Bias and complementarity in species occurrence data from citizen science, museum, and social media collections

Shawan Chowdhury*[1], Uri Roll[2], Sultan Ahmed[3], Md Rokonuzzaman[3], Mifta Ul Munna[3], Morsheda Nur Tisha[4], Mahir Daiyan[5], Bob B. M. Wong[1], Shai Meiri[6,7]

* Corresponding author: shawan.chowdhury@monash.edu

[1]School of Biological Sciences, Monash University, Clayton, Victoria, Australia

[2]Jacob Blaustein Center for Scientific Cooperation, The Jacob Blaustein Institutes for Desert Research, Ben- Gurion University of the Negev, Midreshet Ben- Gurion, Beer- Sheva, Israel

[3]Department of Zoology, University of Dhaka, Dhaka, Bangladesh

[4]Department of Education, Noakhali Science and Technology University, Noakhali, Bangladesh

[5]Faculty of Environmental Science & Disaster Management, Daffodil International University

[6]School of Zoology, Tel Aviv University, Tel Aviv, Israel

[7]School of Biosciences, The University of Melbourne, Parkville, Victoria, Australia

Keywords

Data integration, Facebook data, GBIF, iEcology, museum, online data

Abstract

Effective biodiversity policy and spatial planning depend on species occurrence data. Yet data that informs conservation decisions typically originate from heterogeneous sources that potentially introduce systematic biases that shape downstream ecological inference and policy outcomes. To explore these potential biases, we analysed reptile location data from Bangladesh and India – a mega-diverse region encompassing multiple biodiversity hotspots as a case study. We quantified sampling bias in museum collections, social media, and citizen science platforms across species traits, habitat zonation, conservation status, and environmental gradients. Museum records were associated with higher values for crepuscular activity, aquatic habits, threatened status and small geographic ranges, but lower values for terrestrial substrates, diurnal activity, and an affinity to colder temperatures. Social media records were more associated with built-up environments, Not Evaluated taxa, terrestrial and saxicolous species, active foragers and smaller-ranged species. Citizen science records showed stronger associations with fossorial and semi-aquatic substrates, mixed foraging, larger ranges, and higher human-footprint gradients. Our findings show that biodiversity datasets are not neutral windows to nature, but rather filtered representations shaped by how records are sought, generated, shared and preserved. Strategic integration of heterogeneous data sources can therefore improve the ecological representativeness of biodiversity evidence and reduce blind spots in conservation planning.

Introduction

Environmental changes are reshaping biodiversity status worldwide, driving shifts in species distributions, phenology, traits, and population declines, and causing extinctions (Pereira et al., 2010; Dirzo et al., 2014; Murali et al., 2023). In response, international conservation policies, such as the Kunming–Montreal Global Biodiversity Framework, have increasingly prioritised spatially explicit conservation planning, habitat protection, and restoration (Brooks et al., 2006; CBD, 2022). Central to this framework is a call for coordinated spatial planning to minimise biodiversity loss and restore degraded ecosystems (Targets 1 and 2). In particular, the framework emphasises the use of the best available scientific knowledge to guide decision-making (Target 21; CBD, 2022). Achieving these goals depends critically on biodiversity data that accurately represent where species occur and how they respond to environmental and anthropogenic pressures (Watson et al., 2023).

Species occurrence data underpin conservation prioritisation, spatial planning, and assessments of biodiversity change. Ensuring that such data are robust, representative, and appropriate for spatial decision-making has become a central challenge for ecological science and conservation policy (Hughes et al., 2021). Distribution data can originate from multiple sources, including natural history collections, citizen science platforms, and automated recording systems (Heberling et al., 2021). The rapid expansion of open-access biodiversity infrastructure has dramatically increased data availability, enabling broad access to biodiversity observations through online portals. One such platform is the Global Biodiversity Information Facility (GBIF; https://www.gbif.org/), an intergovernmental research infrastructure that provides open access to biodiversity data and resources for data publication and use (Wüest et al., 2020; Heberling et al., 2021). GBIF currently hosts nearly

3.6 billion species occurrence records, and about 14,000 peer-reviewed studies have used these data.

Most GBIF records are derived from citizen science applications, such as iNaturalist and eBird (Troudet et al., 2017; Mason et al., 2025). Citizen science enables widespread public participation in biodiversity recording and has transformed our understanding of species distributions at unprecedented spatial and temporal scales (Dickinson et al., 2010; Amano et al., 2016; Johnston et al., 2023; Hanson et al., 2026). However, the traditional source of biodiversity data are natural history collections, which contain an estimated 2–4 billion preserved specimens of animals, plants, and fungi globally, spanning over 200 years of collection effort (Johnson et al., 2023). Over the last three decades, thousands of digitisation projects have improved access to natural history collection observations, and many of the world's natural history collections, including some of the largest and most comprehensive, contribute their data to GBIF. Nevertheless, there is still a long way to go as a vast number of museum specimens remain undigitised (Forbes et al., 2025). Furthermore, GBIF museum-based data are heavily biased towards American, European and Australian museum collections (Hughes et al., 2021) whereas in other regions, notably including India and Bangladesh, few museum datasets are digitised, and these are generally not shared via GBIF.

While biodiversity recording and collections have been quite common in the developed world, it is less common in developing countries (Collen et al., 2008; Slade & Ong, 2023). With the recent dominance of smartphones and high-speed internet, members of the public are increasingly sharing biodiversity observations across online platforms, including social media and citizen science platforms (Hausmann et al., 2017; Chamberlain, 2018). If properly harvested, social media data have the potential to reduce data shortfall (Chowdhury et al., 2023), better track range-shifting species (Chowdhury et al., 2026a) and invasive species (Chowdhury et al., 2026b), and even improve conservation assessments (Chowdhury et al., 2024a).

Despite rapid growth in volume, biodiversity data remain systematically biased (both taxonomically and geographically), and structured, standardised surveys remain concentrated in temperate regions (e.g., Europe and North America) (Navarro et al., 2017; Schmeller et al., 2017). Additionally, the democratisation of biodiversity recording means that observation effort is strongly shaped by human preferences, accessibility, and detectability, leading to uneven taxonomic and geographic coverage (Troudet et al., 2017; Petersen et al., 2021). For example, approximately 80% of species occurrence data on GBIF originate from North America and Europe, and 65% of all GBIF data are for birds (GBIF data accessed on 19 January 2026). Observations may also be disproportionately biased towards threatened species (Lin et al., 2022), larger-bodied taxa (Callaghan et al., 2021), and visually distinctive or easily identifiable species (Goldstein et al., 2024). To address these biases, it is critical that we urgently identify where and how existing datasets differ and how they can complement one another.

Here we focused on potential data biases across different source types in reptile records from Bangladesh and India. This region is mega-diverse and encompasses multiple biodiversity hotspots. We further focused on reptiles as this is a highly diverse group that attracts considerable enthusiast and scientific interest – but is much less known than birds and mammals. We collated reptile occurrence records from GBIF, partitioned them into

citizen science and museum data. We further collected social media data from Facebook and quantified how sampling varies across these three data sources. Using a spatially explicit framework and trait-, habitat-, and environment-specific predictors, we tested how different data sources over- or under-represent distinct components of biodiversity. We highlight how integrating contemporary digital data streams can reduce sampling bias and improve the ecological representativeness of biodiversity datasets.

Methods

Species distribution data

We obtained the list of reptile species distributed in Bangladesh and India from the Global Assessment of Reptile Distributions (GARD; http://www.gardinitiative.org/; Caetano et al., 2022) database.

We compiled species distribution data for India and Bangladesh using two sources (Facebook and GBIF). First, we harvested species distribution data from Facebook following Chowdhury et al. (2024b). Specifically, we searched for relevant Facebook groups using combinations of taxa and country names ('Reptile and Bangladesh' and 'Reptile and India'). We identified 11 Facebook groups (Supplementary Table S1). For each group, we searched for species by common and scientific names to obtain a list of relevant posts. From each post, we extracted observation date (day, month, year) and location information (where the photograph was taken). Then, we georeferenced each location using Google Maps (https://www.google.com/maps; a random point within the location). We excluded observations if the photograph was unfit for species-level identification, if locality information was missing, or if the photograph was taken outside the study countries.

We downloaded reptile distribution data from the Global Biodiversity Information Facility (GBIF; GBIF, 2026) using the rgbif R package (Chamberlain et al., 2025). GBIF contains species distribution data from thousands of citizen science applications, including iNaturalist and eBird (Heberling et al., 2021). We filtered the GBIF occurrence records to the study countries. The Facebook search ended in 2024, and the earliest data in this source were from 2013; therefore, we removed all GBIF data (for both museums and citizen science) before 2013 and after 2024. We removed all duplicate records and rows with missing species or location information from all datasets. From the GBIF data, we created two groups: citizen science (human observations) and museum (preserved specimens).

Trait data

We used species-specific traits from the SquamBase (Meiri et al., 2024), ReptTraits (Oskyrko et al., 2024), and GARD databases. We obtained data on body mass (in g, $\log_{10}$), range size (km2, $\log_{10}$), habitat (aquatic, arboreal, cryptic, fossorial, saxicolous, semi-aquatic, terrestrial, arboreal and saxicolous, arboreal and terrestrial, fossorial and terrestrial, and arboreal, saxicolous, and terrestrial), activity pattern (cathemeral, crepuscular, diurnal, and nocturnal), foraging mode (active, mixed, and sit-and-wait), venom (venomous and non-venomous), and IUCN Red List status (Not Evaluated, Data Deficient, Least Concern, Near Threatened, Vulnerable, Endangered, and Critically Endangered). For analytical purposes,

we grouped the IUCN Red List categories into four categories: Not Evaluated, Data Deficient, non-threatened (Least Concern, Near Threatened), and threatened (Vulnerable, Endangered, and Critically Endangered).

Spatial data

We collected data on spatial predictor variables from different sources. Using the geodata R package (Version 0.6.2; Hijmans et al., 2024), we obtained the annual mean temperature, rainfall, and elevation from WorldClim (https://www.worldclim.org/). We downloaded the most recent human footprint index map (for 2020) from the Wildlife Conservation Society (WCS, 2005; https://wcshumanfootprint.org/data-access) and built areas from the Global Human Settlement Database (Pesaresi et al., 2024; https://human-settlement.emergency.copernicus.eu/download.php).

Data cleaning and preparation

Using the terra package (Version 1.8.80; Hijmans, 2025) in R (Version 4.4.3; R Core Team, 2025), we reprojected and cropped all environmental predictor layers to uniform 10 × 10 km cells, using the WGS84 datum reference system and the study region's geographic extent. We used the tidyverse R package (Version 2.0.0; Wickham et al., 2019) for all data processing and the ggplot2 R package (Version 4.0.1; Wickham, 2016) for all visualisations.

To reduce the influence of uneven sampling intensity, we aggregated raw occurrence records onto a 10 × 10 km grid for all three data sources. A grid cell was considered to have reptile records from a given data source if at least one record from that source fell within the cell. Cells with no records from any source were excluded from the analyses. We removed records found outside the study region from the final datasets.

We summarised sampling intensity for each grid cell and data source as the number of unique species recorded. We derived grid-level trait composition for each data source by counting each species once per occupied grid cell, regardless of the number of observations of that species in the cell. Then, for each grid cell × data source combination, we calculated (i) the mean of continuous traits across species (e.g., body mass, range size), and (ii) the proportion of species in each categorical trait category (e.g., activity period, habitat). We considered only grid cells with at least one species for the relevant data source.

Statistical analysis of data-source differences

We quantified data-source differences by comparing grid-level predictor values across citizen science, museum and social media records. For each predictor, we first transformed proportional variables using a logit transformation after bounding values to avoid zeros and ones. We then standardised each predictor to enable comparison across variables with different units and distributions. For each predictor and data source, we calculated the weighted mean standardised value, using the number of species recorded in each grid cell as weights. Approximate 95% confidence intervals were calculated around these weighted means. To assess whether predictor values differed among data sources, we fitted a

separate weighted linear model for each predictor, with data source as a categorical explanatory variable and grid-level predictor value as the response. Statistical significance was assessed from the overall effect of the data source, with $p < 0.05$ considered statistically significant. These analyses were designed to identify source-specific ecological profiles rather than to test deviations from a single reference data source.

Results

Our final dataset contained species occurrence data from 4,539 10x10 km grid cells (citizen science: 4,125; museum: 46; and social media: 794). Citizen science data dominated the dataset and covered 91% of the grid cells, whereas only 10% overlapped with at least one of the other two data sources. Approximately 9% of grid cells were identified only after incorporating social media data, meaning these cells contained no GBIF human-observation or preserved-specimen records, and 18 of the 46 museum grids were unique to museum records (Figure 1).

While museum records were concentrated in a few locations, citizen science and social media data were distributed widely across the study region (Figure 1). Social media records were particularly concentrated in densely populated, accessible regions, especially around major cities, whereas museum records were sparse or absent in several heavily modified areas (Figure 1a). The occurrence grids of citizen science data were widely distributed across the study region. While citizen science data were widely distributed, museum and social media data were generally absent from western and, especially, southern India, but more common in northeast India and Bangladesh.

We revealed source-specific differences in reptile occurrence data when we modelled them across multiple traits, habitat, conservation, and environmental predictors (Figure 2). Museum records showed the strongest positive associations with crepuscular species, aquatic taxa, threatened species and active foragers, and were also associated with smaller-ranged species. Conversely, museum records were proportionally less representative for terrestrial taxa, diurnal species, and non-threatened species and were concentrated in colder regions. However, museum estimates generally had wider confidence intervals than the other sources.

Social media records showed a different ecological profile. They were proportionally more associated with built-up areas, Not Evaluated taxa, active foragers, terrestrial and saxicolous taxa, and smaller geographic range sizes. Social media records were less representative for reptiles that are sit-and-wait foragers, semi-aquatic taxa, non-threatened species, and high temperature. These patterns suggest that social media data capture a distinct subset of reptile occurrence space, particularly records from human-modified environments and visually or socially salient taxa. In contrast, citizen science records showed higher values for fossorial and semi-aquatic taxa, mixed foraging modes, venomous species, broader range sizes, and high human footprint.

Discussion

The quantity, composition and geographic coverage of species occurrence data largely determine outcomes of ecological and conservation assessments. It is therefore essential to understand the potential biases of data sources and how they can complement one another to counter such shortcomings and to better inform management and conservation (Troudet et al., 2017; Hughes et al., 2021; Johnston et al., 2023). We demonstrate that sampling bias in species occurrence data, and the attributes of species more likely to be represented in different dataset types, are fundamentally shaped by the data source. Comparing citizen-science, museum, and social media records, we show that each data source captures a distinct, non-random subset of ecological space, with biases structured by species traits, habitat associations, and environmental and anthropogenic gradients. These biases indicate that no single dataset provides a comprehensive or representative picture of species distributions in isolation.

Museum records appear to reflect both historical collecting practices and targeted sampling of particular taxa. Their stronger associations with aquatic, crepuscular and threatened species may suggest that collections may disproportionately capture taxa that are conservation-relevant, taxonomically important, or historically more likely to be collected. However, museum records were geographically sparse, and estimates of standardised trait and environmental values had wider confidence intervals, highlighting the limitations of relying on digitised museum data alone, especially in regions such as India and Bangladesh where collections remain few, incompletely digitised, or unevenly shared through global biodiversity platforms.

Social media records capture a contrasting component of ecological space. Their stronger association with built-up areas, terrestrial and saxicolous taxa, Not Evaluated species and smaller-ranged species suggests that social media can provide valuable information from human-dominated environments and for species that may otherwise be poorly represented in formal biodiversity repositories. However, these records are also shaped by observer behaviour, accessibility, detectability and social interest, and seemingly represent species that are easier to observe. These patterns likely reflect observer behaviour, accessibility, detectability and social interest rather than true ecological abundance. Similar processes also shape citizen-science datasets, reinforcing that all opportunistic biodiversity records are filtered by human perception and opportunity (Bird et al., 2014).

Citizen science records were the most geographically extensive, yet they were also shaped by source-specific biases. The stronger associations of reptile citizen science records with fossorial and semi-aquatic substrates, venomousness, mixed-foraging taxa, broader-ranged species and human-footprint gradients indicate that citizen science data are not neutral baselines. Rather, they reflect the interplay of species detectability, observer behaviour, platform use and spatial accessibility. This reinforces the need to treat all biodiversity data streams as filtered representations of ecological reality.

While we identify a distinct pattern in species distribution data across three data sources, our results should be interpreted with caution. First, our study focused solely on reptiles, and bias structures may differ across taxa, observer communities, and regions. Our findings align with recent work from Bangladesh showing that social media records are concentrated in built-up areas (Chowdhury et al., 2026b), although relationships with other human-pressure and topographic metrics (e.g., human footprint and elevation) may not generalise across taxa, regions, or methodological choices. Second, we focused on presence-only data

and therefore cannot distinguish between true ecological absence and absences driven by observer behaviour, low detectability, and low accessibility. This limitation is common to most large-scale biodiversity datasets (Goldstein et al., 2024). Third, museum data in the regions are few, and heavily geographically biased – by the existence of collections in different regions, by their degree of digitisation, and by whether they do, or do not, share their data via the GBIF platform. In India, the collections of one of the two major natural history museums, the Bombay Natural History Society, are not digitised. Those of the Zoological Society of India are only partially digitised, and these data are available on GBIF (Chaitanya R., pers. comm.). The major natural history collections in Bangladesh, those of the Bangladesh National Museum, are not digitised and not shared on GBIF. Hence, the region we study is likely to have lower GBIF coverage than in other regions (most notably, but not only, in Europe and North America). Finally, because our analysis did not use an independent, fully representative benchmark, it quantifies relative differences among data sources rather than absolute bias. Accordingly, none of the three sources should be treated as an unbiased baseline. Despite these limitations, the mechanisms underlying the observed biases are pervasive across biodiversity data systems. As such, the qualitative patterns and implications identified here are likely to be broadly relevant beyond the focal region and taxon.

Conclusion

Our findings show that biodiversity data biases are source-dependent, reflecting the mechanisms of data generation, observer behaviour, and historical legacy. Crucially, integrating citizen science, museum, and social media records can reshape the ecological space represented in occurrence datasets by offsetting source-specific extremes and substantially reducing trait- and environment-driven biases. As biodiversity monitoring increasingly relies on diverse digital data streams, recognising and harnessing these complementary strengths offers a practical pathway to more ecologically representative occurrence datasets, strengthening macroecological analyses, species distribution modelling, and conservation planning, and ultimately supporting robust inference and effective conservation action in a rapidly changing world.

Data and code availability

GBIF data (from citizen science and museum specimens) are publicly available (GBIF, 2026). We have provided the Facebook data in the supplementary data file (Supplementary Table S2).

All the R scripts are available in the following GitHub repository: https://github.com/ShawanChowdhury/sm_museum_cs_bias.

Acknowledgements

We thank all the administrators, moderators, and reviewers of social media groups, as well as the citizen scientists who have shared their data publicly.

Author contributions

S.C. conceptualised the idea. S.C., S.A., M.R., M.U.M., M.N.T., and M.D. collected the data. S.C., U.R., and S.M. developed the experimental procedures, and all authors contributed to them. S.C. conducted the analysis, and all authors contributed to the analysis. S.C. wrote the paper, and all authors contributed to the writing of the paper.

Declaration of interests

The authors declare no competing interests.

AI use statement

We have used ChatGPT at different stages to improve writing and analysis in R.

References

Amano, T., Lamming, J. D., & Sutherland, W. J. (2016). Spatial gaps in global biodiversity information and the role of citizen science. *Bioscience*, *66*(5), 393-400.

Bird, T. J., Bates, A. E., Lefcheck, J. S., Hill, N. A., Thomson, R. J., Edgar, G. J., ... & Frusher, S. (2014). Statistical solutions for error and bias in global citizen science datasets. *Biological Conservation*, *173*, 144-154.

Brooks, T. M., Mittermeier, R. A., Da Fonseca, G. A., Gerlach, J., Hoffmann, M., Lamoreux, J. F., ... & Rodrigues, A. S. (2006). Global biodiversity conservation priorities. *Science*, *313*(5783), 58-61.

Chamberlain, S., Barve, V., Mcglinn, D., Oldoni, D., Desmet, P., Geffert, L., Ram, K. (2025). rgbif: Interface to the global biodiversity information facility API. R package version 3.8.1. https://CRAN.R-project.org/package=rgbif.

Chamberlain, J. (2018). Using social media for biomonitoring: how Facebook, Twitter, Flickr and other social networking platforms can provide large-scale biodiversity data. In *Advances in Ecological Research* (Vol. 59, pp. 133-168). Academic Press.

Forbes, O., Young, A. G., & Thrall, P. H. (2025). Global sampling decline erodes science potential of natural history collections. *Nature Communications*, *16*(1), 9255.

GBIF.org (07 January 2026) GBIF Occurrence Download https://doi.org/10.15468/dl.d236jv.

Caetano, G. H. D. O., Chapple, D. G., Grenyer, R., Raz, T., Rosenblatt, J., Tingley, R., Böhm, M., Meiri, S., & Roll, U. (2022). Automated assessment reveals that the extinction risk of reptiles is widely underestimated across space and phylogeny. *PLoS Biology*, *20*(5), e3001544.

354 Callaghan, C. T., Poore, A. G., Hofmann, M., Roberts, C. J., & Pereira, H. M. (2021). Large-
355 bodied birds are over-represented in unstructured citizen science data. *Scientific*
356 *Reports*, *11*(1), 19073.

357 Convention on Biological Diversity (CBD). (2022). Kunming–Montreal Global biodiversity
358 framework. Draft decision submitted by the President CBD/COP/15/L.25, 18 December
359 2022. https://www.cbd.int/conferences/2021-2022/cop-15/documents.

360 Chowdhury, S., Debnath, R., Hawladar, N., Howard, S. R., Hodgins, K. A., Wong, B. B., & Jarić,
361 I. (2026a). Social media data reveal novel habitats for invasive species. *Biological*
362 *Conservation*, *314*, 111668.

363 Chowdhury, S., Hawladar, N., Roy, R. C., Capinha, C., Cassey, P., Correia, R. A., ... & Bonn, A.
364 (2026b). Harnessing social media data to track species range shifts. *Conservation Biology,*
365 *40*(3), e70234.

366 Chowdhury, S., Fuller, R. A., Ahmed, S., Alam, S., Callaghan, C. T., Das, P., ... & Bonn, A.
367 (2024a). Using social media records to inform conservation planning. *Conservation*
368 *Biology*, *38*(1), e14161.

369 Chowdhury, S., Ahmed, S., Alam, S., Callaghan, C. T., Das, P., Di Marco, M., ... & Bonn, A.
370 (2024b). A protocol for harvesting biodiversity data from Facebook. *Conservation*
371 *Biology*, *38*(4), e14257.

372 Chowdhury, S., Aich, U., Rokonuzzaman, M., Alam, S., Das, P., Siddika, A., ... & Callaghan, C.
373 T. (2023). Increasing biodiversity knowledge through social media: A case study from
374 tropical Bangladesh. *BioScience*, *73*(6), 453-459.

375 Collen, B., Ram, M., Zamin, T., & McRae, L. (2008). The tropical biodiversity data gap:
376 addressing disparity in global monitoring. *Tropical Conservation Science*, *1*(2), 75-88.

377 Dickinson, J. L., Zuckerberg, B., & Bonter, D. N. (2010). Citizen science as an ecological
378 research tool: challenges and benefits. *Annual review of ecology, evolution, and*
379 *systematics*, *41*(1), 149-172.

380 Dirzo, R., Young, H. S., Galetti, M., Ceballos, G., Isaac, N. J., & Collen, B. (2014). Defaunation
381 in the Anthropocene. *Science*, *345*(6195), 401-406.

382 Forbes, O., Young, A. G., & Thrall, P. H. (2025). Global sampling decline erodes science
383 potential of natural history collections. *Nature Communications*, *16*(1), 9255.

384 Goldstein, B. R., Stoudt, S., Lewthwaite, J. M., Shirey, V., Mendoza, E., & Guzman, L. M.
385 (2024). Logistical and preference bias in participatory science butterfly data. *Frontiers in*
386 *Ecology and the Environment*, *22*(8), e2783.

387 Hanson, J. O., McCune, J. L., Alamenciak, T., & Bennett, J. R. (2026). Increasing the credibility
388 of conservation plans through citizen science. *Biological Conservation*, *313*, 111552.

389 Hausmann, A., Toivonen, T., Heikinheimo, V., Tenkanen, H., Slotow, R., & Di Minin, E. (2017).
390 Social media reveal that charismatic species are not the main attractor of ecotourists to sub-
391 Saharan protected areas. *Scientific Reports, 7*(1), 763.

Heberling, J. M., Miller, J. T., Noesgaard, D., Weingart, S. B., & Schigel, D. (2021). Data integration enables global biodiversity synthesis. *Proceedings of the National Academy of Sciences*, *118*(6), e2018093118.

Hijmans, R. J., Barbosa, M., Ghosh, A., Mandel, A. (2024). geodata: Download Geographic Data. R package version 0.6-2.

Hughes, A. C., Orr, M. C., Ma, K., Costello, M. J., Waller, J., Provoost, P., ... & Qiao, H. (2021). Sampling biases shape our view of the natural world. *Ecography*, *44*(9), 1259-1269.

Johnson, K. R., Owens, I. F., & Global Collection Group. (2023). A global approach for natural history museum collections. *Science*, *379*(6638), 1192-1194.

Johnston, A., Matechou, E., & Dennis, E. B. (2023). Outstanding challenges and future directions for biodiversity monitoring using citizen science data. *Methods in Ecology and Evolution*, *14*(1), 103-116.

Lin, H. Y., Binley, A. D., Schuster, R., Rodewald, A. D., Buxton, R., & Bennett, J. R. (2022). Using community science data to help identify threatened species occurrences outside of known ranges. *Biological Conservation*, *268*, 109523.

Mason, B. M., Mesaglio, T., Barratt Heitmann, J., Chandler, M., Chowdhury, S., Gorta, S. B. Z., ... & Callaghan, C. T. (2025). iNaturalist accelerates biodiversity research. *BioScience*, *75*(11), 953-965.

Murali, G., Iwamura, T., Meiri, S., & Roll, U. (2023). Future temperature extremes threaten land vertebrates. *Nature*, *615*(7952), 461-467.

Navarro, L. M., Fernandez, N., Guerra, C., Guralnick, R., Kissling, W. D., Londono, M. C., ... & Pereira, H. M. (2017). Monitoring biodiversity change through effective global coordination. *Current Opinion in Environmental Sustainability*, *29*, 158-169.

Oskyrko, O., Mi, C., Meiri, S., & Du, W. (2024). ReptTraits: a comprehensive dataset of ecological traits in reptiles. *Scientific Data, 11*(1), 243.

Pereira, H. M., Leadley, P. W., Proença, V., Alkemade, R., Scharlemann, J. P., Fernandez-Manjarrés, J. F., ... & Walpole, M. (2010). Scenarios for global biodiversity in the 21st century. *Science*, *330*(6010), 1496-1501.

Pesaresi, M., Schiavina, M., Politis, P., Freire, S., Krasnodębska, K., Uhl, J. H., ... & Kemper, T. (2024). Advances on the Global Human Settlement Layer by joint assessment of Earth Observation and population survey data. *International Journal of Digital Earth*, *17*(1), 2390454. https://CRAN.R-project.org/package=geodata.

Meiri, S. (2024). SquamBase—A database of squamate (Reptilia: Squamata) traits. *Global Ecology and Biogeography*, *33*(4), e13812.

Robinson, D., Hayes, A., Couch, S. (2025). broom: Convert Statistical Objects into Tidy Tibbles. R package version 1.0.8. https://CRAN.R-project.org/package=broom.

Schmeller, D. S., Böhm, M., Arvanitidis, C., Barber-Meyer, S., Brummitt, N., Chandler, M., ... & Belnap, J. (2017). Building capacity in biodiversity monitoring at the global scale. *Biodiversity and Conservation*, *26*(12), 2765-2790.

Slade, E. M., & Ong, X. R. (2023). The future of tropical insect diversity: strategies to fill data and knowledge gaps. *Current Opinion in Insect Science*, *58*, 101063.

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

Watson, J. E., Venegas-Li, R., Grantham, H., Dudley, N., Stolton, S., Rao, M., ... & Ward, M. (2023). Priorities for protected area expansion so nations can meet their Kunming-Montreal Global Biodiversity Framework commitments. *Integrative Conservation*, *2*(3), 140-155.

Wickham, H. (2016). Data analysis. In *ggplot2: elegant graphics for data analysis* (pp. 189-201). Springer-Verlag New York.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K. & Yutani H (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686.

Wüest, R. O., Zimmermann, N. E., Zurell, D., Alexander, J. M., Fritz, S. A., Hof, C., ... & Karger, D. N. (2020). Macroecology in the age of Big Data–Where to go from here?. *Journal of Biogeography*, *47*(1), 1-12.

List of Figures

Figure 1.

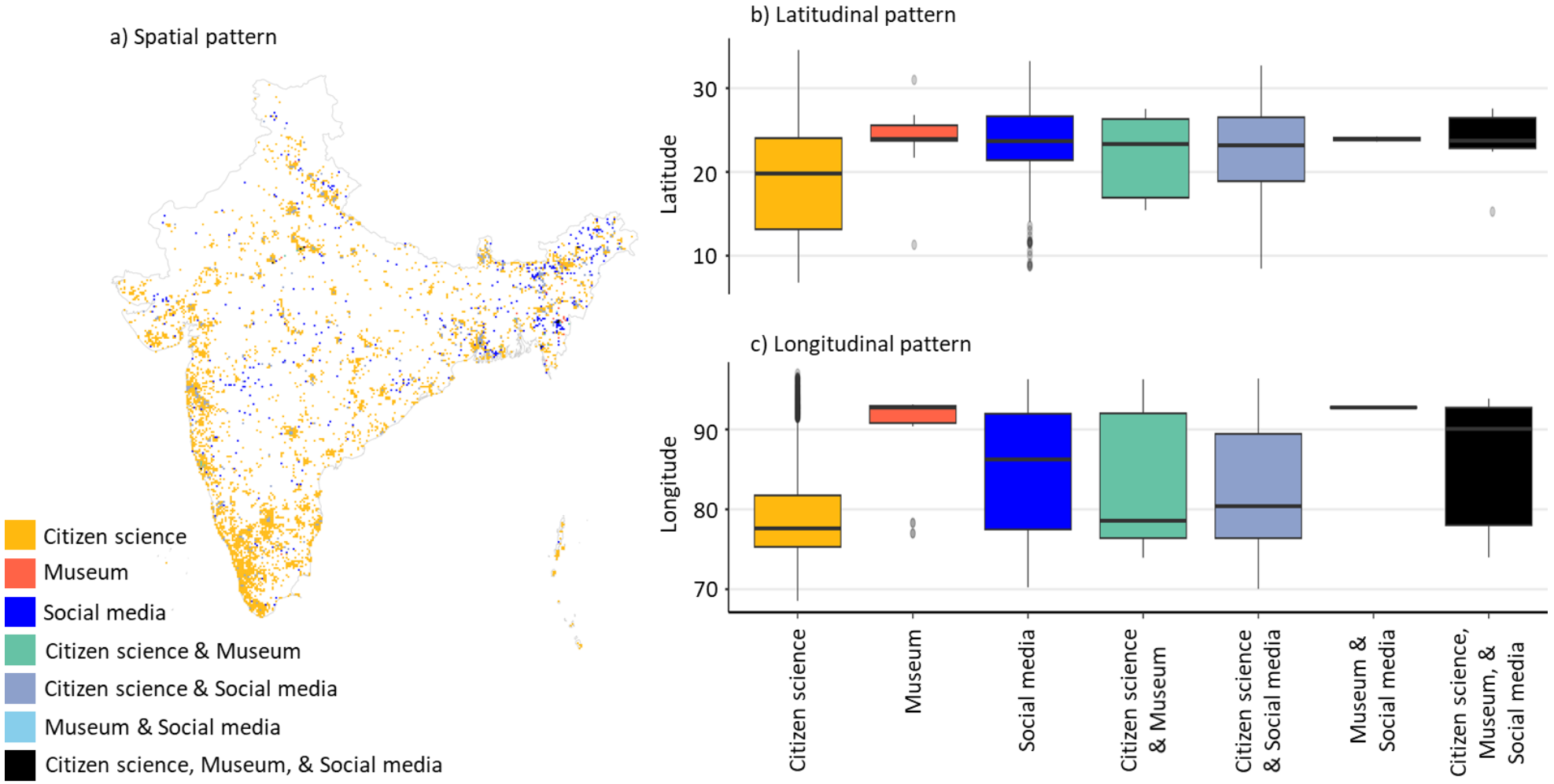
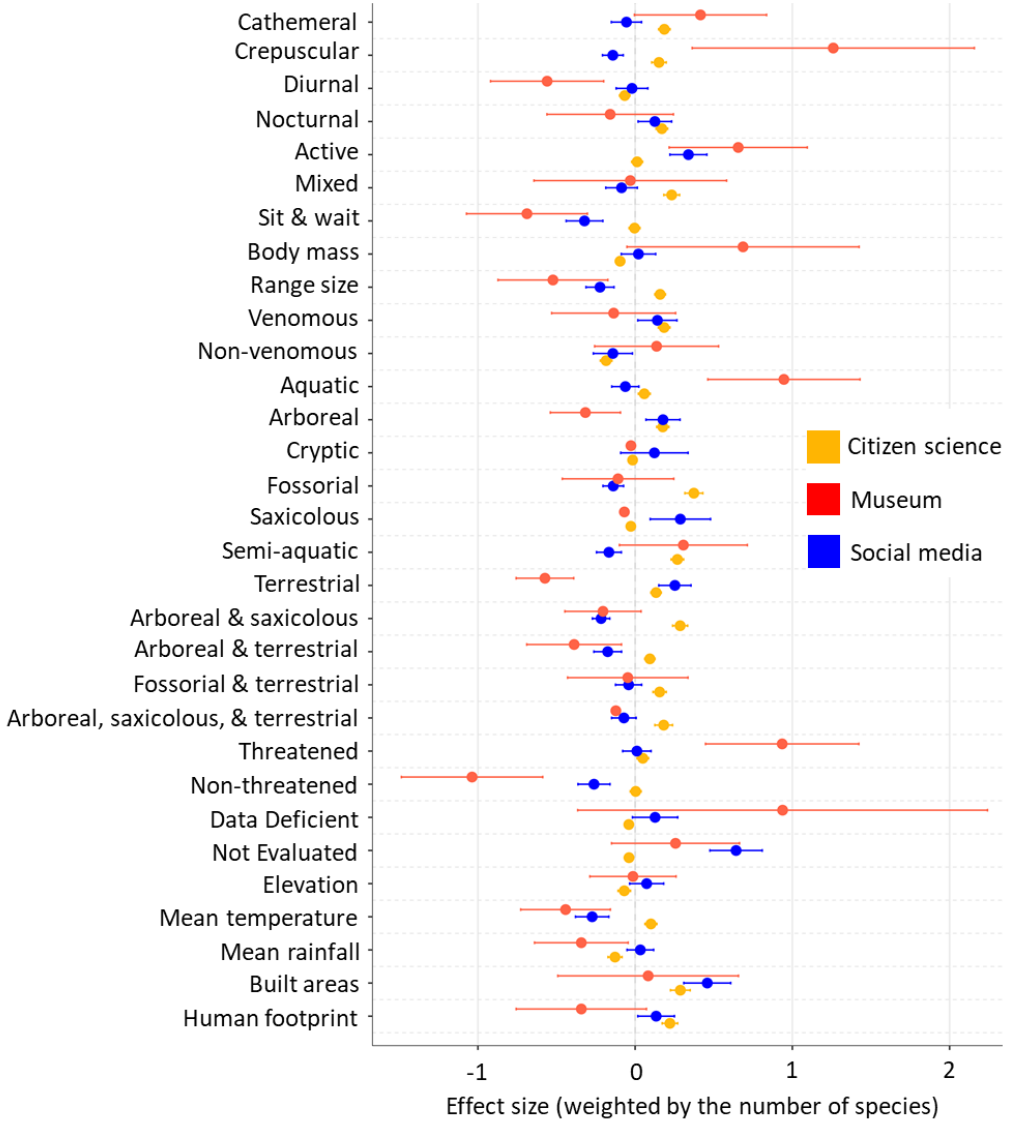

a) Spatial pattern
Citizen science
Museum
Social media
Citizen science & Museum
Citizen science & Social media
Museum & Social media
Citizen science, Museum, & Social media
b) Latitudinal pattern
Latitude
30
20
10
c) Longitudinal pattern
Longitude
90
80
70
Citizen science
Museum
Social media
Citizen science & Museum
Citizen science & Social media
Museum & Social media
Citizen science, Museum, & Social media

Figure 2


Cathemeral
Crepuscular
Diurnal
Nocturnal
Active
Mixed
Sit & wait
Body mass
Range size
Venomous
Non-venomous
Aquatic
Arboreal
Cryptic
Fossorial
Saxicolous
Semi-aquatic
Terrestrial
Arboreal & saxicolous
Arboreal & terrestrial
Fossorial & terrestrial
Arboreal, saxicolous, & terrestrial
Threatened
Non-threatened
Data Deficient
Not Evaluated
Elevation
Mean temperature
Mean rainfall
Built areas
Human footprint
Citizen science
Museum
Social media
-1
0
1
2
Effect size (weighted by the number of species)