Bull Mar Sci. 101(3):1449–1474. 2025
https://doi.org/10.5343/bms.2024.0022

Bulletin of Marine Science
research paper

A critical appraisal of occurrence data in OBIS and GBIF databases: a case study on true mangrove species

Leibniz Centre for Tropical Marine Research (ZMT), Bremen, Germany

** Corresponding author email: <cherrycheung20042005@gmail.com>*

Cheuk Yiu Cheung *
Véronique Helfer

ABSTRACT.—In the context of the biodiversity and climate crises, reliably documenting past and current species distributions is of paramount importance for deciphering the main drivers of species occurrences and range shifts and forecasting those under various global change scenarios. For that purpose, species observation records are essential and, according to the FAIR data principles, should be shared with a broad community of researchers and other stakeholders. Various databases have been created to compile and centralize information about biodiversity in recent years, among which the Ocean Biodiversity Information System (OBIS), dedicated to the marine realm, and the Global Biodiversity Information Facility (GBIF) are the most renowned ones. Here we evaluated how 38 selected "true mangrove" species are represented in those databases and assessed the quality and reliability of the information on geographical location of those observations. While OBIS and GBIF are extremely valuable databases, they still contain erroneous information, highlighting the need for closer communication among scientific experts and database managers together with the implementation of automated validation processes (e.g., using ecosystem distribution maps) to improve the data curation and data quality assurance processes. Further, we showed clear data deficiencies in many regions, including biodiversity hotspots. Many valuable observations are either hidden in publications or in private repositories but not shared with the global community, a practice that should change for the benefit of mangrove conservation and management. We encourage mangrove researchers to be proactive in correcting those data deficiencies by systematically submitting their observations, following the FAIR data principles.

Cartagena Colombia
July | 24-28 | 2023

Proceedings of the 6th Mangrove Macrobenthos and Management meeting

Guest Editors:
Juan Felipe Blanco-Libreros, Maria Fernanda Adame, Gustavo A Castellanos-Galindo, Samantha K Chapman, Karen Diele, José Ernesto Mancera Pineda, Kerrylee Rogers

Handling Editor:
Gustavo A Castellanos-Galindo

Date Submitted: 31 January, 2024.
Date Accepted: 20 September, 2024.
Available Online: 20 September, 2024.

In the face of the current biodiversity and climate crises, identifying species and areas of higher priority for conservation or restoration is a crucial step for efficient species or ecosystem management. In this regard, a reliable and comprehensive documentation of current species distribution is essential. For instance, recent studies in the Bangladesh Sundarbans pointed out that existing protected areas may not be able to effectively cover even the current local biodiversity hotspots or the

core habitats of threatened mangrove species of socioeconomic relevance (Sarker et al. 2016, 2019b). Further, considering the ongoing global change effects, optimal management strategies should be adaptive and consider not only current biodiversity hotspots or core range of target species, but also the locations where these important habitats or species are predicted to be in the future (Hannah et al. 2007, McLeod et al. 2009, Helfer and Zimmer 2018). In this context too, the reliable documentation of past and current species distribution is of paramount importance to decipher the main drivers of species occurrence and use this information to forecast distribution range shifts of species or communities under future climate and development scenarios, using species or community distribution modeling approaches (Rodríguez et al. 2007, Zellmer et al. 2019). Beyond the prediction of species distribution per se, species distribution modeling is also increasingly utilized in trait-based approaches for predicting community structure and ecosystem services under the influence of climate change (Frenette-Dussault et al. 2013, Moor et al. 2015, Green et al. 2022). In the case of mangrove ecosystems, many recent studies on the drivers of their distribution, vulnerability and resilience, and how these will be impacted by anthropogenic pressure and climate change utilized remote sensing data at the ecosystem level (e.g., coverage, above-ground biomass; Duncan et al. 2018, Gouvêa et al. 2022, Amaral et al. 2023). While species composition and functional traits are known to influence ecosystem properties and processes (Symstad et al. 1998, Baeten et al. 2019, Luo et al. 2019), studies at species or community level still heavily depend on field-collected data (e.g., Sarker et al. 2019a,b), as long as remote-sensing techniques cannot unequivocally identify species—this might change in the near future as the fields of remote sensing and artificial intelligence progress rapidly, and are thus far mostly restricted to local or regional studies.

True mangroves, a group of vascular plants specialized morphologically and physiologically to inhabit exclusively the mid and upper intertidal zone of tropical and subtropical coasts, play a major role in structuring mangrove communities (Tomlinson 2016). Understanding how true mangrove species respond to climate change using trait-based approaches (e.g., Li et al. 2022) could help to predict the properties and processes of mangrove ecosystems under climate change. While a functional trait dataset of true mangroves is available (Quadros and Zimmer 2017), the data is not georeferenced and can therefore not be used for distribution modeling. Concurrently while plant trait databases such as the TRY plant trait database (Kattge et al. 2020) and the Global Inventory of Floras and Traits database (GIFT; Weigelt et al. 2020) contain trait measurement of mangrove species, these data are not always georeferenced. For both distribution modeling and trait-based studies, species observation records (i.e., occurrence or presence data) are essential, and studying global issues such as climate change requires a global data input (Turnhout and Boonman-Berson 2011).

Since the late 1990s various biodiversity databases have been created to compile, centralize, and make publicly available information about biodiversity. Among them, The Global Biodiversity Information Facility (GBIF; launched in 2001) is currently the largest and most comprehensive biodiversity database, and the Ocean Biodiversity Information System (OBIS; launched in 2000) is the global biodiversity database dedicated to marine species. Since the publication of the FAIR (Findable, Accessible, Interoperable, Reusable) Guiding Principle (Wilkinson et al. 2016), enormous efforts have been made to share and standardize biological observations with a broad

community of researchers and other stakeholders. In recent years, the potential use of biodiversity data contributed by citizen science projects such as iNaturalist has been explored, and these data are being incorporated into biodiversity databases such as GBIF notably (De Cecco et al. 2021). As a result, the input and use of data in these databases have rapidly increased over the years, including the use in species or community distribution modeling and for predicting those under climate change scenarios (Heberling et al. 2021).

As these species occurrence data are collected across a relatively large time scale by various contributors using different methodologies, and imported from different source repositories, quality and completeness of these data may vary even when data standards (such as Darwin Core; Wieczorek et al. 2012) are followed. While those data quality issues have been reported in various studies focusing on different taxa, quality of data is not always properly considered by biodiversity data users (Ball-Damerow et al. 2019). The most addressed quality issues include incomplete or missing spatial or temporal information (e.g., Vandepitte et al. 2015, Serra-Diaz et al. 2017, Colli-Silva et al. 2020), geospatial errors or uncertainties (e.g., Yesson et al. 2007, Maldonado et al. 2015, Vandepitte et al. 2015, Ribeiro et al. 2022) and taxonomic errors (e.g., Gaiji et al. 2013, Freitas et al. 2020, Ribeiro et al. 2022). Further, duplicates were also found in these biodiversity databases (Mesibov 2013, Moudrý and Devillers 2020), and taxonomic or spatial bias (e.g., Beck et al. 2014, Troudet et al. 2017) and data gaps (e.g., Feeley 2015, Garcia-Rosello et al. 2023) were also highlighted. While the errors or uncertainties mentioned above might be of minimal impact for macroscopic studies (e.g., Queiroz et al. 2021), for other applications, such as abundance-based distribution models aiming at better linking species effect on ecosystem processes (Waldock et al. 2022), this would lead to biased habitat suitability maps that will then lead to wrong assessment of the effect of a species for specific ecosystem processes.

This study assessed the availability and quality of occurrence data of selected true mangrove species from the Global Biodiversity Information Facility (GBIF) and the Ocean Biodiversity Information System (OBIS), the two major global biodiversity databases. Here the following were assessed: (i) availability of occurrence data, (ii) availability of information for the distinct entries, particularly those related to the basis of record and georeferencing, and (iii) quality and reliability of the geographical location of the occurrence data. The results of the current study provides insights on the suitability of true mangrove occurrence data for species distribution and trait modeling analyses.

Materials and Methods

Selection of the Target Species.—From the 73 plant species or hybrids reported to occur in mangrove ecosystems (status Spalding et al. 2010; *see* Appendix 1), we retained only the major true mangroves, i.e., species of true mangroves that can form pure stands (as opposed to species considered minor components, following Tomlinson 2016), summing up to 44 species from nine genera.

The species taxonomy was verified in the World Register of Marine Species (WoRMS, Horton et al. 2021); in case of discrepancy between WoRMS and Spalding et al. (2010), the taxonomy proposed in WoRMS was retained and used for data acquisition from the various databases, as it is frequently updated and verified and it is also used by OBIS and GBIF as a taxonomy backbone (Costello and Appeltans 2008).

As a result, *Rhizophora harrisonii* in OBIS was labeled *R.* × *harrisonii* in this study, and WoRMS-accepted *Avicennia germinans* was retained instead of *Hilairanthus germinans* (accepted name of *A. germinans* on GBIF at the time of data acquisition). *Avicennia rumphiana,* listed as a species in Spalding et al. (2010), is recognized as a subspecies of *Avicennia marina* (*A. marina* subsp. *rumphiana*) in WoRMS and *Lumnitzera* × *rosea*, although accepted at the species rank in WoRMS (but with the taxonomic remark *Lumnitzera littorea* × *Lumnitzera racemosa*), was considered as a synonym of *Lm. racemosa* in GBIF; those two taxa were therefore not retained for data download (but occurrences of *A. marina* subsp. *rumphiana* were contained in the dataset for the parent species; no occurrence of *L.* × *rosea* was recovered from the parent species dataset, although some entries contained "*L. rosea*" for the GBIF attribute scientificName). While *Ceriops australis* was accepted as synonym of *Ceriops tagal* in GBIF at the time of data acquisition, it was considered a valid species elsewhere (in OBIS and in WoRMS) and treated as such within this study. The final number of species retained for data acquisition after taxonomic verification was 42 (*see* Appendix 2).

Data Acquisition.—In the initial phase of this study, several databases were explored to get occurrence data for the target species, including the Global Biodiversity Information Facility (GBIF), the Ocean Biodiversity Information System (OBIS), the Mangrove Reference Database and Herbarium (MRDH; Dahdouh-Guebas 2023), and Tropicos (Missouri Botanical Garden 2023). Only GBIF and OBIS were retained for this study as the two are the most used biodiversity databases in the literature (Ball-Damerow et al. 2019) which provide georeferenced occurrence data for various species, including mangroves. While both MRDH and Tropicos also contain georeferenced data for the selected mangrove species, direct extraction of the data was not available for MRDH, and only brief records with coordinates down to minutes were available on Tropicos, thus these two databases were not further explored.

GBIF (www.gbif.org) is an international organization network which aims at providing free and open access biodiversity data, aggregating and making available information from various sources such as museum specimens, DNA barcodes and published datasets. GBIF also receives public-contributed data by including user-verified, "research-grade" occurrence data from the iNaturalist platform. As of September 2023, GBIF holds over 2.4 billion georeferenced entries. OBIS (www.obis.org) is a marine biogeographic information system managed by an International Committee which receives, maintains quality, publishes and provides free and open access to biogeographic information of marine species, and is one of the earliest Associate Members and largest publishers of data to GBIF (Costello et al. 2007). As of September 2023, OBIS provides about 120 million presence records of over 180,000 marine species, of which 54 million records were already provided to GBIF.

Occurrence data was available for 38 of the 42 selected species (*see* Appendix 1), and were extracted from the OBIS and GBIF databases in December 2021 (the R script for the data acquisition and curation processes is available on Github and Zenodo (https://github.com/Mangroven/Mangroveobs_GBIFOBIS, https://10.5281/zenodo.15836331). Data from OBIS were downloaded using the occurrence function of the official robis R package (OBIS 2021, Provoost et al. 2022; *see* Appendix 3 for individual dataset citations). For GBIF, occurrence data was downloaded directly

from their portal, using the "simple" version of occurrence data option (*see* Appendix 4 for data download DOIs); entries without coordinates, with zero coordinates (of coordinates of 0°N and 0°E, which indicates that errors possibly occurred in recording geographical coordinates; flagged "ZERO_COORDINATE" by GBIF), and/or with coordinates mismatch with country information (flagged "COUNTRY_COORDINATE_MISMATCH") were excluded using filters available on the GBIF portal (for occurrence data of *A. marina*, by mistake, coordinate-related filters were not applied; the entries without coordinates, with zero coordinates or with coordinate mismatch were removed manually after data download from GBIF). Occurrence data for subspecies or varieties of the 38 selected species that were accepted in WoRMS (*A. marina* subsp. *australasica*, *A. m.* subsp. *marina*, *A. m.* subsp. *rumphiana*, *A. m.* var. *intermedia; Laguncularia racemosa* var. *glabriflora*, *Lg. racemosa* var. *racemosa*; *Lm. racemosa* var. *racemosa*) were included in the species downloads (this was verified by extracting those data from GBIF independently from the species data and comparing the results to the species data). The data acquisition process is summarized in Appendix 5.

After acquisition of occurrence data from the two biodiversity databases, the dataset was simplified to retain only those attributes that were relevant for subsequent analysis, particularly those related to taxonomy, geospatial location, collection date, source of data, and quality remarks/flags were retained for further data curation. As a result, 25 out of 51 attributes from GBIF data and 32 out of 124 attributes from OBIS data were retained (*see* Appendix 6). Some additional attributes (such as "nameinlist"; for more details, *see* the data acquisition and curation R script used in this study) were then added to facilitate subsequent data handling and curation.

Data Curation.—*Looking for Duplicates Within and Across Databases.*—Occurrence data acquired from OBIS and GBIF were then combined (an additional attribute, "database", indicating the source of the data was added) and checked for duplicates within and across databases using R version 4.3.1 (R Core Team 2023) within RStudio v2023.6.2.561 (Posit Team 2023), using the "group_by" and "mutate" functions of the dplyr package. Because of the absence (at the time of data acquisition as of now) of any universal unique identifiers for occurrence data across the two databases, identifying truly shared data in GBIF and OBIS data was impossible (as discussed in Moudrý and Devillers 2020).

Complete duplicates (entries that have appeared more than once with identical information except for the attributes "ID" and "database" across database duplicates) were first identified within each database independently, and then across databases. Complete duplicates across and within databases were deduplicated using the "distinct" function in R. Potentially duplicated entries within and across databases were then identified based on the species ("nameinlist"), the location ("decimalLatitude" and "decimalLongitude" attributes), and the time ("eventDate", "day", "month", and "year" attributes) the occurrences were recorded. Only potential duplicates across databases were then further deduplicated after manual inspection, retaining the entry with more nonempty attributes.

Excluding Unsuitable Occurrence Data.—After deduplication, the occurrence data were further examined to exclude entries that were unsuitable for the purpose of the current study (i.e., assessing availability and quality of true mangroves occurrence

data for species or trait modeling). Entries that were (i) referring to fossil records or living specimens (a specimen that is alive, e.g., a living plant in a botanical garden) as specified in the attribute basisOfRecord, (ii) flagged as having potentially unreliable information on geographical location (i.e., "GEODETIC_DATUM_INVALID" and "PRESUMED_SWAPPED_COORDINATE" in the attribute "issueFlag"), (iii) located outside the latitudinal range of mangroves (32°N–39°S; adopted from Saenger et al. 2019), or (iv) without information on year (i.e., empty "year" and "eventDate" attributes) were excluded.

Checking the Reliability of the Spatial Location.—The occurrence data were then imported to QGIS 3.28 (QGIS 2023) for further inspection. The "join by nearest" function was used with the "global biophysical typology of mangroves" map (Worthington et al. 2020) to (i) identify occurrences that were outside known mangrove areas and (ii) calculate distances of these occurrences to the nearest mangrove patch.

The data curation process is summarized in the Supplementary Appendix S7.

Descriptive Statistics and Graphical Visualization of Curated Dataset.—The curated dataset was further explored using R to document the number of (i) entries by species (considering also the biogeographical region or Flora they belong to Indo-West Pacific or Atlantic-East Pacific); (ii) species by number of entries (organized by classes: <10, 10–49, 50–99, 100–999, 1000–9999, >10,000); (iii) entries by countries of record (based on the attributes "country" and "countryCodes" provided by GBIF and OBIS); (iv) entries by basis of record.

The temporal distribution of the data was examined based on the attribute "year" and "eventDate". Besides, the coordinate accuracy and precision of the occurrence data were also examined. Coordinate accuracy was estimated using the number of decimal places of the coordinates (i.e., rounding of the coordinates; *see* Moudrý and Devillers 2020); whenever there was a difference in the number of decimal places between latitude and longitude, the one with the greater number of decimals was considered. Coordinate precision was assessed based on the "coordinateUncertaintyInMeters" or "coordinatePrecision" attributes as provided by GBIF and OBIS.

The curated dataset was also explored using QGIS and species distribution range maps from the IUCN red list assessment to identify (i) entries that were outside the documented distribution range of species, especially species that have reported introduced or invasive populations (such as *Bruguiera gymnorrhiza*, *Lg. racemosa*, *Sonneratia apetala*), and (ii) potential spatial data gaps in mangrove occurrence data. Besides, we plotted the number of occurrences against national mangrove area (following Hamilton and Casey 2016) on R to investigate whether there is a relationship between data availability and mangrove area and detect potential data deficiency in countries with a large national mangrove area.

Finally, we examined the distribution of environmental data, which would be essential in distribution and trait modeling, by extracting those on QGIS for (1) the curated occurrence data using the point sampling tool plugin (extracting values for each points, thus approaching abundance-based modeling), (2) the curated occurrence data after conversion to a raster (extracting only one value per cell, that contained potentially several curated occurrence data, thus approaching occurrence-based modeling) using the raster calculator and raster pixel to points tools, and (3)

the known species distribution (IUCN Red List documented distribution) using the clip raster by mask layer and raster pixel to points tools. We concentrated here only on the annual mean temperature (BIO1) and the annual precipitation (BIO12) of the WorldClim bioclimatic data at 30 s resolution (Fick and Hijmans 2017).

Results

Overall, occurrence data were available for all 38 selected species. GBIF contained data for all 38 species, including three subspecies (*A. marina* subsp. *australasica*, *A. marina* subsp. *marina*, *A. marina* subsp. *rumphiana*) and four varieties (*A. marina* var. *intermedia*; *Lg. racemosa* var. *glabriflora*, *Lg. r.* var. *racemosa*; *Lm. racemosa* var. *racemosa*) accepted in WoRMS. From OBIS, occurrence data could be extracted only for 30 species (with only one subspecies: *A. m.* subsp. *eucalyptifolia* but was already included in the observation data of its parent species *A. marina*). A total of 84,299 entries were downloaded, with 47,655 and 36,644 entries downloaded from GBIF and OBIS respectively (Appendix 2).

Data Curation.—*Looking for Duplicates Within and Across Databases.*—We identified 18,921 entries (of which 95.89%, i.e., 18,130 entries, were from the GBIF database) as complete duplicates (i.e., entries that have appeared more than once with identical information, except entry identifier/ID, and the source database for across database duplicates), which represent 22.45% of the whole dataset. Deduplication of those entries (i.e., removal of repeated entries) resulted in the exclusion of 16,729 entries and a remaining dataset of 67,570 entries. No complete duplicates were found across databases.

While further inspecting the dataset for potential duplicates, we identified four entries as across-database potential duplicates based on species, location and time of collection. Those across-database duplicates were deduplicated and only the GBIF entries (n = 2) were retained, as they contained more non-empty attributes than the corresponding OBIS entries, resulting in a dataset of 67,568 entries. Further inspection conducted within each database separately detected 1998 entries within GBIF and 102 within OBIS as potential duplicates based on species, geographical location and day, month and year of record. As it could not be confirmed whether these potential duplicates were true duplicates, they were retained for subsequent analyses.

Excluding Unsuitable Occurrence Data.—The remaining dataset was further curated, and four and 14 entries (from GBIF) referred to as fossil or living specimens respectively were removed. Additionally, 27 and 828 entries from the GBIF database were suspected to have swapped coordinates (flagged "PRESUMED_SWAPPED_COORDINATE") or an invalid geodetic datum (flagged "GEODETIC_DATUM_INVALID") and were excluded. Another 60 entries (59 from GBIF and one from OBIS) were located outside the latitudinal range of mangroves (32°N–39°S); they were referred to as preserved specimen, human observation or "unknown" in the attribute basisOfRecord; we considered those occurrences as dubious and excluded them from subsequent analyses. Finally, another 1774 entries were removed as no information could be retrieved for the year of observation (empty "year" and

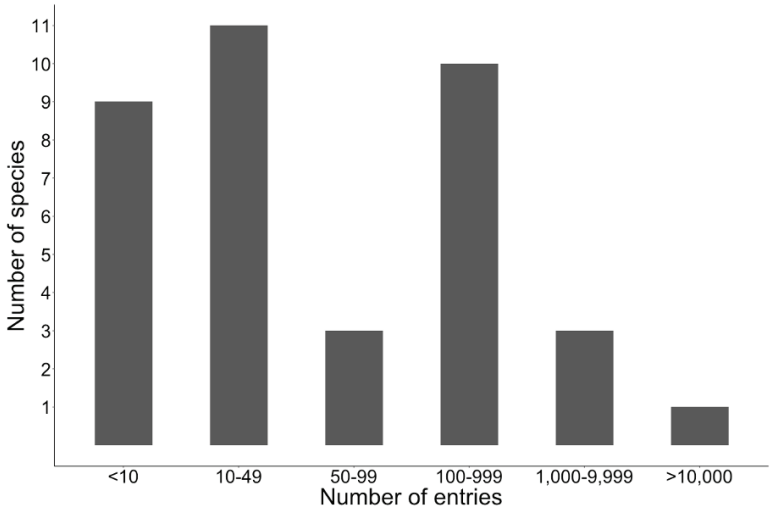

Figure 1. Number of species by number of entries (organized by classes) after data curation (n = 26,805; GBIF and OBIS data combined).

"eventDate" attributes). Removing all unsuitable entries (n = 2707) resulted in 64,861 remaining entries.

Checking the Reliability of the Spatial Location.—The last step of the data curation, consisting in verifying that the occurrences were located within documented mangrove stands (following Worthington et al. 2020), resulted in the exclusion of 38,056 entries and a final curated dataset of 26,805 entries (GBIF: 17.74%, n = 4755; OBIS: 82.26%, n =22,050), documenting the distribution of 37 species (Appendix 2; *Sonneratia griffithii* that had initially seven entries in GBIF was not represented in the curated dataset). Overall, the whole data curation resulted in a loss of 68.20% of the initial data.

Descriptive Statistics and Graphical Visualization of Curated Dataset.—Regarding the taxonomic coverage of the entries retained in our final dataset, we could observe a strong taxonomic bias, with only three species (*Avicennia marina*, *C. tagal,* and *R. stylosa*; all from the IWP region) constituting 82.78% (22,188) of all (26,805) retained entries (Appendix 8A), and about half (20 out of 37) of the species (AEP and IWP combined) were represented in the final dataset with less than 50 entries (Fig. 1). However, when data from Australia (n = 23,247), which dominated the curated dataset, were excluded, three AEP species (*Rhizophora mangle, A. germinans,* and *Lg. racemosa*) dominated the remaining data, constituting 72.57% (2582) of all (3558) remaining entries (Appendix 8B).

Looking at the geographical distribution of the data, the retained entries were recorded from 79 countries / regions (e.g., Hong Kong), with no relevant information for 13 entries (representing 0.05% of the data). There was a strong spatial bias in the data provenance, with 86.73% (n = 23,247) of the retained entries coming from Australia, followed by Mexico (4.2%; n = 1125) and Brazil (2.16%; n = 579); remaining countries represented only 6.87% (n = 1841) of the data (Appendix 9).

Looking at the type of observations (attribute basisOfRecord), a majority of the retained entries were obtained by direct human observation (92.56%; 2908 from GBIF and 21,902 from OBIS), followed by preserved specimens (6.76%; 1664 from GBIF and 148 from OBIS). One entry and 14 entries from GBIF were reported as "machine observation" (i.e., output of a machine observation process, e.g., photograph, a video, an audio recording, or a remote sensing image) and "material sample" (i.e., a material entity, e.g., whole or part of an organism, soil or microbial sample) respectively. A total of 168 entries (0.63% of retained entries, all from GBIF) from several data contributors (as indicated in the attribute institutionCode) were provided with the attribute basisOfRecord as "UNKNOWN" (Appendix 10). A closer inspection of those entries with an unknown (but non-empty) basis of record, revealed that most of them were related to herbarium datasets; as no indication that those observations could be based on fossil or living specimens from artificial environments (e.g., botanical garden) was found, those entries were retained. A more stringent curation of the data could consider excluding those entries from the final dataset, as we cannot ascertain that those observations relate to individuals living in their natural environment.

The temporal distribution of the data, ranging from 1886 to 2021 (year of data acquisition), showed some anomalies, with a huge data input (66.5%; n = 17,833) in 2001 [Appendix 11A, originating from one country (Australia), and another high data input in 1996 (13.05%; n = 3497) with 99.17% (n = 3468) originating from Australia]. Besides these two anomalies, there is an increasing trend in data with a higher increase since 1990 (Appendix 11B).

Regarding the precision of the coordinates, a majority (98.89%, n = 26,501) of the retained data had coordinates with four or more decimal places (Fig. 2A), which corresponds to ≤11-m accuracy at the equator (Moudrý and Devillers 2020). For coordinate precision, 97.08% of the OBIS data had a precision of 0.1–1 km. However, coordinate precision information was missing for about 2% of the OBIS data, and up to 43.53% of the GBIF data did not provide a valid coordinate precision (Fig. 2B).

Examining the retained data, one species was found to have entries located far from its documented distributed range; *R. mangle*, an AEP exclusive species, had an entry from GBIF data obtained from human observation (according to attribute basisOfRecord) located far from the species' known distribution range in Bangladesh (Fig. 3). Further, while its native range is well covered with occurrence data in the Americas (with some exceptions such as in parts of Venezuela, Guiana and Suriname), its distribution range in Africa is not well-covered, showing a data deficiency in Gulf of Guinea (including Nigeria, Cameroon, Gabon, Congo, Ivory Coast, and Liberia) and Sierra Leone.

Looking at spatial distribution of the data for *A. marina*, which had the highest number of retained entries (n = 10,677), most entries (98.28%; n = 10,493) were originating from Australia (Fig. 4). Despite having a very high number of entries, occurrence data were absent from many areas across its native range, including southeast Asia.

Finally, comparing the number of entries recorded from different countries and the national mangrove area, we can observe two outliers (Fig. 5A): (i) Australia with the highest number of entries while having only the fifth largest national mangrove area, constituting 3% of the world's mangroves (Hamilton and Casey 2016); (ii) Indonesia with a very low number of entries, while it is the country with the highest national mangrove area, also known to shelter a very high

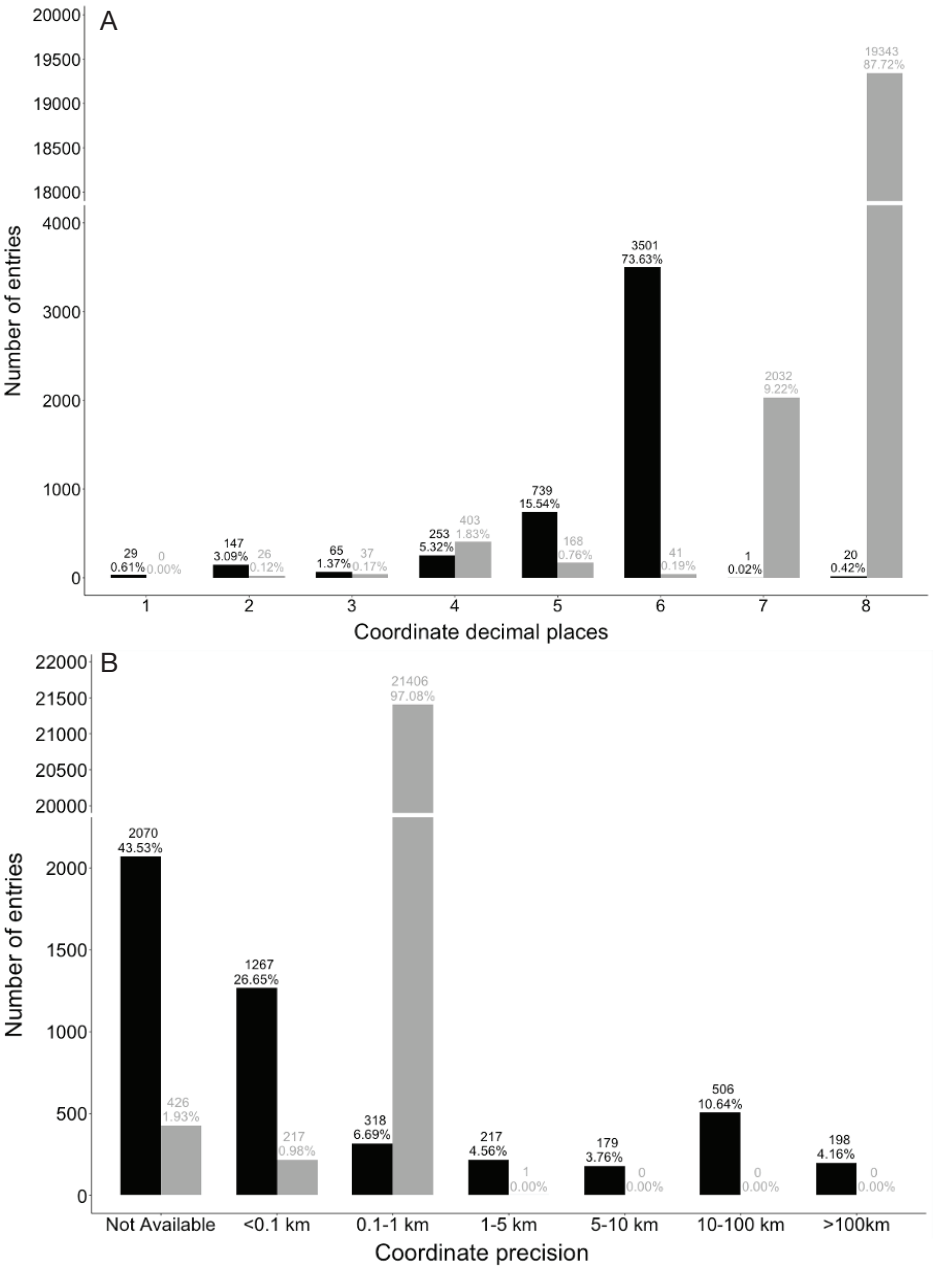

Figure 2. (A) Number of entries by decimal places of coordinates; (B) number of observations by coordinate precision. In black, data from GBIF; in grey, data from OBIS.

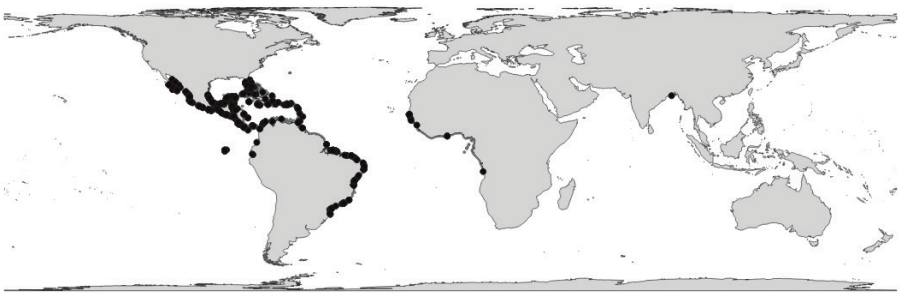

Figure 3. Retained occurrence data of *Rhizophora mangle* (black) and known native range of the species (dark grey; Ellison et al. 2015). Grey circles indicate data deficient areas. Base map data from Natural Earth.

biodiversity, with 43 true mangrove species (FAO 2005). After removing those two outliers, to inspect the data further, we see that Brazil, and to some extent Malaysia and Papua New Guinea, are presenting a similar pattern as Indonesia (relatively low number of entries relative to the national mangrove area), while Mexico and the United States exhibit a similar pattern as Australia (Fig. 5B). When compared against documented mangroves, distribution of the occurrence data within countries were not evenly distributed, even in countries with the highest number of entries in the retained data. For example, for Australia, most of the occurrence data retained were from the northeastern mangroves; mangroves from western Australia were underrepresented (Fig. 6A). In the United States, Brazil and

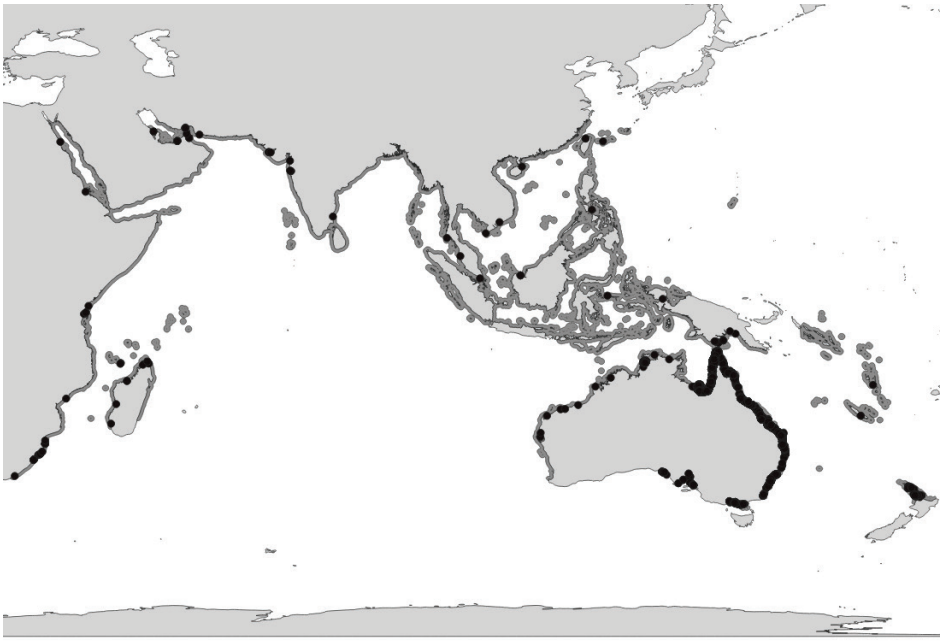

Figure 4. Retained occurrence data of *Avicennia marina* (black) and known native range of the species (dark grey; Duke et al. 2010). Base map data from Natural Earth.

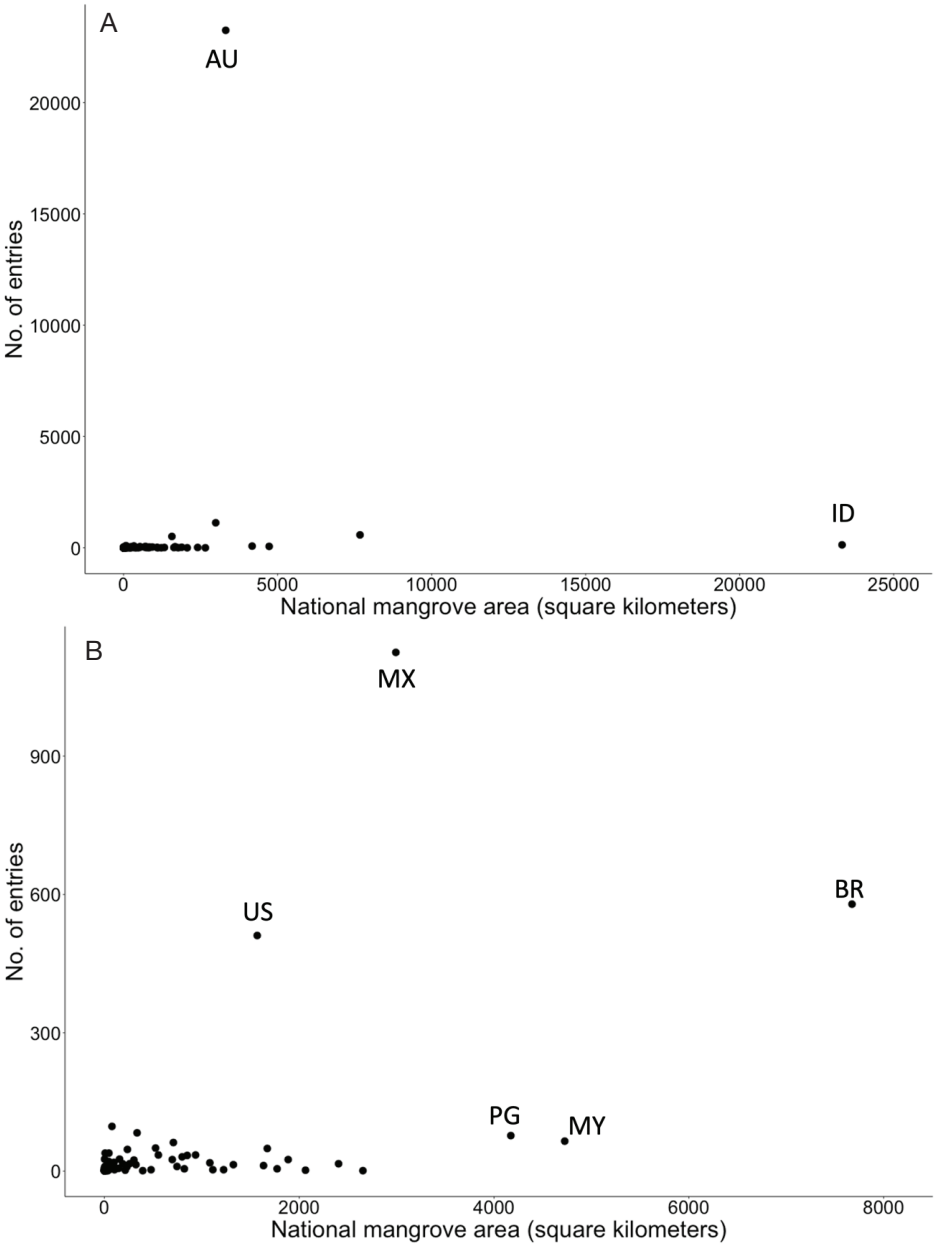

Figure 5. Relation between the number of retained occurrence data and the national mangrove area (following Hamilton and Casey 2016) of (A) all countries with retained occurrence data and (B) with Australia and Indonesia excluded. Country abbreviations: AU = Australia, BR = Brazil, ID = Indonesia, MX = Mexico, MY = Malaysia, PG = Papua New Guinea, US = United States. Base map data from Natural Earth.

Mexico (Fig. 6B, C, and D, respectively), countries from which a high number of entries were retrieved, the data coverage was more even, but data gaps could be observed in some locations. In countries that have a large mangrove national area with a deficient data coverage, occurrence data were distributed across various locations within the country in Indonesia and Malaysia (Fig. 6E and F) but was relatively constrained geographically in several locations in Papua New Guinea (Fig. 6G).

For both *A. marina* and *R. mangle*, who had a high number of retained entries in the curated data whose distribution was not evenly distributed across the species range, the distribution of the climatic data differed when extracted using occurrence data or the known species distribution as mask for the extraction (Fig. 7). For *A. marina*, the distribution of the temperature data (Fig. 7A) was similar for both datasets (extraction from the occurrence data or from the species known distribution), but the distribution of the precipitation data was not (Fig. 7B), with the occurrence data covering areas with low precipitations while the whole species range encompasses a much wider range or precipitation patterns. For *R. mangle*, the distribution of the temperature (Fig. 7C) data was bimodal for the occurrence data, with a mode at ca. 24 °C which is not observed in the distribution over the whole species range (which shows a higher density at ca. 27 °C). For the precipitations (Fig. 7D), the two datasets show clearly different patterns, with lower precipitations in the areas covered by the occurrence dataset, while the dataset extracted from the known species distribution range show higher densities at higher precipitations (with a mode at ca. 2500 mm). For both species and climatic variables, a similar distribution of data was observed for the point-based (curated observation data) and the raster-based (raster cells that contained at least one curated observation data) extractions.

Discussion

Duplicates and Shared Data.—About 42% (18,130 out of 47,655 entries) and 2% (791 out of 36,644 entries) of occurrence data from GBIF and OBIS respectively, for the 38 selected true mangrove species, were identified as either complete or potential duplicates in this study, which overall constituted approximately 24% (18,921 out of 84,299 entries) of the occurrence data acquired from the two biodiversity databases. Slightly lower proportion (37%) of duplicates were reported for marine mammal occurrence data from the GBIF database by Moudrý and Devillers (2020), while much lower proportions were reported for the overall GBIF database (about 10%; Gaiji et al. 2013). For OBIS data, the observed proportion of duplicates is lower than that observed by Moudrý and Devillers (2020) for marine mammals (19%). Overall, the proportions of duplicated data were lower in OBIS than GBIF. Unlike previous studies where duplicates were identified based on taxonomical (species), temporal (date of data collection) and geospatial (coordinates) information only (Gaiji et al. 2013, Moudrý and Devillers 2020), duplicates in this study were dominated by complete duplicates, that were defined as repeated entries with identical information except of entry identifier/ID and were nonetheless found in high proportions, predominantly in the GBIF database.

While the likelihood of independent observation of a species at the same location and time is a priori quite low, such cases can occur especially for the GBIF database that incorporates citizen science-contributed data (e.g., from iNaturalist; input of multiple spatially and temporally identical observations of the same species from

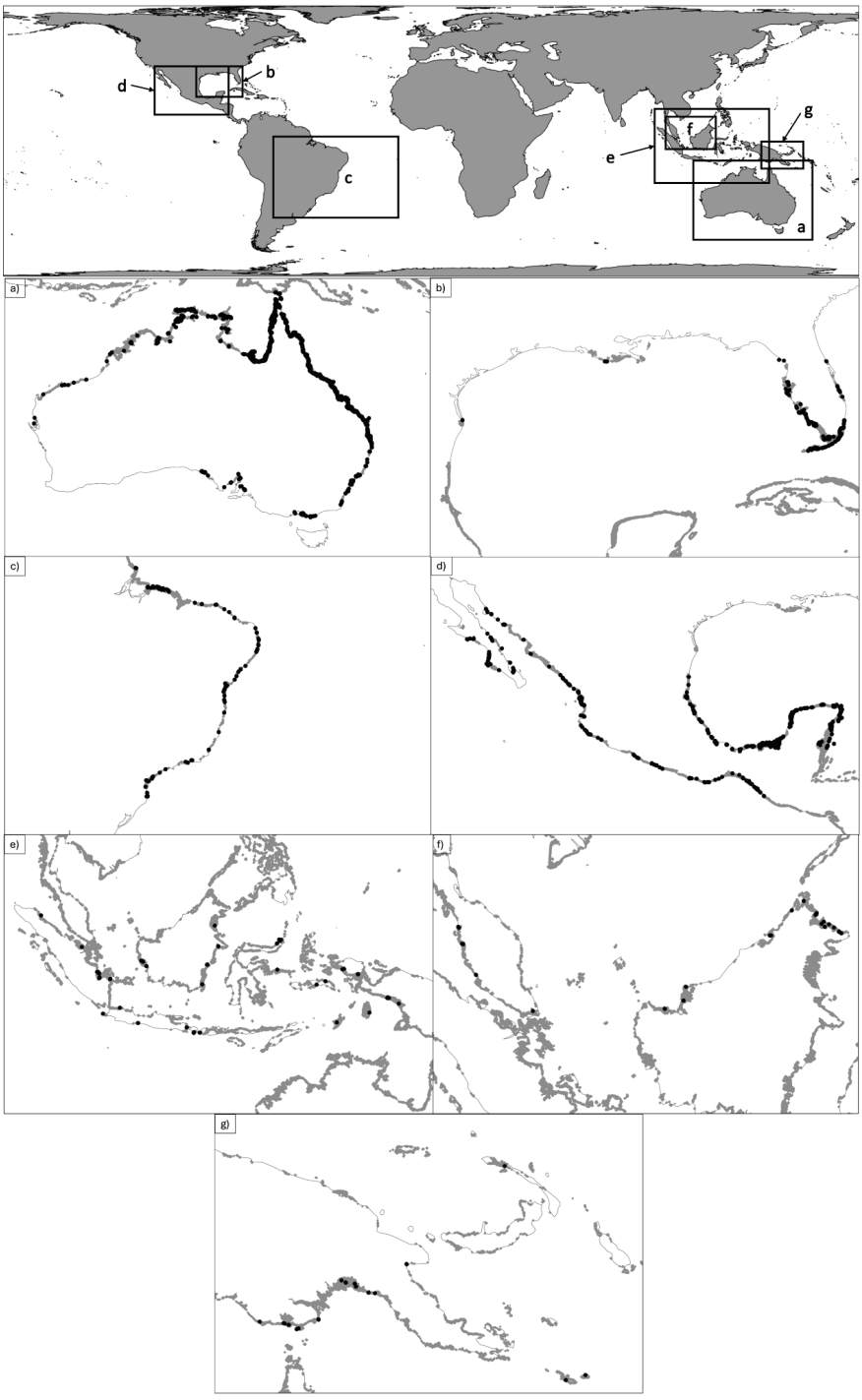

Figure 6. Retained occurrence data (black) and mapped mangrove area (gray; Worthington et al. 2020) of (A) Australia, (B) the United States, (C) Brazil, (D) Mexico, (E) Indonesia, (F) Malaysia, and (G) Papua New Guinea. Base map data from Natural Earth.

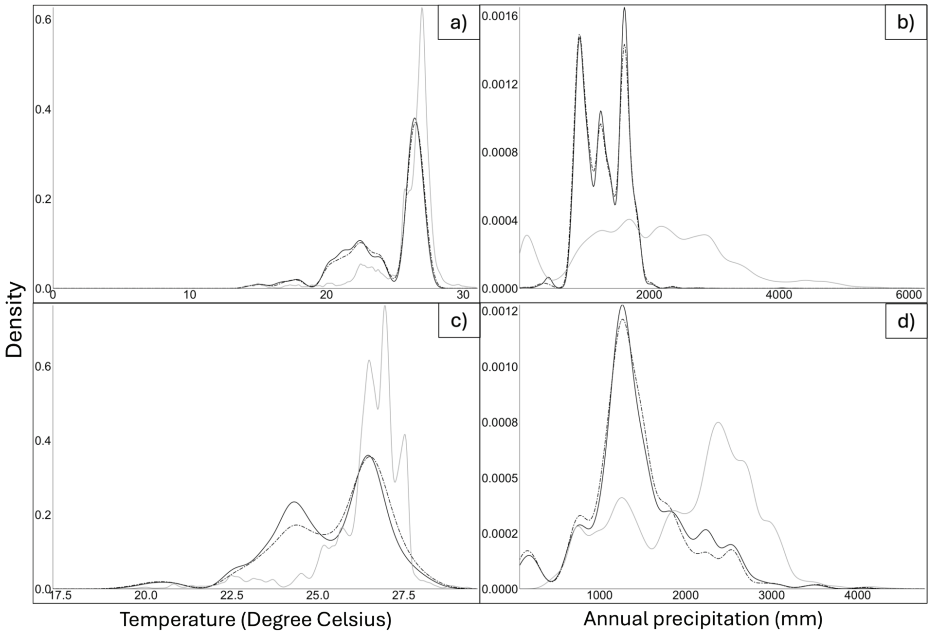

Figure 7. Kernel Density Plots of WorldClim bioclimatic variables at 30 s resolution (Fick and Hijmans 2017) extracted using mangrove curated occurrence data (in black), and species known distribution in IUCN red list (in gray). For the curated occurrence data, the full and the double-dash lines represent the data obtained using the point-based (mimicking abundance-based models) and raster-based (mimicking occurrence-based models) extraction approaches, respectively. (A) Annual mean temperature for *Avicennia marina*; (B) Annual precipitation for *A. marina*; (C) Annual mean temperature for *Rhizophora mangle*; (D) Annual precipitation for *R. mangle*. Plots were generated in R using the ggplot function.

different observers during citizen science events is indeed possible in these publicly contributed biodiversity platforms). Nonetheless, among the GBIF complete duplicates identified within this study, only 0.9% (n = 163) were provided by the iNaturalist platform (as specified for the attribute institutionCode). Besides citizen science-contributed data, there are several alternative causes that could lead to duplicated entries in biodiversity databases. For example, collection of multiple specimens of the same species at the same date and location (Mesibov 2013) could be filed as separated entries in herbarium- or museum-contributed datasets, or field observations of the several individuals of the same species at the same location and date, especially if the location is determined at a sampling unit scale that could encompass several individuals of the same species (plots in forest structure survey) could result in identical data similar to the complete duplicates identified in this study. This could be the case of the current study as observation entries from human observation and preserved specimens (likely from herbarium or museum collections) dominated the dataset (Appendix 10). Duplicates could also be caused by repeated entry of the same observation into a dataset (by mistake), or aggregation of the same observation existing in separate datasets (i.e., submission of the same observation to distinct repositories that are regularly exchanging data). Regarding the latter, we detected only four entries (<0.05% of the acquired occurrence data) as potentially duplicated entries between GBIF and OBIS, suggesting that there was some data sharing between the databases, or that the observer provided their data to the two

databases independently; due to the very low number of occurrences concerned, the last option is more plausible. The proportion of possibly shared data detected for true mangroves was unexpectedly extremely low since the two major biodiversity databases are in ongoing cooperation and over 50 million occurrence records from OBIS have been shared to GBIF (Costello et al. 2007). Moudrý and Devillers (2020) reported about 11% of marine mammal occurrence data being shared between OBIS and GBIF, a proportion that might have been overestimated as they used loser criteria (latitude and longitude coordinates only) used for identifying potentially shared data; in the current study, both spatial (decimalLatitude and decimalLongitude) and temporal (eventDate, day, month, year) attributes were considered. Finally, duplicates and/or shared data could have been underestimated, in this study and in general, due to failed detection caused by missing or incomplete values provided in key attributes related to spatial and/or temporal information, or data heterogeneity (i.e., multiple variants representing the same attribute values, as a result of a lack of standardized vocabulary; Chapman et al. 2020). Identifying shared data between biodiversity databases is challenging, as the exchange of data between these databases is often not clearly documented and has therefore to be deduced by the data users themselves (Feng et al. 2022). While the presence of duplicated data would not be problematic in the case of occurrence-based models, they could provide a biased representation of a species distribution when applying abundance-based species distribution models, that are particularly relevant when trying to relate species effect on ecosystem properties, processes, and services (Waldock et al. 2022).

Missing and Dubious Information Related to Geographic Location.—The retained entries (from the curated dataset) had coordinates provided with a high accuracy; coordinates of nearly all the retained entries were provided with four decimal places or higher, and most (>95%) were at five decimal places or higher (Fig. 2A). This corresponds to an accuracy of one meter or higher at the equator, which would allow for the distinction of individual trees (Moudrý and Devillers, 2020). This accuracy is more than adequate for geographically extensive studies of sessile species such as modeling future distribution of widely distributed mangrove species, where regional or global climatic datasets used as predictors are often of lower resolution. For example, WorldClim bioclimatic variables are available at highest resolution of 30 seconds (approximately 900 meters at equator; Fick and Hijmans 2017). The high coordinate accuracy would also make local studies (e.g., within a mangrove reserve) using those occurrence data possible. Yet, information about coordinate accuracy is not always available in species occurrence data from GBIF or OBIS (Ball-Damerow et al. 2019, Moudrý and Devillers 2020). In the current study, about ten percent of all retained entries (n = 2496 out of 26,805 entries) were missing information about the coordinate precision, especially in the GBIF data (Fig. 2B); in general, data from OBIS had higher coordinate accuracy and precision than those from GBIF.

Less than half (n = 26,805; 31.80%) of the acquired occurrence data for the selected true mangrove species were located within mapped mangrove stands. Mangrove occurrences outside mangrove areas could be because of an incorrect, but accurate and precise, location being provided. For example, for observations from preserved specimens, geographic locations of herbariums where the specimens are stored, instead of locations of specimen collection, were apparently provided for a few of the entries examined in this study (e.g., an *A. marina* stored at Fairchild Tropical

Botanic Garden, United States, with recorded coordinates of the herbarium). For observations from direct observation, such error can be caused by the observer's location instead of the species' location being recorded (Moudrý and Devillers 2020), albeit in theory this should be relatively uncommon for occurrence data of plants where observations are less likely to be made at a far distance from the trees than highly mobile animals. Information such as the basis of record could help identify the source of this type of dubious observation, but this information is sometimes not meaningful. While the field BasisOfRecord is required for data submission to both GBIF and OBIS, it was not clearly specified (was reported with mention "Unknown") in about 4% of the retained mangrove occurrences from GBIF (*see* Appendix 10), while it was properly reported in OBIS for all entries. This might be explained by the additional quality checks performed for OBIS data, for important data fields (OBIS 2021); entries containing fields with dubious or missing values could more likely be identified, thus resulting in observation data with fewer issues related to information availability than observed for GBIF.

On the other hand, true occurrences of mangroves could have been excluded in the curation process in this study. Mangrove trees grow within the intertidal zone, at the intersection of the terrestrial and marine realms, therefore identifying occurrences outside mangrove habitats is not trivial, as application of location-related filters, such as excluding entries flagged with "ON_LAND" for marine species, would not be suitable for mangrove species, especially for those located on the landward edge of mangrove forests. Mangrove ecosystem distribution maps can serve as a useful tool for this purpose and was the approach chosen in this study; however, these maps are often derived using remote sensing approaches of varying resolutions (e.g., Spalding et al. 2010, Worthington et al. 2020, Bunting et al. 2022), which come with known limitations (Kuenzer et al. 2011): mapping accuracy decreases at areas where tree density is low, at narrow or complex landscapes; areas with high disturbances may not be mapped accurately (Ferreira et al. 2009, Bunting et al. 2018); there are known identified data gaps in these maps (Bunting et al. 2018). Inaccuracies (including both under- and overestimation of mangrove coverage) of some of these large-scale mangrove maps (e.g., Bunting et al. 2018) have been detected on the local scale, using higher-resolution satellite or drone images (Hsu et al. 2020). Moreover, as these maps are not in real-time, occurrences from newly established mangrove patches may be mistaken as observations outside a species' range. While the use of a combination of multiple maps from different years could help derive the maximal mangrove extent, creating buffers around mangrove patches would help tackle map resolution issues to a certain extent, those procedures can be demanding in terms of time and computational effort and could result in overlooking erroneous entries. Therefore, the recommendation would be, in case of data usage for species or trait distribution modeling purposes, to define different sets of data, of high versus moderate reliability, based on the distance to the nearest mangrove patch and build models based on the distinct datasets separately to evaluate potential bias.

Occurrence data may also include observations within suitable habitats, but outside the species' known distribution range, especially in the context of climate change leading to species range shifts. In the current study, some observations located within known mangrove stands were found outside the species' documented natural distribution (using e.g., the distribution range provided by the IUCN red list assessments). On the other hand, *R. mangle*, a species from the Atlantic-East Pacific

(AEP), was observed in the Indo-West Pacific (IWP) with one isolated observation recorded in the Indian ocean (Fig. 3). Besides erroneous input of coordinates, these occurrences could also be due to incorrect species identification, or observations related to introduced individuals or newly colonizing (potentially invasive, but not necessarily) species with recently settled populations outside documented natural ranges. Identifying and correcting taxonomic errors would be possible for occurrences derived from preserved materials where specimens can be re-examined but is challenging for human-observed occurrences unless evidence (such as photos) is provided with the occurrence data. In either way, expertise and extensive knowledge of individual mangrove species are required, and such work would be time intensive. Introduced or naturally newly established (invasive or not) populations have been reported for several mangrove species (*Lg. racemosa*, Cheng et al. 2023; *B. gymnorrhiza*, Fourqurean et al. 2010; *S. apetala*, Zhang et al. 2022). For biodiversity data available on databases such as GBIF and OBIS, several Darwin Core standard attributes (such as "degreeOfEstablishment" and "establishmentMeans") allow data providers to provide information on whether the observations were obtained from native or nonnative individuals. We did not quantitatively assess this in the current study, since information is not always provided for these attributes. Thus, it would currently be challenging to identify occurrences with observations from individuals in non-native locations, unless using external reference maps and/or with careful manual inspection of the data by the data users. Dropping these entries with locations far outside known distribution ranges would be a more conservative approach in curating occurrence data by data users.

Unbalanced Taxonomic and Geographical Coverage.—About one-fifth (nine species; 24.32%) of the 37 selected species with occurrence data were represented by only ten entries or less (Fig. 1). Deficiencies of occurrence data in biodiversity databases were previously highlighted by Yesson et al. (2007) and Enquist et al. (2019), showing that a lot of terrestrial plant species were only represented by ten observations or less. Further, suspected data deficient areas were also noticed for true mangrove species represented by high numbers of observations. For instance, *A. marina*, the species with the highest number of entries in the retained occurrence data (with a majority of data coming from Australia), was underrepresented in large parts of its native range in southeast Asia (Fig. 4), where one of the major biodiversity hotspots is located (Myers et al. 2000). Poor data coverage was also found in several countries with the world's largest national mangrove areas, particularly in Indonesia (Figs. 5, 6E). Geographic bias within GBIF data has also been reported for occurrence data of legumes (Yesson et al. 2007) where known hotspots of biodiversity in Africa and Asia were found to be data deficient. This uneven contribution of occurrence data, even when a high number of occurrences is available, should be taken with caution when using these data for distribution modeling. Using the curated occurrence data (both for the point- or raster-based extraction) and known species distribution (IUCN Red List documented distribution) to extract climatic data for *A. marina* and *R. mangle* resulted in distinct distributions of the species along those climatic variables, suggesting that some portions of the climatic niche of the species could be over- or underrepresented in the biodiversity databases (Fig. 7). Spatially biased data due to uneven sampling or data contribution efforts, or to duplicated data, could result in deriving erroneous species-environment relationships as

environmental predictors that are representative to only the areas of higher density of observations instead of the species' observed distribution range, could be selected during the modeling process (Kramer-Schadt et al. 2013). For example, the climatic characteristics of northeastern Australia could be overrepresented when using uncorrected observation data from GBIF and OBIS when building a species distribution model for *A. marina*. We observed a spatial bias, for both the point- and raster-based extraction approaches, suggesting that both occurrence-based or abundance-based models would be affected by this uneven data contribution effort (Fig. 7). These derived erroneous species-environment relationships will lead to biased assessment of the climatic niche of species and potential erroneous predictions of their future distribution under climate change scenarios, or of the current habitat suitability for mangrove forest (re-)establishment.

The current data also failed to cover known exotic populations of some of the true mangroves that have been introduced outside their native range. For example *Lg. racemosa* was introduced in China in 1999 for mangrove afforestation (Gu et al. 2019), where it is suggested to be potentially highly invasive (Cheng et al. 2023). Yet, no occurrences within the known species' introduced range in China were found in this study, even considering the data that were excluded during the curation process. This suggests that occurrence data of known nonnative populations of true mangroves is not well-documented currently in the two biodiversity databases investigated in this study. Poor data coverage could be an indicator of insufficient research resources and efforts. Alternatively, such data exists but is not yet digitized, published, or contributed into biodiversity databases. It is estimated that only one-tenth of biocollections are available in digital form (Ball-Damerow et al. 2019), and only a limited number of herbaria have provided data to GBIF (Yesson et al. 2007).

Other Issues and Limitations.—Missing values in one or more key attributes (e.g date, basis of record, country) and heterogeneity in attribute values reduce the findability, interoperability and ultimately reusability of data (i.e., are not following the FAIR data principles; Wilkinson et al. 2016, Chapman et al 2020). Erroneous occurrence data could impact spatial analyses of species or traits distribution. While duplicated entries may not be related to data quality, duplicates could influence assessment on completeness and data coverage of global databases (Moudrý and Devillers 2020). Erroneous georeference and taxonomic identification can lead to an inaccurate estimation of species richness, which could not be relieved by increasing spatial scales (Maldonado et al. 2015).

In addition, species occurrences from biodiversity databases are often used, and are essential, to derive models for the current distribution of species or species habitat suitability maps that can be used for management plans [e.g., (re-)establishment], or to infer future distributions under climatic (or other) scenarios (e.g., Fazlioglu et al. 2020, Samal et al. 2023). Species observation data from these databases are often used for extracting environmental variable values, which is essential for modeling species-environment relationships (Coro et al. 2024).

Many species distribution models (SDMs) are sensitive to sample size (Wang and Jackson 2023), spatial sampling biases (Kramer-Schadt et al. 2013) and locational errors (Graham et al. 2008). Geospatially and taxonomically erroneous occurrence data, together with duplicated data, uneven sampling or data contribution effort, could lead to the incorrect delineation of the ecological niche of a species. As

a result, these issues could lead to inaccurate models of current distribution and predicted future range shift of species, which could further mislead conservation and management decisions.

Outlook and Recommendations.—GBIF and OBIF, two of the largest biodiversity databases, provide a great amount of occurrence data of various species, including true mangroves, available for public use. While the two are extremely valuable observation databases and constant efforts have been put into improving data availability and quality, they still contain data that are (1) possible duplicates, (2) lacking information for important attributes (e.g., basisOfRecord, year, country or countryCode), (3) erroneous in respect of their geographic location, and suffer from (4) deficient data coverage.

Hence, data users should take great caution when attempting to use mangrove occurrence data from biodiversity databases. Cleansing will be essential to exclude problematic occurrences, and there are several tools (e.g., BioGeo, Robertson et al. 2016; CoordinateCleaner, Zizka et al. 2019) and workflows (BDcleaner, Jin and Yang 2020) developed for this purpose. However, these data cleansing and correcting procedures should be further customized according to the group and geographic area under focus (Zizka et al. 2020), and to sample sizes (Kramer-Schadt et al 2013).

To bring biodiversity data more in line with the FAIR principles, additional input and effort from different stakeholders will be needed. For database managers, in order to improve data interoperability and reusability, especially between collaborating databases which practice data sharing, across-database globally unique identifiers could be implemented to make identification of shared data less complicated, as suggested by Guralnick et al. (2015). Enhancing the enforcement of mandatory or key attributes could help identify entries with other missing information during the data quality assessment process. To reduce data heterogeneity and in turn increase data interoperability, a universal and mandatory standardized list of attribute values (ontologies or "vocabularies of values"; Chapman et al. 2020) could be maintained. For some Darwin Core terms (e.g., degreeOfEstablishment), suggested or mandated lists of controlled vocabulary are currently available, but this is not the case for some attributes. Database managers should also maintain a closer communication with scientific experts (such as the IUCN mangrove specialist group in this specific case) to improve the data curation and data quality assurance processes for biodiversity monitoring and biodiversity research, as the experts could provide valuable, up-to-date knowledge on validity of species observations regarding taxonomy- and location-related issues. This might in turn also be beneficial for the researchers in case of occurrences found outside the known documented species range, which if proven valid, would be important to update the distribution range in a timely manner. With the rapid expansion of these biodiversity databases and the enormous amount of data of varying quality they hold, further automation of data validation using current distribution maps (e.g., datasets from United Nations Environment Programme World Conservation Monitoring Centre and IUCN Red List assessment distribution information as used in this case study on true mangroves) would help quickly identify observations that are potentially unreliable spatially or taxonomically, or those from nonnative populations, based on current knowledge. Such automated validation processes will be even more relevant when occurrence data derived from

eDNA biomonitoring will be integrated in the global biodiversity databases such as GBIF (Berry et al. 2021), a process currently in implementation.

Finally, while FAIR data principles should be common practice nowadays, there are still too many valuable observations, some of decent quality, that are deposited in private or national repositories and not shared with the global community. Notably, there are still areas where data is deficient, including many of the world's biodiversity hotspots. In the context of global challenges (biodiversity, climate, and pollution crises) efforts to document biodiversity information should be global. Database managers have a crucial role to play, notably by increasing the network of collaborating databases for data sharing, including various regional or local databases, thereby ensuring that biodiversity information will have better coverage and greater accessibility at the global scale. Development of application programming interface (API) services for biodiversity database networks (Sterner et al. 2020) is one possible solution to facilitate sharing and synchronization of biodiversity data. Standardization of data format, such as the implementation of the DwC standard, will be pivotal for efficient data synchronization. Increased resources will be needed to collect, digitalize, and publish data, which will require collective efforts from researchers across different fields of research. We encourage the mangrove research community to be proactive in this domain.

Acknowledgments

We express our thanks to Martin Zimmer for his valuable suggestions on species selection and support throughout this study. Thanks to the two anonymous reviewers for their valuable comments and suggestions that contributed to improving the manuscript.

Literature Cited

Amaral C, Poulter B, Lagomasino D, Fatoyinbo T, Taillie P, Lizcano G, Canty S, Silveira JAH, Teutli-Hernández C, Cifuentes-Jara M, et al. 2023. Drivers of mangrove vulnerability and resilience to tropical cyclones in the North Atlantic Basin. Sci Total Environ. 898:165413. https://doi.org/10.1016/j.scitotenv.2023.165413

Baeten L, Bruelheide H, van der Plas F, Kambach S, Ratcliffe S, Jucker T, Allan E, Ampoorter E, Barbaro L, Bastias CC, et al. 2019. Identifying the tree species compositions that maximize ecosystem functioning in European forests. J Appl Ecol. 56(3):733–744. https://doi.org/10.1111/1365-2664.13308

Ball-Damerow JE, Brenskelle L, Barve N, Soltis PS, Sierwald P, Bieler R, LaFrance R, Ariño AH, Guralnick RP. 2019. Research applications of primary biodiversity databases in the digital age. PLOS ONE. 14(9):e0215794. https://doi.org/10.1371/journal.pone.0215794

Beck J, Böller M, Erhardt A, Schwanghart W. 2014. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. Ecol Inform. 19:10–15. https://doi.org/10.1016/j.ecoinf.2013.11.002

Berry O, Jarman S, Bissett A, Hope M, Paeper C, Bessey C, Schwartz MK, Hale J, Bunce M. 2021. Making environmental DNA (eDNA) biodiversity records globally accessible. Environmental DNA. 3(4):699–705. https://doi.org/10.1002/edn3.173

Bunting P, Rosenqvist A, Hilarides L, Lucas RM, Thomas N, Tadono T, Worthington TA, Spalding M, Murray NJ, Rebelo L-M. 2022. Global mangrove extent change 1996–2020: Global Mangrove Watch Version 3.0. Remote Sens. 14(15):3657. https://doi.org/10.3390/rs14153657

Bunting P, Rosenqvist A, Lucas RM, Rebelo L-M, Hilarides L, Thomas N, Hardy A, Itoh T, Shimada M, Finlayson CM. 2018. The Global Mangrove Watch—a new 2010 global baseline of mangrove extent. Remote Sens. 10(10):1669. https://doi.org/10.3390/rs10101669

Chapman AD, Belbin L, Zermoglio PF, Wieczorek J, Morris PJ, Nicholls M, Rees ER, Veiga AK, Thompson A, Saraiva AM, et al. 2020. Developing standards for improved data quality and for selecting fit for use biodiversity data. Biodivers Inf Sci Stand. 4:e50889. https://doi.org/10.3897/biss.4.50889

Cheng C, Ke X, Lang T, Zhong C, Lv X, Zhang M, Yu C, Fang Z, Zhou H, Chen Y. 2023. Current status and potential invasiveness evaluation of an exotic mangrove species, *Laguncularia racemosa* (L.) C.F. Gaertn, on Hainan Island, China. Forests. 14(10):2036. https://doi.org/10.3390/f14102036

Colli-Silva M, Reginato M, Cabral A, Forzza RC, Pirani JR, Vasconcelos TN. 2020. Evaluating shortfalls and spatial accuracy of biodiversity documentation in the Atlantic Forest, the most diverse and threatened Brazilian phytogeographic domain. Taxon. 69(3):567–577. https://doi.org/10.1002/tax.12239

Coro G, Sana L, Bove P. 2024. An open science automatic workflow for multi-model species distribution estimation. International Journal of Data Science and Analytics. https://doi.org/10.1007/s41060-024-00517-w

Costello MJ, Appeltans W. 2008. Taxonomic editors plan a World Register of Marine Species (WoRMS). MarBEF Newsletter. 8:36–37.

Costello MJ, Stocks K, Zhang Y, Grassle FJ, Fautin DG, 2007. About the Ocean Biogeographic Information System. Available from: http://www.iobis.org/about/. [Accessed on 12 August 2024].

Dahdouh-Guebas F, editor. World Mangroves database. 2023. Accessed 13 September, 2023. Available from: https://doi.org/10.14284/460

Di Cecco GJ, Barve V, Belitz MW, Stucky BJ, Guralnick RP, Hurlbert AH. 2021. Observing the observers: How participants contribute data to iNaturalist and implications for biodiversity science. BioScience 71(11):1179–88.https://doi.org/10.1093/biosci/biab093

Duke N, Kathiresan K, Salmo III SG, Fernando ES, Peras JR, Sukardjo S, Miyagi T, Ellison J, Koedam NE, Wang Y, et al. 2010. *Avicennia marina*. The IUCN Red List of Threatened Species 2010: e.T178828A7619457. Accessed 9 January, 2024. https://doi.org/10.2305/IUCN.UK.2010-2.RLTS.T178828A7619457.en

Duncan C, Owen HJF, Thompson JR, Koldewey HJ, Primavera JH, Pettorelli N. 2018. Satellite remote sensing to monitor mangrove forest resilience and resistance to sea level rise. Methods Ecol Evol. 9(8):1837–1852. https://doi.org/10.1111/2041-210X.12923

Ellison A, Farnsworth E, Moore G. 2015. *Rhizophora mangle*. The IUCN Red List of Threatened Species 2015: e.T178851A69024847. Accessed 9 January, 2024. https://doi.org/10.2305/IUCN.UK.2015-1.RLTS.T178851A69024847.en

Enquist BJ, Feng X, Boyle B, Maitner B, Newman EA, Jørgensen PM, Roehrdanz PR, Thiers BM, Burger JR, Corlett RT, et al. 2019. The commonness of rarity: global and future distribution of rarity across land plants. Sci Adv. 5(11):eaaz0414. https://doi.org/10.1126/sciadv.aaz0414

FAO. 2005. Food and Agriculture Organization of the United Nations. Indonesia country profile. *In*: Global forest resources assessment 2005 thematic study on mangroves. Rome, Italy: FAO Forestry Department. Available from: https://www.fao.org/forestry/9010-0a3439fa97f25579ad64b3d2935d63650.pdf

Fazlioglu F, Wan JSH, Chen L. 2020. Latitudinal shifts in mangrove species worldwide: evidence from historical occurrence records. Hydrobiologia. 847(19):4111–4123. https://doi.org/10.1007/s10750-020-04403-x

Feeley K. 2015. Are we filling the data void? An assessment of the amount and extent of plant collection records and census data available for tropical South America. PLOS ONE. 10(4):e0125629. https://doi.org/10.1371/journal.pone.0125629

Feng X, Enquist BJ, Park DS, Boyle B, Breshears DD, Gallagher RV, Lien A, Newman EA, Burger JR, Maitner BS, et al. 2022. A review of the heterogeneous landscape of biodiversity

databases: opportunities and challenges for a synthesized biodiversity knowledge base. Glob Ecol Biogeogr. 31(7):1242–1260. https://doi.org/10.1111/geb.13497

Ferreira MA, Andrade F, Bandeira SO, Cardoso P, Mendes RN, Paula J. 2009. Analysis of cover change (1995–2005) of Tanzania/Mozambique trans-boundary mangroves using Landsat imagery. Aquat Conserv Mar Freshwater Ecosyst. 19(S1):S38–S45. https://doi.org/10.1002/aqc.1042

Fick SE, Hijmans RJ. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int J Climatol. 37(12):4302–4315. https://doi.org/10.1002/joc.5086

Fourqurean JW, Smith TJ, Possley J, Collins TM, Lee D, Namoff S. 2010. Are mangroves in the tropical Atlantic ripe for invasion? Exotic mangrove trees in the forests of South Florida. Biol Invasions 12:2509–22. https://doi.org/10.1007/s10530-009-9660-8

Freitas TMS, Montag LFA, De Marco P, Hortal J. 2020. How reliable are species identifications in biodiversity big data? Evaluating the records of a neotropical fish family in online repositories. Syst Biodivers. 18(2):181–191. https://doi.org/10.1080/14772000.2020.1730473

Frenette-Dussault C, Shipley B, Meziane D, Hingrat Y. 2013. Trait-based climate change predictions of plant community structure in arid steppes. J Ecol. 101(2):484–492. https://doi.org/10.1111/1365-2745.12040

Gaiji S, Chavan V, Ariño AH, Otegui J, Hobern D, Sood R, Robles E. 2013. Content assessment of the primary biodiversity data published through GBIF network: status, challenges and potentials. Biodivers Inform. 8(2). https://doi.org/10.17161/bi.v8i2.4124

Garcia-Rosello E, Gonzalez-Dacosta J, Guisande C, Lobo JM. 2023. GBIF falls short of providing a representative picture of the global distribution of insects. Syst Entomol. 48(4):489–497. https://doi.org/10.1111/syen.12589

Gouvêa LP, Serrão EA, Cavanaugh K, Gurgel CFD, Horta PA, Assis J. 2022. Global impacts of projected climate changes on the extent and aboveground biomass of mangrove forests. Divers Distrib. 28(11):2349–2360. https://doi.org/10.1111/ddi.13631

Graham CH, Elith J, Hijmans RJ, Guisan A, Townsend Peterson A, Loiselle BA, NCEAS Predicting Species Distributions Working Group. 2008. The influence of spatial errors in species occurrence data used in distribution models. J Appl Ecol. 45(1):239–247. https://doi.org/10.1111/j.1365-2664.2007.01408.x

Green SJ, Brookson CB, Hardy NA, Crowder LB. 2022. Trait-based approaches to global change ecology: moving from description to prediction. Proc R Soc Lond B Biol Sci. 289:20220071. https://doi.org/10.1098/rspb.2022.0071

Gu X, Feng H, Tang T, Tam NF-Y, Pan H, Zhu Q, Dong Y, Fazlioglu F, Chen L. 2019. Predicting the invasive potential of a non-native mangrove reforested plant (*Laguncularia racemosa*) in China. Ecol Eng. 139:105591. https://doi.org/10.1016/j.ecoleng.2019.105591

Guralnick RP, Cellinese N, Deck J, Pyle RL, Kunze J, Penev L, Walls R, Hagedorn G, Agosti D, Wieczorek J, et al. 2015. Community next steps for making globally unique identifiers work for biocollections data. ZooKeys. 494:133–154. https://doi.org/10.3897/zookeys.494.9352

Hamilton SE, Casey D. 2016. Creation of a high spatio-temporal resolution global database of continuous mangrove forest cover for the 21st century (CGMFC-21). Glob Ecol Biogeogr. 25(6):729–738. https://doi.org/10.1111/geb.12449

Hannah L, Midgley G, Andelman S, Araújo M, Hughes G, Martinez-Meyer E, Pearson R, Williams P. 2007. Protected area needs in a changing climate. Front Ecol Environ. 5(3):131–138. https://doi.org/10.1890/1540-9295(2007)5[131:PANIAC]2.0.CO;2

Heberling JM, Miller JT, Noesgaard D, Weingart SB, Schigel D. 2021. Data integration enables global biodiversity synthesis. Proc Natl Acad Sci USA. 118(6):e2018093118. https://doi.org/10.1073/pnas.2018093118

Helfer V, Zimmer M. 2018. High-throughput techniques as support for knowledge-based spatial conservation prioritization in mangrove ecosystems. *In*: Makowski C, Finkl CW, editors. Threats to mangrove forests: hazards, vulnerability, and management. Cham: Springer International Publishing.

Horton T, Kroh A, Ahyong S, Bailly N, Bieler R, Boyko CB, Brandão SN, Gofas S, Hooper JNA, Hernandez F, et al. 2021. World Register of Marine Species. Available from http://www.marinespecies.org at VLIZ. Accessed 2021-09-25. https://doi.org/10.14284/170

Hsu AJ, Kumagai J, Favoretto F, Dorian J, Guerrero Martinez B, Aburto-Oropeza O. 2020. Driven by drones: improving mangrove extent maps using high-resolution remote sensing. Remote Sens. 12(23):3986. https://doi.org/10.3390/rs12233986

Jin J, Yang J. 2020. BDcleaner: A workflow for cleaning taxonomic and geographic errors in occurrence data archived in biodiversity databases. Glob Ecol Conserv. 21:e00852. https://doi.org/10.1016/j.gecco.2019.e00852

Kattge J, Bönisch G, Díaz S, Lavorel S, Prentice IC, Leadley P, Tautenhahn S, Werner GD, Aakala T, Abedi M, et al. 2020. TRY plant trait database–enhanced coverage and open access. Glob Change Biol. 26(1):119–188. https://doi.org/10.1111/gcb.14904

Kramer-Schadt S, Niedballa J, Pilgrim JD, Schröder B, Lindenborn J, Reinfelder V, Stillfried M, Heckmann I, Scharf AK, Augeri DM, et al. 2013. The importance of correcting for sampling bias in MaxEnt species distribution models. Divers Distrib. 19(11):1366–1379. https://doi.org/10.1111/ddi.12096

Kuenzer C, Bluemel A, Gebhardt S, Quoc TV, Dech S. 2011. Remote sensing of mangrove ecosystems: A review. Remote Sens. 3(5):878–928. https://doi.org/10.3390/rs3050878

Li X, Wen Y, Chen X, Qie Y, Cao KF, Wee AK. 2022. Correlations between photosynthetic heat tolerance and leaf anatomy and climatic niche in Asian mangrove trees. Plant Biol. 24(6):960–6. https://doi.org/10.1111/plb.13460

Luo Y-H, Cadotte MW, Burgess KS, Liu J, Tan S-L, Zou J-Y, Xu K, Li D-Z, Gao L-M. 2019. Greater than the sum of the parts: how the species composition in different forest strata influence ecosystem function. Ecol Lett. 22(9):1449–1461. https://doi.org/10.1111/ele.13330

Maldonado C, Molina CI, Zizka A, Persson C, Taylor CM, Albán J, Chilquillo E, Rønsted N, Antonelli A. 2015. Estimating species diversity and distribution in the era of Big Data: to what extent can we trust public databases? Glob Ecol Biogeogr. 24(8):973–984. https://doi.org/10.1111/geb.12326

McLeod E, Salm R, Green A, Almany J. 2009. Designing marine protected area networks to address the impacts of climate change. Front Ecol Environ. 7(7):362–370. https://doi.org/10.1890/070211

Mesibov R. 2013. A specialist's audit of aggregated occurrence records. ZooKeys. 293:1–18. https://doi.org/10.3897/zookeys.293.5111

Missouri Botanical Garden. 2023. Tropicos. Accessed 13 September, 2023. Available from: https://tropicos.org

Moor H, Hylander K, Norberg J. 2015. Predicting climate change effects on wetland ecosystem services using species distribution modeling and plant functional traits. Ambio. 44(S1):113–126. https://doi.org/10.1007/s13280-014-0593-9

Moudrý V, Devillers R. 2020. Quality and usability challenges of global marine biodiversity databases: an example for marine mammal data. Ecol Inform. 56:101051. https://doi.org/10.1016/j.ecoinf.2020.101051

Myers N, Mittermeier RA, Mittermeier CG, da Fonseca GAB, Kent J. 2000. Biodiversity hotspots for conservation priorities. Nature. 403(6772):853–858. https://doi.org/10.1038/35002501

OBIS. 2021. Ocean Biodiversity Information System. Intergovernmental Oceanographic Commission of UNESCO. Accessed 13 December, 2021.

Posit Team. 2023. RStudio: integrated development environment for R. v2023.6.2.561. Posit Software. Boston, Massachusetts: PBC. http://www/posit.co/

Provoost P, Bosch S, Appeltans W. OBIS. 2022. robis: Ocean Biodiversity Information System (OBIS) Client. Accessed 25 November, 2023. Available from: https://cran.r-project.org/web/packages/robis/index.html

QGIS 2023. QGIS Geographic Information System. QGIS Association. Available from: http://www.qgis.org

Quadros AF, Zimmer M. 2017. Dataset of "true mangroves" plant species traits. Biodivers Data J. 5:e22089. https://doi.org/10.3897/BDJ.5.e22089

Queiroz N, Humphries NE, Couto A, Vedor M, da Costa I, Sequeira AMM, Mucientes G, Santos AM, Abascal FJ, Abercrombie DL, et al. 2021. Reply to: caution over the use of ecological big data for conservation. Nature. 595:E20–E28. https://doi.org/10.1038/s41586-021-03464-9

R Core Team. 2023. R: a language and environment for statistical computing. v4.3.1. Vienna, Austria: R Foundation for Statistical Computing.

Ribeiro BR, Guidoni-Martins K, Tessarolo G, Velazco SJE, Jardim L, Bachman SP, Loyola R. 2022. Issues with species occurrence data and their impact on extinction risk assessments. Biol Conserv. 273:109674. https://doi.org/10.1016/j.biocon.2022.109674

Robertson MP, Visser V, Hui C. 2016. Biogeo: an R package for assessing and improving data quality of occurrence record datasets. Ecography. 39(4):394–401. https://doi.org/10.1111/ecog.02118

Rodríguez JP, Brotons L, Bustamante J, Seoane J. 2007. The application of predictive modelling of species distribution to biodiversity conservation. Divers Distrib. 13(3):243–251.

Saenger P, Ragavan P, Sheue CR, López-Portillo J, Yong JW, Mageswaran T. 2019. Mangrove biogeography of the Indo-Pacific. *In:* Gul B, Böer B, Khan M, Clüsener-Godt M, Hameed A, editors. Sabkha ecosystems. Tasks for vegetation science. vol 49. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-030-04417-6_23

Samal P, Srivastava J, Charles B, Singarasubramanian SR. 2023. Species distribution models to predict the potential niche shift and priority conservation areas for mangroves (*Rhizophora apiculata, R. mucronata*) in response to climate and sea level fluctuations along coastal India. Ecol Indic. 154:110631. https://doi.org/10.1016/j.ecolind.2023.110631

Sarker SK, Matthiopoulos J, Mitchell SN, Ahmed ZU, Mamun Md BA, Reeve R. 2019a. 1980s–2010s: The world's largest mangrove ecosystem is becoming homogeneous. Biol Conserv. 236:79–91. https://doi.org/10.1016/j.biocon.2019.05.011

Sarker SK, Reeve R, Paul NK, Matthiopoulos J. 2019b. Modelling spatial biodiversity in the world's largest mangrove ecosystem—The Bangladesh Sundarbans: a baseline for conservation. Divers Distrib. 25(5):729–742. https://doi.org/10.1111/ddi.12887

Sarker SK, Reeve R, Thompson J, Paul NK, Matthiopoulos J. 2016. Are we failing to protect threatened mangroves in the Sundarbans world heritage ecosystem? Sci Rep. 6(1):21234. https://doi.org/10.1038/srep21234

Serra-Diaz JM, Enquist BJ, Maitner B, Merow C, Svenning JC. 2017. Big data of tree species distributions: how big and how good? For Ecosyst. 4:1–2. https://doi.org/10.1186/s40663-017-0120-0

Spalding M, Kainuma M, Collins L. 2010. World atlas of mangroves (v3.1). A collaborative project of ITTO, ISME, FAO, UNEP-WCMC, UNESCO-MAB, UNU-INWEH and TNC. London (UK): Earthscan, London. 319 p. https://doi.org/10.34892/w2ew-m835

Sterner BW, Gilbert EE, Franz NM. 2020. Decentralized but globally coordinated biodiversity data. Front. big. Data (Basel). 3:519133. https://doi.org/10.3389/fdata.2020.519133

Symstad AJ, Tilman D, Willson J, Knops JMH. 1998. Species loss and ecosystem functioning: effects of species identity and community composition. Oikos. 81(2):389–397. https://doi.org/10.2307/3547058

Tomlinson PB. 2016. The botany of mangroves. 2nd edition. New York: Cambridge University Press.

Troudet J, Grandcolas P, Blin A, Vignes-Lebbe R, Legendre F. 2017. Taxonomic bias in biodiversity data and societal preferences. Sci Rep. 7(1):9132. https://doi.org/10.1038/s41598-017-09084-6

Turnhout E, Boonman-Berson S. 2011. Databases, scaling practices, and the globalization of biodiversity. Ecol Soc. 16(1):35. https://doi.org/10.5751/ES-03981-160135

Waldock C, Stuart-Smith RD, Albouy C, Cheung WW, Edgar GJ, Mouillot D, Tjiputra J, Pellissier L. 2022. A quantitative review of abundance-based species distribution models. Ecography. 2022(1):e05694. https://doi.org/10.1111/ecog.05694

Vandepitte L, Bosch S, Tyberghein L, Waumans F, Vanhoorne B, Hernandez F, De Clerck O, Mees J. 2015. Fishing for data and sorting the catch: assessing the data quality, completeness and fitness for use of data in marine biogeographic databases. Database 2015: bau125. https://doi.org/10.1093/database/bau125

Wang L, Jackson DA. 2023. Effects of sample size, data quality, and species response in environmental space on modeling species distributions. Landscape Ecol. 38(12):4009–31.

Weigelt P, König C, Kreft H. 2020. GIFT–A global inventory of floras and traits for macroecology and biogeography. J Biogeogr. 47(1):16–43. https://doi.org/10.1111/jbi.13623

Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D. 2012. Darwin Core: an evolving community-developed biodiversity data standard. PLOS ONE. 7(1):e29715. https://doi.org/10.1371/journal.pone.0029715

Wilkinson MD, Dumontier M, Aalbersberg Ij J, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 3(1):160018. https://doi.org/10.1038/sdata.2016.18

Worthington TA, zu Ermgassen PSE, Friess DA, Krauss KW, Lovelock CE, Thorley J, Tingey R, Woodroffe CD, Bunting P, Cormier N, et al. 2020. A global biophysical typology of mangroves and its relevance for ecosystem structure and deforestation. Sci Rep. 10(1):14652. https://doi.org/10.1038/s41598-020-71194-5

Yesson C, Brewer PW, Sutton T, Caithness N, Pahwa JS, Burgess M, Gray WA, White RJ, Jones AC, Bisby FA, Culham A. 2007. How global is the global biodiversity information facility? PLOS ONE. 2(11):e1124. https://doi.org/10.1371/journal.pone.0001124

Zellmer AJ, Claisse JT, Williams CM, Schwab S, Pondella DJ. 2019. Predicting optimal sites for ecosystem restoration using stacked-species distribution modeling. Front Mar Sci. 6:3. https://doi.org/10.3389/fmars.2019.00003

Zhang J, Lin Q, Peng Y, Pan L, Chen Y, Zhang Y, Chen L. 2022. Distributions of the non-native mangrove *Sonneratia apetala* in China: based on Google Earth imagery and field survey. Wetlands. 42(5):35. https://doi.org/10.1007/s13157-022-01556-4

Zizka A, Carvalho FA, Calvente A, Baez-Lizarazo MR, Cabral A, Coelho JFR, Colli-Silva M, Fantinati MR, Fernandes MF, Ferreira-Araújo T, et al. 2020. No one-size-fits-all solution to clean GBIF. PeerJ. 8:e9916. https://doi.org/10.7717/peerj.9916

Zizka A, Silvestro D, Andermann T, Azevedo J, Duarte Ritter C, Edler D, Farooq H, Herdean A, Ariza M, Scharn R, et al. 2019. CoordinateCleaner: standardized cleaning of occurrence records from biological collection databases. Methods Ecol Evol. 10(5):744–751. https://doi.org/10.1111/2041-210X.13152