa and study design. This highlights the need for targeted, taxon-specific approaches (Hughes et al. 2021), as pooled indicators risk perpetuating the dominance of overrepresented taxa and may misinform biodiversity indicators.

Temporal trajectories reinforce these contrasts. Record mobilisation remained modest throughout the 1980s and 1990s, then increased sharply after the mid-2000s with the advent of large-scale digitisation and citizen-science platforms (Heberling et al. 2021; Lajeunesse and Fourcade 2023; Nelson and Ellis 2018; Troudet et al. 2017). However, this expansion is far from uniform: most records since 2010 stem from a few taxa (notably Passeriformes; Figures 5 and S5) and a few heavily sampled regions (Europe, North America, and Australia). Tropical regions, despite harbouring some of the highest levels of global species richness, remain severely underrepresented in biodiversity databases (Collen et al. 2008; Hughes et al. 2021; Meyer et al. 2015, 2016). In addition, geographically remote and logistically inaccessible areas (whether tropical, arid, polar, or mountainous) also show pronounced data deficits (Christie et al. 2021; Kadmon et al. 2004; Meyer et al. 2015). These imbalances reflect not only practical constraints related to accessibility and infrastructure, but also deeper structural drivers. Historical asymmetries in scientific exploration and colonial-era collection practices concentrated museum holdings, taxonomic expertise, and research institutions in Europe and North America, shaping long-term patterns of specimen curation, digitisation, and data mobilisation (Faxon and Chapman 2025; Nelson and Ellis 2018). Contemporary disparities in research funding, institutional capacity, and access to stable internet and mobile infrastructures further limit large-scale biodiversity monitoring and participation in digital citizen-science platforms in many biodiversity-rich countries (Amano and Sutherland 2013; Di Marco et al. 2017). Sociopolitical instability and the absence of sustained monitoring programmes can additionally constrain systematic data mobilisation and open-data publication (Amano and Sutherland 2013; Christie et al. 2021). Together, these factors sustain persistent geographic inequities in global biodiversity archives. Data growth alone, especially when concentrated in already heavily sampled regions and taxa, does not ensure representativeness (Hughes et al. 2021; Troudet et al. 2017). Addressing these imbalances will require targeted investment in local research capacity, digitisation of regional collections, strengthened data infrastructures, and equitable international partnerships prioritising biodiversity-rich but data-poor regions.

### 4.1 | Applications in Bias-Aware Modelling and Other Ecological Uses

The generated multi-resolution sampling effort grids directly support several bias-correction methods in SDMs by providing the necessary spatially and temporally explicit proxies. The pre-generated rasters are intended for users seeking immediate, standardised bias layers, whereas the underlying workflow enables expert users to generate custom rasters tailored to specific taxa, spatial extents, or temporal windows. The selection of the optimal metric depends on the taxon and context; observation count captures raw sampling intensity, whereas species richness

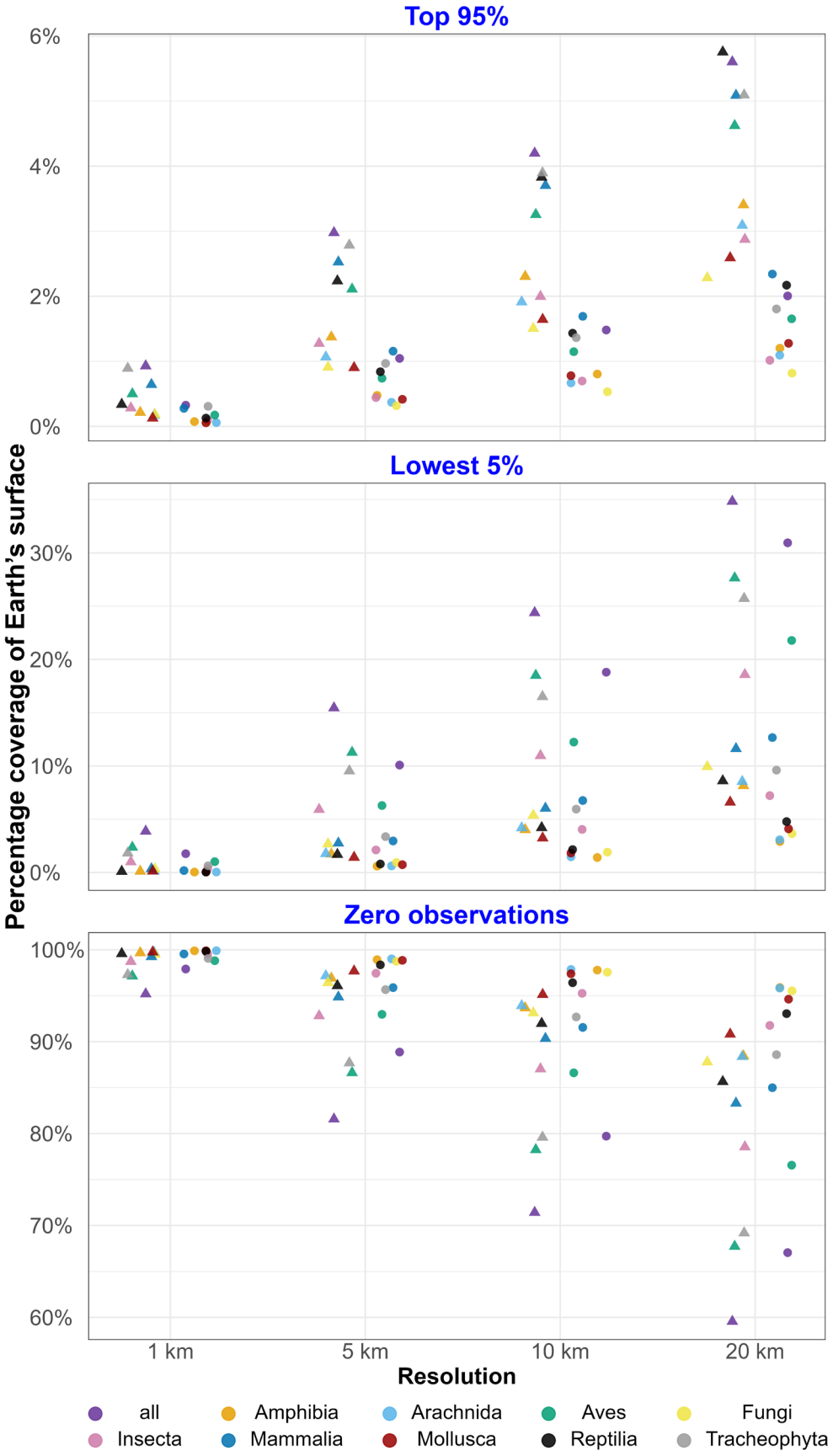


**FIGURE 4** | Legend on next page.

**FIGURE 4** | Spatial concentration and knowledge gaps in global biodiversity data across taxonomic groups and spatial resolutions. Each panel displays percentage coverage of Earth's surface at four spatial resolutions (~1, 5, 10, and 20 km; *x*-axis). Top panel: percentage of surface area containing the cumulative top 95% of all observations (indicating spatial concentration of sampling effort). Middle panel: percentage of surface area containing the cumulative bottom 5% of observations (revealing marginal coverage in data-sparse regions). Bottom panel: percentage of surface area with zero observations (unsampled areas). Filled circles represent values for the entire globe (land and ocean), whilst triangles represent terrestrial areas only. Colours distinguish the nine major taxonomic groups and all data combined (see legend). Points are horizontally jittered to prevent overlapping. Note that high percentages in the top panel and low percentages in the middle panel both indicate poor spatial coverage, whilst high percentages in the bottom panel indicate extensive knowledge gaps. All taxonomic groups show similar patterns of extreme spatial concentration, with 95% of records occupying < 3% of Earth's surface at all resolutions (< 6% for land; top panel), and a large proportion of the globe remaining unsampled (bottom panel). See Table 2 for exact values.


**FIGURE 5** | Temporal trends in GBIF data mobilisation (1980–2024; $\log_{10}$ scale). (a) Annual observations (left panel) and species (right panel) count for nine major taxonomic groups (colour-coded). (b) Annual observations (top row) and species (bottom row) count for descendant-level taxa within each group (separate columns). Each line represents a different descendant taxon. Bird records increased exponentially after 2000, reaching over 270 million in 2024, driven primarily by the Passeriformes. Tracheophyta (vascular plants) and insects exhibit steeper increases in species richness, indicating broader taxonomic mobilisation within these groups. See Figure S5 for the linear-scale version.

reflects taxonomic coverage. Both metrics provide complementary insights into the bias structure, distinguishing between intensive, narrow efforts and broader biodiversity inventories.

Presence-background SDMs, such as MaxEnt (Phillips et al. 2006), contrast presence locations with background information, and so are sensitive to background sampling (Merow et al. 2013; Syfert et al. 2013). Two established tactics are to provide a bias file (rasters weighting background sampling by effort) or to restrict background to a "target-group" that is expected to share sampling processes with the focal species. Bias files (or bias grids) draw background points proportional to sampling effort, reducing spurious environmental contrasts from spatial clustering of records (Elith et al. 2011; Phillips and Dudík 2008; Syfert et al. 2013). The observation count layers presented here provide ready-made, taxon- and year-matched bias files. Users should select descendant-level rasters that best approximate the sampling processes affecting the focal species. No single taxonomic rank is prescribed: instead, users should choose the *lowest taxonomic level with sufficient data coverage* and broadly comparable observation practices (commonly family or order, or higher ranks for sparsely sampled taxa). As with any bias correction, inappropriate taxonomic aggregation (e.g., combining taxa with fundamentally different sampling methods) can reduce effectiveness, and users should evaluate taxonomic coherence and data sufficiency when selecting bias layers.

Target-group background uses occurrence records from taxa that are assumed to share the same sampling process as the focal species (e.g., the same bird family or insect order), restricting background sampling to areas where such records occur (Phillips et al. 2009). This approach aims to equalise environmental bias between presence and background data, thereby reducing the influence of uneven sampling effort and improving model transferability, but requires sufficient target-group data and assumes similar detectability and observer communities (Phillips et al. 2009). The descendant-level sampling-effort rasters presented here provide a direct and scalable way to operationalise this principle. Each raster encodes the spatial footprint of sampling for a candidate target group; grid cells with values greater than zero represent locations where at least one occurrence record from that group has been reported, whereas cells with zero values indicate no recorded sampling effort. For target-group background implementation, users can restrict background point generation to grid cells with non-zero effort for the selected descendant group, thereby aligning the background domain with the realised sampling distribution of the target group. When selecting an appropriate target group, users should balance two constraints: ecological and observational similarity to focal species, including detectability and observer community, and sufficient record density to ensure adequate spatial coverage. As a practical guideline, groups with several thousand records typically provide stable coverage at moderate resolutions; however, sensitivity analyses across descendant levels (e.g., family, order, and class) are recommended to evaluate robustness and optimise model performance.

Bias covariates represent model-based approaches that explicitly incorporate spatial layers quantifying sampling effort into prediction models, and directly model the observation process (Warton et al. 2013). Covariates describing sampling efforts or site accessibility (e.g., distance to roads, cities, or protected areas) are included alongside environmental predictors. For unbiased predictions, bias covariates are set to a common level (e.g., zero distance or optimum effort) across the entire study area to remove the influence of bias (El-Gabbas and Dormann 2017; Warton et al. 2013). This offers flexibility and interpretability, but it depends critically on the quality of the proxy, the selection of adjustment values, and can fail if bias covariates strongly correlate with the environmental predictors (El-Gabbas and Dormann 2017; Fithian et al. 2015). The sampling effort grids presented here can be used as sampling bias covariates, by incorporating observation count or species richness layers as predictor variables in the models. When selecting adjustment levels, it is recommended to avoid extreme values (e.g., maximum effort) that rarely represent typical conditions and distort predictions. Instead, use representative values from the calibration region (e.g., median or 75th percentile). Because effort values span orders of magnitude, log transformation of bias covariates stabilises variance, reduces the influence of extremes, and improves convergence.

When applying the sampling effort grids in downstream analyses, users should ensure that the spatial resolution of the selected raster matches that of the environmental predictors used in the SDM. Fine resolutions (e.g., ~1 km) are appropriate for local to regional analyses and for taxa with dense sampling, whereas coarser resolutions (e.g., ~10–20 km) are generally more robust for sparsely sampled taxa, taxa with high dispersal ability or large home ranges, broad-scale assessments, or analyses covering large extents. Temporal alignment is equally important. Raster years should be matched to the focal study period (e.g., recent years for contemporary SDMs). When modelling historical distributions, trends, or long-lived species, users should consider potential time lags between sampling effort, environmental change, and species responses, as highlighted by Essl et al. (2024).

Beyond SDMs, the stratified observation count and species richness layers are versatile tools for macroecology, conservation planning, and biodiversity monitoring (Johnson et al. 2024). These layers enable visualisation and quantification of where biodiversity information is dense, sparse, or absent, thus enabling objective assessments of spatial completeness and guiding efficient sampling prioritisation. Species richness and observation count rasters, combined, provide a two-dimensional view of sampling completeness, identifying priority inventory areas and guiding resource allocation (Petersen et al. 2021). However, low observation count may reflect genuine ecological differences among environments, as some land-cover types are intrinsically species-poor, and should not be interpreted solely as sampling bias (Petersen et al. 2021). The percentage-based spatial coverage analyses (Figures 3 and 4; Table 2; Appendix S2) and descendant-level rasters (Appendix S2) provide a quantitative measure of ignorance, showing where sampling effort falls below completeness thresholds for each group. Integrating both metrics into planning tools like Zonation (Lehtomäki and Moilanen 2013), e.g., as ignorance/uncertainty layers or cost layers, can help identify under-surveyed high-potential areas that warrant future exploration. Bias-free SDM predictions used in conservation planning can further help maximise species representation and persistence with limited resources (Carvalho

et al. 2017; El-Gabbas et al. 2020; Hortal et al. 2015; Watson et al. 2011).

For conservation monitoring, sampling effort layers quantify the representativeness of protected-area networks or ecological observatories. Overlaying observation count rasters with protected-area boundaries helps identify data-rich reserves (due to accessibility, data sharing, or research institutions) versus under-documented ones, despite their conservation importance. Monitoring agencies can prioritise sampling in underrepresented habitats to improve coverage and trend detection. Multiple resolutions (~1, 5, 10, and 20 km) and time steps enable scale matching to decision contexts: finer resolutions for site-level monitoring, and coarser resolutions for national or continental prioritisation. Time-stratified rasters allow explicit accounting for evolving data gaps, guiding adaptive monitoring and sampling campaigns.

In addition to conservation planning and monitoring applications, the sampling effort grids presented here can also support citizen-science initiatives and coordinated biodiversity monitoring efforts. By explicitly identifying spatial, temporal, and taxonomic gaps in occurrence data, these products can guide targeted data-collection campaigns, prioritise under-sampled regions or taxa, and inform adaptive sampling strategies. In particular, gap maps derived from the sampling effort grids may help design targeted recommendations for contributors, support feedback mechanisms that highlight underrepresented areas, or inform gamification approaches aimed at improving data coverage and balance. Such applications may improve the efficiency and balance of biodiversity data mobilisation without requiring changes to existing citizen-science platforms or workflows.

### 4.2 | Adherence to Open-Science Principles

Although the primary contribution of this study is the publicly available sampling effort raster products, the accompanying workflow is provided to ensure transparency, reproducibility, and long-term extensibility. The workflow and data products adhere to open-science and FAIR principles (findable, accessible, interoperable, and reusable) (Wilkinson et al. 2016). Each processing step is fully scripted in R (https://github.com/elgabbas/global_sampling_efforts/), ensuring a transparent and reproducible transformation from raw GBIF retrieval to final rasters. The modular design facilitates reuse and adaptation for alternative taxa, temporal windows, or resolutions without reliance on proprietary software. The reproducible pipeline allows periodic re-execution to generate updated layers in response to the continued expansion of GBIF data (Heberling et al. 2021; Hobern et al. 2019), as well as ongoing retrospective georeferencing and addition of historical records (Marcer et al. 2022).

All final raster products, code, and metadata are archived in open repositories with persistent DOIs, ensuring long-term citability and traceability. Creative Commons Attribution 4.0 licensing maximises reusability, aligning with GBIF's open-data policy and international commitments. These practices lower entry barriers for researchers in data-limited regions, enabling equitable participation in global biodiversity research and capacity building.

### 4.3 | Limitations and Future Directions

Although extensive and globally standardised, the sampling-effort rasters inevitably inherit the spatiotemporal and taxonomic limitations of GBIF's underlying occurrence data. Approximately 80% of all GBIF records are represented, yet coverage remains uneven. Cryptic, rare, and poorly surveyed taxa are underrepresented, whilst tropical and subtropical regions remain critically data-deficient, reflecting both logistical constraints and longstanding structural inequities in global research capacity, data infrastructure, and digitisation efforts.

Addressing these imbalances requires continued mobilisation of biodiversity data, particularly from under-sampled tropical regions and historically underrepresented taxa. Many occurrence records are curated within national or thematic databases and infrastructures, including Observation.org (https://observation.org/), OBIS for marine taxa, and regional aggregators such as the Atlas of Living Australia (https://www.ala.org.au/), which maintain substantial open-access biodiversity repositories. While GBIF already indexes contributions from several major platforms, including eBird (https://ebird.org/) and iNaturalist (https://inaturalist.org/), data exchange and harmonisation across biodiversity infrastructures remain uneven and taxon- or region-specific. Expanded interoperability with additional national inventories, regional monitoring programmes, and ongoing museum and herbarium digitisation initiatives would substantially enhance temporal continuity, geographic representativeness, and taxonomic completeness (Blades et al. 2025; de Araujo et al. 2022; Nelson and Ellis 2018). Strengthening interoperability would foster a more balanced global evidence base for biodiversity monitoring and modelling.

Extending coverage beyond predominantly terrestrial datasets is another important priority. As this study relies exclusively on occurrence records retrieved from GBIF, which is predominantly terrestrial in scope, sampling-effort estimates for marine environments are inherently incomplete because they lack complementary marine-focused data sources such as OBIS. Nevertheless, GBIF does contain non-negligible numbers of records for aquatic and marine-associated taxa (e.g., marine mammals, seabirds, molluscs, insects, and reptiles), and excluding these records in a generic and reproducible way is not straightforward across taxonomic groups. For this reason, the distributed sampling-effort rasters retain global coverage across both land and marine grid cells to ensure transparency and reproducibility of the underlying processing decisions. For terrestrial applications, users are encouraged to apply a land mask to restrict analyses to terrestrial grid cells where appropriate. Marine-focused analyses should instead rely on dedicated marine biodiversity infrastructures. Importantly, summary statistics restricted to terrestrial grid cells, specifically spatial coverage percentages and completeness-based summaries, are reported separately where relevant (Table 2). This prevents marine underrepresentation in GBIF from biasing terrestrial sampling-effort summaries.

Because the sampling-effort rasters are provided on a geographic (latitude–longitude) grid (WGS84), cell areas vary systematically with latitude, and raw record counts are therefore not directly interpretable as globally comparable densities. While this does

not affect their use as relative sampling bias surfaces or modelling covariates, users should exercise caution when interpreting absolute magnitudes across broad latitudinal gradients. Where density-standardised metrics are required, area correction or equal-area aggregation should be implemented externally.

Beyond spatial sampling bias and spatial coverage limitations, residual data quality issues persist: taxonomic misidentifications, residual coordinate inaccuracies, and duplicated records from overlapping sources. Although the workflow's filtering steps effectively remove explicit spatial outliers and implausible coordinates, taxonomic and temporal inconsistencies remain. Users should therefore conduct additional taxon-specific and contextual validation tailored to their study systems. Numerous taxonomic cleaning and name-harmonisation tools are available, and users may integrate those most appropriate for their focal taxa and research objectives (e.g., de Melo et al. 2024). Such tools are typically taxon-specific and rely on different reference backbones, making a generic integration into a global, multi-taxon workflow challenging. In this study, taxonomic standardisation is based on the GBIF taxonomic backbone, benefiting from its ongoing curation and updates. While residual taxonomic uncertainty may influence species-level analyses, it is unlikely to substantially affect broad-scale sampling-effort patterns in the aggregated rasters. Annual temporal stratification effectively captures long-term mobilisation trends and broad temporal gradients but overlooks finer seasonal or monthly fluctuations. Future versions could incorporate monthly or seasonal stratification, providing more temporally nuanced estimates of sampling effort.

In addition to data limitations, the workflow's global scope and high spatial resolution entail substantial computational demands. The full implementation processes billions of GBIF occurrence records and requires HPC infrastructure, including large-memory nodes, parallel execution, and considerable intermediate storage capacity. Reproducing the complete pipeline, therefore, requires advanced R programming expertise and access to such computational environments. For many users without dedicated computational infrastructure, rerunning the workflow at global extent may not be feasible. For this reason, the study provides pre-calculated, ready-to-use rasters to ensure reproducibility, broad accessibility, and to avoid redundant large-scale recomputation across research groups. These rasters are designed for users seeking standardised representations of sampling effort aligned with commonly used taxonomic groups, spatial resolutions, and temporal aggregations. They are particularly suitable when focal species belong to the provided groups and the default filtering and aggregation procedures are appropriate.

The computational framework is documented to ensure methodological transparency and to support advanced users who may wish to adapt components of the workflow, e.g., modifying taxonomic scope, temporal windows, or applying alternative data-cleaning and filtering criteria (e.g., correction for potential taxonomic misidentifications). Although certain parameters allow spatial or taxonomic restriction, the workflow has not been systematically benchmarked for arbitrary spatial subsets or user-defined extents, and its primary validation and intended use remain at the global scale. Custom recalculations are therefore best regarded as expert-level applications. For most downstream applications, including SDMs and large-scale biodiversity assessments, the pre-computed global rasters are expected to provide a computationally efficient and methodologically consistent solution.

The choice of focal groups included in the pre-calculated products reflects a compromise between global data availability, taxonomic stability within the GBIF backbone, and widespread usage in macroecological and modelling applications. Some taxa, such as bryophytes, fish, or lichen-forming fungi, were not included as separate strata despite their ecological importance. These groups often exhibit distinct taxonomic structures, specialised identification requirements, or uneven global coverage that complicate harmonised large-scale aggregation. Their omission from the distributed products does not represent a limitation of the framework, which can be adapted to alternative taxa where sufficient data quality and volume permit.

The continued evolution of open biodiversity infrastructures, coupled with advances in high-performance and cloud computing, will enable broader participation in large-scale data synthesis. Establishing collaborative, regionally distributed data-processing hubs could decentralise capacity, ensuring future biodiversity informatics efforts are more inclusive, sustainable, and globally representative, and better positioned to support global ecological forecasting and conservation planning.

## 5 | Conclusion

This study delivers a globally archived, high-resolution sampling-effort dataset derived from GBIF occurrence records, generated through a reproducible workflow. Paired layers of observation count and species richness, provided across multiple spatial resolutions and taxonomic hierarchies, offer a globally consistent quantification of sampling effort suitable for bias-aware ecological analyses. By replacing static or all-taxa proxies with taxon-specific and temporally resolved effort layers, the dataset enables clearer separation of ecological signals from artefacts of uneven sampling. When integrated into modelling frameworks such as bias layers or target-group background approaches, the rasters allow explicit representation of sampling heterogeneity, strengthening species distribution inference. Beyond SDMs, the dataset supports macroecological synthesis, biodiversity monitoring, and systematic planning. Open-source implementation and FAIR archiving ensure transparency, accessibility, and long-term reuse. By quantifying the spatial, temporal, and taxonomic distribution of biodiversity records, the dataset provides a structured foundation for gap identification and evidence-informed biodiversity assessment at regional to global scales.

### Author Contributions

A.E.-G. conceived the study, developed the methodology, designed and implemented the computational workflow, curated and processed the data, performed the analyses, generated the outputs, and wrote the manuscript.

### Acknowledgements

The author gratefully acknowledges GBIF and its numerous data publishers for making biodiversity occurrence data openly accessible,

which forms the foundation of this work. Computing resources were provided by the LUMI HPC facility (Finland), whose infrastructure enabled large-scale data processing and raster generation. The workflow relies on open-source R packages, including rgbif (Chamberlain et al. 2023), sf (Pebesma 2018), the tidyverse (Wickham et al. 2019), and terra (Hijmans 2020), whose developers are sincerely thanked for maintaining the tools that make reproducible biodiversity informatics possible. Open Access funding enabled and organized by Projekt DEAL.

### Funding

The author has nothing to report.

### Conflicts of Interest

The author declares no conflicts of interest.

### Data Availability Statement

All raster outputs and associated metadata are archived on OSF (https://osf.io/hz4sy) and Zenodo (El-Gabbas 2026) under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence. Raw occurrence data are freely available through GBIF using the download keys listed in Appendix S1. The complete R workflow, including scripts for GBIF data retrieval, filtering, and rasterisation, is openly available on GitHub (https://github.com/elgabbas/global_sampling_efforts/). Programmatic access to the archived raster products is provided via the *ecokit* R package (El-Gabbas 2025), enabling direct download and local caching of the data.

### Endnotes

[1] https://www.gbif.org/about-species-counts.

### References

Aiello-Lammens, M. E., R. A. Boria, A. Radosavljevic, B. Vilela, and R. P. Anderson. 2015. "spThin: An R Package for Spatial Thinning of Species Occurrence Records for Use in Ecological Niche Models." *Ecography* 38, no. 5: 541–545. https://doi.org/10.1111/ecog.01132.

Amano, T., and W. J. Sutherland. 2013. "Four Barriers to the Global Understanding of Biodiversity Conservation: Wealth, Language, Geographical Location and Security." *Proceedings of the Royal Society B: Biological Sciences* 280, no. 1756: 20122649. https://doi.org/10.1098/rspb.2012.2649.

Araujo, M. B., R. P. Anderson, A. Marcia Barbosa, et al. 2019. "Standards for Distribution Models in Biodiversity Assessments." *Science Advances* 5, no. 1: eaat4858. https://doi.org/10.1126/sciadv.aat4858.

Baker, D. J., I. M. D. Maclean, and K. J. Gaston. 2024. "Effective Strategies for Correcting Spatial Sampling Bias in Species Distribution Models Without Independent Test Data." *Diversity and Distributions* 30, no. 3: e13802. https://doi.org/10.1111/ddi.13802.

Beck, J., M. Böller, A. Erhardt, and W. Schwanghart. 2014. "Spatial Bias in the GBIF Database and Its Effect on Modeling Species' Geographic Distributions." *Ecological Informatics* 19: 10–15. https://doi.org/10.1016/j.ecoinf.2013.11.002.

Blades, B., C. Ronquillo, and J. Hortal. 2025. "Mobilisation of Data From Natural History Collections Can Increase the Quality and Coverage of Biodiversity Information." *Ecology and Evolution* 15, no. 4: e71139. https://doi.org/10.1002/ece3.71139.

Broennimann, O., B. Petitpierre, M. Chevalier, et al. 2021. "Distance to Native Climatic Niche Margins Explains Establishment Success of Alien Mammals." *Nature Communications* 12, no. 1: 2353. https://doi.org/10.1038/s41467-021-22693-0.

Callaghan, C. T., J. E. M. Watson, M. B. Lyons, W. K. Cornwell, and R. A. Fuller. 2021. "Conservation Birding: A Quantitative Conceptual Framework for Prioritizing Citizen Science Observations." *Biological Conservation* 253: 108912. https://doi.org/10.1016/j.biocon.2020.108912.

Carvalho, S. B., G. Velo-Anton, P. Tarroso, et al. 2017. "Spatial Conservation Prioritization of Biodiversity Spanning the Evolutionary Continuum." *Nature Ecology & Evolution* 1, no. 6: 151. https://doi.org/10.1038/s41559-017-0151.

Chamberlain, S., V. Barve, D. Mcglinn, et al. 2023. "rgbif: Interface to the Global Biodiversity Information Facility API." R Package Version 3.7.8. https://CRAN.R-project.org/package=rgbif.

Christie, A. P., T. Amano, P. A. Martin, et al. 2021. "The Challenge of Biased Evidence in Conservation." *Conservation Biology* 35, no. 1: 249–262. https://doi.org/10.1111/cobi.13577.

Collen, B., M. Ram, T. Zamin, and L. McRae. 2008. "The Tropical Biodiversity Data Gap: Addressing Disparity in Global Monitoring." *Tropical Conservation Science* 1, no. 2: 75–88. https://doi.org/10.1177/194008290800100202.

Davis, A., D. Strubbe, and Q. Groom. 2023. "Global Taxonomic Occurrence Grids Using GBIF Data for Species Distribution Models." https://doi.org/10.5281/zenodo.7556851.

de Araujo, M. L., A. C. Quaresma, and F. N. Ramos. 2022. "GBIF Information Is Not Enough: National Database Improves the Inventory Completeness of Amazonian Epiphytes." *Biodiversity and Conservation* 31, no. 11: 2797–2815. https://doi.org/10.1007/s10531-022-02458-x.

de Melo, P. H. A., N. Bystriakova, E. Lucas, and A. K. Monro. 2024. "A New R Package to Parse Plant Species Occurrence Records Into Unique Collection Events Efficiently Reduces Data Redundancy." *Scientific Reports* 14, no. 1: 5450. https://doi.org/10.1038/s41598-024-56158-3.

Di Marco, M., S. Chapman, G. Althor, et al. 2017. "Changing Trends and Persisting Biases in Three Decades of Conservation Science." *Global Ecology and Conservation* 10: 32–42. https://doi.org/10.1016/j.gecco.2017.01.008.

El-Gabbas, A. 2025. "ecokit: Tools for Ecological and General Utilities." *Zenodo*. R Package. https://github.com/elgabbas/ecokit; https://doi.org/10.5281/zenodo.15477684.

El-Gabbas, A. 2026. "Multi-Resolution Sampling-Effort Data for Global Biodiversity Modelling." *[Data set]. Zenodo*. https://doi.org/10.5281/zenodo.17591681.

El-Gabbas, A., and C. F. Dormann. 2017. "Improved Species-Occurrence Predictions in Data-Poor Regions: Using Large-Scale Data and Bias Correction With Down-Weighted Poisson Regression and Maxent." *Ecography* 41, no. 7: 1161–1172. https://doi.org/10.1111/ecog.03149.

El-Gabbas, A., F. Gilbert, and C. F. Dormann. 2020. "Spatial Conservation Prioritisation in Data-Poor Countries: A Quantitative Sensitivity Analysis Using Multiple Taxa." *BMC Ecology* 20, no. 1: 35. https://doi.org/10.1186/s12898-020-00305-7.

Elith, J., and J. R. Leathwick. 2009. "Species Distribution Models: Ecological Explanation and Prediction Across Space and Time." *Annual Review of Ecology, Evolution, and Systematics* 40, no. 1: 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159.

Elith, J., S. J. Phillips, T. Hastie, M. Dudík, Y. E. Chee, and C. J. Yates. 2011. "A Statistical Explanation of MaxEnt for Ecologists." *Diversity and Distributions* 17, no. 1: 43–57. https://doi.org/10.1111/j.1472-4642.2010.00725.x.

Essl, F., A. Garcia-Rodriguez, B. Lenzner, et al. 2024. "Potential Sources of Time Lags in Calibrating Species Distribution Models." *Journal of Biogeography* 51, no. 1: 89–102. https://doi.org/10.1111/jbi.14726.

Faxon, H., and M. Chapman. 2025. "Beyond Spatial Bias: Understanding the Colonial Legacies and Contemporary Social Forces Shaping Biodiversity Data." *Environmental Research Letters* 20, no. 6: 064053. https://doi.org/10.1088/1748-9326/add6b6.

Fithian, W., J. Elith, T. Hastie, and D. A. Keith. 2015. "Bias Correction in Species Distribution Models: Pooling Survey and Collection Data for Multiple Species." *Methods in Ecology and Evolution* 6, no. 4: 424–438. https://doi.org/10.1111/2041-210X.12242.

Gaiji, S., V. Chavan, A. H. Ariño, et al. 2013. "Content Assessment of the Primary Biodiversity Data Published Through GBIF Network: Status, Challenges and Potentials." *Biodiversity Informatics* 8, no. 2: 94–122. https://doi.org/10.17161/bi.v8i2.4124.

Grand, J., M. P. Cummings, T. G. Rebelo, T. H. Ricketts, and M. C. Neel. 2007. "Biased Data Reduce Efficiency and Effectiveness of Conservation Reserve Networks." *Ecology Letters* 10, no. 5: 364–374. https://doi.org/10.1111/j.1461-0248.2007.01025.x.

Green, A. M., M. W. Chynoweth, and Ç. H. Şekercioğlu. 2020. "Spatially Explicit Capture-Recapture Through Camera Trapping: A Review of Benchmark Analyses for Wildlife Density Estimation." *Frontiers in Ecology and Evolution* 8: 563477. https://doi.org/10.3389/fevo.2020.563477.

Guillera-Arroita, G., J. J. Lahoz-Monfort, J. Elith, et al. 2015. "Is My Species Distribution Model Fit for Purpose? Matching Data and Models to Applications." *Global Ecology and Biogeography* 24, no. 3: 276–292. https://doi.org/10.1111/geb.12268.

Guisan, A., R. Tingley, J. B. Baumgartner, et al. 2013. "Predicting Species Distributions for Conservation Decisions." *Ecology Letters* 16, no. 12: 1424–1435. https://doi.org/10.1111/ele.12189.

Heberling, J. M., J. T. Miller, D. Noesgaard, S. B. Weingart, and D. Schigel. 2021. "Data Integration Enables Global Biodiversity Synthesis." *Proceedings of the National Academy of Sciences of the United States of America* 118, no. 6: e2018093118. https://doi.org/10.1073/pnas.2018093118.

Hijmans, R. J. 2020. "terra: Spatial Data Analysis." R Package Version 1.8-80. https://CRAN.R-project.org/package=terra; https://doi.org/10.32614/CRAN.package.terra.

Hirzel, A., and A. Guisan. 2002. "Which Is the Optimal Sampling Strategy for Habitat Suitability Modelling." *Ecological Modelling* 157, no. 2–3: 331–341. https://doi.org/10.1016/s0304-3800(02)00203-x.

Hobern, D., B. Baptiste, K. Copas, et al. 2019. "Connecting Data and Expertise: A New Alliance for Biodiversity Knowledge." *Biodiversity Data Journal* 7: e33679. https://doi.org/10.3897/BDJ.7.e33679.

Hortal, J., F. de Bello, J. A. F. Diniz-Filho, T. M. Lewinsohn, J. M. Lobo, and R. J. Ladle. 2015. "Seven Shortfalls That Beset Large-Scale Knowledge of Biodiversity." *Annual Review of Ecology, Evolution, and Systematics* 46, no. 1: 523–549. https://doi.org/10.1146/annurev-ecolsys-112414-054400.

Hughes, A. C., M. C. Orr, K. Ma, et al. 2021. "Sampling Biases Shape Our View of the Natural World." *Ecography* 44, no. 9: 1259–1269. https://doi.org/10.1111/ecog.05926.

Isaac, N. J. B., and M. J. O. Pocock. 2015. "Bias and Information in Biological Records." *Biological Journal of the Linnean Society* 115, no. 3: 522–531. https://doi.org/10.1111/bij.12532.

Jiménez-Valverde, A., A. T. Peterson, J. Soberón, J. M. Overton, P. Aragón, and J. M. Lobo. 2011. "Use of Niche Models in Invasive Species Risk Assessments." *Biological Invasions* 13, no. 12: 2785–2797. https://doi.org/10.1007/s10530-011-9963-4.

Johnson, T. F., A. P. Beckerman, D. Z. Childs, et al. 2024. "Revealing Uncertainty in the Status of Biodiversity Change." *Nature* 628, no. 8009: 788–794. https://doi.org/10.1038/s41586-024-07236-z.

Kadmon, R., O. Farber, and A. Danin. 2004. "Effect of Roadside Bias on the Accuracy of Predictive Maps Produced by Bioclimatic Models." *Ecological Applications* 14, no. 2: 401–413. https://doi.org/10.1890/02-5364.

Kass, J. M., A. B. Smith, D. L. Warren, et al. 2024. "Achieving Higher Standards in Species Distribution Modeling by Leveraging the Diversity of Available Software." *Ecography* 2025, no. 2: e07346. https://doi.org/10.1111/ecog.07346.

Kramer-Schadt, S., J. Niedballa, J. D. Pilgrim, et al. 2013. "The Importance of Correcting for Sampling Bias in MaxEnt Species Distribution Models." *Diversity and Distributions* 19, no. 11: 1366–1379. https://doi.org/10.1111/ddi.12096.

La Sorte, F. A., J. M. Cohen, and W. Jetz. 2024. "Data Coverage, Biases, and Trends in a Global Citizen-Science Resource for Monitoring Avian Diversity." *Diversity and Distributions* 30, no. 8: e13863. https://doi.org/10.1111/ddi.13863.

Lajeunesse, A., and Y. Fourcade. 2023. "Temporal Analysis of GBIF Data Reveals the Restructuring of Communities Following Climate Change." *Journal of Animal Ecology* 92, no. 2: 391–402. https://doi.org/10.1111/1365-2656.13854.

Lehtomäki, J., and A. Moilanen. 2013. "Methods and Workflow for Spatial Conservation Prioritization Using Zonation." *Environmental Modelling & Software* 47: 128–137. https://doi.org/10.1016/j.envsoft.2013.05.001.

Mair, L., A. C. Mill, P. A. Robertson, et al. 2018. "The Contribution of Scientific Research to Conservation Planning." *Biological Conservation* 223: 82–96. https://doi.org/10.1016/j.biocon.2018.04.037.

Marcer, A., A. D. Chapman, J. R. Wieczorek, et al. 2022. "Uncertainty Matters: Ascertaining Where Specimens in Natural History Collections Come From and Its Implications for Predicting Species Distributions." *Ecography* 2022, no. 9: e06025. https://doi.org/10.1111/ecog.06025.

Merow, C., J. M. Allen, M. Aiello-Lammens, and J. A. Silander. 2016. "Improving Niche and Range Estimates With Maxent and Point Process Models by Integrating Spatially Explicit Information." *Global Ecology and Biogeography* 25, no. 8: 1022–1036. https://doi.org/10.1111/geb.12453.

Merow, C., M. J. Smith, and J. A. Silander. 2013. "A Practical Guide to MaxEnt for Modeling Species' Distributions: What It Does, and Why Inputs and Settings Matter." *Ecography* 36, no. 10: 1058–1069. https://doi.org/10.1111/j.1600-0587.2013.07872.x.

Meyer, C., H. Kreft, R. Guralnick, and W. Jetz. 2015. "Global Priorities for an Effective Information Basis of Biodiversity Distributions." *Nature Communications* 6: 8221. https://doi.org/10.1038/ncomms9221.

Meyer, C., P. Weigelt, and H. Kreft. 2016. "Multidimensional Biases, Gaps and Uncertainties in Global Plant Occurrence Information." *Ecology Letters* 19, no. 8: 992–1006. https://doi.org/10.1111/ele.12624.

Nelson, G., and S. Ellis. 2018. "The History and Impact of Digitization and Digital Data Mobilization on Biodiversity Research." *Philosophical Transactions of the Royal Society, B: Biological Sciences* 374, no. 1763: 20170391. https://doi.org/10.1098/rstb.2017.0391.

Pebesma, E. 2018. "Simple Features for R: Standardized Support for Spatial Vector Data." *R Journal* 10, no. 1: 439–446. https://doi.org/10.32614/RJ-2018-009.

Périquet, S., L. Roxburgh, A. le Roux, and W. J. Collinson. 2018. "Testing the Value of Citizen Science for Roadkill Studies: A Case Study From South Africa." *Frontiers in Ecology and Evolution* 6: 15. https://doi.org/10.3389/fevo.2018.00015.

Petersen, T. K., J. D. M. Speed, V. Grøtan, and G. Austrheim. 2021. "Species Data for Understanding Biodiversity Dynamics: The What, Where and When of Species Occurrence Data Collection." *Ecological Solutions and Evidence* 2, no. 1: e12048. https://doi.org/10.1002/2688-8319.12048.

Phillips, S. J., R. P. Anderson, and R. E. Schapire. 2006. "Maximum Entropy Modeling of Species Geographic Distributions." *Ecological Modelling* 190, no. 3–4: 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026.

Phillips, S. J., and M. Dudík. 2008. "Modeling of Species Distributions With Maxent: New Extensions and a Comprehensive Evaluation."

*Ecography* 31, no. 2: 161–175. https://doi.org/10.1111/j.0906-7590.2008.5203.x.

Phillips, S. J., M. Dudik, J. Elith, et al. 2009. "Sample Selection Bias and Presence-Only Distribution Models: Implications for Background and Pseudo-Absence Data." *Ecological Applications* 19, no. 1: 181–197. https://doi.org/10.1890/07-2153.1.

Pili, A., B. Leroy, and D. Zurell. 2025. "Correcting Environmental Sampling Bias Improves Transferability of Species Distribution Models." *Ecography* 2025: e08002. https://doi.org/10.1002/ecog.08002.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Ranc, N., L. Santini, C. Rondinini, et al. 2017. "Performance Tradeoffs in Target-Group Bias Correction for Species Distribution Models." *Ecography* 40, no. 9: 1076–1087. https://doi.org/10.1111/ecog.02414.

Randin, C. F., T. Dirnböck, S. Dullinger, N. E. Zimmermann, M. Zappa, and A. Guisan. 2006. "Are Niche-Based Species Distribution Models Transferable in Space?" *Journal of Biogeography* 33, no. 10: 1689–1703. https://doi.org/10.1111/j.1365-2699.2006.01466.x.

Rocchini, D., E. Tordoni, E. Marchetto, et al. 2023. "A Quixotic View of Spatial Bias in Modelling the Distribution of Species and Their Diversity." *NPJ Biodiversity* 2, no. 1: 10. https://doi.org/10.1038/s44185-023-00014-6.

Seddon, P. J., P. S. Soorae, and F. Launay. 2005. "Taxonomic Bias in Reintroduction Projects." *Animal Conservation* 8, no. 1: 51–58. https://doi.org/10.1017/s1367943004001799.

Steinke, D., B. Gemeinholzer, E. Martinez-Meyer, D. Noesgaard, A. Young, and D. Schigel. 2025. "Globally Aggregated Biodiversity Data Impact Predictive and Descriptive Research." *Proceedings of the National Academy of Sciences of the United States of America* 122, no. 50: e2519119122. https://doi.org/10.1073/pnas.2519119122.

Syfert, M. M., M. J. Smith, and D. A. Coomes. 2013. "The Effects of Sampling Bias and Model Complexity on the Predictive Performance of MaxEnt Species Distribution Models." *PLoS One* 8, no. 2: e55158. https://doi.org/10.1371/journal.pone.0055158.

Troudet, J., P. Grandcolas, A. Blin, R. Vignes-Lebbe, and F. Legendre. 2017. "Taxonomic Bias in Biodiversity Data and Societal Preferences." *Scientific Reports* 7, no. 1: 9132. https://doi.org/10.1038/s41598-017-09084-6.

Varela, S., R. P. Anderson, R. García-Valdés, and F. Fernández-González. 2014. "Environmental Filters Reduce the Effects of Sampling Bias and Improve Predictions of Ecological Niche Models." *Ecography* 37, no. 11: 1084–1091. https://doi.org/10.1111/j.1600-0587.2013.00441.x.

Warton, D. I., I. W. Renner, and D. Ramp. 2013. "Model-Based Control of Observer Bias for the Analysis of Presence-Only Data in Ecology." *PLoS One* 8, no. 11: e79168. https://doi.org/10.1371/journal.pone.0079168.

Watson, J. E. M., H. S. Grantham, K. A. Wilson, and H. P. Possingham. 2011. "Systematic Conservation Planning: Past, Present and Future." In *Conservation Biogeography*, 136–160. Wiley. https://doi.org/10.1002/9781444390001.ch6.

Wickham, H., M. Averick, J. Bryan, et al. 2019. "Welcome to the Tidyverse." *Journal of Open Source Software* 4, no. 43: 1686. https://doi.org/10.21105/joss.01686.

Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3: 160018. https://doi.org/10.1038/sdata.2016.18.

Wüest, R. O., N. E. Zimmermann, D. Zurell, et al. 2019. "Macroecology in the Age of Big Data—Where to Go From Here?" *Journal of Biogeography* 47, no. 1: 1–12. https://doi.org/10.1111/jbi.13633.

Zizka, A., D. Silvestro, T. Andermann, et al. 2019. "CoordinateCleaner: Standardized Cleaning of Occurrence Records From Biological Collection Databases." *Methods in Ecology and Evolution* 10, no. 5: 744–751. https://doi.org/10.1111/2041-210x.13152.

## Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Appendix S1:** DOIs for GBIF data used to prepare sampling effort grids **Appendix S2:** Global patterns of species and observation distributions per taxonomic group **Figure S1:** Bivariate frequency distributions of species richness (*x*-axis) and observation count (*y*-axis) per grid cell at three spatial resolutions: (a) ~1 km; (b) ~5 km; and (c) ~10 km. Each panel displays a two-dimensional binning (100 × 100 bins) where colour intensity indicates the frequency of cells with each species-observation combination. Separate sub-panels show patterns for each taxonomic group and all data combined. Note that axis ranges differ among panels to optimise visualisation of group-specific patterns. **Figure S2:** Scale-dependent patterns in observation count ($\log_{10}$ scale) for Western Europe and the USA at four spatial resolutions: (A) ~1 km, (B) ~5 km, (C) ~10 km, (D) ~20 km. Warmer colours indicate higher density. Finer resolutions capture strong clustering around urban centres and institutions; coarser resolutions smooth hotspots and show country borders, reflecting national variations in data sharing. Note: each panel uses an independent colour scale; cross-panel comparisons of absolute values are not valid. **Figure S3:** Scale-dependent patterns in species richness ($\log_{10}$ scale) for Western Europe and the USA at four spatial resolutions: (A) ~1 km, (B) ~5 km, (C) ~10 km, (D) ~20 km. Warmer colours indicate higher richness. Patterns are qualitatively similar across resolutions but differ in apparent intensity, with coarser grains smoothing local hotspots. Note: each panel uses an independent colour scale; cross-panel comparisons of absolute values are not valid. **Figure S4:** Observation count patterns for Passeriformes only (top) and all data without Passeriformes (bottom); both at 20 km resolution at $\log_{10}$ scale. Despite Passeriformes contributing ~42% of all records, overall spatial patterns remain unchanged after exclusion, indicating that sampling inequality reflects structural biases shared across taxa rather than dominance by a single lineage. See Appendix S2 for taxon-specific maps. **Figure S5:** Temporal trends in GBIF data mobilisation (1980–2024; linear scale). (A) Annual observations (left) and species counts (right) for nine major taxonomic groups (colour-coded). (B) Annual observations (top row) and species counts (bottom row) for descendant taxa within each group (columns). The linear scale emphasises the dramatic acceleration in bird observations after 2010, particularly among Passeriformes, which peaked at over 150 million records in 2024. See Figure 5 for the $\log_{10}$-scale version, which reveals trends in less frequently recorded groups.