



***marinedb*: An open-source, flexible and scalable Python workflow for curating, standardizing, and integrating marine biodiversity data**

Elena Berhocoïrigoin¹, Philippe Le Noac'h², Martin Laviale², Sakina-Dorothee Ayata^{3,4}, Jérémy Fix⁵, and Cédric Pradalier¹

¹CNRS IRL2958 GeorgiaTech - CNRS, 2 rue Marconi, Metz, 57050, France

²Université de Lorraine, CNRS, LIEC, 57000 Metz, France

³Sorbonne Université, CNRS, MNHN, IRD, Laboratoire d'Océanographie et du Climat : Expérimentation et Approches Numériques, LOCEAN-IPSL, 75005 Paris, France

⁴Institut Universitaire de France, 75005 Paris, France

⁵LORIA, CNRS, CentraleSupélec, Université Paris-Saclay, F-57000 Metz, France

Correspondence: Berhocoïrigoin Elena (eberhocol@georgiatech-metz.fr)

Abstract. The digital era and the emergence of data-aggregation initiatives have greatly expanded access to biodiversity occurrence records, but accessibility alone does not ensure usability for ecological modeling. Marine occurrence data must be cleaned and reconciled across spatial, temporal, and taxonomic dimensions before they can be reliably linked to environmental predictors and used to model species distributions. Here, we introduce *marinedb*, an open-source, modular, and scalable Python package for curating marine occurrence data. The package provides a suite of tools that can be freely combined into user-defined workflows tailored to study objectives and data characteristics. We applied *marinedb* to three major biodiversity resources: the Global Biodiversity Information Facility (GBIF), the Ocean Biogeographic Information System (OBIS), and the Jellyfish Database Initiative (JeDI). Together, these resources span hundreds of millions of occurrence records located at sea and more than one hundred thousand species. First, we used the framework to assess data quality non-destructively across the three sources. The analysis revealed substantial and source-dependent issues, including records located on land, missing depth information, heterogeneous, incomplete, or low-precision temporal information, records lacking species-level information or referring to non-marine taxa, weak or unspecified evidential support, absence records, and strong taxonomic imbalance. Second, we applied a tailored dataset-production workflow to produce a ready-to-use global marine occurrence dataset for large-scale modeling of present-day species distributions, which comprises 158,910,536 records representing 139,844 species. Together, the reusable Python tools and the curated dataset derived from GBIF, OBIS, and JeDI lower the technical barrier for multi-species and deep-learning approaches, both of which require extensive and diverse occurrence records, while preserving flexibility for future use-case-specific curation.



1 Introduction

1.1 Unlocking marine modeling: The need for a turnkey dataset

20 Occurrence data document the presence, and more rarely the absence, of organisms at particular places and times (Peterson et al., 2011). When combined with environmental predictors, they provide the empirical basis for correlative species distribution models (SDMs, hereafter), which estimate species–environment relationships to predict geographic distributions (Elith and Leathwick, 2009). Because robust SDM calibration and evaluation depend on both a sufficient number of records and adequate environmental representativeness (Moudrý et al., 2024), broad-scale marine applications, especially those using artificial neural
25 networks or aiming to model many taxa simultaneously, call for extensive occurrence data. However, compiling and cleaning such data requires time and foundational domain expertise. Data curation depends on prior knowledge of biodiversity-data concepts and common error sources, on the ability to assess how these considerations may affect downstream modeling, and on translating this assessment into appropriate decisions regarding the transformation, annotation, retention, or exclusion of records. Beyond this general expertise, dedicated effort is often required to understand database-specific fields and database-
30 level curation or quality-control procedures. The diversity of APIs, web platforms, and data formats (e.g., CSV, Parquet) used by biodiversity information infrastructures further demands technical proficiency. As data volumes grow, this technical burden extends to big data processing and access to appropriate computing resources. Together, these constraints can significantly impede the development of large-scale marine SDMs. Ready-to-use datasets designed for species distribution modeling can help lower these barriers, as demonstrated by the Geo-LifeCLEF initiative in the terrestrial realm (Botella et al., 2018). Such
35 resources can foster new contributions, especially from machine learning practitioners unfamiliar with the intricacies inherent to biodiversity data. Large-scale multi-species marine distribution modeling could therefore benefit from a sufficiently large, high-quality, and fit-for-purpose dataset.

1.2 Overcoming data limitations: Accessible yet not usable

Biodiversity data have long accumulated across natural history collections, published literature, institutional archives, and local
40 or regional datasets, but their use at broad scales was constrained by non-digital or obsolete digital storage, heterogeneous data types and formats, and fragmented access (Edwards et al., 2000; Hampton et al., 2013). Over the past two decades, the large-scale digitization of legacy data, together with advances in data storage and web-based data sharing, has greatly increased the availability of biodiversity records (Soberón and Peterson, 2004; Nelson and Ellis, 2018). At the same time, community-science observations, partly driven by the spread of mobile devices and online platforms, have further accelerated
45 this expansion (Chandler et al., 2017; Fraisl et al., 2022). Availability alone, however, does not ensure accessibility at scale. Records may still remain scattered across disparate sources that are difficult to discover and integrate. To bridge this gap, aggregating initiatives have emerged: by standardizing and compiling distributed records into open-access, interoperable data infrastructures and resources, they enable unprecedented access to vast volumes of occurrence data (Peterson et al., 2011). Early examples include the Global Biodiversity Information Facility (GBIF), a generalist aggregator, and the Ocean Biogeographic
50 Information System (OBIS), a domain-specialized aggregator, whose founding documents explicitly identified this need for



more readily discoverable and accessible biodiversity data (OECD Megascience Forum, 1999; Grassle, 2000). The Jellyfish Database Initiative (JeDI), a more recent resource focused on gelatinous zooplankton, addressed a similar limitation at a narrower scale: the lack of a global, long-term database for assessing whether gelatinous zooplankton populations were rising worldwide. This case illustrates how the absence of a broad, integrated resource can translate into uncertainty about major ecological questions. In response, JeDI compiled records from previously inaccessible sources, such as unpublished datasets and historical expedition reports, while also building on earlier aggregation efforts such as OBIS (Condon et al., 2012). As of mid-2026, these three initiatives exemplify the scale of biodiversity data aggregation, with approximately 430,000 records in JeDI, 199 million species observations in OBIS, and more than 3.7 billion occurrence records in GBIF. Yet while such initiatives have largely addressed the challenge of quantity and accessibility, they do not necessarily ensure usability: data quality and fitness for purpose remain critical constraints. In particular, several studies have documented persistent quality issues in data provided through aggregator platforms, despite ongoing efforts to detect, correct, and flag problematic records (e.g., Mesibov, 2013; Vandepitte et al., 2015; Maldonado et al., 2015; Mesibov, 2018; Zizka et al., 2019; Moudrý and Devillers, 2020; Ramírez et al., 2022; Bonnet-Lebrun et al., 2023; Cheung and Helfer, 2025). In some cases, these quality-control procedures have also been shown to introduce new issues (Mesibov, 2018; Franz and Sterner, 2018).

As a result, systematic cleaning of observations drawn from these sources remains essential, especially since aggregator-level quality-control procedures and automated curation workflows cannot accommodate the diversity of use cases. There is no universally appropriate configuration of filters or transformations: data curation must be guided by the specific scientific objectives of each study, as those objectives define which records are fit for use (Chapman, 2005b; Vandepitte et al., 2015; Veiga et al., 2017; Gueta and Carmel, 2016; Zizka et al., 2020; Hughes et al., 2024). In 2015–2016, GBIF coordinated expert working groups to assess data fitness for three research domains: agrobiodiversity, species distribution modeling, and alien and invasive species. This work emphasized that GBIF occurrence records were not suitable for SDMs without prior cleaning (Anderson et al., 2016). This limitation, though articulated for GBIF, can reasonably be extended to OBIS and JeDI. More specifically, in the context of large-scale modeling of present-day marine species distributions, several characteristics of these databases pose serious challenges to developing informative and reliable SDMs that are aligned with the objective. These challenges include observations georeferenced on land; taxa reported at supra- or infra-specific ranks; terrestrial species recorded in marine environments; fossil or historical occurrences; records assigned to temporal ranges or bearing invalid dates; observations with erroneous or imprecise coordinates; and pronounced taxonomic, environmental, or spatial biases. Thus, even OBIS, which aggregates marine biodiversity data at a global scale (fitness for purpose), contains hundreds of millions species observations (quantity), and has implemented quality-control procedures (quality), cannot be used as-is for the intended modeling task: substantial data curation remains necessary.

The need for such curation becomes even more pressing when modeling species distributions at broad spatial scales, across multiple species simultaneously, or using artificial neural networks, as such approaches require substantially greater data volume and diversity. In this context, leveraging the complementarity of multiple databases, such as OBIS, GBIF, and JeDI, is preferable to relying on a single data source. Notably, studies evaluating the quality and usability of global marine biodiversity databases found that GBIF and OBIS only partially overlap (Moudrý and Devillers, 2020; Bonnet-Lebrun et al., 2023).



However, combining multiple databases introduces additional challenges and increases the complexity of data curation. Chief among these is the need for taxonomic harmonization, as the reference database may differ for each source. Specifically, GBIF uses a taxonomic backbone built from 105 catalogs, including the World Register of Marine Species (WoRMS; WoRMS Editorial Board, 2026), while OBIS relies exclusively on WoRMS. By contrast, no information is currently available regarding the taxonomic standard used by JeDI. This heterogeneity is further compounded by differences in database update frequency, and consequently by the uneven incorporation of taxonomic revisions: none for JeDI, up to twice a year for GBIF, and within a few weeks of each change for OBIS. Such disparities can reduce the fitness for use of occurrence data for species distribution modeling, which depends on the accurate and consistent grouping of records by species.

To date, no existing dataset can, to the best of our knowledge, be considered "ready to use" for broad-scale modeling of marine species' geographic distributions, especially in deep-learning or multi-species frameworks.

1.3 Bridging the usability gap: The missing toolchain

Many tools are available for the quality control and cleaning of species occurrence data. However, this abundance can make it difficult to identify the most appropriate solutions for a given project. Much of this proliferation stems from the functional and contextual specialization of these tools, some of which exhibit both. Many are built for narrowly defined tasks (functional specialization), such as georeferencing from textual descriptions (e.g., GEOLocate¹), parsing scientific names into taxonomic components (e.g., *GNParser* (Mozzherin et al., 2017)), or correcting spatial errors (e.g., *biogeo* (Robertson et al., 2016)). Others are tailored to particular application domains (contextual specialization), such as marine (e.g., *obistools* (Provoost and Bosch, 2018), Marine Regions²) or terrestrial environments (e.g., *CoordinateCleaner* (Zizka et al., 2019), which is primarily geared toward land-based data). Some tools combine both forms of specialization: *flora* (Carvalho, 2020), for instance, offers taxonomic matching limited to plants, algae, and fungi in Brazil. Importantly, while it contributes to the fragmentation of the tool ecosystem, contextual specialization is not necessarily problematic and may even be desirable. On the one hand, it reflects the distinct ecological contexts and associated requirements these tools are designed to address. On the other hand, a unified solution could introduce new errors due to the coexistence of conflicting functionalities, which leads to greater operational complexity and, ultimately, increases the risk of misuse. This fractured landscape often compels users to combine multiple tools or develop cleaning workflows from scratch—efforts that, in turn, feed into an already saturated ecosystem. The result is a vicious cycle in which tool diversity, far from facilitating use, hinders the standardization of practices, reinforces technical barriers, and undermines the reproducibility of analyses.

In response to these challenges, integrated solutions—defined here as tools that handle at least two of the three core dimensions of occurrence data: spatial, temporal, and taxonomic—have begun to emerge in recent years. Examples include tools combining spatial and temporal curation, such as *scrubr* (Chamberlain, 2016) and, for fossil records, *CoordinateCleaner* (Zizka et al., 2019); tools combining spatial and taxonomic curation, such as *BDcleaner* (Jin and Yang, 2020); and more recent frameworks integrating all three dimensions, such as *bdc* (Ribeiro et al., 2022) and *BeeBDC* (Dorey et al., 2023). However, because

¹<https://www.geo-locate.org/>

²<https://marineregions.org/>



these tools were originally developed for terrestrial applications, several features are ill-suited to marine datasets. For instance, some perform country–coordinate consistency checks based on land-only boundaries or flag locations corresponding to capital cities, biodiversity institutions, administrative centroids, or urban areas. Certain features may even compromise data integrity—for example, by discarding valid offshore observations or altering their coordinates through automatic corrections, such as swapping latitude and longitude or shifting points toward nearby land. This terrestrial bias is further reflected in the absence of functions essential for marine data, such as taxonomic harmonization of marine species and validation of reported depth values. To the best of our knowledge, *obistools* and the LifeWatch & EMODnet Biology QC Tool (De Pooter and Perez-Perez, 2019), which is based on *obistools*, are the only tools specifically developed for marine biodiversity that address the spatial, temporal, and taxonomic dimensions concurrently. Nonetheless, they serve more as quality control frameworks than as automated cleaning tools, their primary role being to facilitate dataset preparation for submission to OBIS or EMODnet. One notable exception to the limitations outlined so far is the GBIF processing pipeline, which stands apart in both scale and scope: it processes massive volumes of terrestrial and marine data while integrating the spatial, temporal, and taxonomic dimensions. Nevertheless, its relevance to the present analysis remains limited by restricted third-party reusability and by the persistence of documented data-quality issues in GBIF-derived data, suggesting that aggregator-level harmonization cannot replace the fine-grained curation required for immediate use in targeted research applications. Moreover, apart from the Java-based GBIF processing pipeline, the tools reviewed above have been built primarily in R, although some *obistools* functionality is also available in Python. This R-centred tooling landscape presents a structural barrier to broader involvement from the artificial intelligence community and to the application of deep-learning methods, which are dominated by the Python ecosystem.

Finally, none of the tools discussed so far integrates all of the following features, which are particularly important for producing occurrence data fit for use in species distribution modeling: standardizing infraspecific taxa to the species level; filtering underrepresented taxa, for which the number of observations may be insufficient to infer geographic distributions reliably, especially at global scales; handling date ranges, which introduce temporal uncertainty and may need to be collapsed, annotated, or filtered before occurrence records can be matched to environmental conditions; and subsampling overrepresented species to reduce uneven taxonomic representation in a way that also mitigates spatial sampling bias for those species. The ability to include or exclude specific taxonomic groups is also valuable for tailoring the dataset to the ecological scope of the study.

Building on these observations, we developed *marinedb*, an open-source, flexible and scalable Python package for curating marine occurrence data. We applied the framework to three major biodiversity resources—GBIF, OBIS, and JeDI—first to assess their fitness for use across spatial, temporal, and taxonomic dimensions, and then to construct a curated global marine dataset suitable for large-scale modeling of present-day species distributions along geographic and environmental gradients.

2 Methods

The curation of marine occurrence data involves addressing heterogeneous issues that span multiple dimensions, including spatial, temporal, and taxonomic. To accommodate this complexity, *marinedb* is designed as a modular framework in which



independent operations can be combined into flexible processing workflows according to user objectives and input data characteristics. The framework provides tools for excluding terrestrial records and detecting common geospatial issues (e.g., out-of-range, zero, or low-precision coordinates). *marinedb* also supports temporal standardization and validation (e.g., date parsing, interval processing, and the identification of internal inconsistencies, non-existent dates, and unlikely dates). Complementary functionalities enable taxonomic harmonization through WoRMS, the exclusion of non-marine, fossil, or supra-specific taxa, and the filtering of underrepresented species or subsampling of overrepresented ones. Additional utilities allow users to select, remove, or annotate records based on data completeness, observation type, occurrence status, or other user-defined criteria. Finally, dataset-restructuring operations, such as field renaming and selection, facilitate the integration of heterogeneous biodiversity datasets. An overview of the available modules is provided in Table 1. The framework was applied to GBIF, OBIS, and JeDI both to assess their fitness for use and to construct a dataset for large-scale modeling of present-day marine species distributions (Fig. 2).

2.1 Data sources

Data were sourced from three publicly available online resources: (1) the Global Biodiversity Information Facility (GBIF), the largest cross-taxon biodiversity repository to date; (2) the Ocean Biogeographic Information System (OBIS), a data aggregator specializing in marine biodiversity; and (3) the Jellyfish Database Initiative (JeDI), a global database dedicated to gelatinous zooplankton. GBIF-processed occurrence data were downloaded from the online portal on September 17, 2024 (GBIF.org, 2024). During retrieval, records without coordinates, dated before 1700 or after 2024, or classified as fossil specimens were excluded, while both terrestrial and marine species were retained. The resulting tabular file, totaling 1.5 TB, contained 2,811,378,853 observations. OBIS quality-controlled presence records, provided as a 19 GB full-export parquet file, were downloaded on February 4, 2025 (OBIS, 2024), and included 132,469,294 records. Finally, the JeDI dataset, comprising 429,462 observations, was downloaded as a CSV file from the Biological and Chemical Oceanography Data Management Office (BCO-DMO) website on August 31, 2023 (Condon et al., 2014).

Despite differences in scope and structure, the three databases share a common set of information characteristic of species occurrence data. They comprise spatial fields (e.g., geographic coordinates, depth, or locality descriptors), temporal fields (e.g., dates and separate year, month, and day components), and taxonomic fields. The latter encompass hierarchical classifications and taxonomic authorship, i.e., the author names and dates associated with the formal description of a taxon. Further metadata characterize the observation itself, typically detailing occurrence status (presence/absence), sampling method (e.g., bottom trawl or acoustic detection), observation type (e.g., specimen, photograph, or sighting), quantitative measurements (e.g., counts or biomass), licensing information, and data quality annotations.

2.2 Modular workflow configuration and execution

marinedb provides a suite of tools that users can freely select, arrange, and configure through a configuration file composed of four parts, as illustrated in Fig. 1. The first part defines the input and output file names and directories. The second part, the *processing* section, defines the curation steps to be applied. This section can be left empty if no intermediate operations



Table 1. Overview of data curation modules implemented in *marinedb*. Modules are categorized by data dimension (temporal, spatial, taxonomic, and general) and operation type (selection, transformation, or dual).

Module	Category	Description
format	Pre-processing: Transformation	Applies dataset-specific transformations to ensure required structure and formatting prior to processing
isin, notisin	General: Selection	Selects, annotates, or excludes records based on specified field values (e.g., taxonomic groups, presence/absence, licensing)
contains, doesnotcontain	General: Selection	Selects, annotates, or excludes records based on whether specified patterns are present within field values (e.g., multi-valued quality flags)
isna	General: Selection	Identifies, annotates, or excludes records with missing, placeholder, or structurally invalid field values based on user-defined rules (e.g., depth, date or scientific-name fields)
isboundedby	General: Selection	Selects, annotates, or excludes records based on whether field values satisfy user-defined bounds (e.g., coordinate values or uncertainty)
marineloc	Spatial: Selection	Identifies and excludes records located on land
islalloninvalid	Spatial: Selection	Identifies, annotates, or excludes records with missing or out-of-range latitude or longitude values
islallonzero	Spatial: Selection	Identifies, annotates, or excludes records with zero-valued latitude or longitude coordinates
doeslateqlon	Spatial: Selection	Identifies, annotates, or excludes records with identical latitude and longitude values
belowminlatlonprecision	Spatial: Selection	Identifies, annotates, or excludes records with coordinate precision below a user-defined threshold (based on decimal places)
isdateinvalid	Temporal: Selection	Identifies, annotates, or excludes records with invalid dates based on format compliance and calendar validity
convertdate type	Temporal: Transformation	Casts year, month, and day fields to appropriate data types, validates their individual values and combined date, and ensures hierarchical consistency between them
parsedate	Temporal: Transformation	Standardizes dates to ISO 8601 format, reconstructs them from year, month, and day fields when possible, and flags inconsistent or invalid values
processdateinterval	Temporal: Dual	Transforms date intervals into single dates, or annotates or excludes them, based on user-defined strategies and duration thresholds
splitdate	Temporal: Transformation	Extracts year, month, and day from date fields, resolves and flags inconsistencies with existing date component fields, and ensures hierarchical consistency between them
temporal	Temporal: Dual	Applies an integrated temporal processing workflow, including date standardization, interval processing, component extraction, and consistency validation
isdateunlikely	Temporal: Selection	Identifies, annotates, or excludes records with dates outside expected temporal ranges, with user-defined or default bounds
createwormsfilters	Taxonomic: Setup	Cleans and standardizes raw scientific names and retrieves WoRMS matches, including accepted species-level counterparts when available
isinworms	Taxonomic: Dual	Harmonizes taxonomic information using WoRMS matches, resolves ambiguities using classification hierarchy and authorship, and identifies, annotates, or excludes records based on match status (e.g., non-marine, fossil, or supra-specific taxa)
resolvetaxamatch	Taxonomic: Dual	Interactively reviews and resolves uncertain taxonomic matches by comparing dataset records with WoRMS entries and applying user decisions
letteronly	Taxonomic: Selection	Identifies, annotates, or excludes records where specified taxonomic fields contain non-alphabetic characters (e.g., molecular identifiers)
taxasubset	Taxonomic: Dual	Identifies, annotates, or excludes underrepresented taxa and undersamples overrepresented ones based on user-defined occurrence thresholds
mapbasisofrecord	Taxonomic: Transformation	Standardizes observation types by mapping heterogeneous source values to the Darwin Core basisOfRecord vocabulary
basisofrecordisin	Taxonomic: Dual	Identifies, annotates, or excludes records based on user-defined observation types, optionally applying prior standardization



are required. The third part, the *variables* section, specifies which fields to retain in the final dataset, along with instructions for renaming and assigning data types if needed. If left empty, all fields are retained by default. The fourth and final part, the *postprocessing* section, is reserved for operations that depend on the outcome of earlier steps, including resolving uncertain taxonomic matches, filtering underrepresented species, and undersampling overrepresented ones, all of which require prior taxonomic harmonization. This section may likewise be omitted if unnecessary. Once configured, the curation process can be initiated using *clean.py*, the main script that orchestrates the workflow through block-wise processing and supports parallel execution for large datasets. At present, *marinedb* requires a Linux environment. Windows users may run it through the Windows Subsystem for Linux, but native cross-platform compatibility is planned for future releases.

The available curation tools fall into two categories: selection modules, which detect and annotate, retain, or exclude records; and transformation modules, which standardize or restructure data by modifying existing values or generating new curated fields. Two operations, however, do not support non-destructive execution: excluding records located on land and undersampling overrepresented species necessarily discard records and cannot be configured to retain flagged entries instead.

The following subsections describe each functionality and its underlying rationale, organized according to the three major dimensions of the data: temporal, spatial, and taxonomic. For each functionality, we also report source-specific statistics from its application to OBIS, GBIF, and JeDI to characterize their overall fitness for use, and we indicate how the identified issues informed the construction of the final modeling dataset. To quantify the extent of each issue, statistics were computed using a non-destructive, non-cumulative approach: cleaning steps were applied independently of one another after the initial exclusion of terrestrial records, rather than sequentially to subsets altered by previous steps. The only exceptions concerned some species-related and temporal metrics, which relied on prior taxonomic harmonization and date standardization, respectively, to ensure meaningful statistics. A summary of the data-quality statistics is provided in Appendix B (Table B1).

2.3 Data preparation module

Compatibility with *marinedb* requires input datasets to contain specific fields in supported formats. Although failure to provide the expected structure triggers execution errors, these prerequisites are deliberately minimal and should be met by most datasets. When necessary, users can implement the appropriate adjustments in the dedicated *format* module, which is always executed first. By isolating these preparatory transformations, *marinedb* maintains a general-purpose core while retaining the flexibility to accommodate heterogeneous datasets. The required input fields and formats are detailed below, along with external tools for preprocessing tasks beyond the scope of *marinedb*.

The package expects two coordinate columns containing latitude and longitude values in decimal degrees and defined in the EPSG:4326 (WGS84) spatial reference system (SRS). If spatial information is expressed in another format, such as degrees, minutes, and seconds (e.g., 44°50'41.0"N 0°34'03.1"W), Well-Known Text geometry strings, geographic names (e.g., "Northern Adriatic Sea"), or free-text location descriptions (e.g., "Archipiélago de San Andrés, Providencia y Santa Catalina", or "Entre Cap-Bon-Ami et Cap-des-Rosiers"), or if a different SRS is used, the data must be converted beforehand. Such con-



```
data:
  # --- Input / output paths ---

  inputfile_path: ...
  inputdir_path: ...
  outputdir_path: ...
  outputfile_path: ...

  # --- Processing workflow ---

  processing:
    # single-column
    - <column>:
      - <step>:
        <param>: <value>
        <param>: [<value>, <value>, ...]
      ...
    - <step>:
      <param>:
        <key>: <value>
        <key>: <value>
      ...
    # multi-column
    - tool:
      - <step>:
        <param>: <column>
        <param>: <column>
        <param>: <value>
      ...
    # additional steps

  # --- Column selection ---
  # --- Column renaming* ---
  # --- Column type casting* ---
  # *: optional

  variables:
    - <column>
    - <column_before>: <column_after>
    - <column_before>:
      <column_after>: <type>
    # additional columns

  # --- Post-processing operations ---

  postprocessing:
    - tool:
      - <step>:
        <param>: <value>
      ...
    # additional steps
```

Figure 1. Structure of the YAML configuration file defining the data-cleaning workflow. The file is organized into four sections: (i) input/output paths, (ii) processing steps, (iii) column selection, renaming, and type casting, and (iv) postprocessing operations. Angle brackets denote placeholders. The key step denotes a curation operation, column a dataset column, param a step parameter, and value user-defined values. key-value pairs denote dictionary entries when required.

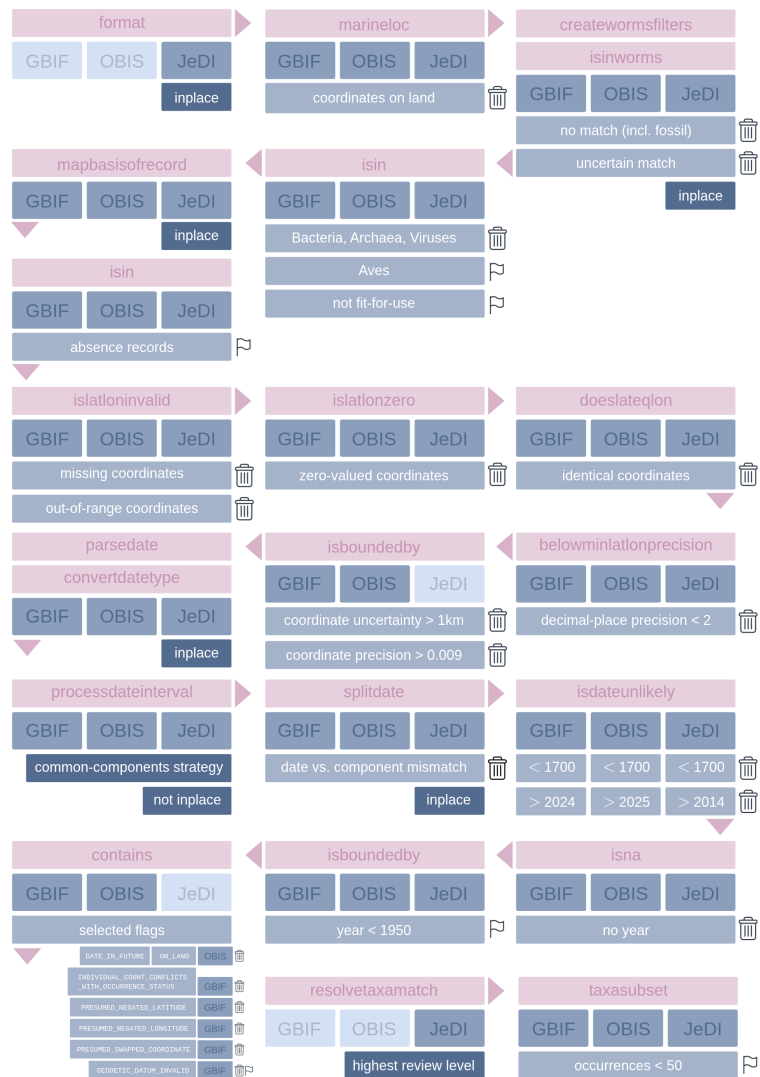


Figure 2. Configuration-based workflow applied to GBIF, OBIS, and JeDI during source-specific dataset curation. Each block corresponds to a *marinedb* module, with database labels indicating the sources to which the module was applied. For selection modules, trash-bin and flag icons indicate whether detected records were removed or annotated and retained. For transformation modules, *inplace* indicates that existing values were modified, whereas *not inplace* indicates that new curated columns were generated.



versions can be performed using online tools (e.g., OBIS Coordinate Conversion Tool³, Marine Regions⁴, GEOLocate⁵, or BELS (VertNet, 2026)), external packages (e.g., *biogeo*, *obistools*, or *sf* (Pebesma, 2018) in R; *BELS* or *geopandas* (Van den Bossche et al., 2026) in Python), or by coding the conversion directly in *format*.

Similarly, *marinedb* requires a column containing full scientific names. In the JeDI dataset, only the specific epithet was provided in the species-related field, so the complete binomial name had to be reconstructed in *format* by prepending the genus. If taxonomic harmonization is enabled, higher-rank columns (e.g., genus, family, order) must also be present. They may be left empty if authorship information is supplied, either in a dedicated field or appended to the scientific name. Likely because JeDI focuses exclusively on taxa within Animalia, it originally lacked a kingdom column, which had to be created to meet this structural requirement.

If the user intends to address the temporal dimension of the data, a column containing observation dates is necessary. This column may be left empty if the dataset includes separate fields for year, month, and day, from which *marinedb* can automatically reconstruct complete dates.

2.4 Spatial modules

Records located on land. *marinedb* provides the *marineloc* module to exclude occurrences located on land. To efficiently process large volumes of data without compromising accuracy, the filtering procedure is divided into two stages (Fig. 3a). The first assigns each observation to one of three categories—land, sea, or coast—based on its coordinates and a precomputed mask. This mask builds on the land–sea mask provided by the *global-land-mask* package (Karin, 2020), which is itself derived from the Global Land One-kilometer Base Elevation (GLOBE) dataset (GLOBE Task Team and others, 1999). To define a third transitional category representing coastal zones, the original binary mask was modified using morphological dilation and erosion. Directly projecting latitude–longitude pairs onto the resulting grid enables the rapid exclusion of records that can be confidently classified before proceeding to the second stage, which is more computationally intensive but considerably more precise. This second step resolves ambiguities associated with coastal observations by leveraging the *basemap* package (Whitaker and contributors, 2025), which draws on high-resolution coastlines from the Global Self-consistent, Hierarchical, High-resolution Geography Database (GSHHG; Wessel and Smith, 1996) (Fig. 3b). Once all records have been classified, those assigned to the land category are excluded from the dataset.

To limit memory usage, the module processes input files iteratively in blocks. This block-wise strategy also provides a useful basis for evaluating the computational benefit of the two-stage approach. For a block of 100,000 records, observations classified entirely during the first stage (i.e., unambiguous land or sea records) were processed in approximately 2 s on average. By contrast, a full block of coastal records requiring second-stage classification took approximately 52 min, making high-resolution spatial processing more than 1,500 times slower than coarse spatial screening. For the GBIF dataset, the first stage

³<https://obis.shinyapps.io/coordinates/>

⁴<https://marineregions.org/>

⁵<https://www.geo-locate.org/web/default.html>



was sufficient to classify 63.2% of the observations. By processing these records early, the mask-based step effectively avoids the computational bottleneck of coastline-based classification for the majority of the dataset.

The exclusion of land-based records can be parallelized either locally (on a single machine) or in a distributed environment (across multiple machines), with each node optionally performing internal parallelization. This combination of intra- and inter-node parallelism makes the approach particularly well suited to large-scale datasets. For the GBIF case described above, approximately one billion records still required second-stage classification after the initial mask-based screening. On a single processor, processing this volume would take roughly one year of continuous computation. By distributing the task across 20 nodes equipped with 10 CPUs each, this timeframe would be reduced to less than two days, assuming near-linear scaling. This comparison illustrates that, without parallelization, applying detailed coastline classification at this scale would be computationally prohibitive. However, distributed execution requires temporarily departing from *marinedb*'s standard configuration-file-based workflow, as detailed in Appendix A2.

Several limitations should nevertheless be considered when applying *marineloc*. First, although GSHHG provides high-resolution global coastline data, it remains a generalized representation and cannot fully capture localized geographic features (Fig. 3b). In addition, some coastal environments are intrinsically difficult to classify: intertidal zones, which lie between high and low tide levels, are alternately exposed and submerged, making any fixed land–sea boundary somewhat arbitrary (Fig. 3b). As a result, some valid coastal records may be erroneously discarded. However, because species distribution modeling is ultimately constrained by the resolution of the environmental layers, such records would likely be excluded at a later stage regardless.

Second, land-based coordinates do not always indicate non-marine observations. The filter may occasionally result in the undue removal of legitimate records, such as observations of marine mammals or birds that spend part of their lives in terrestrial or freshwater habitats. It may also exclude genuine marine occurrences with imprecise coordinates, such as records derived from gridded data or rounded to the nearest minute or degree, where the reported location falls inland despite the observation having occurred at sea. In addition, erroneous coordinates, such as the observer's position being recorded instead of the species' location, can lead to the exclusion of otherwise valid observations. Nevertheless, unlike the removal of ecologically complex but valid data, discarding imprecise or incorrect cases is generally desirable for species distribution modeling, as such occurrences lack the spatial accuracy required for reliable predictions.

The marine filter excluded 2,602,503,097 occurrences (92.6%) from GBIF, 19,393,076 (14.6%) from OBIS, and 10,146 (2.4%) from JeDI. For GBIF, the large removal of records is consistent with its role as a cross-realm biodiversity infrastructure. More broadly, the overwhelming dominance of land-based records in the initial dataset reflects the persistent imbalance in data availability between terrestrial and marine realms. Given their marine focus, the proportion removed from OBIS and JeDI may appear more unexpected, but it should not be interpreted as a direct measure of erroneous terrestrial records in these databases for the reasons discussed above. For the final modeling dataset, records classified as land were excluded.

Common geospatial issues. *marinedb* can detect and either annotate or remove several common geospatial issues: (1) missing coordinates (latitude, longitude, or both), (2) coordinates falling outside valid boundaries ($-90 \leq \text{latitude} \leq +90$; $-180 \leq \text{longitude} \leq +180$), (3) zero coordinates (latitude, longitude, or both), and (4) identical latitude and longitude values. While an

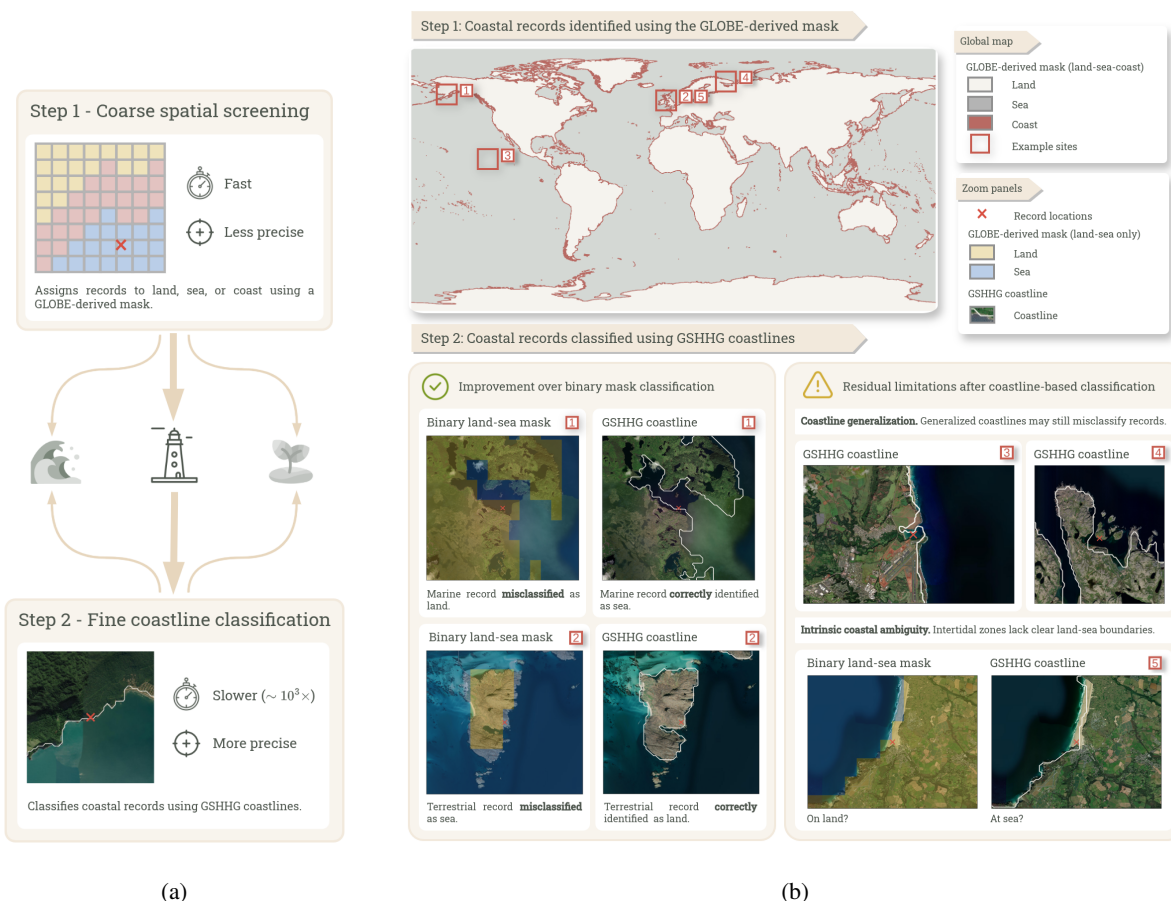


Figure 3. Two-step land–sea filtering of occurrence records. (a) Two-stage *marineloc* workflow. A coarse GLOBE-derived mask is first used to rapidly classify unambiguous land and sea records, while coastal records are passed to a second, more precise GSHHG-based classification stage. (b) Examples of classification improvements and residual limitations. The examples show how coastline-based classification can correct errors produced by the original binary GLOBE-derived land–sea mask, while also illustrating residual limitations related to coastline generalization and intrinsic ambiguity in intertidal zones. Background imagery: Esri World Imagery, used as visual reference for land–sea interpretation.



observation located exactly at (0,0), in the Gulf of Guinea (Atlantic Ocean), or precisely along the equator (latitude = 0) or the Greenwich meridian (longitude = 0) is theoretically possible, such cases are highly unlikely in practice. These anomalies typically result from data entry errors or from missing values automatically replaced with zeros. Similarly, identical latitude and longitude values usually reflect erroneous input rather than a real location. Each of these issues is addressed by a dedicated module: missing and out-of-bounds coordinates are handled by *islatloninvalid* (or alternatively by *isna* and *isboundedby* if users require separate flags for latitude and longitude), zero-valued coordinates by *islatlonzero* (or by *iszero* for separate flagging), and identical coordinates by *doeslateqlon*.

Two additional issues frequently encountered in georeferenced datasets are swapped coordinates and incorrect coordinate sign attribution. These errors can often be identified through inconsistencies between the reported coordinates and associated metadata or external reference data, such as spatial outliers (e.g., geographically isolated records or mismatches between country and coordinates), environmental outliers (e.g., records falling outside the expected temperature, salinity, or depth range of the species), or taxonomic outliers (e.g., species absent from the relevant national checklist). Although *marinedb* does not currently address these cases, several external tools can assist in their detection. For instance, the GBIF data processing pipeline includes a check for coordinate–country consistency; its online data validator can therefore be used to flag potentially problematic records (though its treatment of occurrences in international waters remains unclear). The *biogeo* package offers various outlier detection functions across spatial and environmental dimensions, based on user-supplied information. It can be combined with resources listed by the LifeWatch E-lab⁶, such as the Marine Regions gazetteer service for retrieving geographic names from coordinates. However, most of the *biogeo* outlier detection tools are geared toward manual quality control and visualization rather than automated cleaning, and may not scale well to very large datasets. Similarly, while not suitable for large-scale processing, visual inspection of observation points using, for example, the mapping function provided by *obistools*, can help identify and flag or remove obvious geospatial errors. Future development of *marinedb* could address coordinate-order and sign-attribution errors directly by integrating spatial, environmental, or taxonomic consistency checks.

No records with missing or out-of-range coordinates were found in GBIF, OBIS, or JeDI after the *marineloc* curation step, as such records are systematically removed during land–sea filtering. For GBIF and OBIS, this result was also expected from the input data: records lacking valid coordinates had been excluded during the GBIF download, while OBIS quality control prevents publication of records with missing or out-of-range latitude or longitude values. OBIS also removes records for which both coordinates are zero, meaning that the zero-valued cases reported below correspond to records where only one of the two coordinates equals zero. In GBIF, records with coordinates at (0,0) are flagged but retained. *marinedb* identified 961,751 records in GBIF (0.46%), 29,843 in OBIS (0.03%), and 691 in JeDI (0.16%) with at least one zero-valued coordinate, as well as 971,610 records in GBIF (0.46%), 1,483 in OBIS (< 0.01%), and 6 in JeDI (< 0.01%) with identical latitude and longitude values. Although low in proportion, the absolute number of affected GBIF records was substantial, approaching one million for each anomaly. A threshold-related discrepancy with the GBIF zero-coordinate flag was also observed: 23,365 records flagged by GBIF as zero-coordinate records, representing two distinct near-origin locations, were not flagged by *marinedb*. This reflects different operational thresholds for zero-valued coordinates: $|\text{coordinate}| \leq 10^{-5}$ in *marinedb*, compared with an

⁶<https://lifewatch.be/e-lab>



315 apparent GBIF threshold of $\sim 10^{-3}$ inferred from the discrepant records. By contrast, OBIS and JeDI contained comparatively few records exhibiting these two geospatial issues. Together, the OBIS and GBIF cases illustrate the complementarity of downstream, fitness-for-use checks: source-level quality control may remove clearly invalid configurations, such as coordinates at (0,0), or retain them with quality flags, while leaving other suspicious patterns, such as a single zero-valued coordinate or identical latitude–longitude values, to be detected during use-case-specific processing. For the final modeling dataset, records with missing, out-of-range, zero-valued, or identical coordinates were excluded. Additionally, leveraging GBIF quality flags for geospatial issues not yet addressed by *marinedb*, records flagged for failed coordinate reprojection or presumed coordinate swapping or sign inversion were also excluded.

Positional uncertainty. Positional uncertainty can affect species distribution model performance by distorting estimated species–environment relationships. When explicit positional metadata are available, the *isboundedby* module can annotate or exclude records for which coordinate uncertainty exceeds a user-defined maximum, or coordinate precision falls below a user-defined minimum. In the absence of such metadata, decimal-place precision can be used as a proxy for positional uncertainty. Decimal-place precision refers to the number of decimal places recorded in the coordinates, retaining the higher precision when latitude and longitude differ. Accordingly, *marinedb* provides the *belowminlatlonprecision* utility to identify records whose decimal-place precision falls below a user-defined threshold. In practice, this threshold is often set to two decimal places, although the appropriate value depends on the spatial resolution and autocorrelation of environmental predictors, as well as species ecology such as body size, mobility, and niche breadth (Moudrý et al., 2024). At the equator, two decimal places correspond to approximately 1.11 km, with this distance decreasing toward the poles due to meridian convergence (e.g., ~ 0.79 km at 45° latitude). Decimal-place precision should nevertheless be interpreted cautiously because it reflects coordinate formatting rather than actual positional uncertainty. Apparent precision may be inflated by trailing-zero padding or floating-point serialization artifacts, whereas coordinates may be deliberately generalized to protect sensitive taxa (Chapman, 2020). Filtering on this criterion alone may therefore retain records with falsely high apparent precision while disproportionately excluding generalized records from sensitive taxa, potentially introducing taxonomic or conservation-related biases.

OBIS and GBIF include two Darwin Core fields that can inform positional uncertainty: `coordinateUncertaintyInMeters`, which expresses the radius of uncertainty around the reported coordinates, and `coordinatePrecision`, which describes coordinate precision in decimal degrees. In GBIF, both fields are subject to basic validity checks: non-numeric values are flagged and removed, as are values outside plausible ranges, defined as 0–1 for `coordinatePrecision` and 0–20,037,509 m for `coordinateUncertaintyInMeters`. In OBIS, documented quality-control procedures are limited to `coordinateUncertaintyInMeters`: values that do not fall within 0–10,000,000 m are flagged but retained. However, values were absent for most records in both fields: `coordinatePrecision` was missing for 96.10% of GBIF records (200,727,142 records) and 91.10% of OBIS records (103,012,687 records), while `coordinateUncertaintyInMeters` was missing for 60.39% of GBIF records (126,144,548 records) and 56.62% of OBIS records (64,026,344 records). When a value was available for the latter field, uncertainty was below 1 km for 66.16% of GBIF records and 91.32% of OBIS records. JeDI contains no equivalent fields. Records from GBIF and OBIS with `coordinateUncertaintyInMeters` greater than 1 km were excluded from the final modeling dataset. A comparable threshold was applied to `coordinatePrecision`,



with records exceeding 0.009 decimal degrees removed. For OBIS, this required an additional standardization step outside the configuration-based workflow, in which a small number of non-standard values, including strings containing explicit units (e.g., "1000m"), were converted to a common decimal-degree representation when interpretable, or treated as missing otherwise (e.g., "from gazetteer"). Values greater than 1 decimal degree were additionally considered implausible and ignored for this assessment.

The limited availability of explicit positional metadata underscores the value of decimal-place precision as an alternative criterion that can be applied consistently across databases. For the present data quality assessment, we used a threshold of two decimal places. Widely used global marine environmental products are commonly provided at spatial resolutions of several kilometres or coarser, including Bio-ORACLE at 0.05° (Assis et al., 2024), World Ocean Atlas products at 0.25° to 1° (Reagan et al., 2023), and Copernicus Marine global reanalysis products at up to 0.083° (E.U. Copernicus Marine Service Information (CMEMS), 2025). The two-decimal-place threshold is therefore consistent with recommendations to keep positional uncertainty lower than the spatial resolution of the predictors and, ideally, lower than half that resolution (Moudrý et al., 2024). However, applying a single threshold remains a simplification in a multi-species framework, as discussed above. Using this criterion, *marinedb* flagged 5,042,045 records in GBIF (2.41%), 923,155 in OBIS (0.82%), and 24,078 in JeDI (5.74%). Although the absolute number of low-precision records was substantial in GBIF and OBIS, their small proportion relative to the size of each dataset suggests that their exclusion would not significantly reduce data availability for broad-scale multi-species modeling. As with any filtering step that removes a large number of records, however, the impact on downstream model performance depends on the distribution of excluded observations across taxa, regions, and environmental conditions. Records with coordinates reported to fewer than two decimal places were excluded from the final modeling dataset.

Depth. Missing depth can reduce the ecological realism of downstream models by preventing records from being matched to the appropriate environmental conditions, and may affect model performance in vertically structured marine environments (Klaassen et al., 2026). Depth itself may also carry predictive information for some taxa when it acts as a direct geographic constraint, as in the case of deep-diving cetaceans, making depth completeness relevant beyond environmental matching (Klaassen et al., 2026). Even when depth is not the primary descriptor of the organism's position, Darwin Core guidelines indicate that an explicit depth value may still be needed to define the reference surface from which vertical distance is measured, as for occurrences above the sea surface, such as flying birds, or for taxa detected in sediment-core samples taken from the seabed. *marinedb* can annotate or exclude records lacking depth information using the *isna* module. *isna* is a general-purpose utility for identifying records with missing values and can optionally standardize heterogeneous encodings of missingness before processing. When enabled, standardization recodes empty fields and predefined placeholders (e.g., "-1.#IND", "#n/a", "unknown") to a common internal representation. Users can also supply dataset-specific placeholders, such as "nd" in JeDI, and enforce minimal format rules requiring a field to contain at least one digit, one letter, or either. Records that do not meet the selected rule are treated as uninformative for the targeted field and marked as missing. For example, a depth field can be required to contain at least one digit, allowing non-standard but potentially valid entries such as "10 m" to be retained while excluding values such as "?". The utility does not, however, assess whether reported depths are valid. Such checks can instead be performed with external tools such as *obistools*, which by default compares reported depths with



385 bathymetry from the General Bathymetric Chart of the Oceans (GEBCO Bathymetric Compilation Group 2026, 2026) and
 verifies, when intervals are specified, that maximum depths exceed minimum depths. When datasets already include quality
 flags derived from depth-validation procedures, as in OBIS and GBIF, the *isin* and *notisin* modules can be used to annotate
 or exclude records based on these flags. Future versions of *marinedb* could integrate similar checks directly, building on the
 approach implemented in *obistools* or GBIF. Once depth values have been validated, the *isboundedby* module can classify
 390 observations according to user-defined thresholds, enabling records to be assigned to depth-based categories such as epipelagic
 or mesopelagic zones.

Depth information differs in structure across databases. OBIS and JeDI encode depth using minimum and maximum depth
 fields, which may be equal. GBIF, by contrast, provides a single interpreted depth field, which represents the midpoint between
 minimum and maximum depth when both bounds are supplied and valid, together with a depth accuracy field corresponding
 395 to half of the interval width. When only one bound is provided or can be parsed, that value is retained as the interpreted depth.
 GBIF depth-related quality flags, however, refer to the input minimum and maximum depths. In OBIS, 37.22% of records
 lacked values in both depth fields (42,083,645 records), compared with 41.01% in JeDI (171,976 records). Depth completeness
 was substantially lower in GBIF, where 74.17% of records had no interpreted depth value (154,914,642 records). Incomplete
 depth information is therefore a major limitation across the three databases.

400 Among records with depth information, the most frequent issue identified by OBIS quality flags was reported depth ex-
 ceeding reference bathymetry, affecting 1,050,639 records (1.48%). OBIS documentation notes, however, that records may
 occasionally be flagged incorrectly when reported depths are more accurate than the gridded GEBCO bathymetry used for
 the validity check. Internal depth inconsistencies were less frequent, with 93,455 records (0.17%) for which minimum depth
 exceeded maximum depth. Among GBIF records with evidence of depth information, defined as records with an interpreted
 405 depth value or at least one depth-related quality flag, 44,877 records (0.08%) were flagged as having non-numeric depths, of
 which 23,801 (53.04%) contained non-metric depth information. Additional flags identified 155,785 records (0.29%) with
 unlikely depths (i.e., negative or greater than 11,000 m), and 288,912 (0.54%) with potentially swapped minimum and max-
 imum depth values. All flagged records nevertheless retained an interpreted depth value, likely reflecting GBIF's ability to
 convert non-metric units to metres, reinterpret negative depths as positive values, and correct swapped depth bounds. How-
 410 ever, the GBIF ingestion workflow can introduce artefacts into interpreted depth values when parsing succeeds for only one
 of two supplied bounds, causing part of the original depth information to be ignored. More critically, incorrect identification
 and conversion of depth units can generate erroneous depth values, as observed in cases where the preposition "in" was in-
 terpreted as an inch marker (e.g., "in 6 m depth"). Because JeDI contained no comparable source-level flags, depth quality
 was assessed by direct inspection. All non-missing JeDI depth values were numeric, and no records fell outside GBIF's range
 415 for unlikely depths except entries coded as -999, a common missing-value placeholder in oceanographic datasets. However,
 among JeDI records with depth information, 55,670 records (22.51%) had minimum depths exceeding maximum depths, the
 highest proportion of depth inconsistencies observed among the three databases. Given the high proportion of missing values,
 the heterogeneous quality of depth information within and across sources, and the lack of dedicated depth-validation and stan-



standardization procedures in the current version of *marinedb*, depth-related fields were not retained in the integrated modeling dataset.

2.5 Temporal modules

Date format. Date standardization is challenging because occurrence datasets contain heterogeneous and ambiguous date representations, including differences in delimiters and date-component order (e.g., "31.12.1997" versus "12/31/1997"), variation in date precision (e.g., "2001" or "12-1997"), date intervals (e.g., "1995-11-04 / 1995-11-06"), textual dates in multiple languages (e.g., "December 1st 1997" or "1. Dezember 1997"), and time-zone information (e.g., "2007-04-05T12:30-02:00"). To support robust interpretation of verbatim dates, the *parsedate* module standardizes dates to ISO 8601 format using the date-parsing component of the GBIF Parsers library (GBIF.org, 2025), originally developed for the GBIF ingestion pipeline and repackaged as a standalone JAR file for direct use in Python. *parsedate* also discards time-of-day information and supports parallel execution. If conversion fails or the date is missing, *parsedate* attempts to reconstruct it from the year, month, and day fields, provided that these fields exist in the input data, a valid year is available, and the available components form a valid calendar date without clearly conflicting with any existing but unparseable date. Otherwise, the date is replaced with a missing value. For example, "02-03-2007" is an ambiguous non-ISO 8601 date, intentionally left unparsed by the GBIF parser because the first two components could represent either day-month or month-day order. If the separate fields indicate "2007" for the year, "02" for the month, and "21" for the day, reconstruction is aborted because the day field clearly contradicts the verbatim date. Any issues detected during processing are recorded in a separate field.

Some records may nevertheless be incorrectly flagged as inconsistent due to code limitations, similar to those of the GBIF date parser, such as unsupported languages (e.g., "8. veebruar 2007" in Estonian) or formats (e.g., week-based formats like "2005-W52-1"). Such misclassifications can also occur in edge cases where the separate date components fall within an unparseable date range (e.g., "2021", "04", "06" for year, month, and day, with a range string such as "2021-01-21 | 2021-05-05", where the vertical bar is not supported as a range delimiter). Conversely, some mismatches may go undetected, particularly in cases of unparsed date intervals or inherent ambiguity. For instance, "06/10/1997" could represent either October 6 or June 10, depending on the intended format. If the year and month fields contain "1997" and "06" while the day is missing, the apparent consistency may conceal an actual mismatch. In such situations, the correct interpretation cannot be determined with certainty, necessitating a design choice on how to proceed. Overall, *marinedb* adopts a "reasonable doubt" strategy: unless a clear conflict is detected, data are assumed to be internally consistent, and the component order implied by the available separate fields is treated as authoritative, acknowledging that this approach may occasionally produce undetected (and likely undetectable) mismatches of the types described above.

In OBIS, 4,130,996 records (3.65%) lacked a date. Of these, 1,703,491 (41.24%) were reconstructed from the separate year, month, and day fields. The remaining missing-date records could not be reconstructed either because no year value was available (2,417,752 records; 99.60%) or because the year value was invalid (9,753 records; 0.40%), including ranges, zero values, or implausible numeric entries. Among OBIS records with an initial date value, 108,405,042 (99.50%) contained a well-formatted and valid date. The remaining 540,180 dated records (0.50%) could not be parsed because of formatting issues



or invalid dates, including malformed intervals, unsupported or ambiguous delimiters, ambiguous two-digit years, ambiguous month–day ordering, invalid components, and swapped interval endpoints. Among these records, 369,939 (68.48%) were
 455 reconstructed from the separate year, month, and day fields. Most unreconstructed cases lacked a year value (167,696 records; 98.51%), while *parsedate* detected internal inconsistencies between the date field and the separate components in 2,541 records (1.49%). Manual inspection, however, showed that a small subset of the latter cases represented false mismatches caused by unparseable date intervals rather than genuine conflicts.

For JeDI, 23,270 records (5.55%) lacked a date, but all were reconstructed from the separate year, month, and day fields.
 460 These cases corresponded to records for which only year or year and month information was available, likely explaining why the date field was empty. Among JeDI records with an initial date value, 395,542 (99.87%) could not be parsed because their dates included two-digit years and, in some cases, ambiguous month–day ordering. All of these dates were reconstructed from the separate fields, as no missing or invalid components and no internal inconsistencies were detected.

For GBIF, a temporal filter was applied when downloading the data, retaining only records within the range 1700–2024 and
 465 thereby excluding records without dates. The absence of missing dates in the analysed GBIF subset should therefore not be interpreted as evidence of complete date coverage in GBIF as a whole. Among the retained records, no formatting or validity issues were found in the date fields. However, GBIF quality flags indicated that 1,995,569 source records (0.96%) had such issues, despite being associated with a date value in the downloaded dataset. This illustrates both the presence of date-related problems in source data and the effect of GBIF’s quality-control workflow, which produces cleaner interpreted date fields
 470 through parsing, validation, and reconstruction from separate year, month, and day fields where possible.

Overall, temporal information was largely available before processing, but its representation and quality differed among databases. In OBIS, 95.87% of records had a date with no formatting or validity issue, and *parsedate* recovered 44.39% of missing, invalid, or poorly formatted dates. In JeDI, date-format problems affected 94.33% of records, but all missing or un-
 475 parseable dates were reconstructed and no invalid components or internal inconsistencies were detected. In GBIF, the results primarily reflected the quality of GBIF-interpreted date fields, with no remaining formatting or validity issues detected in the analysed subset. For the final modeling dataset, dates were standardized in place using *parsedate*, and records lacking a usable date after processing were excluded.

Date interval. Date intervals introduce temporal uncertainty, as an occurrence can correspond to any date within the reported range. The consequences for downstream modeling depend on how much the relevant environmental conditions vary within that
 480 interval and on how this variability is represented in the environmental data. For predictors that are effectively time-invariant at the temporal scales typically considered in species distribution modeling, such as bathymetry, uncertainty in the observation date is unlikely to affect the environmental value assigned to a record. By contrast, predictors such as sea surface temperature or chlorophyll concentration can vary substantially over short timescales. When fine-resolution environmental data are used, different dates within a long occurrence interval may therefore correspond to markedly different environmental values. At coarser
 485 temporal resolutions, the same environmental value may be assigned to multiple dates, reducing the additional uncertainty introduced by the interval, but at the cost of smoothing potentially relevant environmental variability. The potential impact of a date interval on occurrence–environment matching therefore depends jointly on its duration, the temporal variability of the



predictor, and the temporal resolution of the environmental data. Appropriate handling strategies may include assigning the environmental value from a single date within the interval, summarizing environmental values across the interval, or excluding records whose temporal uncertainty exceeds a predefined threshold.

To address these challenges, the *processdateinterval* module allows users to annotate or exclude records containing intervals, or to collapse intervals to a single temporal value. Collapsing follows one of three strategies: retaining the start date, retaining the end date, or preserving only the components common to both (e.g., the year, if the two dates differ in month). For the first two strategies, users can restrict processing to intervals shorter than a specified threshold, expressed in days, months, or years, and replace dates with missing values when the threshold is exceeded. Unlike retaining the start or end date, the common-components strategy avoids assigning the occurrence to an arbitrary position within the interval. Instead, it preserves the finest temporal precision shared by both interval endpoints, which can then guide subsequent exclusion of records or aggregation of environmental data.

Date intervals were present in GBIF and OBIS but absent from JeDI. In GBIF, 5,999,994 records (2.87% of parsed dates, defined here as dates retained or reconstructed by *parsedate*) contained date intervals. No interval exceeded one year, although this pattern may be a consequence of the temporal filter applied during download. Nevertheless, 1,942,106 interval records (32.37%) spanned more than one month, indicating that a substantial fraction of GBIF intervals may require aggregation or exclusion when matching to environmental data at fine temporal resolution. In OBIS, 2,815,869 records (2.55% of parsed dates) contained date intervals. Of these, 153,797 (5.46%) spanned more than one year and 330,915 (11.75%) spanned more than one month. For subsequent temporal assessments and in the final modeling dataset, date intervals were collapsed in an additional field using the common-components strategy, while the standardized date field was retained unchanged.

Year, month and day. The *splitdate* module extracts year, month, and day components from dates. When the input data also include separate year, month, or day fields, as in the three databases analysed here, these fields are first processed with *convertdatatype*, which checks for hierarchical inconsistencies (e.g., a month without a year or a day without a month), validates each component individually, and verifies the combined date they form. The components extracted from the date field are then compared with the corresponding values in the separate fields. When inconsistencies are detected, *splitdate* applies one of two user-defined strategies. The first removes inconsistent components from both the date field and the separate fields, while enforcing hierarchical integrity by also removing dependent components. For example, if the date is “2002-10-07” but the separate month field contains “09”, both the month and day are removed, yielding “2002” in the date field and missing values in the month and day fields. The second strategy overwrites inconsistent values in the separate fields with the components extracted from the date field. Giving precedence to the date field is a design choice rather than an indication that this field is inherently more reliable. Under this strategy, the same example would retain “2002-10-07” in the date field and replace the separate month value with “10”. Regardless of the chosen strategy, any issues detected during processing are recorded in a separate field.

For the final modeling dataset, year, month, and day fields were derived from the processed date field, with inconsistent components removed from both the processed date and pre-existing date-component fields.



In OBIS, before component extraction, 63,989,550 records (56.59%) lacked at least one year, month, or day value in the separate date-component fields. Of these, 86.09% had all three values missing. A further 69,188 records (0.12%) exhibited hierarchical inconsistencies, invalid components, or components forming invalid dates. During extraction, *splitdate* detected internal inconsistencies between the processed date field and the separate components in 33,302 records (0.03% of parsed dates), and these inconsistent components were removed under the mismatch-removal strategy. After extraction, the number of records missing at least one separate date component decreased markedly to 8,352,244 (7.39%). At the end of the full temporal workflow, 2,810,313 OBIS records (2.49%) still lacked date information, whereas 94.97% of records with date information were resolved to day level (104,723,974 records). The remaining records had either month-level (1,334,380 records; 1.21%) or year-level precision (4,207,551 records; 3.82%).

For JeDI, 23,273 records (5.55%) initially lacked at least one separate date component, but no record lacked all three components. No hierarchical inconsistencies, invalid components, or components forming invalid dates were found. During extraction, *splitdate* detected no internal inconsistencies between the processed date field and the separate components. This absence should be interpreted cautiously, as 99.88% of JeDI records had their dates reconstructed from the separate fields during date standardization. Thus, only the remaining 503 records provided an independent comparison between an existing, valid date and separate year, month, and day values, all of which were consistent. Overall, component extraction did not change the number of records missing at least one separate component: recovered dates contained no information beyond that already present in the separate fields, and there were no date intervals or internal inconsistencies that would have caused component loss. At the end of the full temporal workflow, date information was available for all JeDI records. Of these, 94.45% were resolved to day level (396,043 records), while 7,197 records (1.72%) had month-level precision and 16,076 records (3.83%) had only year-level precision.

For GBIF, 15,192,517 records (7.27%) initially lacked at least one separate date component, but no record lacked all three components, consistent with the temporal filter applied during download. No hierarchical inconsistencies, invalid components, components forming invalid dates, or internal inconsistencies between the processed date field and separate components were detected in the analysed dataset, reflecting the prior GBIF quality-control workflow. GBIF quality flags nevertheless indicated that 17,843 source records (0.01%) had inconsistencies between date fields before interpretation. Because date fields had already been cleaned and harmonized during GBIF ingestion, the number of records missing at least one separate component remained unchanged after *splitdate* component extraction. At the end of the full temporal workflow, date information was available for all GBIF records. The vast majority were resolved to day level (193,683,239 records; 92.73%), while smaller fractions had month-level (5,817,463 records; 2.79%) or year-level precision (9,375,054 records; 4.49%).

Missing, invalid or unlikely date. *marinedb* can detect and either annotate or exclude observations whose dates are missing (*isna* or *isdateinvalid*), invalid (*isdateinvalid*), or unlikely (*isdateunlikely*). A date is considered invalid if it does not conform to the ISO 8601 format, specifies a non-existent month or day, or represents an impossible date, accounting for leap years (e.g., "2003-04-31", "1900-02-29"). Depending on user settings, missing dates may also be treated as invalid. A date is considered unlikely if it falls before 1700, predating modern taxonomy, or after the current year. These bounds are user-configurable. In the analyses presented here, the upper bound was adjusted to the temporal scope of each dataset: the download year was used



for GBIF and OBIS (2024 and 2025, respectively), whereas the dataset version year reported by BCO-DMO was used for JeDI (2014). All of the temporal-dimension functionalities described in this section so far, except for the handling of unlikely dates, are consolidated in the *temporal* module, streamlining their use and facilitating integration into processing workflows.

560 Because date validity, final date completeness, and temporal precision were described above, the following results focus on unlikely dates. No unlikely dates were detected in GBIF, consistent with the GBIF quality-control workflow, which removes unlikely source dates, and with the temporal filter applied during download, which excluded records left without date information after GBIF interpretation. No unlikely dates were detected in JeDI when using 2014 as the upper temporal bound. In OBIS, *isdateunlikely* flagged 4,440 records ($< 0.01\%$). OBIS quality flags also identify dates outside expected temporal bounds, but
 565 these flags captured only a subset of the records detected by *marinedb*. This difference mainly reflects the fact that, in this analysis, unlikely dates were assessed after date standardization and reconstruction, whereas OBIS quality-control checks are applied to the original date field and fail when the value is missing or unparsable. Additional discrepancies arose from differences in the temporal bounds used by *marinedb* and OBIS. A few edge cases also involved year-range strings such as “2012-2013”, which were not interpreted as intervals by the GBIF parser used in *marinedb* and therefore led to erroneous unlikely-date flags.
 570 Conversely, six records flagged by OBIS as future were not flagged by *marinedb*, illustrating a time-dependent limitation of using the current or download year as the upper bound: a date flagged as future at a given processing time may no longer fall beyond the upper bound when the dataset is reprocessed later. Nevertheless, this dependence remains traceable in *marinedb*, because the output field name records the temporal bounds used for the check. This also illustrates the complementary value of exploiting source-level quality flags, which may preserve information that is not directly recoverable from subsequent re-
 575 processing. For the final modeling dataset, records with unlikely dates were excluded using the dataset-specific upper bounds defined above. OBIS records carrying the *DATE_IN_FUTURE* quality flag were additionally filtered out.

Old or out-of-range records. Finally, the *isboundedby* module can annotate or exclude observations outside a user-defined temporal range, including records predating a specified year. Historical records require particular attention because they may combine several sources of uncertainty. First, older occurrences are more prone to spatial imprecision or georeferencing error, especially when coordinates have been inferred retrospectively from locality descriptions (Chapman, 2005a, b; Soberón and Peterson, 2004). Second, the taxonomic identity of older records may be less reliable because of incorrect or unverifiable historical identifications or names of doubtful application, and may be difficult to reconcile with current taxonomy because of outdated synonyms (Soberón and Peterson, 2004). Third, historical occurrences may introduce temporal mismatch that can affect species distribution model performance when they are paired with contemporary environmental predictors, particularly
 585 for variables that change substantially through time (Bracken et al., 2022). However, excluding observations can also reduce sample size and environmental representativeness, especially for poorly sampled taxa or regions. The appropriate threshold therefore depends on the trade-off between reducing spatial uncertainty, taxonomic uncertainty, and temporal mismatch while retaining sufficient occurrence data. Common thresholds include 1900, 1945, and 1950, but *marinedb* allows users to define analysis-specific lower and upper bounds. Among records with a year value, 2,399,735 GBIF records (1.15%), 963,676 OBIS
 590 records (0.87%), and 4,497 JeDI records (1.07%) were dated before 1950. Records predating 1950 were flagged rather than



excluded from the final modeling dataset, because the appropriate historical threshold depends strongly on the temporal scope of the analysis, particularly with respect to the environmental data used.

2.6 Taxonomic modules

Taxonomic harmonization. A major strength of *marinedb* lies in its ability to harmonize scientific names and classifications.

595 This step is essential not only for identifying records assigned to terrestrial taxa despite being georeferenced at sea, whether because of identification errors or spatial data-quality issues, but also for reconciling records across taxonomic reference databases. Taxonomic names are not one-to-one identifiers: the same name can designate different taxa (homonymy), whereas multiple names can refer to the same taxon (synonymy). In practice, discrepancies can arise both because different reference databases may classify the same organisms in conflicting ways and because the same reference database may yield distinct
 600 results when applied at different times, as taxonomies evolve. These inconsistencies are further compounded by the uneven update frequencies of occurrence databases, which result in differential incorporation of taxonomic revisions. For instance, OBIS relies exclusively on WoRMS and incorporates changes within weeks, whereas GBIF builds its backbone taxonomy from 105 catalogs, including WoRMS, and updates it at most twice a year. By contrast, the taxonomic standard used by JeDI is undocumented, and because the BCO-DMO version of the dataset used in this study has not been updated since its
 605 release in 2014, its taxonomic information has likewise remained unchanged. These discrepancies have direct consequences for species distribution modeling: without harmonization, occurrences of the same species may be fragmented across multiple names, whereas occurrences of different species may be incorrectly aggregated under a shared name, thereby weakening model reliability and ecological realism.

The taxonomic harmonization process begins with *createwormsfilters*, which first prepares scientific names for querying the
 610 World Register of Marine Species (WoRMS), an authoritative and comprehensive reference for marine taxon names. Cleaning and standardization routines are applied to increase the likelihood of successful name matching. Rather than relying solely on predefined assumptions, these routines were formulated by analyzing the wide range of irregularities in GBIF verbatim scientific names, from which the examples below are drawn. They standardize capitalization to the expected nomenclatural format, in which only the genus name begins with an uppercase letter (e.g., “Myocastor COYPUS BONARIENSIS”, “Graneledone
 615 Eledoninae verrucosa”), handle special characters (e.g., “Spirorbis (Spirorbis) tridentatus”, “Beroe cucumis | cydippid larva”), and fix encoding errors (e.g., “F.SchÄ¼tt”). They also strip authors and dates (e.g., “Ceramium virgatum Roth”, “Gonyaulax spinifera (Clap. & J.Lachm.) Diesing, 1866”) as well as numeric substrings (e.g., “BOLD:AEF8294”). In addition, they remove nomenclatural qualifiers (e.g., “Coprinellus micaceus s.l., incl. pallidissimus, saccharinus, truncorum”, “Megaselia sp. BIOUG27368”) and contextual qualifiers (e.g., “endosymbiont of Lamellibrachia barhami”, “Calathura brachiata juv.”). This
 620 removal facilitates taxonomic matching but does not preserve the uncertainty conveyed by qualifiers such as “cf.”, “aff.”, or “s. lat.”. Consequently, records that could warrant exclusion as unreliable identifications in conservative species distribution modeling applications are retained without being explicitly flagged as uncertain. They also detect composite strings containing multiple taxa, for which only the first taxon is retained for subsequent matching (e.g., “Clupea harengus / Sprattus sprattus”, “Haematopus ostralegusNumenius arquata”). Because hybrid names and formulas may require dedicated representation



625 and validation (Chapman, 2005a), they are excluded from automated matching in this general-purpose workflow (e.g., “Rhizophora × harrisonii”, “Branta hutchinsii x Branta leucopsis”). Before querying WoRMS, *createwormsfilters* also excludes likely supraspecific entries, such as one-word taxonomic strings.

In a second step, the remaining taxon names are submitted to the WoRMS web service. As taxonomic names in occurrence databases may be affected by typographical errors, misspellings, or spelling variants, WoRMS uses fuzzy matching to identify
 630 taxonomic entries whose names closely resemble, but need not exactly match, the submitted name. The service then returns zero, one, or multiple matches. Beyond name matching, these queries retrieve WoRMS-derived taxonomic information, including the full classification, authority, nomenclatural status, accepted name when available, and, optionally, stable identifiers such as LSIDs. The returned metadata also include extinction status and environmental flags indicating whether a taxon is marine, brackish, freshwater, or terrestrial. For each synonym or infraspecific name, *createwormsfilters* uses this information to resolve
 635 the entry to the currently accepted scientific name when available; otherwise, it retains the unaccepted but still recognized WoRMS entry. To avoid artificially reducing the pool of possible matches, queries are performed without restricting candidate entries to extant or marine taxa. Filtering based on these attributes is applied at a later stage. This step can be parallelized, subject to WoRMS’ limitation of two concurrent queries.

The next stage, handled by *isinworms*, consolidates taxonomic matching and resolves ambiguities among multiple matches
 640 by leveraging the remaining classification hierarchy and, when available, authorship information. Comparing the full classification helps rule out false matches caused by homonyms belonging to different taxonomic groups—for instance, the marine fossil *Cyathus striatus* Cooper, 1946, versus the terrestrial fungus *Cyathus striatus* (Hudson) Willdenow, 1787—and can reduce the set of candidate matches. In many cases, however, this comparison alone is not sufficient: multiple potential matches may remain after filtering, results may be unreliable when higher-rank information is missing or incomplete, and matching may even
 645 fail despite strong agreement at the kingdom and species levels, potentially indicating taxonomic revisions. In such situations, *isinworms* also compares taxonomic authorship, when this information is available. Users may specify multiple columns to maximize the chance of retrieving authorship data, if this information is redundantly stored in different fields—for example, both in a field containing the full scientific name with authorship included and in another field dedicated solely to authorship. For both higher-rank classifications and taxonomic authorities, fuzzy matching is applied to account for spelling errors and
 650 orthographic variants, as well as alternative abbreviations in authority strings. String similarity is assessed using the Levenshtein distance, a metric that measures the minimum number of single-character insertions, deletions, or substitutions required to transform one string into another. Because authority strings are particularly complex and heterogeneous, they undergo additional pre-processing before the Levenshtein comparison is applied: *isinworms* handles encoding errors and special characters (e.g., “Pham-Hoàng Hô”, “(Strøm, 1768)”) and removes common nomenclatural connectors, interpretive qualifiers, and name
 655 particles (e.g., “ex”, “sec.”, “von”). Authorship matching also relies on further rules to accommodate the structure of authority strings: *isinworms* tolerates reasonable discrepancies in dates, with a default threshold of two years, processes entries listing multiple authors and up to two dates (e.g., “(AH Clark, 1912) non Carpenter, 1888”), and accommodates incomplete records in which only some authors or dates, or none, are provided.

Once taxonomic matching has been resolved, *isinworms* uses the WoRMS environmental flags to exclude taxa identified as



660 non-marine, defined here as taxa explicitly flagged as neither marine nor brackish. When neither flag is positive but at least one of the two flags is missing, marine status is treated as uncertain, and the taxon is annotated and retained. Users can also choose, through a dedicated parameter, to exclude records identified as fossil based on the WoRMS extinction flag. This option is useful, for example, when the objective is to model present-day species distributions. When extinction status is unavailable, the record is retained.

665 At the end of the process, taxonomic harmonization can yield three outcomes: success, failure, or uncertainty. Success occurs when the scientific name corresponds to a marine species recognized by WoRMS. Failure is triggered when the scientific name is missing, non-canonical, unindexed in WoRMS, or assigned to a non-marine taxon. Records assigned to fossil taxa are also classified as failures when fossil exclusion is enabled. Uncertainty arises when the available information is insufficient to decide between multiple WoRMS candidates, when marine status is uncertain, or when an apparent mismatch may stem
 670 from limitations of the matching procedure rather than from the candidates representing distinct taxa. The boundary between failure and uncertainty can be adjusted through three user-defined uncertainty levels. The first level, which is the strictest and default setting, includes unresolved multiple-candidate matches, matches with uncertain marine status, and situations where higher-rank matching fails despite agreement at the species and kingdom levels, provided that no authorship mismatch is detected, most often because authorship information is absent. These situations may reflect taxonomic revisions. The second level
 675 extends uncertainty to cases with consistent higher-rank classification but mismatched authorship, when processing indicates that part of the authority string may not have been accounted for, suggesting parsing limitations. The third and broadest level further includes cases with both inconsistent higher-rank classification and mismatched authorship, while retaining the same requirement that authority-string processing appears incomplete. Increasing the uncertainty level therefore increases the number of cases treated as uncertain rather than failed, allowing users to retain and optionally review records that may represent
 680 false mismatches. Ultimately, users can choose to either annotate or remove records for which matching is uncertain or has failed. An overview of the taxonomic harmonization workflow is provided in Fig. 4. In the configuration file, users can simply specify *isinworms*, which automatically calls *createwormsfilters*.

As an optional final step, *resolvetaxamatch* allows the user to manually address the subset of uncertain cases that may represent false mismatches between the submitted scientific name and the WoRMS candidates. Eligibility for review depends
 685 on the review level chosen by the user and follows the same three-level hierarchy described above: higher-rank mismatches with agreement at the species and kingdom levels; authorship mismatches with otherwise consistent classifications, when suspected incompleteness in authority-string processing suggests that the automated workflow may not have fully exploited the available information; and cases combining higher-rank and authorship mismatches under the same condition. To avoid repeated decisions, review is performed on unique combinations of classification and verbatim fields rather than on every
 690 occurrence record. For each selected case, *marinedb* presents the information available in the occurrence dataset alongside the corresponding entries retrieved from WoRMS. The user can then decide interactively whether one of the WoRMS candidates should be applied or whether no candidate is suitable. In the latter case, the reviewed combination is reclassified as no match, and the annotation or removal rule defined in *isinworms* for failed matches is applied. The resulting decision is propagated to all records sharing the same reviewed combination.



695 At the end of taxonomic harmonization, 74,639,557 OBIS records (66.01%) were matched to marine species recognized
 by WoRMS, compared with 113,232,170 records (54.21%) in GBIF and 198,214 records (47.27%) in JeDI. After excluding
 ambiguous non-exact matches and records whose WoRMS status indicated an unresolved, unstable, not formally established,
 obsolete, or unassessed taxonomic identity, 73,746,589 OBIS records (98.80% of matched records; 65.22% of all records),
 112,576,174 GBIF records (99.42%; 53.90%), and 198,026 JeDI records (99.91%; 47.23%) were considered taxonomically
 700 suitable for downstream modeling. This included some non-accepted names with no accepted species-level counterpart that
 nevertheless supported biologically coherent species-level grouping. The retained records represented 129,604 species in
 OBIS, 139,799 in GBIF, and 559 in JeDI. Among records excluded based on nomenclatural status, WoRMS entries with
 unassessed status involved 812 scientific names in OBIS, 1,043 in GBIF, and only 2 in JeDI. They represented only 1.02% of
 matched records in OBIS (759,400 records), 0.46% in GBIF (515,825 records), and < 0.01% in JeDI (7 records), suggesting
 705 that names indexed by WoRMS but not yet taxonomically assessed were not a major source of record loss in this analysis.

The causes of failed matching differed among databases. In JeDI and OBIS, no-match records were overwhelmingly driven
 by the absence of species-level information: 206,178 JeDI records were affected, representing 49.17% of the full dataset
 (95.04% of no-match records), as were 37,354,831 OBIS records, representing 33.04% of the full dataset (99.16% of no-match
 records). In GBIF, most no-match records fell into three categories. First, the absence of species-level information affected
 710 29,147,987 records, accounting for 13.95% of all records (35.29% of no-match records). Second, 26,540,170 records were
 classified as non-marine, accounting for 12.71% of the full dataset (32.13% of no-match records), either because a non-marine
 WoRMS entry was selected or because all candidate matches were non-marine. Although rare, records assigned to non-marine
 taxa were also found in OBIS (14,794 records). Third, no WoRMS candidate was available for 26,041,048 GBIF records,
 either because no candidate was retrieved or because the corresponding WoRMS entry was quarantined or deleted, accounting
 715 for 12.47% of all records (31.53% of no-match records). Manual inspection of a subset suggested that this category was
 dominated by names referring to non-marine taxa, with marginal cases involving molecular or sample-derived identifiers (e.g.,
 "marine gamma_proteobacterium_HTCC2188"), rare or historical names absent from WoRMS (e.g., *Cynoscion thalassinus*),
 and occasional limitations of automated name cleaning. Fossil records were uncommon, with none in JeDI and only 265,287
 records in OBIS (0.23% of all records; 0.70% of no-match records) and 498,335 in GBIF (0.24%; 0.60%). The low GBIF
 720 proportion should be interpreted in light of the download settings, which excluded fossil specimens, and should therefore not
 be taken as evidence of low representation of marine fossil records in GBIF as a whole. The remaining no-match cases were
 much less frequent across all three databases and involved inconsistencies between occurrence-record and WoRMS-candidate
 classifications or authorship information.

Uncertain cases were comparatively rare in OBIS and JeDI but more frequent in GBIF, affecting 766,280 OBIS records
 725 (0.68%), 4,164 JeDI records (0.99%), and 13,053,643 GBIF records (6.25%). The nature of uncertainty differed among
 databases. In GBIF, most uncertain records reflected incomplete WoRMS environmental flags, with 12,342,630 records
 (94.55% of uncertain records) having uncertain marine status. In OBIS, uncertainty was more often caused by insufficient in-
 formation in the occurrence records to discriminate among candidate matches (489,301 records; 63.85% of uncertain records),
 whereas unresolved marine status accounted for 170,151 records (22.20%). In JeDI, almost all uncertain records reflected



insufficient discriminating information (4,072 records; 97.79% of uncertain records). At the broadest review level, records eligible for manual review with *resolvetaxmaxatch* formed only a small subset of uncertain cases: 92 records in JeDI (2.21% of uncertain records), 106,828 in OBIS (13.94%), and 318,733 in GBIF (2.44%). These records corresponded to only 10 unique combinations of classification and verbatim values in JeDI, 796 in OBIS, and 3,869 in GBIF, representing 0.56%, 0.16%, and 0.28% of all unique combinations, respectively. These restricted review sets illustrate that *resolvetaxmaxatch* is calibrated to target small, potentially resolvable fractions of the data, limiting manual burden while still allowing users to recover records that would otherwise be classified as failed matches.

Manual inspection of no-match and uncertain cases revealed diverse causes, illustrating the complexity of taxonomic harmonization with highly heterogeneous occurrence data. As outlined above, most no-match cases reflected expected data-quality or scope-related issues, including records without species-level information (e.g., "Bonnemaisonia C.Agardh", "Ammodiscus spp."), references to non-marine taxa or apparently erroneous taxon names (e.g., "Annatiara rugosum"), non-taxonomic identifiers (e.g., "Marine virus AFVG_250M600"), or internal inconsistencies. The clearest internal inconsistency found during manual inspection was an OBIS record in which the name and authorship of the marine mammal *Arctocephalus tropicalis* were associated with the name and classification of the microalga *Giraudyopsis stellifera*.

However, other cases reflected the difficulty of reconciling occurrence data against a taxonomic reference system. Some submitted names were not retrieved by WoRMS because they were non-canonical, even though they referred to marine taxa and could often be resolved using SeaLifeBase or the Catalogue of Life. These included names with explicit subgeneric indicators (e.g., "Acetabularia subg. Acicularia farlowii"), gender-agreement variants (e.g., *Telopathes magnus* versus *Telopathes magna*), and alternative generic combinations (e.g., *Pseudocella brachychaites* versus *Thoracostoma (Pseudocella) brachychaites*). A smaller number involved rare or historical names currently omitted from major taxonomic registers, such as *Pyrosoma hybridum*, or hybrid formulas not supported by *marinedb* and therefore not submitted to WoRMS.

Similarly, classification mismatches did not necessarily indicate conflicts between distinct taxa. In a substantial fraction of inspected records where both classifications referred to the same species, the discrepancies instead reflected profound taxonomic changes. For example, some microbial eukaryotes classified in the source data under Protozoa or Plantae, such as *Hermesinum adriaticum*, are currently placed by WoRMS within Chromista, producing apparent higher-rank mismatches despite referring to the same species. Other classification or authorship mismatches arose from desynchronization between the scientific name used for WoRMS retrieval and the verbatim name, classification, or authorship used to resolve ambiguities among multiple matches. One set of fields referred to the accepted species (e.g., *Helicopedinella tricostata* as the query name), while another referred to a synonym or infraspecific form (e.g., *Pseudopedinella tricostata* and "(M.I. Roukhiyajnen) H.A. Thomsen 1988" as the verbatim name and authorship, respectively).

Authorship mismatches further highlighted heterogeneous nomenclatural citation practices, including abbreviations or spelling variants (e.g., "Rozanov" versus "Rosanoff"), bibliographic extensions (e.g., "Hwass in Bruguière, 1792" versus "Hwass, 1792"), alternative authorship attributions (e.g., "(Hedw.) Schimp." versus "W.P. Schimper, 1860" for *Anomodon rostratus*), and two-level authority strings combining species and infraspecific authorship (e.g., "Lactuca viminea (L.) J.Presl & C.Presl subsp. ramosissima (All.) Bonnier"). Additionally, uncertain matches exposed a distinct limitation of authority-string pro-



765 censing: although standardization generally facilitated matching, it also obscured nomenclaturally meaningful differences in some cases. For example, removing interpretive qualifiers made "(Grube, 1863)" indistinguishable from "Auctt. non Grube, 1863" for *Harmothoe longisetis*, while removing parentheses caused "(McIntosh, 1885)" and "McIntosh, 1885" to be treated as equivalent for *Polycirrus kerguelensis*. As a result, some records appeared to lack sufficient information to discriminate among candidate matches, even though manual inspection showed that the original authority strings contained useful information.

770 Together, these examples highlight the limits of a fully automated, general-purpose workflow: conservative matching rules can reject plausible correspondences when occurrence records and taxonomic registries encode different historical or nomenclatural perspectives and practices. *resolvetaxamatch* addresses part of this problem by enabling targeted manual review of cases that are potentially resolvable from the available information, but other cases would require information from external sources, taxon-specific expertise, or future improvements to authority parsing, authority matching, and classification matching.

775 For the final modeling dataset, scientific names and classifications were harmonized in place using *isinworms*. Uncertain JeDI matches were additionally reviewed manually using *resolvetaxamatch* at the broadest review level. Records classified as no match or uncertain were excluded, as were fossil taxa because the target application concerned present-day distributions. Records associated with WoRMS statuses incompatible with reliable species-level grouping were flagged as not fit for use.

Missing scientific name. The absence of a scientific name can be addressed in two ways. First, during taxonomic harmonization with *isinworms*, users may choose to annotate or exclude records that fail to match any WoRMS taxon at the end of the process. This unmatched group naturally encompasses records lacking a standard, resolvable scientific name, including missing, supraspecific, or malformed names, as well as non-canonical molecular, strain, or environmental-sequence identifiers (e.g., "Megaselia sp. BOLD:AAG3266", "Coccoid stramenopile_CCMP1144", or "Marine group_II_euryarchaeote_REDSEA-S30_B12"). This approach was applied to GBIF, OBIS, and JeDI, with the corresponding statistics reported in the previous subsection. Alternatively, if taxonomic harmonization is not performed, users may apply two independent filters. The *isna* utility can be used to identify blank entries in the species-related field, although it does not assess content validity and therefore overlooks cases such as supraspecific names. The *lettersonly* module removes records where the user-specified taxonomic rank contains non-alphabetic characters, thereby capturing many of the identifier formats described above. However, it may also capture legitimate entries, including non-standardized scientific names with authorship dates, and valid taxa from Bacteria (e.g., the class "CG2-30-54-11"), Viruses (e.g., the marine species "Pilayella littoralis virus 1"), or Archaea. In practice, excluding these microbial groups can nonetheless be appropriate in large-scale multi-species distribution modeling, where the focus typically lies on macroscopic taxa.

Insufficient basis of record. *marinedb* provides the *mapbasisofrecord* module, which standardizes the vocabulary used to describe the nature of observations and ensures a consistent representation of the type and strength of evidence supporting each record. The process relies on a mapping file, largely derived from a GBIF resource, that translates heterogeneous source terminology into the standardized categories defined by Darwin Core and adopted by OBIS and GBIF. This classification distinguishes specimen-based records, subdivided into preserved specimens (e.g., herbarium material, tissue stored in ethanol), living specimens (e.g., organisms maintained in aquaria), and fossil specimens (e.g., body fossils, ichnofossils); literature-based records (specimens cited in scholarly publications); sampling-event records (e.g., material samples, environmental DNA); ma-

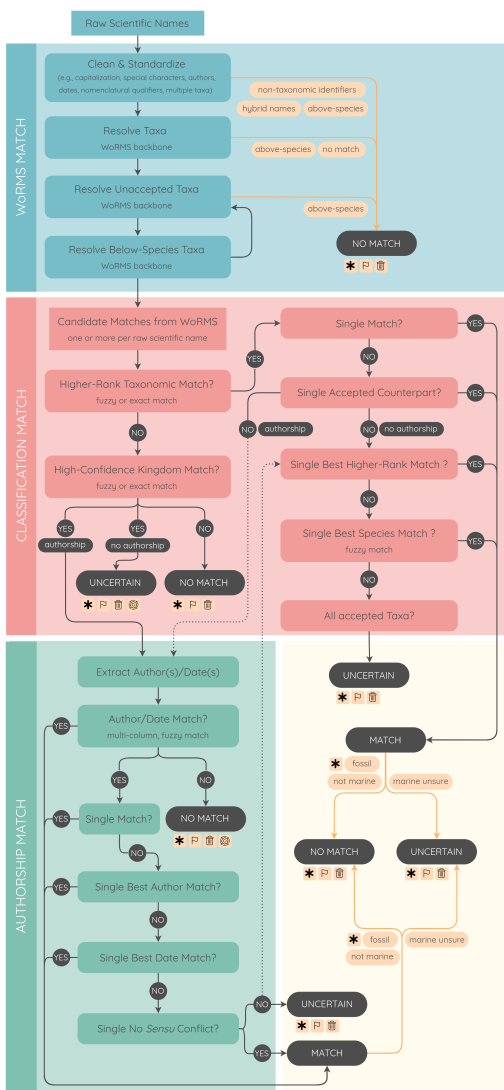


Figure 4. Taxonomic harmonization workflow in *marinedb*: (1) scientific name matching via WoRMS, (2) taxonomic classification matching, and (3) taxonomic authorship matching. A star denotes a user-defined choice, a trash represents removal, a flag indicates annotation, and a gear denotes additional processing, specifically the optional manual resolution of uncertain matches.

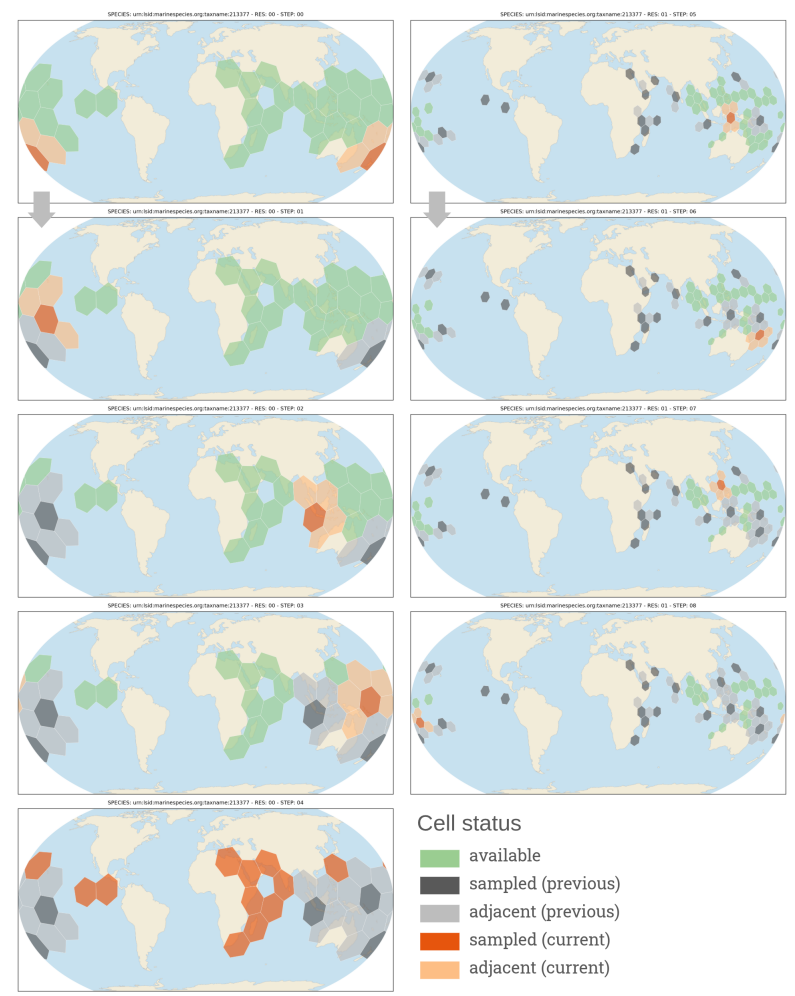


Figure 5. Simplified illustration of the H3-based subsampling procedure used by *taxasubset*, shown for *Echinothrix calamaris*. The occurrence threshold was set to 20 to keep the process visually interpretable. Panels should be read from top to bottom in the left column, then from top to bottom in the right column. Green cells are available for selection; dark and light grey cells indicate cells sampled or excluded in previous steps; dark and light orange cells indicate cells sampled or excluded at the current step. When the number of available cells remaining at a given resolution is lower than the number of locations still to be sampled, all are selected before moving to the next finer resolution, as shown in the bottom-left panel.



chine observations (e.g., photographs, audio recordings); human observations lacking physical or digital evidence (e.g., bird sightings); and records with unrecognized or missing values. The mapping file is released with *marinedb* and is fully customizable. We extended it to accommodate the JeDI dataset, which does not natively comply with Darwin Core vocabulary.

Although optional, this terminological standardization facilitates the identification of records that may be insufficiently reliable for modeling due to limited supporting evidence. Occurrence records linked to specimens are generally regarded as highly reliable: specimens provide tangible proof of the observation and allow for later re-examination. This traceability enables the correction of misidentifications or the revision of classifications in light of taxonomic advances (Chapman, 2005a). By contrast, records lacking physical or digital evidence remain inherently unverifiable and are therefore less reliable. Accordingly, *basisofrecordisin* can be used, with or without prior standardization, to annotate or exclude records based on user-specified criteria, so that the final dataset meets the desired standards of evidence.

Specimen-based records accounted for 7.53% of OBIS records (8,517,483 records), 8.49% of GBIF records (17,731,465 records), and 3.07% of JeDI records (12,888 records). By contrast, records with unspecified observation conditions or no physical or digital evidence represented 61.30% of OBIS records (69,285,441 records), 75.82% of GBIF records (158,369,199 records), and 21.30% of JeDI records (89,330 records). In OBIS and GBIF, this latter category was largely driven by human observations, the most represented observation type overall (39.10% and 74.56% of records respectively). In JeDI, material samples were the dominant category, accounting for 60.39% of the dataset. For the final modeling dataset, observation-type terminology was standardized using *mapbasisofrecord*, and fossil specimens were excluded consistently with the focus on present-day distributions.

Absence records. Because demonstrating absence requires systematic and repeated sampling, such records are rare compared to presence data and often reflect non-detection rather than unequivocal absence (Peterson et al., 2011). Their ambiguous interpretation and scarcity have encouraged the development of presence-only SDM, thereby requiring the exclusion of absence records from training data to avoid misapplication. Yet they remain highly valuable for evaluation, since presence–absence data provide the most robust basis for assessing predictive performance, justifying their annotation for later use. To accommodate these different uses, *isin* (and its complement *notisin*) can be used to annotate or remove species absence observations.

Although OBIS indicates that absence records are excluded from the full-export file, inspection of the non-standardized `occurrenceStatus` field identified 16,839 records (0.01%) with values suggesting absence ("not observed", "Absence", "pseudo-absent", "Ausent"). Because these records were nevertheless interpreted by OBIS as presences, they indicate an internal inconsistency between the non-standardized `occurrenceStatus` field and the OBIS-interpreted presence–absence flag. In GBIF, 17,922,100 records were identified as absences, representing a substantial fraction of the dataset (8.58%). In JeDI, absence records were unusually frequent, with 191,674 records accounting for 45.71% of the dataset. Additional internal inconsistencies were detected in both GBIF and JeDI: GBIF flagged 10,308 records with contradictory occurrence-status and individual-count information, while 126 JeDI absence records were associated with non-zero quantitative measurements.

For the final modeling dataset, absence records were flagged rather than excluded, except when occurrence-status information was internally inconsistent. In such cases, the affected records were removed, with JeDI inconsistencies handled manually outside the configuration-based workflow.



835 **Underrepresented species.** *marinedb* further provides the *taxasubset* module, which enables the annotation or exclusion of underrepresented species based on a user-defined occurrence threshold. The module aims to identify taxa for which the number of observations is likely insufficient to reliably infer geographic distribution, especially at global scales. Limited sample sizes constrain the ability of models to learn generalizable patterns and increase sensitivity to sampling bias and noise, potentially leading to unstable or misleading predictions (Guisan et al., 2017). For these reasons, excluding poorly represented taxa is a
 840 common practice in species distribution modeling. The occurrence threshold typically ranges from 50 to 150 records, although the appropriate value depends on species ecology, model complexity, and data limitations (Moudrý et al., 2024).

Using the subset previously defined as taxonomically suitable for downstream modeling as the basis for this and the following taxonomic analyses, the proportion of species with fewer than 50 occurrences was 78.29% in OBIS (101,469 species, corresponding to 1.22% of records), 76.08% in GBIF (106,353 species, corresponding to 0.90% of records), and 60.11% in
 845 JeDI (336 species, corresponding to 1.51% of records). By combining these sources, 18,501 species that would have fallen below the minimum occurrence threshold in at least one standalone database could be retained. This estimate does not account for potential duplicates across databases and may therefore overstate the gain in record numbers, a caveat that is especially relevant given the known overlap among these sources, including OBIS records incorporated in JeDI and GBIF, as well as JeDI-derived records published through OBIS. For the final modeling dataset, a flag was added to identify species represented
 850 by at least 50 records.

Overrepresented species. Complementarily, *taxasubset* enables the subsampling of overrepresented species based on a user-defined occurrence threshold. This functionality primarily aims to mitigate imbalances in species representation, which can lead to uneven model performance across taxa, while also reducing geographic sampling bias. As model performance has been shown to plateau beyond a certain sample size (e.g., Gábor et al., 2020), subsampling highly represented species can limit
 855 redundancy without substantially affecting predictive performance.

To this end, users can specify a maximum number of occurrences per taxon. For species exceeding this threshold, occurrence locations are discretized using the H3 hierarchical hexagonal grid system (Uber Technologies, 2026) at a user-defined maximum resolution. The default resolution corresponds to cells of approximately 0.74 km^2 , with a maximum within-cell distance of about 1.06 km. This scale is broadly consistent with the two-decimal-place coordinate-precision threshold used previously.
 860 If the number of occupied cells does not exceed the occurrence threshold, all locations are retained. Otherwise, a subset of cells is selected to maximize spatial coverage while avoiding overrepresentation of spatially clustered records. Sampling proceeds from coarser to finer spatial resolutions, assigning higher selection probabilities to larger cells while enforcing spatial dispersion by excluding neighboring cells from subsequent draws. This strategy increases the likelihood of capturing a representative portion of the species' geographic and environmental range. A simplified example of the cell-selection procedure is illustrated
 865 in Fig. 5. Once the target number of locations is reached, occurrences are subsampled to balance the number of observations across selected cells as evenly as possible, subject to local data availability. This approach limits the dominance of densely sampled areas and species, thereby helping to reduce a source of bias that could affect the robustness of downstream models. In the current implementation, *taxasubset* is not parallelized and requires the full input dataset to fit in memory, which limits its scalability for very large datasets. Future releases will address this limitation.



870 As a point of reference, while only 3.88% of species in OBIS (5,023 species) exceeded 1,000 occurrences, they collectively accounted for 67,749,405 records, representing 91.87% of the dataset. This trend was mirrored in GBIF, where 4.44% of species (6,207 species) accounted for 93.78% of records (105,575,736 records), and in JeDI, where 10.73% of species (60 species) accounted for 76.13% of records (150,755 records). Applying a maximum of 1,000 records per species would therefore substantially reduce dataset size, leaving 11,020,923 records in OBIS, a loss of 85.06%; 13,222,800 records in GBIF, a loss of 88.25%; and 107,551 records in JeDI, a loss of 45.69%.

Taxonomic selection. The *isin* and *notisin* modules facilitate the selection, annotation, or exclusion of taxonomic groups, allowing large, multi-species datasets to be restricted to a scope consistent with the study objectives. Even within multi-species modeling frameworks, limiting the range of taxa may still be necessary. The inclusion of ecologically disparate groups can hinder model training and ultimately compromise predictive performance, as higher taxonomic diversity typically results in more heterogeneous responses to environmental drivers.

Following this rationale, Archaea, Bacteria, and Viruses were excluded from the final modeling dataset to ensure biological and methodological consistency within a global, multi-species framework focused on contemporary distributions. The divergences in taxonomic definition, ecological determinants (nature and scale), and fundamental biological traits preclude a unified processing framework for both eukariotic and non-eukariotic taxa. In practice, these groups accounted for 103,807 records in OBIS (0.14%), 91,597 in GBIF (0.08%), and none in JeDI, consistent with JeDI's focus on gelatinous zooplankton.

While cumulative misalignments may justify the exclusion of some groups, others require a more nuanced approach based on flagging rather than outright removal. Unlike strictly marine taxa, seabirds are obligate terrestrial breeders and act as central-place foragers during the breeding season. Their spatial distributions can therefore be strongly influenced by land-based constraints (e.g., distance to nesting colonies) and atmospheric drivers (e.g., wind patterns, surface pressure) that are not captured by standard oceanographic grids. Furthermore, integrating seabirds into a global multi-species framework increases ecological heterogeneity and may lead models to rely on indirect or spurious associations. For example, subsurface variables such as dissolved nutrients or oxygen at depth may be important predictors for many aquatic organisms but have limited direct ecological relevance for seabirds, although they may covary with trophic or oceanographic processes that affect their distribution. Flagging these species enables users to make informed methodological choices: they may either treat seabirds within a dedicated framework that accounts for the land–sea–air interface or retain them in a multi-species model using sufficiently flexible architectures, such as deep neural networks, that can accommodate heterogeneous species responses. Overall, seabird records represented 15.18% of OBIS entries (11,196,245 records) and were even more prominent in GBIF, where they accounted for 43.13% of the dataset (48,555,694 records). No seabird records were found in JeDI, consistent with its focus on gelatinous zooplankton. Seabird records were flagged in the final modeling dataset, allowing users to retain or exclude them depending on the ecological scope of their analyses.

2.7 General-purpose modules

The *contains*, *doesnotcontain*, *isin*, *notisin*, *isboundedby*, and *isna* modules provide general-purpose utilities that can be applied beyond the use cases considered in the preceding sections.



For example, using *contains* to process OBIS quality flags, we observed that 4,377,381 observations (3.87%) remained
 905 flagged with the `ON_LAND` issue despite the marine-based spatial filtering. This apparent contradiction arises from differences
 in the coastlines used to distinguish land from sea: *marinedb* relies on GSHHS coastlines, whereas OBIS uses land polygons
 derived from OpenStreetMap⁷. This observation illustrates the limitations discussed in Sect. 2.4 regarding GSHHS. While
 GSHHS provides globally stable, consistent, and topologically robust coastlines suitable for reproducible large-scale automated
 processing, it relies on historical shoreline datasets and may lack fine-scale spatial detail or exhibit positional inaccuracies in
 910 some regions. Conversely, OpenStreetMap-derived polygons can offer finer and more up-to-date local coastline delineation, but
 their completeness, positional accuracy, and consistency may vary spatially due to uneven contributor coverage, heterogeneous
 data sources, and variable mapping practices associated with the collaborative nature of the project (e.g., Philipp et al., 2022,
 Chen et al., 2023). Overall, these differences illustrate the methodological trade-offs between global consistency and fine-
 scale coastline representation. Future versions of *marinedb* could integrate OBIS's *pyxylookup* package (Bosch, 2026) and
 915 allow users to select the shoreline reference best suited to their application. Another option would be to build on the global
 high-resolution products recently released by Mudiyanse et al. (2025), which provide not only global coastlines but also
 water-probability maps and intertidal-zone delineations at 2 m resolution. For the final modeling dataset, OBIS records carrying
 the `ON_LAND` quality flag were excluded.

More generally, *contains* and *doesnotcontain* can be used to annotate or exclude records based on any source-level quality
 920 flag. The `GEODETIC_DATUM_INVALID` flag in GBIF provides another informative example. Manual inspection of records
 carrying this flag revealed that some original `geodeticDatum` values referred to valid non-WGS 84 reference systems,
 including SIRGAS2000 and NAD83. The GBIF WGS 84 fallback assumption, applied when the original datum is missing
 or cannot be interpreted, may therefore assign erroneous locations to records whose coordinates were originally expressed
 in another reference system. GBIF explicitly documents this risk. Accordingly, records combining an invalid geodetic datum
 925 with a country- or continent-coordinate mismatch were systematically excluded from the final modeling dataset. Records car-
 rying `GEODETIC_DATUM_INVALID` without an associated geographic-mismatch flag were examined separately outside the
 configuration-based workflow. Among these, records whose original `geodeticDatum` indicated an unknown or unrecorded
 datum (e.g., "Unknown" or "datum not recorded"), an invalid code or string (e.g., "WGS157" or "nGis"), or a valid non-WGS 84
 reference system were also excluded.

930 Similarly, *isin* can be used to identify and exclude records not licensed for specific uses, provided that licensing metadata
 are available in the source dataset, as is the case for OBIS and GBIF. In OBIS, licensing information was missing for 75.77%
 of observations (85,683,833 records). Among records with available metadata, 18.51% (5,069,568 records) were associated
 with licenses imposing restrictions on reuse, including 617 observations identified as copyrighted. By contrast, GBIF provided
 complete and standardized licensing metadata, with 79.23% of records (165,503,196 records) distributed under licenses per-
 935 mitting unrestricted reuse or reuse with attribution. The remaining records were available for non-commercial use only. These
 findings highlight the importance of license-aware dataset filtering in ensuring that data usage complies with licensing con-
 straints. For the final modeling dataset, OBIS licensing metadata were standardized using a dedicated script distributed with

⁷<https://www.openstreetmap.org/>



marinedb but not currently integrated into the configuration-based workflow. Because the objective was to publicly share data modified by the processing workflow, records prohibiting redistribution of modified content or requiring explicit authorization for reuse were excluded, and the original license field was replaced with the standardized version.

2.8 Construction of the integrated modeling dataset

The same GBIF, OBIS, and JeDI exports used for the fitness-for-use assessment were used to produce a dataset tailored to large-scale modeling of present-day marine species distributions, including multi-species and deep-learning approaches. Unlike the non-destructive assessment, dataset construction combined selected operations cumulatively, using record exclusion or in-place transformation where required. Processing was first performed separately for each source, allowing database-specific structures and content to be handled before integration. Following the source-specific curation described in the preceding sections, additional modeling-oriented operations were applied using standalone scripts distributed with *marinedb* but not integrated into the standard configuration-file-based workflow. These operations encoded standardized `basisOfRecord` values numerically, generated identifier-based taxonomic classifications, and excluded records previously flagged as taxonomically not fit for use.

Additionally, cross-source overlap was partially addressed using institutional, collection, dataset, and occurrence identifiers. Duplicate records may arise from repeated publication of the same record by different data providers whose sources overlap, physical duplication of specimens across institutions, creation of derived records from specimen subsampling, such as DNA sequence records, or residual catalogue entries following specimen renumbering or transfer to another museum (Mesibov, 2013; Zizka et al., 2020). Their presence can inflate record numbers, introduce biases, and increase computational burden, particularly in large-scale multi-species modeling applications. However, reliable duplicate detection is difficult because identical species, coordinates, and dates may correspond either to true duplicate records or to distinct observations collected at the same place and time, while incomplete temporal information, low coordinate precision, inconsistent identifiers, and conflicting taxonomic or locality information can prevent true duplicates from being recognized (Mesibov, 2013; Moudrý and Devillers, 2020; Zizka et al., 2020; de Melo et al., 2024). For this reason, *marinedb* does not currently implement general automated deduplication. The reduction of cross-source overlap during construction of the integrated modeling dataset was therefore deliberately conservative and was not intended as comprehensive deduplication. Instead, the curated source-level datasets released alongside the package retain provenance fields, including institutional and collection identifiers, together with dataset- and record-level identifiers where available, so that users can perform more extensive deduplication subsequently, for example using dedicated external tools such as the duplicate-detection function implemented in *BeeBDC* (Dorey et al., 2023).

The three sources were then harmonized to a common schema. Coordinates were rounded to five decimal places, depth-related fields were omitted from the integrated dataset as they had not undergone dedicated validation and standardization, and only fields retained for modeling or for linking records back to their source and source-level products were kept. After integration, species-representation flags were added to identify taxa represented by at least 10, 50, and 100 records.



970 3 Results & Discussion

A summary of the fitness-for-use assessment of GBIF, OBIS, and JeDI presented in the preceding sections is provided in Appendix B, Table B1. The dataset-production process generated a hierarchy of reusable data products corresponding to different levels of transformation and distinct trade-offs between information richness and ease of use. These products comprise marine-restricted datasets retaining the original source structure after land-record exclusion, curated source datasets incorporating source-specific cleaning while preserving detailed metadata, and an integrated modeling dataset adopting a simplified, harmonized schema. The products were published separately so that users can either work directly with the integrated product or reuse less transformed intermediate datasets. Table 2 summarizes the published products, their transformation levels and intended reuse, and provides their associated DOIs.

Table 2. Published data products generated during dataset production, including their source, transformation level, intended reuse, number of records, and Concept DOI. Citations refer to the specific dataset versions used in this study.

Data product	Transformation level	Intended reuse	Number of records	Concept DOI
Marine-restricted source dataset GBIF (Berhocoïrigoin, 2026c)	Land-record exclusion; original fields retained	Starting point for alternative marine occurrence-data curation	208,875,757	10.5281/zenodo.22746724
Marine-restricted source dataset OBIS (Berhocoïrigoin, 2026f)	Land-record exclusion; original fields retained	Starting point for alternative marine occurrence-data curation	113,076,219	10.5281/zenodo.22729618
Curated source dataset GBIF (Berhocoïrigoin, 2026b)	Source-specific curation; enriched and selected fields retained	Reuse requiring detailed source-level information or alternative downstream filtering	101,462,899	10.5281/zenodo.22746817
Curated source dataset OBIS (Berhocoïrigoin, 2026e)	Source-specific curation; enriched and selected fields retained	Reuse requiring detailed source-level information or alternative downstream filtering	64,097,370	10.5281/zenodo.22746794
Curated source dataset JeDI (Berhocoïrigoin, 2026d)	Source-specific curation; enriched and selected fields retained	Reuse requiring detailed source-level information or alternative downstream filtering	189,713	10.5281/zenodo.22746887
Integrated modeling dataset GBIF + OBIS + JeDI (Berhocoïrigoin, 2026a)	Cross-source filtering, standardization, and integration; selected fields retained	Direct use in large-scale modeling of present-day marine species distributions	158,910,536	10.5281/zenodo.22746916

The final modeling dataset comprises 158,910,536 occurrence records representing 139,844 species. GBIF contributes the largest share of records (60.07%), followed by OBIS (39.84%), whereas JeDI contributes a small but taxonomically focused fraction (0.09%). The represented species span five kingdoms, with Animalia accounting for 89.91% of species (125,739), followed by Chromista (7,631; 5.46%), Plantae (5,810; 4.15%), Fungi (505; 0.36%), and Protozoa (159; 0.11%) (Table 3).

Species representation is highly imbalanced. Nearly half of the species (68,251; 48.81%) have ten records or fewer, while only 0.87% of species (1,218) account for 75% of all records. Consequently, the number of species available for modeling declines rapidly with stricter minimum-occurrence requirements: for example, increasing the threshold from 10 to 50 records reduces this number from 74,294 (53.1%) to 39,065 (27.9%) (Fig. 6). Among the most frequent species are several seabirds



Table 3. Taxonomic diversity across kingdoms and major taxonomic ranks.

Kingdom	Phylum	Class	Order	Family	Genus	Species
Animalia	31	88	455	3,845	22,947	125,739
Chromista	13	73	231	649	2,118	7,631
Plantae	7	26	103	251	1,166	5,810
Fungi	5	25	62	139	286	505
Protozoa	8	18	34	51	92	159
Total	64	230	885	4,935	26,609	139,844

and commercially exploited fishes and crustaceans (Table 4). Record volume thus appears to reflect observational biases toward large, conspicuous, or economically important taxa rather than biological abundance alone.

Table 4. Top 15 species with the highest numbers of occurrence records.

Species	Vernacular name	Organism type	Number of records	Percentage of total records
<i>Larus argentatus</i>	European herring gull	Bird	2,578,907	1.62%
<i>Mirounga leonina</i>	Southern elephant seal	Mammal	2,084,443	1.31%
<i>Larus fuscus</i>	Lesser black-backed gull	Bird	2,000,957	1.26%
<i>Gadus morhua</i>	Atlantic cod	Fish	1,938,645	1.22%
<i>Pandalus jordani</i>	Oregon pink shrimp	Crustacean	1,718,275	1.08%
<i>Fulmarus glacialis</i>	Northern fulmar	Bird	1,587,470	1.00%
<i>Morus bassanus</i>	Northern gannet	Bird	1,552,171	0.98%
<i>Melanogrammus aeglefinus</i>	Haddock	Fish	1,500,387	0.94%
<i>Uria aalge</i>	Common murre	Bird	1,394,675	0.88%
<i>Haematopus ostralegus</i>	Eurasian oystercatcher	Bird	1,336,831	0.84%
<i>Clupea harengus</i>	Atlantic herring	Fish	1,209,617	0.76%
<i>Larus marinus</i>	Great black-backed gull	Bird	1,175,791	0.74%
<i>Eudyptula minor</i>	Little penguin	Bird	1,026,838	0.65%
<i>Rissa tridactyla</i>	Black-legged kittiwake	Bird	901,517	0.57%
<i>Somateria mollissima</i>	Common eider	Bird	838,059	0.53%

Spatial coverage is global but highly uneven (Fig. 7). At H3 resolution 3 (mean cell area $\sim 12,393 \text{ km}^2$), records are concentrated along coastlines and across the Northern Hemisphere, whereas large areas of the open ocean contain relatively few or no observations. Antarctic and subantarctic waters also include several areas of high record density, although these are often associated with comparatively lower observed species richness. Yet a high number of records and species does not necessarily imply balanced species-level representation: in a substantial fraction of cells meeting both criteria, at least 80% of the species

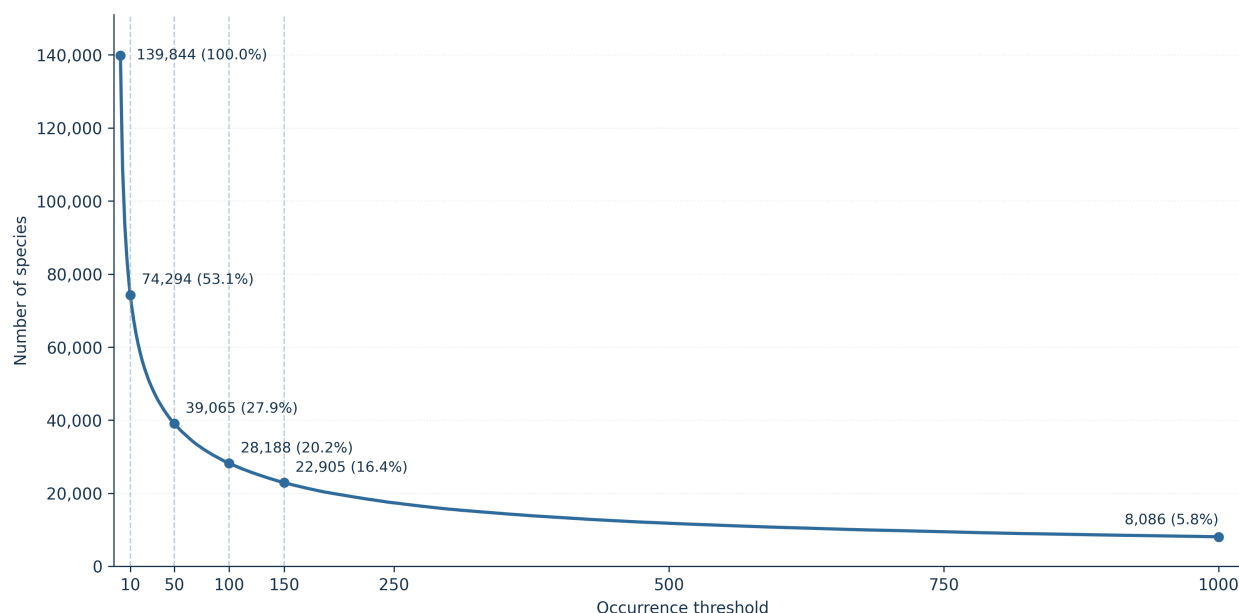


Figure 6. Number of species represented by at least a given number of occurrence records. Labels indicate the number and percentage of species meeting selected occurrence thresholds.

occurring in the cell total no more than ten records. At the species level, observed geographic coverage was similarly restricted: only 41.48% of species (58,002 species) occur in at least 10 cells, 30.36% (42,459) in at least 20 cells, 18.81% (26,307) in at least 50 cells, and 12.47% (17,444) in at least 100 cells. Among species with at least 10 records and occupying more than one H3 cell, occurrences are nevertheless distributed relatively evenly across occupied cells, with a median Pielou evenness of 0.95 and an interquartile range of 0.91–0.98.

Records span 1700–2025, but temporal coverage is strongly concentrated toward recent decades: 90% of dated records occur after 1983 and 75% after 1997 (Fig. 8a). The monthly distribution is also uneven, although less strongly than the annual distribution (Fig. 8b). Among records with month information, 54.61% occur between May and September, while December and January are comparatively underrepresented (Appendix Fig. C2). When year and month are considered jointly, records are most strongly concentrated between 2004 and 2023 and from May to October (Appendix Fig. C1). Month information is unavailable for 4.23% of records overall, with missing values concentrated among older observations but still present in more recent records (Appendix Fig. C2).

The integrated dataset contains 13,965,502 absence records (8.79%), distributed across 7,450 species. Among these species, 75.96% have at least 15 absence records, suggesting that absence information may provide sufficient replication for downstream evaluation in many of the taxa for which absences are available (Moudrý et al., 2024). Overall, human observations constitute the dominant basis of record (106,543,619 records; 67.05%), followed by generic occurrences with unknown observation conditions (18,508,296; 11.65%), machine observations (16,874,694; 10.62%), preserved specimens (10,468,673;

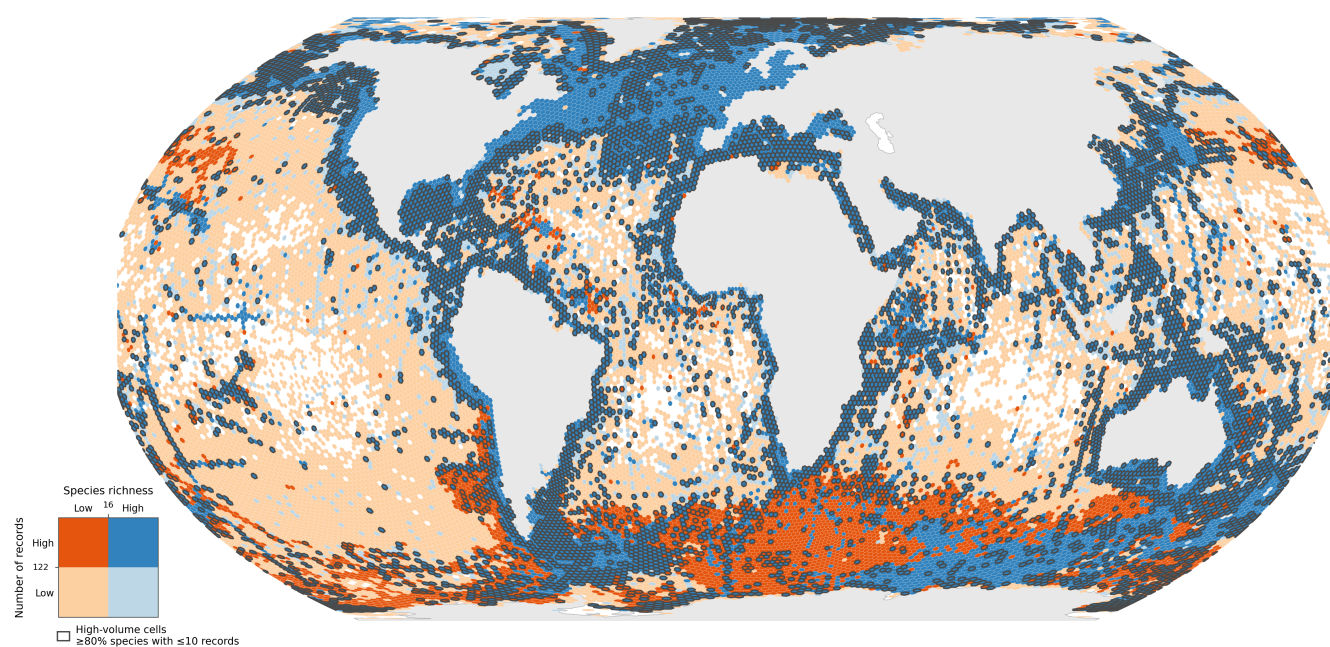


Figure 7. Spatial distribution of species richness and occurrence-record volume. Occurrence records were aggregated into H3 resolution-3 cells (average area $\sim 12,393 \text{ km}^2$). Cell colors jointly represent species richness and number of occurrence records, with the median of each metric defining low and high values. Orange-to-blue hues indicate increasing species richness, while light-to-dark shades indicate increasing record volume. Dark blue therefore represents cells with both high species richness and high record volume, whereas light orange represents cells with both low species richness and low record volume. Uncolored cells indicate areas with no occurrence records. Outlined cells indicate high-volume cells in which $\geq 80\%$ of species are represented by ≤ 10 records.

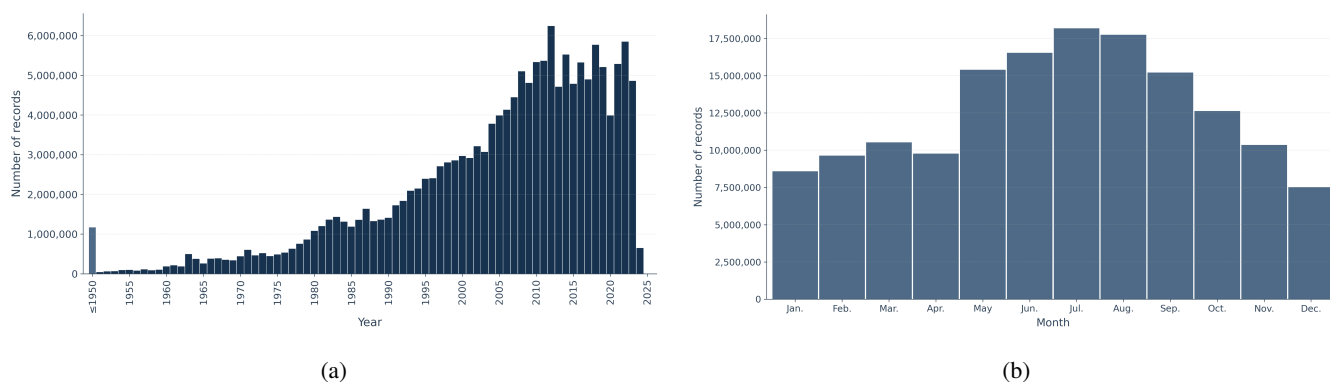


Figure 8. Temporal distribution of occurrence records. (a) Annual distribution of records, with records from 1950 and earlier grouped into a single category (≤ 1950). (b) Monthly distribution of records.



6.59%), material samples (6,405,309; 4.03%), living specimens (63,917; 0.04%), and material citations (46,028; 0.03%). The dominance of human observations is consistent with the strong representation of coastal regions and several highly observed taxonomic groups described above.

4 Conclusions

1015 The digital era and the emergence of data-aggregation initiatives have greatly expanded access to occurrence records, but accessibility alone does not ensure immediate usability. Occurrence records must be curated and reconciled across spatial, temporal, and taxonomic dimensions before they can be reliably linked to environmental predictors and used to model present-day marine species distributions. In this article, we introduced *marinedb*, an open-source, modular, and scalable Python package for curating marine occurrence data. By automating and documenting curation operations that can be freely combined into
 1020 user-defined workflows, *marinedb* provides a reproducible basis for adapting marine occurrence-data processing to different objectives, rather than prescribing a single universally appropriate strategy.

The application of *marinedb* to GBIF, OBIS, and JeDI illustrated the importance of developing a framework that is both scalable and flexible. First, support for parallel execution enabled the processing of these three databases, which together comprise hundreds of millions of occurrence records located at sea and more than one hundred thousand species. Second, the
 1025 quality assessment showed that the nature and prevalence of data-quality issues varied among source datasets and that their relevance depended on the intended use. GBIF, as a broad cross-realm aggregator, was overwhelmingly dominated by terrestrial records, while even marine-focused resources such as OBIS and JeDI contained records classified as land. Missing depth information was widespread across all three databases, and temporal information required substantial standardization in OBIS and JeDI. Taxonomic harmonization further revealed that large fractions of records either lacked species-level information,
 1030 were associated with names not recognized in WoRMS, or referred to non-marine taxa, and were therefore unsuitable for marine species distribution modeling. Among the other prominent issues were records with weak or unspecified evidential support, absence records in GBIF and JeDI, substantial representation of seabirds in GBIF and OBIS, and strong taxonomic imbalance, with most species represented by few occurrences while a small minority accounted for most records. Together, these results confirm that aggregator-level quality control, although essential, cannot replace dataset- and use-case-specific
 1035 curation.

Building on this quality assessment, we applied a tailored workflow to produce a ready-to-use global marine occurrence dataset intended for large-scale modeling of present-day species distributions across geographic and environmental gradients. This dataset is designed to lower the technical barrier associated with assembling and cleaning marine occurrence data, particularly for applications involving many taxa or deep-learning approaches. By providing both curated datasets and an extensible
 1040 Python framework, this work contributes to making marine biodiversity data more directly usable for large-scale ecological modeling and offers a foundation for future developments at the interface between biodiversity informatics, marine ecology, and artificial intelligence.



Code and data availability. All data products generated in this study are publicly available through Zenodo. These comprise six complementary datasets corresponding to different levels of transformation and integration: DOI 10.5281/zenodo.22746724 (Berhocoïrigoin, 2026c),
1045 10.5281/zenodo.22729618 (Berhocoïrigoin, 2026f), 10.5281/zenodo.22746817 (Berhocoïrigoin, 2026b), 10.5281/zenodo.22746793 (Berhocoïrigoin, 2026e), 10.5281/zenodo.22746887 (Berhocoïrigoin, 2026d), and 10.5281/zenodo.22746916 (Berhocoïrigoin, 2026a). The *marinedb* source code is openly available on GitHub at <https://github.com/marinesdm/marinedb/tree/master>. The repository contains the complete package implementation, including standalone scripts distributed with *marinedb* but not integrated into the standard configuration-file-based workflow, as well as the code used to generate the visualizations presented in the Results & Discussion section. Full package documentation,
1050 including installation instructions, API documentation, and usage examples, is available at <https://marinesdm.github.io/marinedb/>.

Appendix A: Technical specifications

A1 Requirements

The package was developed using Python 3.12 and Bash 5.2, and depends on a range of standard and external Python libraries. Built-in modules include: *sys*, *os*, *glob*, *pathlib*, *importlib*, *contextlib*, *inspect*, *operator*, *itertools*, *functools*, *collections*, *shutil*,
1055 *socket*, *subprocess*, *argparse*, *copy*, *math*, *http*, *urllib*, *concurrent*, *multiprocessing*, *gzip*, *re*, *random*, *datetime*, *time*, *difflib*, *bisect*, *warnings*, and *json*. Third-party libraries include: *pyyaml*, *pyproj*, *joblib*, *psutil*, *Levenshtein*, *Unidecode*, *tqdm*, *pyarrow*, *pandas*, *geopandas*, *geodatasets*, *h3pandas*, *h3*, *numpy*, *dask*, *suds*, *pillow*, *basemap*, *basemap-data-hires*, *cartopy*, *antimeridian*, *shapely*, *descartes*, *fiona*, *opencv-python*, *matplotlib*, and *seaborn*. All these requirements with their versions are specified in the package `pyproject.toml` file and are installed automatically when *marinedb* is installed.

1060 A2 Distributed processing

When terrestrial-record filtering needs to be distributed across machines, the integrated script *marineloc.py*, which normally chains together all stages of land–sea filtering, cannot be used directly. Users must instead perform the initial steps manually: splitting the input file with *split_pandas_parquet.py*, recomputing the spatial mask if needed with *createmask.py*, and assigning land or sea status to each split file with *island.py*. This last script, which labels records according to their location, can be
1065 distributed across machines using the *parallel_island.sh* shell script. The distributed step produces a folder containing the processed split files, each augmented with a column indicating whether the coordinates are located on land or at sea. Once classification is complete, users can return to the configuration-based workflow by specifying, in the *marineloc.py* parameters, the paths to both the original and the processed split-file folders. *marineloc.py* then executes the remaining land–sea filtering steps and passes the resulting dataset automatically to the next stage of the pipeline.

1070 For distributed execution, all machines must have access to the input split files and to the folder where processed files are written. Input files may either be stored in a shared directory or copied locally to each machine, but the output folder must be shared so that each machine can determine which files have already been completed. Each machine shuffles the list of files before processing to reduce the probability that multiple machines select the same file simultaneously. This risk is not entirely eliminated, however, because a file is marked as completed only once its processed output has been written. This design favours



robustness to failure: if a machine stops before completing a file, that file remains absent from the processed folder and can therefore be picked up by another machine later. As a result, occasional redundant processing is possible, especially near the end of the file pool, but the risk of leaving files unprocessed is reduced. Alternatively, users can provide a file listing the split files to be processed by a given machine, allowing manual assignment of files across machines. This gives more control but reduces fault tolerance if one machine fails before completing its assigned files.

Appendix B: Data-quality metrics for GBIF, OBIS, and JeDI

Table B1. Summary of data-quality metrics reported for GBIF, OBIS, and JeDI. Metrics are grouped by dimension: spatial in red, temporal in blue, and taxonomic in green. In the taxonomic section, progressively darker shades differentiate successful matches, uncertain cases, mismatch-related cases, failed matches, and complementary taxonomic or observation statistics.

Metric	GBIF	OBIS	JeDI
Coordinates on land	2,602,503,097 (92.6%)	19,393,076 (14.6%)	10,146 (2.4%)
Missing coordinates	0*	0*	0*
Out-of-range coordinates	0*	0*	0*
Zero-valued coordinates	961,751 (0.46%)	29,843 (0.03%)	691 (0.16%)
Identical coordinates	971,610 (0.46%)	1,483 (< 0.01%)	6 (< 0.01%)
Missing positional metadata	PRECISION	PRECISION	no metadata
	200,727,142 (96.10%)	103,012,687 (91.10%)	
	UNCERTAINTY	UNCERTAINTY	
	126,144,548 (60.39%)	64,026,344 (56.62%)	
Coordinate uncertainty > 1 km	27,992,897 (33.83% ^{†a})	4,254,878 (8.67% ^{†a})	—
Decimal-place precision < 2	5,042,045 (2.41%)	923,155 (0.82%)	24,078 (5.74%)
Missing depth	154,914,642 (74.17%)	42,083,645 (37.22%)	171,976 (41.01%)
Non-numeric depth	0	0	0
Depth exceeding bathymetry	—	1,050,639 (1.48% ^{†b})	—
Depth < 0 m or > 11,000 m	0*	—	0
Minimum depth exceeding maximum depth	0	93,455 (0.17% ^{†b})	55,670 (22.51% ^{†b})
Missing dates (before processing)	0*	4,130,996 (3.65%)	23,270 (5.55%)

Continued on next page



Table B1. Summary of data-quality metrics reported for GBIF, OBIS, and JeDI. (*continued*)

Metric	GBIF	OBIS	JeDI
Invalid dates (value or format)	0	540,180 (0.48%)	395,542 (94.33%)
Date intervals	5,999,994 (2.87% ^{†c})	2,815,869 (2.55% ^{†c})	0
Date intervals > 1 year	0*	153,797 (5.46% ^{†d})	0
Date intervals > 1 month	1,942,106 (32.37% ^{†d})	330,915 (11.75% ^{†d})	0
Missing date components (before processing)	15,192,517 (7.27%)	63,989,550 (56.59%)	23,273 (5.55%)
Invalid or inconsistent date components	0	69,188 (0.12%)	0
Date vs. component mismatch	0	33,302 (0.03% ^{†c})	0
Missing date components (after processing)	15,192,517 (7.27%)	8,352,244 (7.39%)	23,273 (5.55%)
Missing dates (after processing)	0*	2,810,313 (2.49%)	0
Dates with day precision (after processing)	193,683,239 (92.73% ^{†e})	104,723,974 (92.61% ^{†e})	396,043 (94.45% ^{†e})
Dates with month precision (after processing)	5,817,463 (2.79% ^{†e})	1,334,380 (1.18% ^{†e})	7,197 (1.72% ^{†e})
Dates with year precision (after processing)	9,375,054 (4.49% ^{†e})	4,207,551 (3.72% ^{†e})	16,076 (3.83% ^{†e})
Unlikely dates	0	4,440 (< 0.01% ^{†e})	0
Year < 1950	2,399,735 (1.15% ^{†e})	963,676 (0.87% ^{†e})	4,497 (1.07% ^{†e})
WoRMS-recognized marine species	113,232,170 (54.21%)	74,639,557 (66.01%)	198,214 (47.27%)
Taxonomically fit-for-use	112,576,174 (53.90%) 139,799 species	73,746,589 (65.22%) 129,604 species	198,026 (47.23%) 559 species
Insufficient information to resolve match	392,280 (0.19%)	489,301 (0.43%)	4,072 (0.97%)
Marine status uncertain	12,342,630 (5.91%)	170,151 (0.15%)	0
Classification or authorship mismatch*	725,739 (0.35%)	137,345 (0.12%)	92 (0.02%)
No species-level information	29,147,987 (13.95%)	37,354,831 (33.04%)	206,178 (49.17%)
Non-marine taxa	26,540,170 (12.71%)	14,794 (0.01%)	0
No WoRMS match	26,041,048 (12.47%)	4,964 (< 0.01%)	10,760 (2.57%)

Continued on next page



Table B1. Summary of data-quality metrics reported for GBIF, OBIS, and JeDI. (*continued*)

Metric	GBIF	OBIS	JeDI
Fossil taxa	498,335 (0.24%)	265,287 (0.23%)	0
Specimen-based records	17,731,465 (8.49%)	8,517,483 (7.53%)	12,888 (3.07%)
Limited-evidence records	158,369,199 (75.82%)	69,285,441 (61.30%)	89,330 (21.30%)
Absence records	17,922,100 (8.58%)	16,839 (0.01%)*	191,674 (45.71%)
Underrepresented species < 50 occurrences	76.08% ^{†f} of species	78.29% ^{†f} of species	60.11% ^{†f} of species
Overrepresented species > 1000 occurrences	4.44% ^{†f} of species → 93.78% ^{†f} of records	3.88% ^{†f} of species → 91.87% ^{†f} of records	10.73% ^{†f} of species → 76.13% ^{†f} of records
Archaea, Bacteria, and Viruses	0.08% ^{†f} of species	0.14% ^{†f} of species	0
Aves	43.13% ^{†f} of species	15.18% ^{†f} of species	0
Restriction on reuse?	non-commercial	non-commercial, no-derivatives, copyrighted	no metadata

* Interpretation requires caution; see the main text.

^{†a} Among records with positional metadata.

^{†b} Among records with depth.

^{†c} Among records with a date after parsing.

^{†d} Among records with a date interval.

^{†e} Among records with a date after curation.

^{†f} Among records and species taxonomically fit-for-use.

Appendix C: Supplementary Results

Author contributions. **Elena Berhocoïrigoin:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing (original draft preparation); **Philippe Le Noac'h:** Conceptualization, Data Curation, Writing (review and editing); **Martin Laviale:** Conceptualization, Writing (review and editing), Funding acquisition; **Sakina-Dorothee Ayata:** Conceptualization, Writing (review and editing), Funding acquisition; **Jérémy Fix:** Conceptualization, Software, Writing (review and editing), Supervision, Project administration, Funding acquisition; **Cédric Pradalier:** Conceptualization, Software, Writing (review and editing), Supervision, Project administration, Funding acquisition.

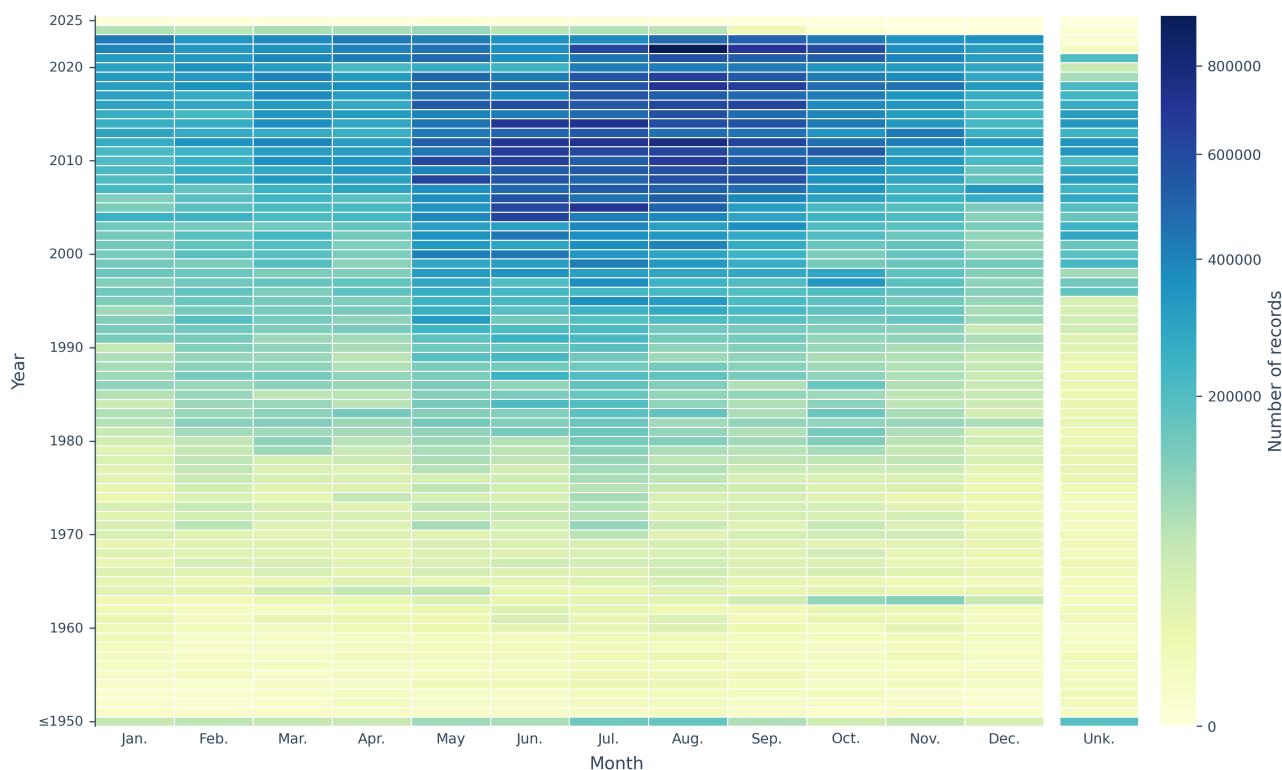


Figure C1. Yearly and monthly distribution of occurrence records. Cell colors indicate the number of records for each year–month combination. Records from 1950 and earlier are grouped into a single category (≤ 1950), and records with unknown month are shown separately (Unk.). High record counts are concentrated in the heatmap region corresponding to 2004–2023 and May–October.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. I thank the GBIF Helpdesk, and especially Markus Döring and Esteban Marentes Herrera, for their prompt and highly informative responses to multiple technical questions, including guidance on the GBIF processing workflow and pointers to relevant sections of the source code. I also thank the OBIS Helpdesk for their assistance with data access and for recommending the use of Parquet exports. I am grateful to the WoRMS Data Management Team, and especially to Bart Vanhoorne, for their support during the development of the taxonomic harmonization workflow, including guidance on WoRMS API outputs, advice on taxonomic interpretation, appropriate query rates, and the resolution of API-access issues. ChatGPT (GPT-5.5 Thinking, OpenAI, <https://chatgpt.com>) was used to assist with manuscript structuring, English-language editing, and iterative revision of selected passages. All scientific analyses, interpretations, and final editorial decisions were made by the authors. This work was supported by the ANR Challenge IA-Biodiv SmartBiodiv project (ANR-21-AAFI-0002, 2022–2025).

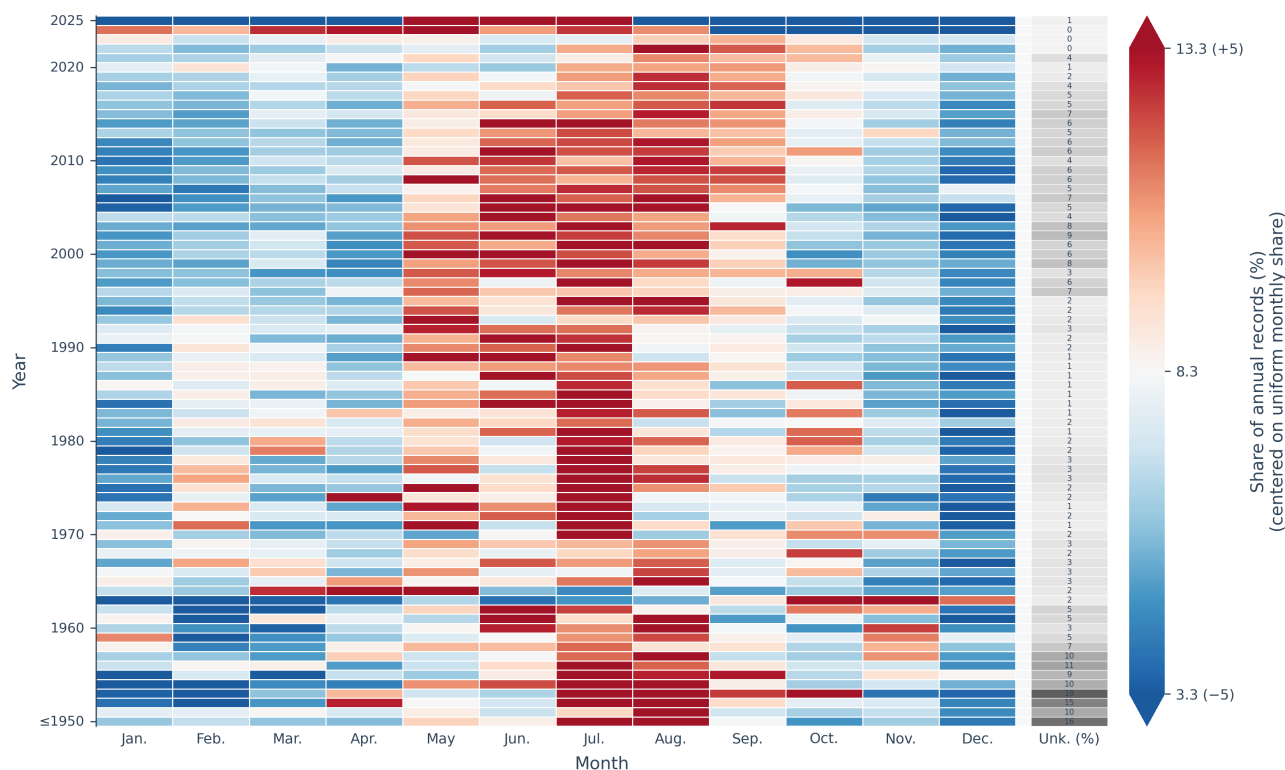


Figure C2. Within-year distribution of occurrence records across months. For each year, colors represent the percentage of annual records occurring in each month, centered on the uniform monthly share (8.3%). Red indicates months overrepresented relative to a uniform distribution and blue indicates underrepresented months. The color scale is limited to ± 5 percentage points around the uniform share and values exceeding these limits are shown at the corresponding color-scale extremes. Records from 1950 and earlier are grouped into a single category (≤ 1950). The percentage of records with unknown month is shown separately in gray (Unk.).

References

- Anderson, R. P., Araújo, M., Guisan, A., Lobo, J. M., Martínez-Meyer, E., Peterson, A. T., and Soberón, J.: Final Report of the Task Group on GBIF Data Fitness for Use in Distribution Modelling: Are Species Occurrence Data in Global Online Repositories Fit for Modeling Species Distributions? The Case of the Global Biodiversity Information Facility (GBIF), Tech. rep., GBIF Secretariat, 2016.
- Assis, J., Fernández Bejarano, S. J., Salazar, V. W., Schepers, L., Gouvêa, L., Fragkopoulou, E., Leclercq, F., Vanhoorne, B., Tyberghein, L., Serrão, E. A., Verbruggen, H., and De Clerck, O.: Bio-ORACLE v3.0. Pushing marine data layers to the CMIP6 Earth System Models of climate change research, *Global Ecology and Biogeography*, 33, e13 813, <https://doi.org/10.1111/geb.13813>, e13813 GEB-2023-0425.R2, 2024.
- Berhocoïrigoin, E.: Global integrated marine occurrence dataset for species distribution modeling., <https://doi.org/10.5281/zenodo.22746917>, 2026a.
- Berhocoïrigoin, E.: Global curated GBIF marine occurrence dataset., <https://doi.org/10.5281/zenodo.22746818>, 2026b.



- Berhocoïrigoin, E.: Global GBIF occurrence records restricted to marine locations, <https://doi.org/10.5281/zenodo.22746725>, 2026c.
- Berhocoïrigoin, E.: Global curated JeDI marine occurrence dataset., <https://doi.org/10.5281/zenodo.22746888>, 2026d.
- 1110 Berhocoïrigoin, E.: Global curated OBIS marine occurrence dataset., <https://doi.org/10.5281/zenodo.22746794>, 2026e.
- Berhocoïrigoin, E.: Global OBIS occurrence records restricted to marine locations, <https://doi.org/10.5281/zenodo.22729619>, 2026f.
- Bonnet-Lebrun, A.-S., Sweetlove, M., Griffiths, H. J., Sumner, M., Provoost, P., Raymond, B., Ropert-Coudert, Y., and Van de Putte, A. P.: Opportunities and limitations of large open biodiversity occurrence databases in the context of a Marine Ecosystem Assessment of the Southern Ocean, *Frontiers in Marine Science*, Volume 10 - 2023, <https://doi.org/10.3389/fmars.2023.1150603>, 2023.
- 1115 Bosch, S.: pyxylookup: Python Client for the OBIS xylookup API, <https://github.com/iobis/pyxylookup>, 2026.
- Botella, C., Bonnet, P., Munoz, F., Monestiez, P. P., and Joly, A.: Overview of GeoLifeCLEF 2018: location-based species recommendation, in: *CEUR Workshops Proceedings*, vol. 2125 of *CEUR Workshops Proceedings*, CEUR-WS, Avignon, France, <https://hal.science/hal-01913238>, 2018.
- Bracken, J. T., Davis, A. Y., O'Donnell, K. M., Barichivich, W. J., Walls, S. C., and Jezkova, T.: Maximizing species distribution model performance when using historical occurrences and variables of varying persistency, *Ecosphere*, 13, e3951, <https://doi.org/10.1002/ecs2.3951>, 2022.
- 1120 Carvalho, G.: flora: Tools for Interacting with the Brazilian Flora 2020, <https://doi.org/10.32614/CRAN.package.flora>, r package version 0.3.4, 2020.
- Chamberlain, S.: scrubr: Clean biological occurrence records, <https://github.com/ropensci-archive/scrubr>, archived R package, 2016.
- 1125 Chandler, M., See, L., Copas, K., Bonde, A. M., López, B. C., Danielsen, F., Legind, J. K., Masinde, S., Miller-Rushing, A. J., Newman, G., Rosemartin, A., and Turak, E.: Contribution of citizen science towards international biodiversity monitoring, *Biological Conservation*, 213, 280–294, <https://doi.org/10.1016/j.biocon.2016.09.004>, sI:Measures of biodiversity, 2017.
- Chapman, A. D.: Principles and Methods of Data Cleaning: Primary Species and Species-Occurrence Data, Tech. rep., Global Biodiversity Information Facility, 2005a.
- 1130 Chapman, A. D.: Principles of Data Quality, Tech. rep., Global Biodiversity Information Facility, 2005b.
- Chapman, A. D.: Current Best Practices for Generalizing Sensitive Species Occurrence Data, Tech. rep., GBIF Secretariat, Copenhagen, <https://doi.org/10.15468/doc-5jp4-5g10>, 2020.
- Chen, Y., Zhao, S., Zhang, L., and Zhou, Q.: Quality Assessment of Global Ocean Island Datasets, *ISPRS International Journal of Geo-Information*, 12, <https://doi.org/10.3390/ijgi12040168>, 2023.
- 1135 Cheung, C. Y. and Helfer, V.: A critical appraisal of occurrence data in OBIS and GBIF databases: a case study on true mangrove species, *Bulletin of Marine Science*, 101, 1449–1474, <https://doi.org/10.5343/bms.2024.0022>, 2025.
- Condon, R. H., Graham, W. M., Duarte, C. M., Pitt, K. A., Lucas, C. H., Haddock, S. H., Sutherland, K. R., Robinson, K. L., Dawson, M. N., Decker, M. B., Mills, C. E., Purcell, J. E., Malej, A., Mianzan, H., Uye, S.-i., Gelcich, S., and Madin, L. P.: Questioning the Rise of Gelatinous Zooplankton in the World's Oceans, *BioScience*, 62, 160–169, <https://doi.org/10.1525/bio.2012.62.2.9>, 2012.
- 1140 Condon, R. H., Lucas, C. H., Duarte, C. M., and Pitt, K. A.: Jellyfish Database Initiative: Global records on gelatinous zooplankton for the past 200 years, collected from global sources and literature (Trophic BATS project), <https://doi.org/10.1575/1912/7191>, accessed 31 August 2023, 2014.
- de Melo, P. H. A., Bystriakova, N., Lucas, E., and Monro, A. K.: A New R package to parse plant species occurrence records into unique collection events efficiently reduces data redundancy, *Scientific Reports*, 14, 5450, <https://doi.org/10.1038/s41598-024-56158-3>, 2024.



- 1145 De Pooter, D. and Perez-Perez, R.: EMODnetBiocheck: LifeWatch & EMODnet Biology QC Tool, <https://github.com/EMODnet/EMODnetBiocheck>, 2019.
- Dorey, J. B., Fischer, E. E., Chesshire, P. R., Nava-Bolaños, A., O'Reilly, R. L., Bossert, S., Collins, S. M., Lichtenberg, E. M., Tucker, E. M., Smith-Pardo, A., Falcon-Brindis, A., Guevara, D. A., Ribeiro, B., de Pedro, D., Pickering, J., Hung, K.-L. J., Parys, K. A., McCabe, L. M., Rogan, M. S., Minckley, R. L., Velazco, S. J. E., Griswold, T., Zarrillo, T. A., Jetz, W., Sica, Y. V., Orr, M. C., Guzman, L. M., Ascher,
1150 J. S., Hughes, A. C., and Cobb, N. S.: A globally synthesised and flagged bee occurrence dataset and cleaning workflow, *Scientific Data*, 10, 747, <https://doi.org/10.1038/s41597-023-02626-w>, 2023.
- Edwards, J. L., Lane, M. A., and Nielsen, E. S.: Interoperability of Biodiversity Databases: Biodiversity Information on Every Desktop, *Science*, 289, 2312–2314, <https://doi.org/10.1126/science.289.5488.2312>, 2000.
- Elith, J. and Leathwick, J. R.: Species Distribution Models: Ecological Explanation and Prediction Across Space and Time, *Annual Review
1155 of Ecology, Evolution, and Systematics*, 40, 677–697, <https://doi.org/10.1146/annurev.ecolsys.110308.120159>, 2009.
- E.U. Copernicus Marine Service Information (CMEMS): Global Ocean Physics Reanalysis, <https://doi.org/10.48670/moi-00021>, accessed on 08 Aug 2026, 2025.
- Fraisl, D., Hager, G., Bedessem, B., Gold, M., Hsing, P.-Y., Danielsen, F., Hitchcock, C. B., Hulbert, J. M., Piera, J., Spiers, H., Thiel, M., and Haklay, M.: Citizen science in environmental and ecological sciences, *Nature Reviews Methods Primers*, 2, 64,
1160 <https://doi.org/10.1038/s43586-022-00144-4>, 2022.
- Franz, N. M. and Sterner, B. W.: To increase trust, change the social design behind aggregated biodiversity data, *Database*, 2018, bax100, <https://doi.org/10.1093/database/bax100>, 2018.
- Gábor, L., Moudrý, V., Barták, V., and Lecours, V.: How do species and data characteristics affect species distribution models and when to use environmental filtering?, *International Journal of Geographical Information Science*, 34, 1567–1584,
1165 <https://doi.org/10.1080/13658816.2019.1615070>, 2020.
- GBIF.org: GBIF Occurrence Download, <https://doi.org/10.15468/dl.9gsamd>, 2024.
- GBIF.org: GBIF Parsers, <https://github.com/gbif/parsers>, software library, 2025.
- GEBCO Bathymetric Compilation Group 2026: The GEBCO_2026 Grid - a continuous terrain model for oceans and land at 15 arc-second intervals, <https://doi.org/10.5285/4f68d5c7-45eb-f999-e063-7086abc036fa>, 2026.
- 1170 GLOBE Task Team and others: The Global Land One-kilometer Base Elevation (GLOBE) Digital Elevation Model, Version 1.0, <http://www.ngdc.noaa.gov/mgg/topo/globe.html>, 1999.
- Grassle, J. F.: The Ocean Biogeographic Information System (OBIS): An On-line, Worldwide Atlas for Accessing, Modeling and Mapping Marine Biological Data in a Multidimensional Geographic Context, *Oceanography*, 13, 5–7, <https://doi.org/10.5670/oceanog.2000.01>, 2000.
- 1175 Gueta, T. and Carmel, Y.: Quantifying the value of user-level data cleaning for big data: A case study using mammal distribution models, *Ecological Informatics*, 34, 139–145, <https://doi.org/10.1016/j.ecoinf.2016.06.001>, 2016.
- Guisan, A., Thuiller, W., and Zimmermann, N. E.: *Habitat Suitability and Distribution Models: With Applications in R*, Ecology, Biodiversity and Conservation, Cambridge University Press, <https://doi.org/10.1017/9781139028271>, 2017.
- Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W. K., Budden, A. E., Batcheller, A. L., Duke, C. S., and Porter, J. H.: Big data and
1180 the future of ecology, *Frontiers in Ecology and the Environment*, 11, 156–162, <https://doi.org/10.1890/120103>, 2013.
- Hughes, A. C., Dorey, J. B., Bossert, S., Qiao, H., and Orr, M. C.: Big data, big problems? How to circumvent problems in biodiversity mapping and ensure meaningful results, *Ecography*, 2024, e07115, <https://doi.org/10.1111/ecog.07115>, 2024.



- Jin, J. and Yang, J.: BDcleaner: A workflow for cleaning taxonomic and geographic errors in occurrence data archived in biodiversity databases, *Global Ecology and Conservation*, 21, e00852, <https://doi.org/10.1016/j.gecco.2019.e00852>, 2020.
- 1185 Karin, T.: toddkarin/global-land-mask: Release of version 1.0.0, <https://doi.org/10.5281/zenodo.4066722>, 2020.
- Klaassen, M., Marques, T. A., Alves, F., and Fernandez, M.: Trends in marine species distribution models: a review of methodological advances and future challenges, *Ecography*, 2026, e07702, <https://doi.org/10.1002/ecog.07702>, 2026.
- Maldonado, C., Molina, C. I., Zizka, A., Persson, C., Taylor, C. M., Albán, J., Chilquillo, E., Rønsted, N., and Antonelli, A.: Estimating species diversity and distribution in the era of Big Data: to what extent can we trust public databases?, *Global Ecology and Biogeography*, 24, 973–984, <https://doi.org/10.1111/geb.12326>, 2015.
- 1190 Mesibov, R.: An audit of some processing effects in aggregated occurrence records, *ZooKeys*, 751, 129–146, <https://doi.org/10.3897/zookeys.751.24791>, 2018.
- Mesibov, R. E.: A specialist's audit of aggregated occurrence records, *ZooKeys*, 293, 1–18, <https://doi.org/10.3897/zookeys.293.5111>, 2013.
- Moudrý, V. and Devillers, R.: Quality and usability challenges of global marine biodiversity databases: An example for marine mammal data, *Ecological Informatics*, 56, 101051, <https://doi.org/10.1016/j.ecoinf.2020.101051>, 2020.
- 1195 Moudrý, V., Bazzichetto, M., Remelgado, R., Devillers, R., Lenoir, J., Mateo, R. G., Lembrechts, J. J., Sillero, N., Lecours, V., Cord, A. F., Barták, V., Balej, P., Rocchini, D., Torresani, M., Arenas-Castro, S., Man, M., Prajzlerová, D., Gdulová, K., Prošek, J., Marchetto, E., Zarzo-Arias, A., Gábor, L., Leroy, F., Martini, M., Malavasi, M., Cazzolla Gatti, R., Wild, J., and Šímová, P.: Optimising occurrence data in species distribution models: sample size, positional uncertainty, and sampling bias matter, *Ecography*, 2024, e07294, <https://doi.org/10.1111/ecog.07294>, 2024.
- 1200 Mozzherin, D. Y., Myltsev, A. A., and Patterson, D. J.: "gnparser": a powerful parser for scientific names based on Parsing Expression Grammar, *BMC Bioinformatics*, 18, 279, <https://doi.org/10.1186/s12859-017-1663-3>, 2017.
- Mudiyanselage, S. D., Dai, C., Howat, I. M., Larour, E., and Husby, E.: A Global High Resolution Coastline Database from Satellite Imagery, *Scientific Data*, 12, 812, <https://doi.org/10.1038/s41597-025-05180-9>, 2025.
- 1205 Nelson, G. and Ellis, S.: The history and impact of digitization and digital data mobilization on biodiversity research, *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374, 20170391, <https://doi.org/10.1098/rstb.2017.0391>, 2018.
- OBIS: Ocean Biodiversity Information System, <https://obis.org>, full export downloaded on 4 February 2024, 2024.
- OECD Megascience Forum: Final report of the working group on biological informatics, <https://doi.org/10.5281/zenodo.3383187>, 1999.
- Pebesma, E.: Simple Features for R: Standardized Support for Spatial Vector Data, *The R Journal*, 10, 439–446, <https://doi.org/10.32614/RJ-2018-009>, 2018.
- 1210 Peterson, A. T., Soberón, J., Pearson, R. G., Anderson, R. P., Martínez-Meyer, E., Nakamura, M., and Araújo, M. B.: *Ecological Niches and Geographic Distributions*, Princeton University Press, <https://doi.org/10.23943/princeton/9780691136868.001.0001>, 2011.
- Philipp, M., Dietz, A., Ullmann, T., and Kuenzer, C.: Automated Extraction of Annual Erosion Rates for Arctic Permafrost Coasts Using Sentinel-1, Deep Learning, and Change Vector Analysis, *Remote Sensing*, 14, <https://doi.org/10.3390/rs14153656>, 2022.
- 1215 Provoost, P. and Bosch, S.: obistools: Tools for data enhancement and quality control, Tech. rep., Ocean Biogeographic Information System. Intergovernmental Oceanographic Commission of UNESCO, <https://cran.r-project.org/package=obistools>, 2018.
- Ramírez, F., Sbragaglia, V., Soacha, K., Coll, M., and Piera, J.: Challenges for Marine Ecological Assessments: Completeness of Findable, Accessible, Interoperable, and Reusable Biodiversity Data in European Seas, *Frontiers in Marine Science*, Volume 8 - 2021, <https://doi.org/10.3389/fmars.2021.802235>, 2022.



- 1220 Reagan, J. R., Boyer, T. P., García, H. G., Locarnini, R., Baranova, O. K., Bouchard, C., Cross, S. L., Mishonov, A. V., Paver, C. R., Seidov, D., Wang, Z., and Dukhovskoy: World Ocean Atlas 2023, <https://doi.org/10.25921/va26-hv25>, 2023.
- Ribeiro, B. R., Velazco, S. J. E., Guidoni-Martins, K., Tassarolo, G., Jardim, L., Bachman, S. P., and Loyola, R.: bdc: A toolkit for standardizing, integrating and cleaning biodiversity data, *Methods in Ecology and Evolution*, 13, 1421–1428, <https://doi.org/10.1111/2041-210X.13868>, 2022.
- 1225 Robertson, M. P., Visser, V., and Hui, C.: Biogeo: an R package for assessing and improving data quality of occurrence record datasets, *Ecography*, 39, 394–401, <https://doi.org/10.1111/ecog.02118>, 2016.
- Soberón, J. and Peterson, T.: Biodiversity informatics: managing and applying primary biodiversity data, *Philosophical Transactions of the Royal Society B: Biological Sciences*, 359, 689–698, <https://doi.org/10.1098/rstb.2003.1439>, 2004.
- Uber Technologies: H3: Uber’s Hierarchical Hexagonal Geospatial Indexing System, <https://github.com/uber/h3-py>, python package, 2026.
- 1230 Van den Bossche, J., Jordahl, K., Fleischmann, M., Richards, M., McBride, J., Wasserman, J., Garcia Badaracco, A., Snow, A. D., Roggemans, P., Ward, B., Gerard, J., Tratner, J., Perry, M., Taves, M., Farmer, C., Hjelle, G. A., Tan, N. Y. S., Bell, R., ter Hoeven, E., et al.: geopandas/geopandas: Version 1.1.3, <https://doi.org/10.5281/zenodo.18934477>, 2026.
- Vandepitte, L., Bosch, S., Tyberghein, L., Waumans, F., Vanhoorne, B., Hernandez, F., De Clerck, O., and Mees, J.: 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>, 2015.
- 1235 Veiga, A. K., Saraiva, A. M., Chapman, A. D., Morris, P. J., Gendreau, C., Schigel, D., and Robertson, T. J.: A conceptual framework for quality assessment and management of biodiversity data, *PLoS ONE*, 12, e0178731, <https://doi.org/10.1371/journal.pone.0178731>, 2017.
- VertNet: BELS: Biodiversity Earth Locality Services, <https://github.com/VertNet/bels>, gitHub repository, accessed 12 June 2026, 2026.
- Wessel, P. and Smith, W. H. F.: A global, self-consistent, hierarchical, high-resolution shoreline database, *Journal of Geophysical Research: Solid Earth*, 101, 8741–8743, <https://doi.org/10.1029/96JB00104>, 1996.
- 1240 Whitaker, J. and contributors: Basemap 2.0.0: Plot data on map projections with matplotlib, <https://github.com/matplotlib/basemap>, python package, 2025.
- WoRMS Editorial Board: World Register of Marine Species, <https://doi.org/10.14284/170>, accessed 19 June 2026, 2026.
- 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.: CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases, *Methods in Ecology and Evolution*, 10, 744–751, <https://doi.org/10.1111/2041-210X.13152>, 2019.
- 1245 Zizka, A., Antunes Carvalho, F., Calvente, A., Rocio Baez-Lizarazo, M., Cabral, A., Coelho, J. F. R., Colli-Silva, M., Fantinati, M. R., Fernandes, M. F., Ferreira-Araújo, T., Gondim Lambert Moreira, F., Santos, N. M. C., Santos, T. A. B., dos Santos-Costa, R. C., Serrano, F. C., Alves da Silva, A. P., de Souza Soares, A., Cavalcante de Souza, P. G., Calisto Tomaz, E., Vale, V. F., Vieira, T. L., and Antonelli, A.: No one-size-fits-all solution to clean GBIF, *PeerJ*, 8, e9916, <https://doi.org/10.7717/peerj.9916>, 2020.
- 1250