

ECOGRAPHY

Software Note

'EcoCleanR': enhancing data quality of biogeographic ranges with application for marine invertebrates

Priyanka Soni¹✉, Austin J. W. Hendy² and David J. Bottjer¹

¹Department of Earth Sciences, University of Southern California, Los Angeles, CA, USA

²Natural History Museum of Los Angeles County, Los Angeles, CA, USA

Correspondence: Priyanka Soni (sonip@usc.edu)

Ecography

2026: e08203

doi: [10.1002/ecog.08203](https://doi.org/10.1002/ecog.08203)

Subject Editor:

Michael Krabbe Borregaard

Editor-in-Chief: Miguel Araújo

Accepted 3 January 2026



Published distribution data, while invaluable for understanding species' biogeography, often suffer from limitations such as dated and static representations of ranges, a bias toward latitudinal information, and lack of resolution in sampling frequency and variation in abundance throughout a species' distribution. Extensive open-source biodiversity data now allow us to construct biogeographic ranges with more modern observations, which can be useful in conservation, evolution, and ecological studies. However, data quality remains a persistent challenge, hampering data reliability and usability. We introduce 'EcoCleanR', an R package that integrates existing tools with new functionalities to address data integration and quality assessment through a systematic, step-by-step approach for marine occurrence data. This package enhances the process of identifying and resolving common issues in biodiversity data, including taxonomy and georeferencing errors. It provides: 1) example scripts to guide users, 2) functionalities to flag problematic occurrence records from multiple databases, and 3) outputs that include species-specific distribution ranges and their corresponding environmental conditions, to facilitate accurate biogeographic and ecological analyses.

Keywords: biogeography, conservation biology, population ecology

Introduction

Species-level geographic range data can be obtained from published books, scientific journals, surveys, expert knowledge, or unpublished museum sources among others. These ranges, which may be expressed as geographic areas, latitudinal limits, or specific environments, are derived from collections and observations spanning more than a century – hereafter referred to as historical ranges (Shoo et al. 2006, Tingley and Beissinger 2009, Lawlor et al. 2024). Reproducibility is frequently constrained by limited documentation of sources and methods, with many range estimates based on outdated observations. Several studies highlight the importance of these historical observations as invaluable baselines to monitor species range dynamics, extinction risk, and population isolation to propose future conservation strategies (Donlan



www.ecography.org

© 2026 The Author(s). Ecography published by John Wiley & Sons Ltd on behalf of Nordic Society Oikos

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

2005, Nogu   et al. 2017, Darrigran et al. 2023, Waqar et al. 2024). These outdated historical geographic ranges have also been used as modern distribution limits by paleontologists as proxies for determining paleoenvironmental and paleoceanographic conditions (Woodring and Popenoe 1945, Valentine 1973, Muhs et al. 2023). Although these historical observations provide valuable insights into species' latitudinal ranges, they have limitations for understanding biotic responses in a changing world. Most obviously, they represent a static range defined by two end points, lacking temporal resolution as well as detail variations in habitat preferences or spatial gaps across the range. Additionally, localities within these ranges may exhibit varying species abundance and frequency, suggesting some areas are more habitable than others or host established populations as opposed to ephemeral appearances. Lastly, traditional geographic ranges or latitudinal limits do not permit the exploration of environmental variables (e.g. salinity, productivity, or temperature) that might be responsible for limiting a species' distribution. Recognizing the importance of encompassing all available data points, whether published or unpublished, from diverse sources, there is a growing need to develop methods that offer a real-time perspective on species distribution. Such dynamic approaches are capable of refining species geographic ranges as new observations are added, surpassing the limitations of static ranges, providing a more comprehensive understanding of the spatial distribution of species. Incorporating environmental variables further enables the delineation of species' biogeographic ranges (Peterson et al. 2012, Lomolino et al. 2017).

Biodiversity tools have evolved to include advanced computational and informatic methods, allowing for better storage and easy access of vast amounts of data collected from natural history museums and citizen science observations across geological, historical, and more recent time scales (Bisby 2000, Graham et al. 2004, Sober  n and Peterson 2004, Heberling et al. 2021). To date, more than 3 billion occurrence records across taxa, time, and space have been digitized and made available through biodiversity aggregators such as the Global Biodiversity Information Facility (GBIF), Integrated Digitized Biocollections (iDigBio), and the Ocean Biodiversity Information System (OBIS). These aggregators have been used to model biogeographic ranges and identify rare, threatened, endemic, endangered and invasive species across various ecosystems, supporting the implementation of conservation efforts (Richardson et al. 2011, Novoa et al. 2015, Alhajer and Fourcade 2019, Panter et al. 2020, Gervazoni et al. 2023, Cordier et al. 2024). However, their benefits are often accompanied by challenges and limitations. Utilizing these online tools to define dynamic species ranges is not straightforward, as these records may be associated with contextual data of variable detail and quality, contain inaccuracies in geographic and taxonomic information, include erroneous entries, or are themselves the product of sampling biases, raising concerns about the reliability of research conducted using these data (Chapman 2005, Mesibov 2013, Meyer et al. 2016, Nic Lughadha et al. 2019, Zizka et al. 2019, Panter et al. 2020).

The most common type of error in spatial occurrence data points is geographic coordinate assignment. A best-practice workflow for maintaining the quality of georeferencing, as defined by Chapman (2005), Chapman and Wicczorek (2020) and Marcer et al. (2022), has been widely accepted to improve georeferencing processes. Over time, utilizing these standards have significantly enhanced the quality of georeferenced data. However, despite these improvements, errors can still occur due to missing or transposed coordinate information or the substitution of coordinates using the centroids of political units (e.g. cities, states, countries), generalized localities, or coordinate rounding and truncation. Even with accurate locality descriptions, challenges may arise from ambiguous or duplicate locality names that exist in different geographical regions, leading to misassigned coordinates. For example, in marine datasets, these errors can result in records plotted in the wrong ocean, points located offshore for species restricted to coastal habitats, or occurrences placed on land or in the continental interiors (e.g. Fig. 1; Yesson et al. 2007, Robertson 2008, Maldonado et al. 2015, Assis et al. 2020, Moudr  y and Devillers 2020, Mancinelli et al. 2021). Our analysis of the complete GBIF dataset (as of August 2025) reveals that only 33% of digitized records contain a minimally usable uncertainty value (defined here as within a 100 km radius from the actual observation). An uncertainty around the location of a specimen gives spatial precision and helps to estimate better suitable habitats (Smith et al. 2023). Historical collections from museums often lack this critical information, further limiting the reliability and accuracy of spatial data (Marcer et al. 2022).

Another common issue faced by biodiversity aggregators is taxonomy error. Biodiversity aggregators develop taxonomy backbones to connect accepted names, synonyms, and utilize the classifications of various data providers (e.g. World Register of Marine Species (WoRMS)) following globally recognized standards established by the Taxonomic Databases Working Group (TDWG). Such taxonomic data providers themselves are still being populated with both new and historic data, and conflicting taxonomic opinions, which can significantly impact biodiversity aggregators (Mesibov 2018, Feng et al. 2022). Despite extensive efforts to utilize a single authoritative list of global taxa, biodiversity aggregators continue to face persistent taxonomic challenges (Farley et al. 2018, Gren    et al. 2023, Sandall et al. 2023). Some common taxonomic errors include: name changes, where a species is reclassified under a new name; obsolete names, where outdated nomenclature persists in databases; and taxonomic splits, where a single species is divided into multiple distinct taxa. Furthermore, there are more complex issues that are difficult to address, such as observations being incorrectly tagged or misidentified as the species of interest (Vecchione et al. 2000, Crall et al. 2011, Fuccillo et al. 2015, Freitas et al. 2020).

Despite addressing georeferencing and taxonomic errors, challenges persist due to apparent spatial outliers (Chen et al. 2008, Aggarwal 2017). While obvious errors, such as points located on land or offshore for a coastal species, are easily

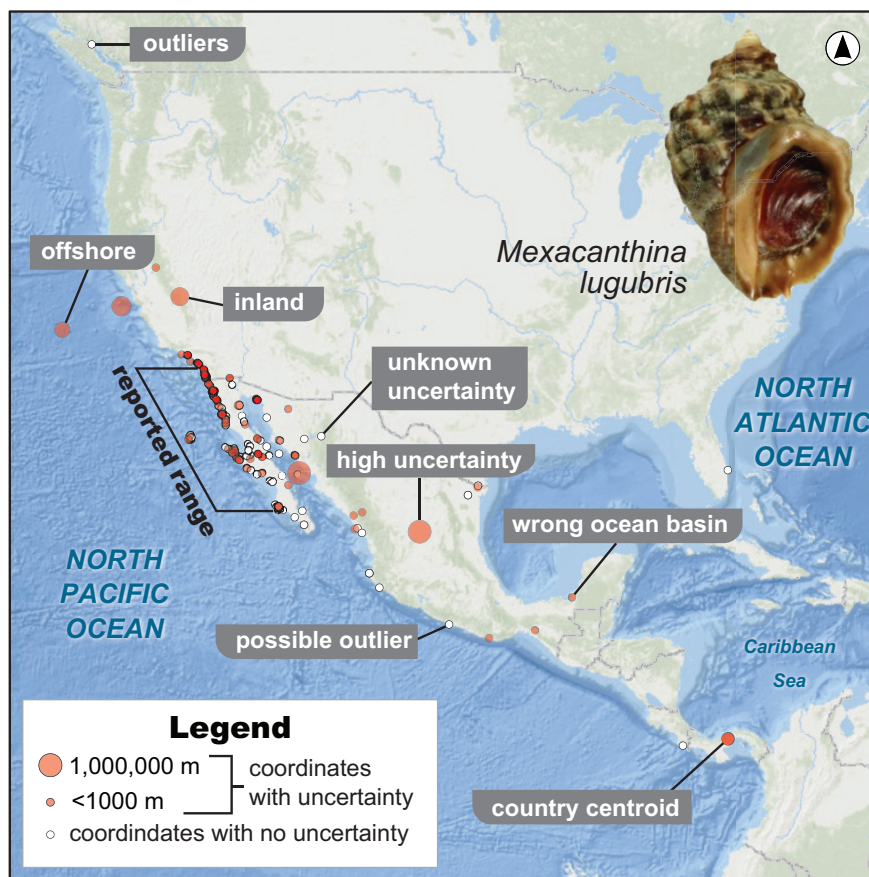


Figure 1. Examples of data quality issues that may impact digitally accessible occurrence data. Occurrence data points of *Mexacanthina lugubris* with known uncertainty error radius in red and occurrence points without uncertainty radius in white, extracted from various data sources (GBIF, iDigBio, OBIS and InvertEBase). Reported range of *M. lugubris* as published in Keen (1971).

detectable, identifying issues becomes more difficult when data points are geographically close to valid records, especially for species with small biogeographic ranges or relatively sparse sampling. Even after applying various data cleaning protocols, aggregator-derived data often contain outliers that can falsely extend species' range, contributing to erroneous inputs for conservation and management planning (Benhadi-Marín 2018).

Here, we present a user-friendly biodiversity package ('EcoCleanR') that standardizes and cleans marine species occurrence data to generate a dynamic habitat range that can evolve as new data are incorporated. Doing so provides the temporal granularity that allows the documentation of when and where a species range is expanding or contracting. We have built a workflow that utilizes new functions developed under 'EcoCleanR' and existing functions from 'CoordinateCleaner' (Zizka et al. 2019), enabling controlled execution of each step outlined in Table 1. A detailed, step-by-step description of the workflow is provided in package vignettes. The workflow can be adjusted based on the desired filters after data integration and provides biogeographic ranges as well as the most suitable environmental conditions for the species' habitability. 'EcoCleanR' has novel functionalities, most notably the identification of outliers based on

both spatial and non-spatial attributes and with a probability index to describe the extent to which a data point can be considered an outlier. This tool provides control to the user for making decisions to remove those datapoints.

Description of product

The package, 'EcoCleanR', is built on R (www.r-project.org) and provides a complete suite of automated functions and guidance on manual operations. The functions encompass data merging and cleaning, in addition to manual steps for improving the quality of georeferencing by assigning coordinates and uncertainty using GEOLocate (Rios and Bart 2010). This package doesn't provide a one-click automated solution. Instead, it provides flexibility to track the results at every step of data cleaning, generating a map view to visualize how the quality of data distribution changes following each iteration of the workflow.

Online documentation is available on GitHub (<https://github.com/sonipri/EcoCleanR>) and in CRAN (<https://CRAN.R-project.org/package=EcoCleanR>) with a step-by-step script and example datasets. The main features of the package are listed below.

Table 1. Software architecture and workflow of 'EcoCleanR' – including functions from 'CoordinateCleaner'. These steps can be executed as defined sequences or skipped/modified based on data processing requirements.

Steps	Algorithms/functions	Description	Package dependencies	Source package
Data extraction	Vignette: <i>data_merging</i>	Extract data from various sources: GBIF, iDigBio, OBIS local input data files	rgbif, ridigbio, robis	EcoCleanR
Data merging	<i>ec_db_merge</i>	Merge data from GBIF, iDigBio, OBIS, and local input data files	dplyr (Wickham et al. 2022)	EcoCleanR
Data cleaning	<i>ec_rm_duplicate</i>	Remove duplicate records after merging databases	dplyr	EcoCleanR
	<i>ec_worms_synonym</i>	Check if the taxa list in the merge file has any bad taxon. Lookup to WoRMs database	taxize	EcoCleanR
	<i>ec_flag_with locality</i> <i>ec_filter_by_uncertainty</i>	Check if records have localities information but no coordinates or higher uncertainty – flag them, generate csv file to look into GEOLocate tool	dplyr	EcoCleanR
	<i>ec_flag_precision</i>	Check the precision of coordinates – if they are low precision or rounding – flag/remove them		EcoCleanR
	<i>cc_val</i>	Identify Invalid lat/long coordinates – flag/remove them	terra (Hijmans 2025a), geosphere (Hijmans 2010)	CoordinateCleaner
	<i>cc_equ</i>	Identify records with identical lat/long – flag/remove them		CoordinateCleaner
	<i>cc_zero</i>	Identify records lacking or with invalid coordinates but containing locality information	terra	CoordinateCleaner
	<i>cc_cen</i>	Identify coordinates in vicinity of country and province centroids	terra, geosphere	CoordinateCleaner
	<i>cc_sea</i>	Check coordinates offshore within the ocean/sea. Note: It flags island data points as bad data points, so it is not helpful for coastal species	terra, rnaturalearth (Massicotte and South 2025), dplyr	CoordinateCleaner
	<i>ec_flag_non_region</i>	Select ranges specific to a certain region or ocean; It removes data points located in the wrong ocean or sea and inland outside the buffer distance	sf (Pebesma 2018, Pebesma and Bivand 2023), mregions2	EcoCleanR
	<i>ec_extract_env_layers</i>	Adding environmental parameters for each data record extracted from BioOracle	sdmpredictors, terra	EcoCleanR
	<i>ec_impute_env_values</i>	Imputing environmental parameters if there is no assignment in BioOracle	geosphere	EcoCleanR
	<i>ec_flag_outlier</i>	Check if any abnormal patterns exist in spatial and non-spatial attributes to find outliers	geosphere	EcoCleanR
	<i>ec_var_summary</i>	Provide a summary table of suitable ranges	dplyr	EcoCleanR
Illustration	<i>ec_geographic_map</i> <i>ec_geographic_map_w_flag</i>	Create maps with data points – pre/post operations	terra, geodata (Hijmans 2025b), sf, ggplot2 (Wickham 2016)	EcoCleanR
	<i>ec_plot_var_range</i>	A plot with species suitable ranges	patchwork (Pedersen 2025), ggplot2, tidyr (Wickham et al. 2025)	EcoCleanR

Data merging

The package vignette *data_merging* describes an algorithm that retrieves and compiles occurrence data for a given species from online data sources, GBIF, iDigBio, OBIS, in addition to local input files. This algorithm utilizes existing packages

(‘rgbif’ for GBIF (Chamberlain et al. 2024), ‘ridigbio’ for iDigBio (Michonneau et al. 2024) and ‘robis’ for OBIS (Provoost and Bosch 2022) to extract data and converting the field names into DwC (Darwin Core) standard if they are different among the databases. Function *ec_db_merge* helps

to merge these databases and removes records with an occurrenceStatus of 'ABSENT' and basisOfRecords categorized as 'FOSSIL_SPECIMEN', 'fossilspecimen' or 'MATERIAL_SAMPLE' to filter modern and 'present only' occurrences effectively. Despite this, it should be noted that fossil occurrences may be incorrectly categorized as 'Preserved specimen' by some data providers; such occurrences can be manually recognized from verbatim locality if stated. The merged table will contain duplicates for the records that are present in more than one database; our function *ec_rm_duplicate* deals with those duplicates based on unique catalog numbers. If there is an identical record (based on the same catalog number) present in two databases, but one has abundance/individual counts, this function selects that record as the preferred record rather than one that does not have any abundance information.

Data cleaning

Taxonomy error

While downloading data from many sources, there is a possibility that an old/unaccepted name of a taxon could be extracted as a synonym. The *ec_worms_synonym* function is built on the 'taxize' package (Chamberlain et al. 2025) and provides a tabular output that includes accepted names and synonyms retrieved from WoRMS, along with the unique taxon names extracted from the merged dataset. This allows users to compare and manually filter species not listed in the WoRMS list. Currently, this function is designed to support WoRMS only, which is widely used as a reference for marine taxonomy (Costello et al. 2013).

Georeferencing records with no coordinates or uncertainty

The function *ec_flag_with_locality* provides an optional check for records that do not have coordinates but have information about the locality (e.g. verbatim locality description) as well as those records that have coordinates but no uncertainty. This function filters those records for further georeferencing via the GEOLocate tool. This tool groups records based on the similarity of locality strings and provides a proposed coordinate/uncertainty value that can be confirmed after manual verification. These georeferenced records can be uploaded back into the main data file through R.

Coordinate precision or rounding issue

The next function, *ec_flag_precision*, checks the precision of the coordinates. In some cases, the coordinates have only one decimal precision (e.g. 1.3°S), giving an error of approximately 11 km in both latitude and longitude (at the equator) from the true location. However, the presentation of latitude and longitude values rounded to the nearest degree (e.g. 1°S, 90°W) or even half-degree not only places coordinates even further from their presumed true location but also reduces confidence. This function flags the records based on two checkpoints: 1) flag both coordinates independently if they have < 2 decimal points; 2) check if any rounding to the nearest 0.5° in either of those coordinates to flag and remove.

Occurrences in the wrong ocean/sea

A common problem in large marine biodiversity datasets is that there are records that are located, either through poor georeferencing or other human errors, in the wrong ocean basin. To address this issue, 'EcoCleanR' provides function *ec_flag_non_region* that integrates with the 'mregions2' package (Fernández Bejarano and Pohl 2024) to take the ocean basin as an input variable to extract marine region gazetteer layers (www.marineregions.org). A buffer distance can be used to extend the borderline of the ocean polygon inland, allowing the inclusion of adjacent coastal features. This polygon is used as a spatial filter, with points outside its extent flagged and excluded as erroneous records. We acknowledge that currently, this function is limited to marine layers, specifically with input variables such as 'North Atlantic Ocean', 'North Pacific Ocean', 'South Atlantic Ocean', and 'South Pacific Ocean'. This function has the potential to be upgraded with more marine layers and terrestrial layers for terrestrial organisms in future versions.

Identify outliers

Outlier detection is a crucial task in data mining to identify abnormal patterns within large datasets. Despite multiple data cleaning steps, some anomalies may persist, making them difficult to detect, particularly when working with various data aggregators. In the context of spatial data, an outlier is a local anomaly where the nonspatial attribute values significantly deviate from those of its neighboring data points (Cerioli and Riani 1999). To detect outliers in *ec_flag_outlier*, attribute space is divided into two components: spatial attributes, such as coordinates, which determine spatial relationships among neighboring data points, and nonspatial attributes, such as environmental variables used to identify abnormal observations in space. We load the environmental data using the 'sdmprediction' package (Bosch and Fernandez 2023), which enables the extraction of environmental layers from sources including Bio-ORACLE (Assis et al. 2018) and MARSPEC (Sbrocco and Barber 2013) for marine organisms, and WORLDCLIM (Fick and Hijmans 2017) for terrestrial taxa (see *list_layer\$layer_code* for sources and layer names).

As a first step, we employed a traditional cluster-based method that uses spatial attributes to define clusters based on the Haversine distance (Robusto 1957, Prasetya et al. 2020). Then, we used nonspatial attributes to calculate the Mahalanobis distance (Mahalanobis 1930, McLachlan 1999) in the multivariate space of each cluster, which accounts for correlations among attributes and scales the distances accordingly. Lastly, a quantile-based threshold was applied to both distances to detect outliers. For a detailed description of the methodology and formulation, see the Supporting information.

Empirical example

We present a step-by-step demonstration of 'EcoCleanR' to derive biogeographic range of *Mexacanthina lugubris*

(Sowerby 1822), a coastal rocky intertidal gastropod found from southern California to southern Baja California (Marko and Vermeij 1999) but subsequent changes in its range have been noted by Fenberg et al. (2014) and in recent iNaturalist observations. In our example, modern 'present only' occurrence data for this species was gathered from biodiversity aggregators (GBIF, iDigBio, OBIS, and InvertEBase) and supplemented with museum observation records. All datasets were merged for analysis.

As of April 2025, our merged database contained 2163 occurrence records, including potential duplicate catalog entries from multiple sources. After removal of duplicate records, we retained 1125 unique records (Fig. 2A). To identify potential taxonomic errors in the merged dataset, we ran *ec_worms_synonym* to check for incorrectly assigned taxa during data extraction. Notably, we detected *Morula lugubre*, which was considered a synonym of *Mexacanthina lugubris* in GBIF. However, as recognized by the WoRMS database,

Morula lugubre is not an accepted synonym, but rather a distinct taxon, *Trachypollia lugubris* (Adams 1852). Based on this verification, we filtered out those erroneous records.

To test the dynamic range outcome from 'EcoCleanR', under varying amount of data input, we applied three filtering scenarios to the remaining merged dataset using the built-in functions to address missing coordinates and coordinate uncertainty values. Scenario A involved removing all records without coordinates and excluding the 5% most extreme uncertainty values, resulting in 66% data retention. Scenario B – we utilized the GEOLocate tool to improve the georeferencing of records that had missing coordinates and uncertainty values but included locality information, achieving a 70% data retention. Scenario C – we removed all records that lacked coordinates or uncertainty values, resulting in a 40% retention of the original data.

Next, we assessed the precision of coordinate values to remove records with poor precision or rounded coordinates

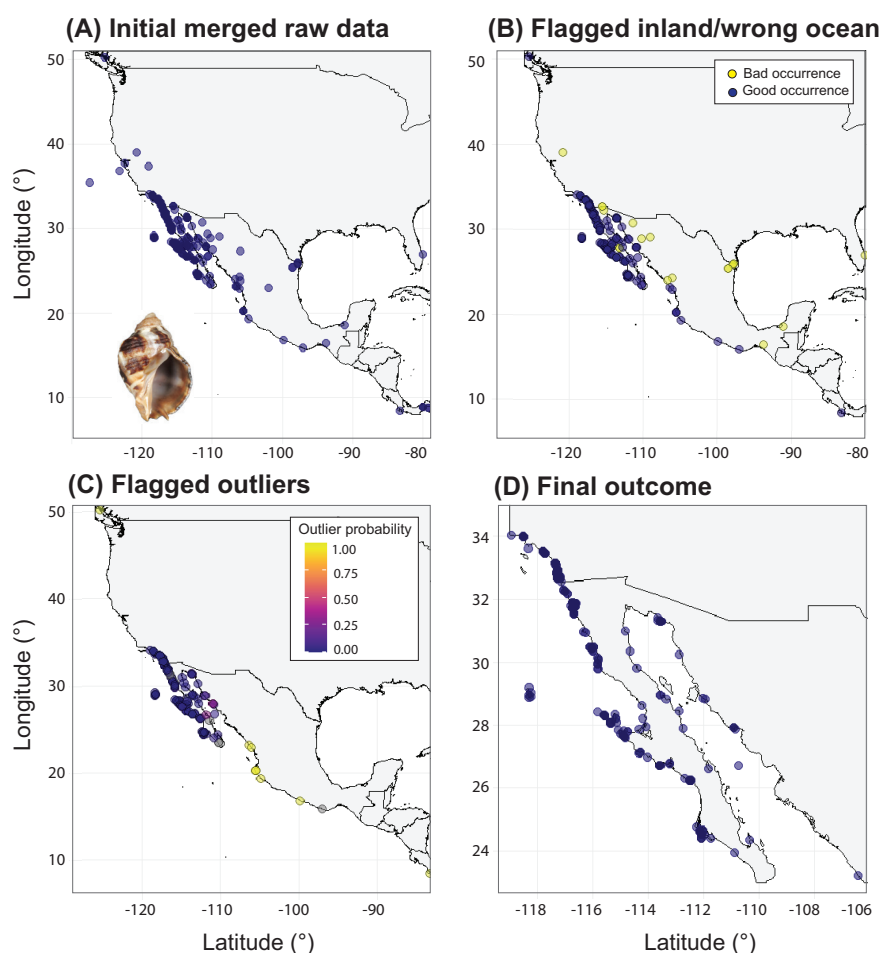


Figure 2. Step by step map view of the data cleaning process in Scenario A: (A) Occurrence points after removing duplicates from the merged data of GBIF, OBIS and iDigBio and InvertEBase, considered here as the raw data. (B) Flagged data points associated with the wrong ocean/sea or inland. (C) Illustrating varying levels of outlier probability. Warmer colors indicate higher outlier probabilities, calculated using a 0.99 threshold based on both Haversine and Mahalanobis distances. Grey points represent records for which environmental variable values could not be retrieved. (D) Final cleaned occurrence points which can be used to define a species range. These points were considered suitable for range estimation after removing 5% of records with the highest outlier probability.

(1% of remaining records). Many biodiversity aggregators have already flagged and rectified common georeferencing issues, such as invalid coordinates, duplicate or zero latitude/longitude values, and occurrences mistakenly assigned to museums or institutions. No such errors were present in our remaining dataset. Next, we addressed issues where data points were incorrectly assigned to the wrong ocean and coastline, finding 2 to 3% erroneous records in the Atlantic shoreline and others located more than 25 km inland from the coastline (Fig. 2B).

This clean data was utilized to extract climate layers for identifying outliers. Since temperature alone can explain 53 to 99% of the present-day distribution and biogeographic ranges of shallow marine benthic fauna (Belanger et al. 2012), we selected mean, maximum, and minimum sea surface temperature values extracted from Bio-Oracle (Assis et al. 2018) to detect abnormal patterns within a multivariate space. One to two percent of the records were identified as outliers due to extremely high Haversine and Mahalanobis distances. Fig. 2C shows a distribution map for Scenario A with outlier probability, and Fig. 2D depicts the final cleaned occurrences. The maximum and minimum values for spatial and nonspatial attributes suitable for species distribution are defined in Table 2 and Fig. 3, for example, species under Scenario A. These values provide a refined biogeographic and environmental range for a species and can be useful for future biodiversity studies. Fig. 4 and the Supporting information illustrate the improvement in data quality following the cleaning steps of 'EcoCleanR' to the initial merged data. Our results demonstrate that as the number of occurrences change, these ranges dynamically adjust (see Scenarios B and C; Supporting information), providing improved estimates compared to static historical ranges. This adaptive behavior ensures that suitable habitable ranges are continuously refined and improved as biodiversity occurrence datasets grow. See the Supporting information, for data-cleaning workflows of two other example species *Stramonita biserialis* and *Saxidomus gigantea*, respectively.

Comparison with other software

In the past decade, many studies have designed tools that aimed to improve biodiversity data quality (Zizka et al. 2019, Ribeiro et al. 2022, de Lima et al. 2023, Dorey et al. 2023, Patten et al. 2024). R packages such as 'CoordinateCleaner', 'bdc' (Ribeiro et al. 2022), 'bdcleaner' (Jin and Yang 2020), 'gatoRs' (Patten et al. 2024), 'biogeo' (Robertson et al. 2016), and 'flexsdm' (Velazco et al. 2022) have been introduced more recently to standardize and clean biodiversity data which becomes further useful in modeling applications. Among these, 'CoordinateCleaner' implements distance-based geographic outlier detection (e.g. IQR/MAD on interpoint distances and nearest-neighbor thresholds). 'bdc', 'bdcleaner' and 'gatoRs' emphasize rule-based quality controls – taxonomic and metadata checks, coordinate plausibility, and standardization than statistical spatial outlier detection. 'biogeo' does not extract environmental layers; instead, it offers visual inspection of environmental variables already present in the input to enable manual flagging of suspect records. 'flexsdm' extends these capabilities by implementing supervised and unsupervised machine-learning algorithms for outlier detection that focus solely on environmental attributes. 'EcoCleanR' builds on these advances by unifying environmental and spatial attributes and applying cluster based distance matrices to identify anomalous occurrences in either context. Unlike tools that generate range maps through model- or simulation-based approaches – which often overestimate species occupancy and are prone to false positives (Hurlbert and White 2005, Rotenberry and Balasubramaniam 2020) – our method derives species ranges directly from observed point locations compiled from diverse sources. We emphasize that our goal is not to discredit modeling approaches or promote one method over another. Rather, our objective is to demonstrate that carefully cleaned occurrence data can effectively represent species ranges and also serve as a reliable baseline to verify model-based outcomes.

Table 2. Biogeographic and environmental range data for three molluscan species generated using the 'EcoCleanR' package under Scenario A. Note: the Supporting information present the data-cleaning workflows for *Stramonita biserialis* and *Saxidomus gigantea*, respectively.

Species	Attributes	Maximum	Minimum	Mean	Range from other resources
<i>Mexacanthina lugubris</i>	Latitude (°)	34.0	22.9		San Diego, California-Bahia Magdalena, Baja California Sur (32.8–24.5°) (Marko and Vermeij 1999)
	Longitude (°)	–106.1	–118.9		
	Mean SST (°C)	29.0	16.2	18.0	
	Min SST (°C)	25.0	11.4	14.4	
	Max SST (°C)	32.7	18.8	22.5	
<i>Stramonita biserialis</i> (Supporting information)	Latitude (°)	33.7	–13.7		Cedros Island, Baja California-Chile (28.4°–?) (Keen 1971)
	Longitude (°)	–76.4	–118.4		
	Mean SST (°C)	30.7	16.8	26.2	
	Min SST (°C)	29.0	13.3	22.8	
	Max SST (°C)	33.8	20.3	29.7	
<i>Saxidomus gigantea</i> (Supporting information)	Latitude (°)	61.0	38.3		Seward Peninsula, Alaska-Capitola, California (65.6–37.0°) (Coan et al. 2000)
	Longitude (°)	–122.3	–162.6		
	Mean SST (°C)	12.4	5.4	10.0	
	Min SST (°C)	11.0	–1.7	5.5	
	Max SST (°C)	19.0	10.5	14.8	

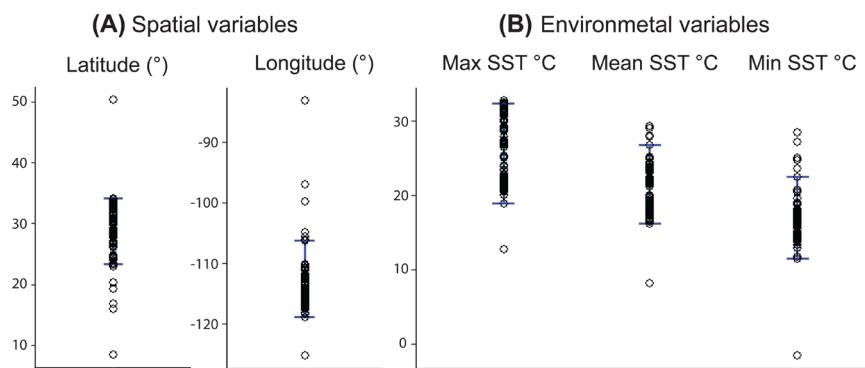


Figure 3. Plot showing (A) spatial (latitude and longitude) and (B) environmental variables (in this case sea surface temperature (SST) in °C which is an average value from 2000–2014) represented by open circles, including all occurrence points. Final cleaned occurrence points (see Fig. 2D for map view of spatial variables) used to derive species ranges are bounded by blue bars.

Summary and future directions

The ‘EcoCleanR’ workflow currently focuses on systematic improvement of marine species occurrences data and introduces several novel functionalities that improve traceability and quantitative assessment of data quality across all steps. Key features include: 1) map visualization and 2) coordinate assignment – a function that extracts records containing locality information but lacking spatial data to georeference using the GEOLocate tool. 3) Georeferencing error detection – a function that helps identify records georeferenced to incorrect ocean basins/seas using marine layers. 4) Outlier detection – a function that identifies outliers by incorporating spatial and nonspatial attributes and estimates outlier probability. 5) Comprehensive species habitat range – unlike

traditional approaches that only provide spatial range estimates, our package also provides a range of suitable environmental conditions. ‘EcoCleanR’s objective is to provide a user-friendly workflow consisting of new and existing functions that can be called based on the user’s requirements. In future versions, we plan to incorporate additional marine region gazetteers, extend functionality to terrestrial species range mapping using continental-scale gazetteers, include the Integrated Taxonomic Information System (ITIS) for taxonomy validation, and integrate georeferencing services directly within ‘EcoCleanR’.

To cite ‘EcoCleanR’ or acknowledge its use, cite this Software note as follows, substituting the version of the application that you used for ‘version 1.0’:

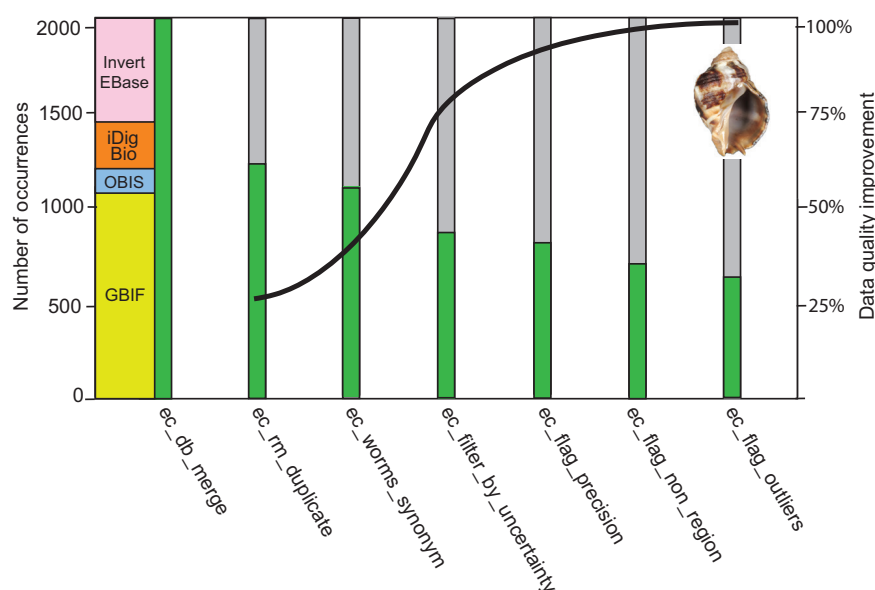


Figure 4. Refinement of occurrence records of *M. lugubris* at each step of the ‘EcoCleanR’ data cleaning workflow after merging datasets from GBIF, iDigBio, OBIS, and InvertEBase. Green bars indicate the number of records retained after cleaning. The line plot illustrates the cumulative improvement in data quality (Supporting information), assuming 100% data cleanliness following the *ec_flag_outliers*.

Soni, P. et al. 2026. 'EcoCleanR': enhancing data quality of biogeographic ranges with application for marine invertebrates. – *Ecography* 2026: e08203 (ver. 1.0).

Acknowledgements – We thank GBIF, OBIS, iDigBio, Invertebrate, and natural history museums for their efforts in digitizing biodiversity data and making it publicly accessible. We are grateful to Binhao Wang for insightful discussions on statistical analyses, Vijay Barve for feedback on R coding and package development, and Nelson Rios for his guidance on georeferencing using the GEOLocate tool. We also extend our sincere thanks to the anonymous reviewers and editors, whose constructive feedback greatly improved both the package and the manuscript.

Funding – The authors received no specific funding for this work.

Conflict of interest – The authors declare no conflict of interest.

Author contributions

Priyanka Soni: Conceptualization (lead); Formal analysis (lead); Investigation (lead); Methodology (lead); Project administration (equal); Software (lead); Supervision (equal); Validation (lead); Visualization (equal); Writing – original draft (lead); Writing – review and editing (equal). **Austin J. W. Hendy:** Conceptualization (supporting); Formal analysis (supporting); Investigation (supporting); Methodology (supporting); Project administration (equal); Supervision (equal); Validation (supporting); Visualization (supporting); Writing – review and editing (supporting). **David J. Bottjer:** Conceptualization (supporting); Formal analysis (supporting); Investigation (supporting); Methodology (supporting); Project administration (equal); Supervision (equal); Validation (supporting); Visualization (supporting); Writing – review and editing (supporting).

Transparent peer review

The peer review history for this article is available at <https://www.webofscience.com/api/gateway/wos/peer-review/eco.g.08203>.

Data availability statement

Data are available from the Dryad Digital Repository: <https://doi.org/10.5061/dryad.fj6q57488> (Soni et al. 2026).

Supporting information

The Supporting information associated with this article is available with the online version.

References

- Adams, C. B. 1852. Catalogue of shells collected at Panama, with notes on their synonymy, station, and geographical distribution. – *Ann. Lyceum Nat. Hist. N. Y.* 5: 229–549.
- Aggarwal, C. C. 2017. An introduction to outlier analysis. – In: *Outlier analysis*. – Springer, pp. 1–34.
- Alhajeri, B. H. and Fourcade, Y. 2019. High correlation between species-level environmental data estimates extracted from IUCN expert range maps and from GBIF occurrence data. – *J. Biogeogr.* 46: 1329–1341.
- Assis, J., Tyberghein, L., Bosch, S., Verbruggen, H., Serrão, E. A. and De Clerck, O. 2018. Bio-ORACLE v2.0: extending marine data layers for bioclimatic modelling. – *Global Ecol. Biogeogr.* 27: 277–284.
- Assis, J., Fragkopoulou, E., Frade, D., Neiva, J., Oliveira, A., Abecasis, D., Faugeron, S. and Serrão, E. A. 2020. A fine-tuned global distribution dataset of marine forests. – *Sci. Data* 7: 119.
- Belanger, C. L., Jablonski, D., Roy, K., Berke, S. K., Krug, A. Z. and Valentine, J. W. 2012. Global environmental predictors of benthic marine biogeographic structure. – *Proc. Natl Acad. Sci. USA* 109: 14046–14051.
- Benhadi-Marín, J. 2018. A conceptual framework to deal with outliers in ecology. – *Biodivers. Conserv.* 27: 3295–3300.
- Bisby, F. A. 2000. The quiet revolution: biodiversity informatics and the internet. – *Science* 289: 2309–2312.
- Bosch, S. and Fernandez, S. 2023. sdmpredictors: species distribution modelling predictor datasets. – R package, <https://cran.r-project.org/web/packages/sdmpredictors>.
- Cerlioli, A. and Riani, M. 1999. The ordering of spatial data and the detection of multiple outliers? – *J. Comput. Graph. Stat.* 8: 239–258.
- Chamberlain, S., Oldoni, D. and Waller, J. 2024. rgbif: interface to the Global Biodiversity Information Facility API. – R package, <https://cran.r-project.org/web/packages/rgbif>.
- Chamberlain, S., Szoecs, E., Foster, Z. and Arendsee, Z. 2025. taxize: taxonomic information from around the web. – R package, <https://cran.r-project.org/web/packages/taxize>.
- Chapman, A. D. 2005. Principles of data quality. – Global Biodiversity Information Facility, <https://doi.org/10.15468/doc.jrgg-a190>.
- Chapman, A. D. and Wicczorek, J. R. 2020. Georeferencing best practices, ver. 1.0. – Global Biodiversity Information Facility, <https://doi.org/10.15468/doc-gg7h-s853>.
- Chen, D., Lu, C. T., Kou, Y. and Chen, F. 2008. On detecting spatial outliers. – *Geoinformatica* 12: 455–475.
- Coan, E. V., Valentich Scott, P. and Bernard, F. 2000. Bivalve seashells of western North America. Marine mollusks from Arctic Alaska to Baja California. – Santa Barbara Museum of Natural History.
- Cordier, J. M., Osorio-Olvera, L., Huais, P. Y., Tomba, A. N., Villalobos, F. and Nori, J. 2024. Capability of big data to capture threatened vertebrate diversity in protected areas. – *Conserv. Biol.* 39: e14371.
- Costello, M. J., Bouchet, P., Boxshall, G., Fauchald, K., Gordon, D., Hoeksema, B. W., Poore, G. C. B., van Soest, R. W. M., Stöhr, S., Walter, T. C., Vanhoorne, B., Decock, W. and Appeltans, W. 2013. Global coordination and standardisation in marine biodiversity through the World Register of Marine Species (WoRMS) and related databases. – *PLoS One* 8: e51629.
- Crall, A. W., Newman, G. J., Stohlgren, T. J., Holfelder, K. A., Graham, J. and Waller, D. M. 2011. Assessing citizen science data quality: an invasive species case study. – *Conserv. Lett.* 4: 433–442.
- Darrigran, G. et al. 2023. Species movements within biogeographic regions: exploring the distribution of transplanted mollusc species in South America. – *Biol. Invas.* 25: 673–691.
- de Lima, R. A. F., Sánchez-Tapia, A., Mortara, S. R., ter Steege, H. and de Siqueira, M. F. 2023. plantR: an R package and workflow for managing species records from biological collections. – *Methods Ecol. Evol.* 14: 332–339.
- Donlan, J. 2005. Re-wilding North America. – *Nature* 436: 913–914.
- Dorey, J. B. et al. 2023. A globally synthesised and flagged bee occurrence dataset and cleaning workflow. – *Sci. Data* 10: 1–27.

- Farley, S. S., Dawson, A., Goring, S. J. and Williams, J. W. 2018. Situating ecology as a big-data science: current advances, challenges, and solutions. – *BioScience* 68: 563–576.
- Fenberger, P. B., Posbic, K. and Hellberg, M. E. 2014. Historical and recent processes shaping the geographic range of a rocky intertidal gastropod: phylogeography, ecology, and habitat availability. – *Ecol. Evol.* 4: 3244–3255.
- Feng, X. et al. 2022. A review of the heterogeneous landscape of biodiversity databases: opportunities and challenges for a synthesized biodiversity knowledge base. – *Global Ecol. Biogeogr.* 31: 1242–1260.
- Fernández Bejarano, S. J. and Pohl, L. 2024. mregions2: access data from marine regions: gazetteer and data products. – R package, <https://cran.r-project.org/web/packages/mregions2>.
- Fick, S. E. and Hijmans, R. J. 2017. WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. – *Intl. J. Climatol.* 37: 4302–4315.
- Freitas, T. M. S., Montag, L. F. A., De Marco, P. and 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: 181–191.
- Fuccillo, K. K., Crimmins, T. M., de Rivera, C. E. and Elder, T. S. 2015. Assessing accuracy in citizen science-based plant phenology monitoring. – *Int. J. Biometeorol.* 59: 917–926.
- Gervazoni, P., Minuti, G., Fuentes-Rodríguez, D., Coetzee, J., Sosa, A., Sabater, L. and Franceschini, C. 2023. Citizen science improves the known and potential distribution of a strong wetland invader: implications for niche modeling and invasion management. – *Environ. Manage.* 71: 1176–1187.
- Graham, C. H., Ferrier, S., Huettman, F., Moritz, C. and Peterson, A. T. 2004. New developments in museum-based informatics and applications in biodiversity analysis. – *Trends Ecol. Evol.* 19: 497–503.
- Grenié, M., Berti, E., Carvajal-Quintero, J., Dädlow, G. M. L., Sagouis, A. and Winter, M. 2023. Harmonizing taxon names in biodiversity data: a review of tools, databases and best practices. – *Methods Ecol. Evol.* 14: 12–25.
- Heberling, J. M., Miller, J. T., Noesgaard, D., Weingart, S. B. and Schigel, D. 2021. Data integration enables global biodiversity synthesis. – *Proc. Natl Acad. Sci. USA* 118: e2018093118.
- Hijmans, R. J. 2010. geosphere: spherical trigonometry. – R package, <https://cran.r-project.org/web/packages/geosphere>.
- Hijmans, R. J. 2025a. terra: spatial data analysis. – R package, <https://cran.r-project.org/web/packages/terra>.
- Hijmans, R. J. 2025b. geodata: download geographic data. – R package, <https://cran.r-project.org/web/packages/geodata>.
- Hurlbert, A. H. and White, E. P. 2005. Disparity between range map- and survey-based analyses of species richness: patterns, processes and implications. – *Ecol. Lett.* 8: 319–327.
- Jin, J. and Yang, J. 2020. BDcleaner: a workflow for cleaning taxonomic and geographic errors in occurrence data archived in biodiversity databases. – *Global Ecol. Conserv.* 21: e00852.
- Keen, A. M. 1971. Sea shells of tropical West America: marine mollusks from Baja California to Peru, 2nd edn. – Stanford Univ. Press.
- Lawlor, J. A. et al. 2024. Mechanisms, detection and impacts of species redistributions under climate change. – *Nat. Rev. Earth Environ.* 5: 351–368.
- Lomolino, M. V., Riddle, B. R., Whittaker, R. J. and Brown, J. H. 2017. Biogeography, 5th edn. – Oxford Univ. Press.
- Mahalanobis, P. C. 1930. A statistical study of certain anthropometric measurements from Sweden. – *Biometrika* 22: 94–108.
- Maldonado, C., Molina, C. I., Zizka, A., Persson, C., Taylor, C. M., Albán, J., Chilquillo, E., Rønsted, N. and Antonelli, A. 2015. Estimating species diversity and distribution in the era of Big Data: to what extent can we trust public databases? – *Global Ecol. Biogeogr.* 24: 973–984.
- Mancinelli, G., Bardelli, R. and Zenetos, A. 2021. A global occurrence database of the Atlantic blue crab *Callinectes sapidus*. – *Sci. Data* 8: 111.
- Marcen, A., Chapman, A. D., Wiczeorek, J. R., Xavier Picó, F., Uribe, F., Waller, J. and Ariño, A. H. 2022. Uncertainty matters: ascertaining where specimens in natural history collections come from and its implications for predicting species distributions. – *Ecography* 2022: 1–17.
- Marko, P. B. and Vermeij, G. J. 1999. Molecular phylogenetics and the evolution of labral spines among eastern Pacific Ocenebrine gastropods. – *Mol. Phylogenet. Evol.* 13: 275–288.
- Massicotte, P. and South, A. 2025. rnaturalearth: world map data from Natural Earth. – R package, <https://docs.ropensci.org/rnaturalearth>.
- McLachlan, G. J. 1999. Mahalanobis distance. – *Resonance* 4: 20–26.
- Mesibov, R. 2013. A specialist's audit of aggregated occurrence records. – *ZooKeys* 293: 1–18.
- Mesibov, R. 2018. An audit of some processing effects in aggregated occurrence records. – *ZooKeys* 751: 129–146.
- Meyer, C., Weigelt, P. and Kreft, H. 2016. Multidimensional biases, gaps and uncertainties in global plant occurrence information. – *Ecol. Lett.* 19: 992–1006.
- Michonneau, F., Collins, M. and Gaynor, M. L. 2024. ridigbio: interface to the iDigBio data API. – R package, <https://cran.r-project.org/web/packages/ridigbio>.
- Moudry, V. and Devillers, R. 2020. Quality and usability challenges of global marine biodiversity databases: an example for marine mammal data. – *Ecol. Inform.* 56: 101051.
- Muhs, D. R., Groves, L. T., Simmons, K. R., Schumann, R. R. and Minor, S. A. 2023. A complex record of last interglacial sea-level history and paleozoogeography, Santa Rosa Island, Channel Islands National Park, California, USA. – *Palaeogeogr. Palaeoclimatol. Palaeoecol.* 610: 111328.
- Nic Lughadha, E. M., Grazielle Staggemeier, V., Vasconcelos, T. N. C., Walker, B. E., Canteiro, C. and Lucas, E. J. 2019. Harnessing the potential of integrated systematics for conservation of taxonomically complex, megadiverse plant groups. – *Conserv. Biol.* 33: 511–522.
- Nogué, S., De Nascimento, L., Froyd, C. A., Wilmshurst, J. M., De Boer, E. J., Coffey, E. E. D., Whittaker, R. J., Fernández-Palacios, J. M. and Willis, K. J. 2017. Island biodiversity conservation needs palaeoecology. – *Nat. Ecol. Evol.* 1: 181.
- Novoa, A., Le Roux, J. J., Robertson, M. P., Wilson, J. R. U. and Richardson, D. M. 2015. Introduced and invasive cactus species: a global review. – *AoB Plants* 7: 1–14.
- Panther, C. T., Clegg, R. L., Moat, J., Bachman, S. P., Klitgård, B. B. and White, R. L. 2020. To clean or not to clean: cleaning open-source data improves extinction risk assessments for threatened plant species. – *Conserv. Sci. Prac.* 2: 1–14.
- Patten, N. N., Gaynor, M. L., Soltis, D. E. and Soltis, P. S. 2024. Geographic and taxonomic occurrence R-based scrubbing (gatoRs): an R package and workflow for processing biodiversity data. – *Appl. Plant Sci.* 12: e11575.
- Pebesma, E. 2018. Simple features for R: standardized support for spatial vector data. – *R J.* 10: 439–446.

- Pebesma, E. and Bivand, R. 2023. Spatial data science: with applications in R. – Chapman and Hall/CRC, 314 p.
- Pedersen, T. L. 2025. patchwork: the composer of plots. – R package, <https://cran.r-project.org/web/packages/patchwork>.
- Peterson, A., Soberón, J., Pearson, R., Anderson, R., Martínez-Meyer, E., Nakamura, M. and Araújo, M. 2012. Ecological niches and geographic distributions. – Princeton Univ. Press.
- Prasetya, D. A., Nguyen, P. T., Faizullin, R., Iswanto, I. and Armay, E. F. 2020. Resolving the shortest path problem using the haversine algorithm. – Crit. Rev. 7: 62–64.
- Provoost, P. and Bosch, S. 2022. robis: Ocean Biodiversity Information System (OBIS) client. – R package, <https://cran.r-project.org/web/packages/robis>.
- Ribeiro, B. R., Velazco, S. J. E., Guidoni-Martins, K., Tessarolo, G., Jardim, L., Bachman, S. P. and Loyola, R. 2022. bdc: a toolkit for standardizing, integrating and cleaning biodiversity data. – Methods Ecol. Evol. 13: 1421–1428.
- Richardson, D. M., Carruthers, J., Hui, C., Impson, F. A. C., Miller, J. T., Robertson, M. P., Rouget, M., Le Roux, J. J. and Wilson, J. R. U. 2011. Human-mediated introductions of Australian acacias – a global experiment in biogeography. – Divers. Distrib. 17: 771–787.
- Rios, N. E. and Bart, H. L. 2010. GEOLocate (ver. 3.22) [computer software]. Tulane University Museum of Natural History.
- Robertson, D. R. 2008. Global biogeographical data bases on marine fishes: caveat emptor. – Divers. Distrib. 14: 891–892.
- Robertson, M. P., Visser, V. and Hui, C. 2016. Biogeo: an R package for assessing and improving data quality of occurrence record datasets. – Ecography 39: 394–401.
- Robusto, C. C. 1957. The cosine-haversine formula. – Am. Math. Mon. 64: 38–40.
- Rotenberry, J. T. and Balasubramaniam, P. 2020. Connecting species' geographical distributions to environmental variables: range maps versus observed points of occurrence. – Ecography 43: 897–913.
- Sandall, E. L., Maureaud, A. A., Guralnick, R., McGeoch, M. A., Sica, Y. V., Rogan, M. S., Booher, D. B., Edwards, R., Franz, N., Ingenloff, K., Lucas, M., Marsh, C. J., McGowan, J., Pinkert, S., Ranipeta, A., Uetz, P., Wieczorek, J. and Jetz, W. 2023. A globally integrated structure of taxonomy to support biodiversity science and conservation. – Trends Ecol. Evol. 38: 1143–1153.
- Sbrocco, E. J. and Barber, P. H. 2013. MARSPEC: ocean climate layers for marine spatial ecology. – Ecology 94: 979–979.
- Shoo, L. P., Williams, S. E. and Hero, J. M. 2006. Detecting climate change induced range shifts: where and how should we be looking? – Austral Ecol. 31: 22–29.
- Smith, A. B., Murphy, S. J., Henderson, D. and Erickson, K. D. 2023. Including imprecisely georeferenced specimens improves accuracy of species distribution models and estimates of niche breadth. – Global Ecol. Biogeogr. 32: 342–355.
- Soberón, J. and Peterson, A. T. 2004. Biodiversity informatics: managing and applying primary biodiversity data. – Phil. Trans. R. Soc. B 359: 689–698.
- Soni, P., Hendy, A. J. W. and Bottjer, D. J. 2026. Data from: 'Eco-CleanR': enhancing data quality of biogeographic ranges with application for marine invertebrates. – Dryad Digital Repository, <https://doi.org/10.5061/dryad.f6q57488>.
- Sowerby, J. and Sowerby, G. B. 1820. The genera of recent and fossil shells; for the use of students in conchology and geology; v.1. – Legare Street Press.
- Tingley, M. W. and Beissinger, S. R. 2009. Detecting range shifts from historical species occurrences: new perspectives on old data. – Trends Ecol. Evol. 24: 625–633.
- Valentine, J. W. 1973. Evolutionary paleoecology of the marine biosphere. – Prentice-Hall.
- Vecchione, M., Mickevich, M. F., Fauchald, K., Collette, B. B., Williams, A. B., Munroe, T. A. and Young, R. E. 2000. Importance of assessing taxonomic adequacy in determining fishing effects on marine biodiversity. – ICES J. Mar. Sci. 57: 677–681.
- Velazco, S. J. E., Rose, M. B., de Andrade, A. F. A., Minoli, I. and Franklin, J. 2022. Flexsdrm: an R package for supporting a comprehensive and flexible species distribution modelling workflow. – Methods Ecol. Evol. 13: 1661–1669.
- Wickham, H. 2016. ggplot2: elegant graphics for data analysis. – Springer, <https://ggplot2.tidyverse.org>.
- Wickham, H., François, R., Henry, L. and Müller, K. 2022. dplyr: a grammar of data manipulation. – R package, <https://dplyr.tidyverse.org>.
- Wickham, H., Vaughan, D. and Girlich, M. 2025. tidyr: tidy messy data. – R package, <https://tidyr.tidyverse.org>.
- Woodring, W. P. and Popenoe, W. P. 1945. Paleocene and Eocene stratigraphy of the northwestern Santa Ana Mountains, Orange County, California. – US Geological Survey Oil and Gas Investigations, Preliminary Chart 12. 1: 103448.
- Waqar, U., Mahmood, T., Mushtaq, M., Murtaza, G., Kashaf, M. and Sheraz, A. 2024. Historical and current distribution ranges of the Asiatic black bear (*Ursus thibetanus*). – Sci. Rep. 14: 2505.
- Yesson, C., Brewer, P. W., Sutton, T., Caithness, N., Pahwa, J. S., Burgess, M., Gray, W. A., White, R. J., Jones, A. C., Bisby, F. A. and Culham, A. 2007. How global is the Global Biodiversity Information Facility? – PLoS One 2: e1124.
- Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Edler, D., Farooq, H., Herdean, A., Ariza, M., Scharn, R., Svantesson, S., Wengström, N., Zizka, V. and Antonelli, A. 2019. CoordinateCleaner: standardized cleaning of occurrence records from biological collection databases. – Methods Ecol. Evol. 10: 744–751.