Species richness variation in marine and terrestrial fauna across wide-spread, fragmented territories: assessing inherent challenges of data scarcity at local and regional scales

Kilian Barreiro[1], Laura Benestan[1], Charlotte Moritz[2], Simon Ducatez[3], Jérémy Le Luyer[1], and Cristián Monaco[1]

[1]IFREMER
[2]CMOANA Consulting
[3]IRD

January 09, 2025

Abstract

The ongoing biodiversity crisis calls for a complete biodiversity inventory of marine and terrestrial ecosystems. The task is particularly challenging for fragmented island territories, where baseline biodiversity information is often difficult to procure. By centralising information from different sources (museums, research institutions, citizen scientists), 'big-data' platforms provide an opportunity to evaluate species biodiversity information of understudied regions. Using data from the Global Biodiversity Information Facility (GBIF), we curated the first biogeographic dataset for both marine and terrestrial animal species in French Polynesia, a large territory composed of 124 islands and atolls that belongs to the Central Pacific region, a marine biodiversity hotspot facing conservation challenges. The dataset revealed heterogeneous species richness across archipelagos and islands, prompting an investigation into potential sampling biases (institutional, taxonomic, spatial) as well as an assessment of island-specific accessibility biases. We estimated that the archipelagos and islands had an inventory completeness rate that ranges from 12 to 85%, suggesting that a large proportion of the studied area remains poorly documented. Spatial and temporal sampling biases were partly explained by accessibility constraints (proximity to airports, roads or ports), and inventory completeness was higher for marine than terrestrial species. The biases quantified here challenge our ability to conduct biogeographic analyses that integrate the land-sea meta-ecosystem. Our database allows identifying taxa and sampling locations that require urgent attention, as well as comprehensively recorded species that can serve as indicators for environmental degradation. Explicitly acknowledging the inherent biases of biodiversity datasets is the first step towards a more comprehensive characterization of species diversity across fragmented territories. This information is crucial for guiding sound adaptive-management and conservation planning strategies.

Introduction

Humans are driving an unprecedented erosion of marine and terrestrial biodiversity, fundamentally altering the structure and functioning of ecosystems, and in return threatening the beneficial contributions that nature provides (IPBES, 2019; Ceballos & Ehrlich, 2023; Gorman *et al.* , 2023). Implementing conservation actions to confront this crisis requires comprehensive and spatially explicit baseline information on species diversity across the planet (Singh, 2002). Ultimately, these data are essential for guiding conservation management based on a sound understanding of the ecological and evolutionary processes that drive spatial and temporal patterns of species distribution across ecosystems (Newmark *et al.* , 2017; Pilowsky *et al.* , 2022).

Thanks to the concerted efforts from museums, research institutions, citizen scientists, and 'big-data' platforms facilitating the integration of information, biodiversity records are increasingly available (Farley *et*

al. , 2018; Kays *et al.* , 2020; Heberling *et al.* , 2021). Over the last two decades, many initiatives to centralise species occurrence data have emerged, notably some online repositories including the Global Biodiversity Information Facility (GBIF, https://www.gbif.org/) and the Ocean Biodiversity Information System (OBIS, https://obis.org/). By adhering to the FAIR principles (Findability, Accessibility, Interoperability, and Reusability) and the metadata-sharing standards, such as the Darwin Core (DwC) (Wieczorek *et al.* , 2012), EML (Fegraus *et al.* , 2005), and BioCASe (Güntsch *et al.* , 2007), these intergovernmental research infrastructures promise to expedite the study of biodiversity across ecosystems. GBIF and OBIS are the largest open-access occurrence data portals for terrestrial and marine species, both being routinely used to inform resource management and conservation programs (e.g., Levin *et al.* , 2014; Amano *et al.* , 2016; Underwood *et al.* , 2018; Lin *et al.* , 2022; Takashina & Kusumoto, 2023).

Despite their growing popularity, open-access biodiversity databases have been criticised on the grounds of poor data quality, potentially limiting their scope and applicability (Hortal *et al.* , 2015). Important shortfalls that are often cited include standardisation issues during sampling (Troia & McManamay, 2016; Zizka *et al.* , 2020), incomplete and/or incorrect records (e.g., species misidentification) and sampling biases, either spatial/temporal (i.e., unbalanced sampling efforts across space/time), taxonomic (i.e., skewed sampling favouring certain taxa), or both (Troudet *et al.* , 2017; Zizka *et al.* , 2020; García-Roselló *et al.* , 2023; Rocchini *et al.* , 2023). While cleaning and filtering methods allow readily correcting for incomplete and/or incorrect entries, sampling biases are difficult to diagnose and require special attention (Schiesari *et al.* , 2007). The spatial sampling bias, considered one of the main challenges limiting our comprehensive understanding of large-scale biodiversity patterns (Wüest *et al.* , 2020), can be partly explained by socio-economic reasons (e.g., wealthy zones are more likely to be surveyed Beck *et al.* , 2014), a scientific bias towards certain taxa (Troudet *et al.* , 2017), differences in sampling standards (König *et al.* , 2019), and/or by logistical difficulties to access certain locations (Kadmon *et al.* , 2004; Engemann *et al.* , 2015).

Remote oceanic islands are likely to show sampling gaps due to their geographical isolation, which ultimately results in patchy and poorly representative data for the study region. The difficulties and high costs associated with organising monitoring campaigns further exacerbate these biases. As a result, some islands are underrepresented in long-term monitoring schemes (Stephenson *et al.* , 2017), and, aside from a few exceptions (e.g., Hachich *et al.* , 2015), comprehensive biodiversity studies across widespread archipelagos remain rare. This paucity of information for islands and atolls is particularly detrimental because they are *a priori* highly vulnerable ecosystems that potentially harbor high levels of endemism due to their isolation (Simberloff, 2000; Russell & Kueffer, 2019). Additionally, fragmented archipelagos are unique natural laboratories that provide opportunities for studying the ecological and evolutionary processes driving biodiversity patterns, dispersal potential, endemism and extinction rates, for both marine and terrestrial organisms. However, a proper understanding of these biogeographical processes first requires robust baseline information on species distribution (Warren *et al.* , 2015; Whittaker *et al.* , 2017).

With 124 high islands and atolls spread across five archipelagos covering 4.8 million km^2 (Andréfouët & Adjeroud, 2019; Galzin & Meyer, 2024), French Polynesia represents the epitome of a fragmented territory. The large number of islands, their relative isolation, and the sheer variation in geomorphological characteristics they exhibit complicate efforts to survey the entire region or avoid sampling biases. Indeed, the marine and terrestrial biogeography of French Polynesia has only been partly studied, with a remarkable skew towards specific taxonomic groups. In the marine realm, targeted investigations have mainly focused on marine molluscs, brown seaweeds (*Phaeophyceae*) and reef fishes (Kulbicki, 2007; Salvat, 2009; Tröndlé & Boutet, 2009; Delrieu-Trottin *et al.* , 2015, 2019; Salvat & Tröndlé, 2017; Boutet *et al.* , 2020; Vieira *et al.* , 2021, 2023; see references therein). In the terrestrial realm, data compilations include a checklist of the recorded land and fresh-water arthropods (Ramage, 2017), a biogeographic atlas of birds (Thibault & Cibois, 2017), and an inventory of the vascular flora (Florence, 1997, 2004; Chevillotte *et al.* , 2019), as well as some rare studies focusing on the phylogeographic origins of specific terrestrial biota (e.g., Gillespie *et al.* , 2008; Hembry, 2018). Overall, the lack of a centralised, complete, and unbiased dataset for the region prevents an exhaustive analysis of the biogeographical status of marine and terrestrial species across French Polynesia. As a model of a highly fragmented island system, improving our fundamental understanding of

French Polynesian biogeography is not only critical for cataloguing the existing fauna of the region, but also for contributing to our general comprehension of the ecological processes driving the current biodiversity crisis in isolated systems (Russell & Kueffer, 2019; Fernandez-Palacios *et al.* , 2021).

Using data originally downloaded from open-access portals (GBIF, OBIS), we compiled and curated the first biogeographic dataset for both marine and terrestrial animal species in French Polynesia. We used these data to: (1) provide a baseline characterization of the number of species in the region; (2) identify taxonomic groups that might require further investigation, as well as comprehensively recorded species that can serve as indicators for environmental degradation; (3) identify poorly- and well-surveyed islands; and (4) quantify island-specific accessibility biases leading to heterogeneous sampling efforts.

Materials and Methods

Data collection

We downloaded occurrence data from the GBIF portal (http//gbif.org; https://doi.org/10.15468/dl.gaxgr7) on May 24, 2023, covering French Polynesia (polygon spanning between 5°S and 30°S, and 134°W and 155°W). Species occurrences are defined as records of a particular species (or other taxonomic rank), with a geographic location and timestamp. These raw data were treated following the Darwin Core (Wieczorek *et al.* , 2012), Ecological Metadata Language (Fegraus *et al.* , 2005), and BioCASe (Güntsch *et al.* , 2007) standards. A pre-filtration of the data was done to exclude records missing geographic location and/or taxonomic classification (e.g., not available or zeros), yielding 297,789 occurrences (Fig. 1). Because GBIF and OBIS signed a data-sharing agreement which was effective at the time we downloaded the data, the marine data from OBIS was also contained in our GBIF data. The coastline shapefiles used to analyse the region included 120 geographical structures, most of which were atolls and high islands. Hereafter, we refer to all geographical structures as "islands". Each record retrieved from the GBIF dataset was assigned to its nearest island based on geographic distances estimated using the function *st_nearest_feature* available in the *sf* package v.1.0-15 (Pebesma & Bivand, 2023) in R (R Core Team, 2024).

Validation of the taxonomic information

To clean, homogenise, and validate the taxonomic information in the dataset, we assumed that misidentifications would occur at the species level. To ensure taxonomic reliability we validated each species name using *ad hoc* taxonomic data repositories. We first validated the species name of each recognized taxon with WoRMS (*World Register of Marine Species,* https://www.marinespecies.org/) using the *wm_records_name* function from the R package *worrms* (Chamberlain & Vanhoorne, 2023). We then assigned a taxonomic status (i.e., "accepted", "doubtful", "synonym") to each record following the criteria outlined by the GBIF Backbone taxonomy (see https://doi.org/10.15468/39omei, https://hosted-datasets.gbif.org/datasets/backbone/). Taxa that were assigned as either "doubtful" or "synonym" were replaced by the updated taxonomic name provided by WoRMS. Taxa not recognized by the WoRMS repository were further examined using the *gnr_resolve* function from the R package *taxize* (Chamberlain *et al.* , 2020), which provides a means to validate species names by accessing several additional repositories via specific Application Programming Interfaces (e.g., ITIS: *Integrated Taxonomic Information System* , https://www.itis.gov/; *CoL* : Catalogue of Life, https://www.catalogueoflife.org/; bold: *Barcode of Living Data,* https://www.boldsystems.org/). Taxa that were not recognized neither by WoRMS nor Taxize were submitted to TAXREF (taxonomic reference curated by the French National Museum of Natural History, https://inpn.mnhn.fr/programme/referentiel-taxonomique-taxref) using the *rt_taxa_search* function from the *rtaxref* R package (Grenié & Gruson, 2022). A final manual check was done for records that could not be identified in the aforementioned taxonomic repositories.

Habitat classification and biogeographical status

Habitat classification for marine and terrestrial species were verified using WoRMS and Taxref, respectively. Habitat information was split into four categories (i.e., marine, brackish, freshwater and terrestrial) according to the classification scheme favoured by WoRMS. Missing habitat information was completed using the

TaxRef database. For our analyses of terrestrial and marine ecosystems, we focused on species that were classified as exclusively "marine" or exclusively "terrestrial". Species classified as amphibious or those inhabiting both terrestrial and marine environments at different life stages (e.g., seabirds like *Gygis alba* or *Sula sula*) or during specific phases of their life cycle (e.g., insects with aquatic larval stages) were included in the cleaned dataset (labeled as "Mixed" in Fig. 1) but excluded from further analyses.

Data filtration sequence

Because geographic, taxonomic and accuracy standards have changed over time (Maldonado *et al.* , 2015; Zizka *et al.* , 2020), and notorious errors were detected in older records, we retained entries dating from 1950 onwards and excluded those without timestamps (Fig. 1). Subsequently, we removed all absence data to rule out potential biases due to false negatives and no-observation data (Bonnet-Lebrun *et al.* , 2023). We then restricted occurrences to those described in the *basisOfRecords* column as: "human observation", "machine observation", "material sample", "material citation" and "preserved specimen" according to recommendations by Smith et al. 2018. Cross-checking values corrected real duplicates in *decimalLatitude, decimalLongitude, ScientificName, Year, Month* and *Day* categories. Finally, species records lacking *Habitat* information were removed from the dataset (Fig. 1).

Taxonomic biases: Identifying under- and over-represented groups

We estimated the taxonomic bias at the *Class* level based on its over- or under-representation, relative to an "ideal sampling effort index". The ideal number of records for a given class was estimated based on the hypothetical scenario where each species received the same number of records, and therefore each class received a number of records directly proportional to its number of species (Troudet *et al.* , 2017), according to:

$$\text{Ideal} = N_{rec} * (N_{sp\_group} / N_{sp\_tot})$$

where N_{rec} = total number of records, N_{sp_group} = number of distinct species within the taxonomic group, and N_{sp_tot} = total number of species present in the whole dataset. Taxonomic bias was assessed based on the difference between the ideal and observed sampling efforts, calculated for each class with more than 100 records in marine habitats and more than ten records in terrestrial habitats. To highlight values that deviated significantly from the ideal, we applied an inverse hyperbolic sine transformation to the data. We also identified the top ten most representative species for each habitat.

Spatial and temporal heterogeneity in the sampling effort

To examine spatial heterogeneity in the sampling effort, we first mapped the number of records and species estimated for each island within each archipelago. The agreement between the number of records and species per island ($\log_{10}$-transformed) was evaluated based on the Pearson correlation coefficients and its statistical significance. Additionally, to quantify the prevalence of heterogeneous sampling effort across space, we assessed the proportion of species recorded only once at each island, a common method in biodiversity studies to evaluate sampling completeness and detect potential under-sampling biases (Lim*et al.* , 2012). The parameter uniqueness—species that have only been collected once—is widely recognized as an indicator of incomplete sampling (Chao *et al.* , 2020; Montes *et al.* , 2021), allowing researchers to infer the adequacy of the sampling effort and identify areas that may require further investigation. We considered q_k as the number of species documented in k sampling-effort units, so that the number of species observed in a single sampling-effort unit is q_1 (i.e., unique), the number of duplicates is q_2, and so on.

Inventory completeness

To investigate the degree of inventory completeness in the dataset for both marine and terrestrial ecosystems at the scale of the archipelago and the island, we estimated the species inventory completeness percentage (C), calculated as:

$$C_{(i)} = (Sobs_{(i)} / Sest_{(i)}) * 100$$

where i = each island or archipelago, *Sobs* = number of species observed, and *Sest* = number of species estimated at each archipelago and island (Soberón *et al.* , 2007). To estimate*Sest* , we used the Species Accumulation Curve (SAC) approach that describes the relationship between species richness and sampling effort, i.e., the number of records available in a grid cell (Deng *et al.* , 2015). To derive the SACs, we split the area in 0.05° (~25 km^2) grid cells. We described the SACs using the *specaccuml* function (method = "exact") available in the R package *vegan* v.2.6-4 (Oksanen *et al.* , 2024). We fitted the Michaelis-Menten model with the *fitspecaccum* function (method = "michaelis-menten") to provide estimates of the number of species likely to be present (i.e., *Sest, which corresponds to* the asymptotic richness, or parameter *Vm* in the Michaelis-Menten equation) and the number of records required to capture 50% (K) of the estimated number of species predicted by the model (Chao, 1984, 1987; Colwell & Coddington, 1994). Because biodiversity assessments can be biased by grid cells with extremely low species records, we considered a minimum threshold of ten observations to run the SACs, as was done in previous studies (Soberón *et al.* , 2007; De Araujo*et al.* , 2022).

While we aimed to calculate the SACs for each island and archipelago based on grid cells, as recently done in several studies on macroecology using GBIF datasets (De Araujo *et al.* , 2022; Ramírez *et al.* , 2022; Chanachai *et al.* , 2024), only seven islands (i.e., Anaa, Huahine, Moorea, Nuku Hiva, Raiatea-Tahaa, Rangiroa and Tahiti) were sufficiently large to yield enough cells (greater or equal to 10) to fit the Michaelis-Menten model. We therefore generated archipelago-scale models based on 5-km grid cells, while for the island-scale models we used the geographic coordinates associated with the species records. A preliminary comparison between these two approaches (0.05-degree resolution grid cells *vs.* records) revealed a significant correlation between them for the archipelago scale ($R^2 = 0.64$). Therefore, we only presented SACs based on records for both archipelagos and islands. To evaluate inventory completeness, we determined the total number of islands with more than 100 records and C greater or equal to 80%, meaning that at least 80% of the species have been sampled (Soberón & Peterson, 2004; Chanachai *et al.* , 2024). We then examined the correlation between the number of records and C to test whether these proxies of sampling effort and reliability were associated. We used a Spearman correlation test for non-parametric data. Statistical significance was evaluated based on [?] = 0.05.

Sampling bias due to accessibility

To explore the influence of accessibility constraints on these sampling biases, we used a Bayesian approach to estimate how sampling rates vary with proximity to several common anthropic accessibility factors (i.e., rivers, roads, cities, airports, and ports). Using the*calculate_ bias* function from the *sampbias* R package v. 2.0.0 (Zizka *et al.* , 2021), we estimated the bias weights (w), which quantify the impact of each accessibility factor on sampling rates. These weights are calculated assuming an exponential decline in sampling rates as distance from accessibility factors increases. This package also provides spatially explicit estimates of the number of records (i.e., expected records) using a Poisson sampling process while accounting for the influence of the accessibility factors. Because the geospatial data contained by default in the *sampbias* package is incomplete for French Polynesia (Natural Earth Data, https://www.naturalearthdata.com/), we manually inputted vector data for rivers, roads, cities with > 1,000 inhabitants, airports, and ports. These data were provided by the French Polynesian agency for marine resources, the *Direction des Ressources Marines* . We defined a grid (*inp_ raster* parameter) contained within the same polygon used for downloading the GBIF data, with 0.05 degrees resolution (~ 5.5 km). This was done for consistency with the SAC analyses. Each grid cell was assigned to the nearest island based on geographic distances estimated using the function*st_ nearest_ feature* available in the *sf* R package v.1.0-15 (Pebesma & Bivand, 2023).

Code availability

Analyses were done in R version 4.4.3 (R Core Team, 2024). The analyses scripts are available in GitHub (https://github.com/KilianBARREIRO/biogeography_datadiv). The data are available in SEANOE (https://www.seanoe.org/data/00878/99018/).

Results

Curated GBIF dataset for French Polynesian marine and terrestrial species

From the original 297,789 records included in the GBIF dataset, we removed 20,967 records that were either dated before 1950, or which did not have a time stamp (Fig. 1). A total of 107,200 records (35.9%) were identified as duplicated, 24.7% of which originated from *citizen science* sources (e.g., iNaturalist research-grade observations). We excluded 12,825 records with no or unclear habitat information, 45.4% of which were sourced from research institutions or peer-reviewed datasets. The number of records accessible via GBIF per year has increased over time since 1950, reaching maximum values in 2011, 2006, and 2009, with 19,770, 11,056, and 9,397 records, respectively (Fig. S1). This increase in records was mainly explained by the punctual contribution of two out of 121 publishers: OBIS-SEAMAP and UMS PatriNat (OFB-CNRS-MNHN, Paris). The mean number of records per species was 22 (median = 3), ranging from 1 to 12,339. The records produced by citizen scientists represented 20.6 % (32,155) of those records. Data collected by citizen scientists were also the main source of data for 17 islands, and the only source (100%) for six islands (Fangatau, Hiti, Marutea nord, Motu Nao, Pinaki, and Vairaatea). *Human observation* , including institutional and citizen science publishers, was the most frequently used recording method, with 71.1% (111,260 records) of total records. *Preserved specimen* and *material sample* categories accounted for 23.8% and 4.6% of records, respectively. WoRMS validated the taxonomy of 82.3% of total records and 97% of non-terrestrial records. Only 268 species lacked information on their habitat, which we completed manually. The resulting cleaned dataset was composed of 156,380 records including 111,889 marine, 15,979 terrestrial and 28,512 mixed records for 5,863 marine, 1,045 terrestrial and 201 mixed species, collected from 1950 to 2022 (Figs. 1 and S1).

Taxonomic composition and biases

The number of recorded species was ˜5.6 times higher for marine than terrestrial ecosystems, with 5,863 marine and 1,045 terrestrial species, respectively. For marine taxa, the dataset included 18 phyla, with three major groups: Mollusca (2,303 species), Chordata (1,713 species), and Arthropoda (1,144 species), accounting for over 94.5% of marine records (105,778 records). Five classes alone accounted for 91.8% of the observations: Teleostei (75,313 records, 1,522 species), Gastropoda (15,416 records, 1,999 species), Malacostraca (5,678 records, 1,070 species), Bivalvia (3,470 records, 270 species) and Mammalia (2,807 records, 25 species; Fig. 2). The most represented marine species were *Tridacna maxima* (small giant clam, 1,310 records), *Megaptera novaeangliae* (humpback whale, 1,011 records), and *Ctenochaetus striatus* (striated surgeonfish, 941 records; Fig. 2). A total of 90% of the marine species had 27 or fewer records, and 30% were unique records.

The terrestrial taxa comprised five phyla, including Arthropods (766 species), Mollusca (202 species), Chordata (73 species), Platyhelminthes (3 species) and Nematoda (1 species). The five most recorded classes were Aves (8,921 records, 53 species), Insecta (3,569 records, 688 species), Gastropoda (2,295 records, 202 species), Arachnida (528 records, 61 species), and Squamata (442 records, 10 species), representing 98.6% of all terrestrial species records. A total of 90% of the terrestrial species had 23 records or fewer, and 30% were unique records. Three introduced bird species, *Geopelia striata* (zebra dove), *Acridotheres tristis* (common myna), *Pycnonotus cafer* (red-vented bulbul), were the most recorded terrestrial species, with 1,174, 1,140 and 957 occurrences (Fig. 2), of which 93.1% were provided by the "Cornell Lab of Ornithology".

Spatial and temporal heterogeneity in sampling effort and the number of recorded species

We observed a significant and strong correlation between the log-10 number of records per island (i.e., a proxy for sampling effort) and the number of species per island for both marine ($\rho = 0.729$, $P < 0.001$) and terrestrial ($\rho = 0.890$, $P < 0.001$) ecosystems (Fig. S2). This analysis excluded islands that lacked records in both marine and terrestrial habitats.

Our dataset included marine species records for 119 out of 120 islands. The number of records per island was heterogeneous (Fig. 3), ranging from 1 to 54,761, with a mean of 940 records (median = 70). The number of species present was also highly heterogeneous across space, ranging from 1 to 2,759 species per island, with a mean of 194 species (median = 56) per island. The Society Archipelago (13 islands) held 63.3% of all marine-species records, 92.2% of which were observed in Moorea (54,761 records), Tahiti (6,721

records), and Raiatea-Tahaa (3,829 records; Fig. 4). Considering the other four archipelagos, the islands that exhibited the highest number of records were Rapa (3,731 records) in the Austral islands (11 islands), Fakarava (3,739 records) in the Tuamotu (69 islands), Nuka Hiva (2,741 records) in the Marquesas (17 islands), and Mangareva (2,490 records) in the Gambier (11 islands; Fig. 3). Gambier was the least sampled archipelago, accounting for 3.6% of all marine records, and for only 13.4% of all marine species identified.

Considering the terrestrial habitat, our dataset identified 64 islands with at least one species record, and 56 islands with no records. As for the marine habitat, the number of terrestrial species records per island was heterogeneous (Fig 3), ranging from 1 to 4,700, with a mean of 250 records per island (median = 11.5). The number of species identified per island ranged from 1 to 388, with a mean of 36 species per island (median = 5). The Society Archipelago held 74.3% of all terrestrial species records, 85.2% of which were registered in the trio Moorea (4,341 records, 388 species), Tahiti (4,700 records, 306 species), and Raiatea-Tahaa (1,070 records, 204 species; Fig. 4). Considering the other four archipelagos, the islands showing the highest number of records were Anaa (694 records) in the Tuamotu, Rurutu (470 records) in the Austral, Nuku Hiva (528 records) in the Marquesas, and Mangareva (113 records) in the Gambier (Fig. 4). As for the marine database, the Gambier archipelago had the lowest number of terrestrial records, representing only 4.7% of all terrestrial species identified.

Inventory completeness

Considering the archipelago scale, the SAC analysis showed that the number of species recorded increased with sampling effort. Although the curves for both marine and terrestrial datasets exhibited a plateau, they did not reach a clear saturation point (Fig. S3). Our calculations suggest that marine inventory completeness was comparable among archipelagos, with 75.59%, 76.99%, 71.42%, 80.41% and 79.01% for the Austral, Gambier, Marquesas, Society and Tuamotu Archipelagos, respectively, indicating that at least 70% of the species were detected overall. According to the asymptote values based on the Michaelis-Menten model (Vm), marine species richness was lowest at the Gambier (1,150 species) and the highest at the Society (5,916 species) archipelagos. The Austral, Marquesas and Tuamotu Archipelagos showed similar asymptote values of 2,901, 2,860 and 2,270 expected species, respectively (Table 1).

Inventory completeness for terrestrial species was highly heterogeneous across archipelagos, ranging from 48.10% for the Marquesas, the northernmost and most remote archipelago, to 84.50% in the Tuamotu, the largest archipelago. Inventory completeness for terrestrial species was higher than for marine species in the Society (C = 84.04%) and Tuamotu Archipelagos (C = 84.50%), but lower in the Gambier (C = 64.06%) and Marquesas (C = 48.10%). In the Austral Islands, inventory completeness was similar between marine and terrestrial species (C = 72.65%). For terrestrial species, the asymptote values based on the Michaelis-Menten model (Vm) ranged from 89 (Gambier) to 839 species (Society), with 283, 275, 102 species estimated for the Austral islands, Marquesas and Tuamotu, respectively.

At the island scale and for the marine dataset, we fitted SACs for 84 out of 119 islands having at least 10 records (Table 2). Inventory completeness was highly heterogeneous, ranging from 10.89% (Faaite Atoll, Tuamotu, 104 records) to 85.85% (Tenararo Atoll, Tuamotu, 17 records), with an average (± SD) of 49.50 % (±16.91%). Assuming a threshold of C [?] 80% and at least 100 records, only two islands were classified as well-sampled: Moorea (54,761 records, C = 84.13%) and Fakarava (3,739 records, C = 83.20%). Among the islands with the highest number of records, we identified low to moderate inventory completeness for Tahiti (1,337 records, C = 69.40%) and Raiatea-Tahaa (3,829 records, C = 34.30%) in the Society, Rapa (3,731 records, C = 67.83%) and Rimatara (1,664 records, C = 31.80%) in the Austral Islands, and Nuku Hiva (2,741 records, C = 61.96%) and Fatu Hiva (1,103 records,C = 27.01%) in the Marquesas. The correlation between inventory completeness and the number of records per island was low (R2 = 0.28; P-value < 0.001).

For the terrestrial dataset, 32 islands had sufficient records (> 10 records) to fit SACs (Table 3). Inventory completeness ranged from 8.54% for the Tikey Atoll (13 records, Tuamotu) to 97.39% for Anaa Atoll (694 records, Tuamotu). Other well-sampled islands (C [?] 80% and 100 records) included Ua Huka (114 records, C = 89.59%) in the Marquesas, Raivavae (201 records, C = 84.64%) and Rimatara (225 records, C = 80.

04%) in the Austral. Islands with the highest number of records, including Moorea (4,341 records, C = 70.56%) and Tahiti (4,699 records, C = 63.23%, respectively) were nearly well-sampled. Terrestrial species inventory completeness and sampling effort were not correlated across these islands (R2 = 0.07; P-value > 0.05).

Some islands exhibited contrasting patterns between terrestrial and marine inventories. For instance, Rimatara was well sampled for terrestrial species (225 records, C = 80.04%) but only moderately sampled for marine species (1,664 records, C = 31.80%). Fakarava showed the opposite trend, with an almost complete marine inventory (C = 83.20%), while its terrestrial inventory was sparse (23 records, C = 25.07%).

Islands for which we were unable to fit SACs were classified as either "neglected islands" (i.e., with no data at all) or "poorly-documented islands" (i.e., with not enough data). For the marine and terrestrial data, we identified two and 56 neglected islands, respectively. The problem of missing data was prevalent across archipelagos, but less important in the Society and Austral archipelagos (Fig. 5). We found 36 and 32 poorly-documented islands for marine and terrestrial ecosystems, respectively. The data scarcity was particularly pronounced in the largest archipelago, the Tuamotu, as well as in the southernmost archipelago, the Gambier (Fig. S4).

Sampling bias due to human accessibility

For marine species, we found that sampling effort was strongly associated with the presence of roads (w = 0.047), and moderately affected by the presence of ports or airports (w = 0.020). The presence of cities and rivers (waterbodies) contributed very little to the sampling bias of marine species (cities' w = 0.002, waterbodies' w = 0.001) (Fig. 6).

Similar results were found for the terrestrial data, where the presence of roads contributed the most to the accessibility bias (w = 0.062). The effect of airports and ports was moderate (w = 0.031) while the influence of cities and water bodies was negligible (cities'w = 0.004, waterbodies' w = 0.001) (Fig. 6). The model also revealed a low number of marine and terrestrial records (Table S1), even after correcting for accessibility biases, in the Tuamotu and Gambier Archipelagos, except for Mangareva, Hao, and Arutua Islands. In contrast, most islands in the Society Archipelago were oversampled relative to the overall sampling effort across French Polynesia.

Discussion

Our study compiles the most comprehensive open-source database on animal biodiversity in French Polynesia, illuminating regional- and island-scale biodiversity patterns of marine and terrestrial fauna across this vast and fragmented territory. While our results highlight significant disparities in sampling effort across islands, this work offers valuable quantitative insights into completeness of taxonomic and spatial data throughout French Polynesia. This work also highlights understudied areas and taxonomic groups, providing a practical tool for conservation planners to guide future sampling strategies and enhance biodiversity representativeness. We argue that this integrative approach is essential for explicitly addressing the inherent biases often present in large-scale biodiversity studies (Rocchini *et al.*, 2023).

Building an accurate GBIF dataset

While open-source biodiversity datasets offer unique opportunities for studying macroecological processes, global repositories face criticism due to significant variation in data quality and quantity, depending on geographic, temporal, and taxonomic factors (Garcia-Rosello *et al.*, 2023). Ignoring these caveats can lead to erroneous conclusions. However, when carefully considered, they can enhance the utility of open-source data by highlighting critical biodiversity knowledge gaps (e.g., Meyer *et al.*, 2016; Cornwell *et al.*, 2019; Moudry & Devillers, 2020). Addressing uncertainties in the data first requires acknowledging that open-source biogeographic datasets are likely to be incomplete (Wuest *et al.*, 2020), especially in vast and fragmented regions and for specific groups of organisms. Secondly, standardised taxonomic repositories (e.g., WORMS) offer workflows for cleaning data retrieved from open-source platforms while adhering to FAIR data-sharing principles. Here, by applying previously validated filtering protocols (Bonnet-Lebrun *et al.*, 2023), we enhanced

the geographic and taxonomic accuracy of GBIF records for French Polynesia, closely matching recent expert taxonomic assessments.

Our database contains a total of 7,109 species, including 1,876 vertebrates and 5,233 invertebrates. Regarding vertebrates, we found that every known marine mammal (26 out of 26 species) and a large number of birds (129 out of 175 species) previously documented in the region are represented (Clements et al. 2024). Our database includes 2,303 marine molluscs out of 3,022 referenced in a recently published checklist and identification guide (Boutet *et al.*, 2020) and the Teleostei class included 1,523 species, which is more than the 1,310 reported in the most complete identification guides for the region (Bacchet *et al.*, 2017; Siu *et al.*, 2017). While the taxonomic coverage is reassuring for marine species, it remains relatively limited for terrestrial species. For example, our records include only 688 out of 2,497 insect species (Insecta) and 61 out of 365 spider species (Arachnida) described in the region (Ramage, 2017). Data scarcity for insects is a global issue, and in some regions, it is partly driven by species extinction rates that outpace discovery rates (Porch *et al.*, 2020; Rocha-Ortega *et al.*, 2021). Islands, which harbour approximately 20% of the world's terrestrial biodiversity, are critical reservoirs of fragile and threatened biodiversity (Fernandez-Palacios *et al.*, 2021). This highlights the urgent need to document the exceptional biodiversity of insular countries like French Polynesia, where some taxonomic groups, such as ground beetles, contribute significantly to global biodiversity (Liebherr, 2012; Fernandez-Palacios *et al.*, 2021). Our study provides an efficient framework for identifying poorly sampled species, which can be extended to other taxonomic groups in French Polynesia (e.g., plants or algae) and applied more broadly to other regions.

Linnean shortfall

The Linnean shortfall—i.e., only a fraction of the planet's species has been described—is a major gap in our understanding of biodiversity (Hortal *et al.* , 2015), limiting our ability to effectively address the ongoing extinction crisis (Ceballos & Ehrlich, 2023). The Linnean shortfall is partly driven by taxonomic sampling biases, where societal preferences influence which groups are more frequently recorded (Troudet *et al.* , 2017). This explains why patterns of sampling efforts are often represented by homogeneously-sampled taxonomic groups such as marine mammals (Moudry & Devillers, 2020), fishes (Mora*et al.* , 2008) or insects (Sanchez-Fernandez *et al.* , 2021). Notably, our taxonomic bias analysis revealed a significant under-representation of non-charismatic invertebrate species such as Gastropoda, Malacostraca, Anthozoa, Bivalvia, Polychaeta in the marine environment, as well as Insecta, Gastropoda, Arachnida, Malacostraca, in terrestrial ecosystems. This finding aligns with Troudet et *al.* (2017) who also identified biases against these classes at the global scale. Conversely, vertebrates were well-represented, with the humpback whale (*Megaptera novaeangliae*) being one of the most frequently recorded species. This discrepancy often stems from the aesthetic appeal of certain species, which influences both public interest and scientific focus (Stokes, 2007; Ducarme *et al.* , 2013; De Pinho *et al.* , 2014). Furthermore, studies have effectively shown that visual appeal shapes the perception and prioritisation of species in research and conservation (Langlois *et al.* , 2022). To address these biases and enhance biodiversity inventories in French Polynesia, our dataset can help guide future research priorities, focusing on the underrepresented invertebrates and terrestrial species identified. By addressing these gaps, we can move towards a more comprehensive and balanced understanding of biodiversity, which is crucial for developing effective conservation strategies.

Wallacean shortfall

Another significant gap in our understanding of biodiversity is the incomplete knowledge of species' geographic distribution, also known as the Wallacean shortfall (Lomolino, 2004; Wuest *et al.* , 2020). Despite extensive efforts, biodiversity sampling remains a resource-intensive, time-consuming and costly process, often resulting in substantial gaps in the spatial coverage of species records. Short-term projects frequently fail to capture the full spectrum of species within an assemblage because many species can be cryptic, rare or elusive, ultimately leading to incomplete assessments of global biodiversity patterns. However, these data gaps and uncertainties can be gauged and possibly mitigated through robust modelling approaches (Rocchini *et al.* , 2023). In our study, marine inventory completeness was consistently moderate across French Polynesia's archipelagos, being up to 71% of known species at the regional scale. Furthermore, none

of the species accumulation curves for the archipelagos reached saturation, indicating that species richness predictions require more sampling to improve accuracy. Statistical methods to correct these biases (e.g., Chao *et al.* , 2020) could be used for comparing community assemblages among archipelagos, as has been recently done with woody plants (Kusumoto *et al.* , 2023). Another strategy is to focus on well-documented groups, with complete inventories, enabling the description of their spatial distribution patterns (Shirey *et al.* , 2021).

For terrestrial species, we found that inventory completeness was more variable than that of marine species. The Marquesas archipelago was especially under-surveyed, as only half of the total estimated animal species have been documented. Owing to their geographical isolation and intricate topography, the Marquesas Islands harbour a high level of floral and faunal endemism, with many native and endemic arthropod species probably yet to be discovered (Hembry, 2018). Indeed, many studies have highlighted the uniqueness of this archipelago in terms of species assemblages (Delrieu-Trottin *et al.* , 2015; Galzin*et al.* , 2016) and genetic diversity (Reisser *et al.* , 2019). This biological distinctiveness, combined with the underrepresentation of terrestrial studies compared to marine ones, likely accounts for the discrepancy with other archipelagos, despite the strong interest that scientists have expressed for this biodiversity hotspot (Mittermeier *et al.* , 2005). Prioritising terrestrial biodiversity research in the Marquesas is crucial for establishing reliable comparisons across the land-to-sea continuum in this archipelago. Similarly, a more sustained sampling effort is much needed in the Gambier and Tuamotu Archipelagos, where a significant number of islands remain insufficiently inventoried. This is an urgent call because, while scientific expeditions could potentially discover new species (e.g., Williams *et al.* , 2012), other species could become extinct before being documented (e.g., Zimmerman *et al.* , 2009; Richling & Bouchet, 2013).

Sampling-effort biases can obscure the true spatial distribution of biodiversity, complicating the identification of biodiversity hotspots and the quantification of biodiversity loss (Hughes *et al.* , 2021). This study contributes to addressing this gap by pinpointing overlooked locations of the Polynesia-Micronesia biodiversity hotspot. For instance, Raiatea and Tahaa, which together form the largest lagoon in the Society Archipelago, may host a particularly high level of biodiversity not fully reflected in current GBIF data. This hypothesis is supported by research showing that 26 of the 32 marine sponges recorded across French Polynesia were found in Raiatea-Tahaa (Hall*et al.* , 2013). Similarly, our findings confirmed that the island of Rapa harbours remarkable marine diversity, as evidenced by studies on coral-reef and terrestrial communities, including taxa unique to this island (Meyer & Claridge, 2014; Adjeroud *et al.* , 2016). However, despite being one of the best documented islands in the archipelago, Rapa's inventory completeness remains just behind the global threshold of 80%, suggesting that further sampling efforts are necessary to fully capture this island's biodiversity.

Conservation science is often compelled to assist in decision-making based on limited and incomplete data (Soule, 1985). The spatial heterogeneity in sampling effort that we identified for both marine and terrestrial fauna in French Polynesia is considerable, with up to 70% of islands lacking data on their terrestrial environments. This striking data deficiency was also evidenced by another study using GBIF data to analyse species diversity in a remote region (Bonnet-Lebrun *et al.* , 2023). An additional challenge, particularly for vast and fragmented territories such as French Polynesia, is the need for data at a sufficiently high spatial resolution to capture island-wide variation. We identified 56 islands that either lacked digital data entirely or were poorly documented, likely due to their remoteness. To fill the spatial gaps in biodiversity data for French Polynesia, we recommend that future sampling efforts prioritise these islands, while also considering the disparity in data coverage between marine and terrestrial ecosystems.

The marine-terrestrial sampling bias

Marine and terrestrial ecosystems are often studied separately, partly due to historical, cultural, or practical reasons (Raffaelli *et al.* , 2005; Munguia & Ojanguren, 2015). However, because the land-sea continuum operates as an integrated meta-ecosystem, this research divide hampers our ability to fully understand and effectively protect interconnected ecosystems (Alvarez-Romero *et al.* , 2011; Hughes*et al.* , 2021). Maintaining a healthy land-sea ecosystem is particularly crucial in small-island territories, where biodiversity

is vulnerable to human activities (Russell & Kueffer, 2019; Fernandez-Palacios *et al.* , 2021), and where the wellbeing of local populations heavily depends on local natural resources, especially through fishing and tourism. French Polynesia is no exception, with tourism as its primary economic activity and fish and invertebrates as staples in the local diet (Gillett & Tauati, 2018). Unlike the global trend (Hughes *et al.* , 2021), our data show that French Polynesian biodiversity is better documented in marine ecosystems than in terrestrial ones. This discrepancy is partly due to the focus of scientific research and exploration on marine environments (e.g., the oldest of the two major ecology research units in French Polynesia, the CRIOBE, is entirely focused on marine environments) and to the inaccessibility of the mountainous regions (Gillespie *et al.* , 2008) and seamounts (Hanafi-Portier & Samedi, 2024). The gap is also likely influenced by the huge difference in surface area between land (4,167 km2) and sea ($2.5*10^6$ km2), which may also explain why the marine habitats host 20 times more species than terrestrial ones. While surface-area differences are a factor to consider, our records indicate that the disparity is also driven by a lack of terrestrial data for over 56 islands, compared to just two islands with missing marine data. The observed imbalance in marine versus terrestrial data coverage is not only due to the inherent differences between these ecosystems but also reflects underlying biases in sampling practices, exacerbated by the accessibility factors.

Sampling bias is partly influenced by accessibility factors

The accessibility bias hypothesis posits that more accessible areas tend to be surveyed more frequently than less accessible zones (Zizka*et al.*, 2021). This can significantly impact the global understanding of natural communities (Mangiacotti *et al.*, 2013; Hughes *et al.*, 2021). Our database revealed a pronounced geographic bias in species records, with the most accessible islands (i.e., Tahiti and Moorea in the Society Archipelago, Fakarava in the Tuamotu) being heavily sampled. In contrast, less accessible islands (e.g., Tureia, Napuka and Tenarunga in the Tuamotu, Motu One and Motu Nao in the Marquesas) are poorly documented. However, Rapa Island stands out as an exception, having attracted significant attention from the scientific community due to its hosting of several threatened endemic plant and animal species (Gillespie *et al.*, 2008; Meyer & Claridge, 2014; Adjeroud *et al.*, 2016; Barrett *et al.*, 2021). The sampling bias in Tahiti and Moorea is also likely related to the presence of local research institutions (e.g., CNRS-EPHE-Universite de Perpignan CRIOBE station, Ifremer, IRD, UC-Berkeley Gump station, University of French Polynesia) there. While Tahiti's international airport contributes to the sampling bias observed in the Society Islands, our accessibility bias analysis indicated that the distance from 'airports and ports' was not the main anthropogenic factor explaining the variance in sampling effort at the scale of French Polynesia. Overall, our accessibility bias analysis showed that sampling efforts in both marine and terrestrial datasets are predominantly skewed towards areas near roads and, to a lesser extent, airports/ports. This aggregation pattern around roads is well-documented in the literature for both terrestrial and marine species (Reddy & Davalos, 2003; Hughes*et al.*, 2021), particularly in studies based on citizen-science data (Mair & Ruete, 2016).

Accessibility biases can vary depending on geographic and taxonomic contexts (Mair & Ruete, 2016), highlighting the importance of considering situations in a case-by-case basis. For instance, Freitag et al. (1998) found that records of smaller species in African terrestrial ecosystems were minimally affected by accessibility biases, whereas larger species were disproportionately represented in protected areas. Similarly, Cardoso et al. (2024) identified various accessibility-bias factors for marine species in the western Atlantic Ocean, including proximity to the coastline, research institutions, ports, protected areas, and urban centres. Recognizing and understanding the nuances underlying these various biases is crucial for enhancing the accuracy and comprehensiveness of biodiversity datasets.

Institutional bias in open-source databases

While accessibility factors provide important insights into sampling patterns, they are not the sole source of bias impacting our biodiversity records. Institutional biases, particularly those associated with open-source databases, might also play a crucial role. The unevenness in data contributions often stems from disparities in funding, data-sharing policies, and digitization efforts across different regions and institutions. The soaring popularity of GBIF data worldwide is reflected in our dataset for French Polynesia, where the number of

records per year increased from 10 in 1950 to 1,878 in 2022. We anticipate that the dataset will continue to grow with the engagement of additional contributors, thereby enhancing its reliability (Ivanova & Shashkov, 2021), if institutions continue to adhere to standardisation protocols (Wieczorek *et al.*, 2012). Interestingly, the surge in data during 2006, 2009, and 2011, which constitutes the bulk of the dataset, was driven by the digitization of the French Museum of Natural History dataset (managed by PatriNat) and a major field sampling campaign by Cornell University (USA). The patchiness in data contributions to global open-source databases can be attributed to differences in funding and data-sharing policies across countries, inadequate efforts in digitalising local and national databases, and the sporadic and spatially heterogeneous nature of formal research campaigns (Beck *et al.*, 2014). However, combining GBIF records with national databases can yield more complete inventories, as demonstrated by de Araujo et al. (2022) for Amazonian epiphytes. In the case of French Polynesia, engaging local research institutions, private entities, government agencies and developing a citizen science network to compile and share existing (but often inaccessible) information would significantly reduce biases and strengthen the database. The use and adaptation of existing portals such as FauneFrance (https://www.faune-france.org/) or iNaturalist (https://www.inaturalist.org/) to local flora and fauna could for example be advocated to further centralise and favour the collection and compilation of local naturalist data.

Capitalising from citizen science while reducing biases in GBIF datasets

Addressing biases and shortfalls in GBIF datasets is crucial to ensure their efficiency and accuracy in describing species distribution patterns. Citizen science has been increasingly recognized as an effective method for filling gaps in biodiversity information, especially in areas where formal scientific campaigns are limited or sporadic (Isaac *et al.*, 2014; Amano *et al.*, 2016). In our database for French Polynesia, we observed an increase in species records driven by citizen science initiatives, in agreement with the global trend (Heberling *et al.*, 2021). Indeed, a substantial 23% of records originated from participatory science efforts. While citizen scientists may not always adhere to standard scientific protocols, their contributions provide valuable insights into broader trends, which can then be rigorously analysed. To minimise taxonomic and geographic biases, the involvement of taxonomic experts remains crucial (Maldonado*et al.*, 2015).

Conclusions and perspectives

Centralising biodiversity information from museums, research institutions, and citizen scientists into big-data platforms offers a transformative opportunity for evaluating species biodiversity in understudied regions. These platforms enable comprehensive data analysis, facilitate global collaboration, engage the public in science, and ultimately contribute to more informed conservation strategies and biodiversity management. Our study provides significant insights into the biodiversity patterns of both marine and terrestrial fauna across the vast and fragmented territory of French Polynesia. We found that while marine inventory completeness is relatively high, averaging up to 70% of known species at the regional scale, terrestrial biogeography remains underexplored, particularly in the Marquesas and Gambier Archipelagos. The analysis indicates a notable skew in the data toward specific taxonomic groups, highlighting the urgent need for comprehensive surveys to fill these gaps. Furthermore, our findings underscore the value of citizen science initiatives, which have contributed to 23% of species records in our database, demonstrating their potential to enhance biodiversity knowledge in regions where formal scientific efforts are limited. Overall, this research not only emphasises the richness of biodiversity in French Polynesia but also calls for collaborative efforts to centralise and analyse biodiversity data. These efforts are crucial for aiding in conservation strategies and improving management of the unique ecosystems in the Indo-Pacific region, a global biodiversity hotspot that includes Micronesia, Polynesia, and Fiji (Fan *et al.* , 2023). By providing a reliable, spatially resolved biodiversity dataset, this study lays the foundations for future macroecological research in French Polynesia that will help respond to both fundamental and applied environmental questions.

References

Adjeroud, M., Wallace, C.C., Bosserelle, P., Payri, C., Menou, J. & Pichon, M. (2016) Reefs at the edge: coral community structure around Rapa, southernmost French Polynesia. *Marine Ecology* , **37** , 565–575.

Alvarez-Romero, J.G., Pressey, R.L., Ban, N.C., Vance-Borland, K., Willer, C., Klein, C.J. & Gaines, S.D. (2011) Integrated Land-Sea Conservation Planning: The Missing Links. *Annual Review of Ecology, Evolution, and Systematics* , **42** , 381–409.

Amano, T., Lamming, J.D.L. & Sutherland, W.J. (2016) Spatial Gaps in Global Biodiversity Information and the Role of Citizen Science.*BioScience* , **66** , 393–400.

Andrefouet, S. & Adjeroud, M. (2019) *Chapter 38 - French Polynesia . World Seas: an Environmental Evaluation* , pp. 827–854.

Bacchet, P., Zysman, T. & Lefevre, Y. (2017) *Guide des poissons de Tahiti et ses iles* , Quatrieme edition. Editions Au Vent des Iles, Tahiti (Polynesie Francaise).

Barrett, R.L., Taputuarai, R., Meyer, J.-Y.H., Bruhl, J.J. & Wilson, K.L. (2021) Reassessment of the taxonomic status of Cyperaceae on Rapa Iti, Austral Islands, French Polynesia, with a new combination, Morelotia involuta. *Telopea* , **24** , 171–187.

Beck, J., Boller, M., Erhardt, A. & Schwanghart, W. (2014) Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics* , **19** , 10–15.

Bonnet-Lebrun, A.-S., Sweetlove, M., Griffiths, H.J., Sumner, M., Provoost, P., Raymond, B., Ropert-Coudert, Y. & Van De Putte, A.P. (2023) 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* , **10** , 1150603.

Boutet, M., Gourguet, R. & Letourneux, J. (2020) *Marine Molluscs of French Polynesia / Mollusques Marins de Polynesie Francaise* , Au Vent Des Iles, Tahiti, Polynesie francaise.

Cardoso, M.N.M., Azevedo, F., Dias, A., Sousa de Almeida, A.C., Senna, A.R., Marques, A.C., Rezende, D., Hajdu, E., Alves Pereira Lopes-Filho, E., Bettini Pitombo, F., Moura de Oliveira, G., Doria, J.G., Carraro, J.L., Campos De-Paula, J., Bahia, J., Magalhaes de Araujo, J., Paresque, K., Manzoni Vieira, L., Fernandes, L.M., Santos, L.N., Souza Miranda, L., Lorini, M.L., Klautau, M., Pagliosa, P.R., Braga Clerier, P.H., de Moura, R.B., da Rocha Fortes, R., Neves, R.A.F., Moreira da Rocha, R., Stampar, S.N., Salani, S., Miranda, T.P., Pinheiro, U., Venekey, V. & Oliveira, U. (2024) Causes and effects of sampling bias on marine Western Atlantic biodiversity knowledge. *Diversity and Distributions* , **30** , e13839.

Ceballos, G. & Ehrlich, P.R. (2023) Mutilation of the tree of life via mass extinction of animal genera. *Proceedings of the National Academy of Sciences* , **120** , e2306987120.

Chamberlain, S., Szoecs, E., Foster, Z., Arendsee, Z., Boettiger, C., Ram, K., Bartomeus, I., Baumgartner, J., O'Donnell, J., Oksanen, J., Greshake Tzovaras, B., Marchand, P., Tran, V., Salmon, M., Li, G. & Grenie, M. (2020) taxize: Taxonomic information from around the web. R package.

Chamberlain, S. & Vanhoorne, B. (2023) worrms: World Register of Marine Species (WoRMS) Client. R package.

Chanachai, J., Asamoah, E.F., Maina, J.M., Wilson, P.D., Nipperess, D.A., Esperon-Rodriguez, M. & Beaumont, L.J. (2024) What remains to be discovered: A global assessment of tree species inventory completeness.*Diversity and Distributions* , e13862.

Chao, A. (1987) Estimating the Population Size for Capture-Recapture Data with Unequal Catchability. *Biometrics* , **43** , 783.

Chao, A. (1984) Nonparametric Estimation of the Number of Classes in a Population. *Scandinavian Journal of Statistics* , **11** , 265–270.

Chao, A., Kubota, Y., Zeleny, D., Chiu, C., Li, C., Kusumoto, B., Yasuhara, M., Thorn, S., Wei, C., Costello, M.J. & Colwell, R.K. (2020) Quantifying sample completeness and comparing diversities among assemblages. *Ecological Research* , **35** , 292–314.

Chevillotte, H., Ollier, C. & Meyer J.-Y. (2019) *Base de donnees botaniques Nadeaud de l'Herbier de la Polynesie francaise (PAP). Institut Louis Malarde, Delegation a la Recherche, Papeete, Tahiti* . http://nadeaud.ilm.pf

Clements, J. F., Rasmussen, P.C., Schulenberg, T. S., Iliff, M.J., Fredericks, T.A., Gerbracht, J.A., Lepage, D., Spencer, A., Billerman, S.M., Sullivan, B. L., Smith, M. & Wood C.L. (2024). The eBird/Clements checklist of Birds of the World: v2024. Downloaded from https://www.birds.cornell.edu/clementschecklist/download/

Colwell, R.K. & Coddington, J.A. (1994) Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* ,**345** , 101–118.

Cornwell, W.K., Pearse, W.D., Dalrymple, R.L. & Zanne, A.E. (2019) What we (don't) know about global plant diversity. *Ecography* ,**42** , 1819–1831.

De Araujo, M.L., Quaresma, A.C. & Ramos, F.N. (2022) GBIF information is not enough: national database improves the inventory completeness of Amazonian epiphytes. *Biodiversity and Conservation* , **31** , 2797–2815.

De Pinho, J.R., Grilo, C., Boone, R.B., Galvin, K.A. & Snodgrass, J.G. (2014) Influence of Aesthetic Appreciation of Wildlife Species on Attitudes towards Their Conservation in Kenyan Agropastoralist Communities. *PLoS ONE* , **9** , e88842.

Delrieu-Trottin, E., Williams, J.T., Bacchet, P., Kulbicki, M., Mourier, J., Galzin, R., Lison De Loma, T., Mou-Tham, G., Siu, G. & Planes, S. (2015) Shore fishes of the Marquesas Islands, an updated checklist with new records and new percentage of endemic species. *Check List* ,**11** , 1758.

Delrieu-Trottin, E., Williams, J.T., Pitassy, D., Driskell, A., Hubert, N., Viviani, J., Cribb, T.H., Espiau, B., Galzin, R., Kulbicki, M., Lison De Loma, T., Meyer, C., Mourier, J., Mou-Tham, G., Parravicini, V., Plantard, P., Sasal, P., Siu, G., Tolou, N., Veuille, M., Weigt, L. & Planes, S. (2019) A DNA barcode reference library of French Polynesian shore fishes. *Scientific Data* , **6** , 114.

Deng, C., Daley, T. & Smith, A. (2015) Applications of species accumulation curves in large-scale biological data analysis.*Quantitative Biology* , **3** , 135–144.

Ducarme, F., Luque, G.M. & Courchamp, F. (2013) What are "charismatic species" for conservation biologists? *BioSciences Master Reviews* .

Engemann, K., Enquist, B.J., Sandel, B., Boyle, B., Jorgensen, P.M., Morueta-Holme, N., Peet, R.K., Violle, C. & Svenning, J. (2015) Limited sampling hampers "big data" estimation of species richness in a tropical biodiversity hotspot. *Ecology and Evolution* , **5** , 807–820.

Fan, H., Huang, M., Chen, Y., Zhou, W., Hu, Y. & Wei, F. (2023) Conservation priorities for global marine biodiversity across multiple dimensions. *National Science Review* , **10** , nwac241.

Farley, S.S., Dawson, A., Goring, S.J. & Williams, J.W. (2018) Situating Ecology as a Big-Data Science: Current Advances, Challenges, and Solutions. *BioScience* , **68** , 563–576.

Fegraus, E.H., Andelman, S., Jones, M.B. & Schildhauer, M. (2005) Maximizing the Value of Ecological Data with Structured Metadata: An Introduction to Ecological Metadata Language (EML) and Principles for Metadata Creation. *The Bulletin of the Ecological Society of America* , **86** , 158–168.

Fernandez-Palacios, J.M., Kreft, H., Irl, S.D.H., Norder, S., Ah-Peng, C., Borges, P.A.V., Burns, K.C., de Nascimento, L., Meyer, J.-Y., Montes, E. & Drake, D.R. (2021) Scientists' warning – The outstanding biodiversity of islands is in peril. *Global Ecology and Conservation* **31** , e01847.

Florence, J. (1997) *Flore de la Polynesie francaise* , IRD edition/MNHN, Paris, France.

Florence, J. (2004) *Flore de la Polynesie francaise* , IRD Editions/MNHN, Paris, France.

Freitag, S., Hobson, C., Biggs, H.C. & Van Jaarsveld, A.S. (1998) Testing for potential survey bias: the effect of roads, urban areas and nature reserves on a southern African mammal data set. *Animal Conservation* , **1** , 119–127.

Galzin, R., Duron, S.-D. & Meyer, J.-Y. eds. (2016) *Biodiversite terrestre et marine des iles Marquises, Polynesie francaise* , Societe francaise d'Ichtyologie. Paris.

Galzin, R. & Meyer, J.-Y. (2024) *Les 124 iles de la Polynesie francaise : types, superficies, noms et occupation humaine. Bulletins de la Societe des Etudes Oceaniennes* **362** : 123-136.

Garcia-Rosello, E., Gonzalez-Dacosta, J. & Lobo, J.M. (2023) The biased distribution of existing information on biodiversity hinders its use in conservation, and we need an integrative approach to act urgently.*Biological Conservation* , **283** , 110118.

Gillespie, R.G., Claridge, E.M. & Goodacre, S.L. (2008) Biogeography of the fauna of French Polynesia: diversification within and between a series of hot spot archipelagos. *Philosophical Transactions of the Royal Society B: Biological Sciences* , **363** , 3335–3346.

Gillett, R. & Tauati, M.I. (2018) *Fisheries of the Pacific Islands. Regional and national information. FAO Fisheries and Aquaculture Technical Paper 625* , Food and Agriculture Organization of the United States, Apia, Samoa.

Gorman, C.E., Torsney, A., Gaughran, A., McKeon, C.M., Farrell, C.A., White, C., Donohue, I., Stout, J.C. & Buckley, Y.M. (2023) Reconciling climate action with the need for biodiversity protection, restoration and rehabilitation. *Science of The Total Environment* ,**857** , 159316.

Grenie, M. & Gruson, H. (2022) rtaxref: An R Client for TAXREF the French Taxonomical Reference API. R package.

Guntsch, A., Berendsohn, W.G. & Mergen, P. (2007) The BioCASE Project - a Biological Collections Access Service for Europe.

Hachich, N.F., Bonsall, M.B., Arraut, E.M., Barneche, D.R., Lewinsohn, T.M. & Floeter, S.R. (2015) Island biogeography patterns of marine shallow-water organisms in the Atlantic. *Journal of Biogeography* ,**42** , 1871–1882.

Hall, K.A., Sutcliffe, P.R., Hooper, J.N.A., Alencar, A., Vacelet, J., Pisera, A., Petek, S., Folcher, E., Butscher, J., Orempuller, J., Maihota, N. & Debitus, C. (2013) Affinities of sponges (Porifera) of the Marquesas and Society Islands, French Polynesia. *Pacific Science* , **67** , 493–511.

Hanafi-Portier, M. & Samedi, S. (2024) *Les monts sous-marins de Polynesie Francaise, etat des lieux des connaissances et recommandations scientifiques* , OFB, Office Francais de la Biodiversite ; Museum national d'Histoire Naturelle, Paris, France.

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

Hembry, D.H. (2018) Evolutionary biogeography of the terrestrial biota of the Marquesas Islands, one of the world's remotest archipelagos.*Journal of Biogeography* , **45** , 1713–1726.

Hortal, J., De Bello, F., Diniz-Filho, J.A.F., Lewinsohn, T.M., Lobo, J.M. & Ladle, R.J. (2015) Seven Shortfalls that Beset Large-Scale Knowledge of Biodiversity. *Annual Review of Ecology, Evolution, and Systematics* , **46** , 523–549.

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

IPBES (2019) *Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services* , IPBES secretariat. (ed. by E.S. Brondizio,

J. Settele, S. Diaz, and H.T. Ngo) Bonn, Germany.

Isaac, N.J.B., Van Strien, A.J., August, T.A., De Zeeuw, M.P. & Roy, D.B. (2014) Statistics for citizen science: extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution* ,**5** , 1052–1060.

Ivanova, N.V. & Shashkov, M.P. (2021) The Possibilities of GBIF Data Use in Ecological Research. *Russian Journal of Ecology* ,**52** , 1–8.

Kadmon, R., Farber, O. & Danin, A. (2004) Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models.*Ecological Applications* , **14** , 401–413.

Kays, R., McShea, W.J. & Wikelski, M. (2020) Born-digital biodiversity data: Millions and billions. *Diversity and Distributions* ,**26** , 644–648.

Konig, C., Weigelt, P., Schrader, J., Taylor, A., Kattge, J. & Kreft, H. (2019) Biodiversity data integration—the significance of data resolution and domain. *PLOS Biology* , **17** , e3000183.

Kulbicki, M. (2007) Biogeography of reef fishes of the French territories in the south pacific. *Cybium* , **31** , 275–288.

Kusumoto, B., Chao, A., Eiserhardt, W.L., Svenning, J.-C., Shiono, T. & Kubota, Y. (2023) Occurrence-based diversity estimation reveals macroecological and conservation knowledge gaps for global woody plants.*SCIENCE ADVANCES* .

Langlois, J., Guilhaumon, F., Baletaud, F., Casajus, N., De Almeida Braga, C., Fleure, V., Kulbicki, M., Loiseau, N., Mouillot, D., Renoult, J.P., Stahl, A., Stuart Smith, R.D., Tribot, A.-S. & Mouquet, N. (2022) The aesthetic value of reef fishes is globally mismatched to their conservation priorities. *PLOS Biology* , **20** , e3001640.

Levin, N., Coll, M., Fraschetti, S., Gal, G., Giakoumi, S., Goke, C., Heymans, J., Katsanevakis, S., Mazor, T., Ozturk, B., Rilov, G., Gajewski, J., Steenbeek, J. & Kark, S. (2014) Biodiversity data requirements for systematic conservation planning in the Mediterranean Sea. *Marine Ecology Progress Series* , **508** , 261–281.

Liebherr, J. (2012) The first precinctive Carabidae from Moorea, Society Islands: new Mecyclothorax spp. (Coleoptera) from the summit of Mont Tohiea. *ZooKeys* , **224** , 37–80.

Lim, G.S., Balke, M. & Meier, R. (2012) Determining Species Boundaries in a World Full of Rarity: Singletons, Species Delimitation Methods.*Systematic Biology* , **61** , 165–169.

Lin, H., Caley, M.J. & Sisson, S.A. (2022) Estimating global species richness using symbolic data meta-analysis. *Ecography* ,**e05617** .

Lomolino, M.V. (2004) *Conservation biogeography* . *Frontiers of biogeography: new directions in the geography of nature* (ed. by M.V. Lomolino) and L.R. Heaney), pp. 293–296. Sinauer Associates, Sunderland, MA.

Mair, L. & Ruete, A. (2016) Explaining Spatial Variation in the Recording Effort of Citizen Science Data across Multiple Taxa.*PLOS ONE* , **11** , e0147796.

Maldonado, C., Molina, C.I., Zizka, A., Persson, C., Taylor, C.M., Alban, J., Chilquillo, E., Ronsted, N. & Antonelli, A. (2015) Estimating species diversity and distribution in the era of B ig D ata: to what extent can we trust public databases?*Global Ecology and Biogeography* , **24** , 973–984.

Mangiacotti, M., Scali, S., Sacchi, R., Bassu, L., Nulchis, V. & Corti, C. (2013) Assessing the Spatial Scale Effect of Anthropogenic Factors on Species Distribution. *PLoS ONE* , **8** , e67573.

Meyer, J.-Y. & Claridge, E.M. eds (2014) Terrestrial Biodiversity of the Austral Islands, French Polynesia. Museum national d'Histoire naturelle, Collection Patrimoines Naturels 72, Paris, 144 pp.

Meyer, C., Weigelt, P. & Kreft, H. (2016) Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters* , **19** , 992–1006.

Mittermeier, R.A., Gil, P.R., Hoffmann, M., Pilgrim, J., Brooks, T., Mittermeier, C.G., Lamoreux, J. & Fonseca, G.A.B. (2005) *Hotspots revisited: Earth's Biologically Richest and Most Endangered Terrestrial Ecoregions* , The University of Chicago Press, Chicago, IL.

Montes, E., Lefcheck, J., Guerra Castro, E., Klein, E., De Azevedo Mazzuco, A.C., Bigatti, G., Cordeiro, C., Simoes, N., Macaya, E., Moity, N., Londono-Cruz, E., Helmuth, B., Choi, F., Soto, E., Miloslavich, P. & Muller-Karger, F. (2021) Optimizing Large-Scale Biodiversity Sampling Effort: Toward an Unbalanced Survey Design. *Oceanography* ,**34** , 80–91.

Mora, C., Tittensor, D.P. & Myers, R.A. (2008) The completeness of taxonomic inventories for describing the global diversity and distribution of marine fishes. *Proceedings of the Royal Society B: Biological Sciences* , **275** , 149–155.

Moudry, V. & Devillers, R. (2020) Quality and usability challenges of global marine biodiversity databases: An example for marine mammal data.*Ecological Informatics* , **56** , 101051.

Munguia, P. & Ojanguren, A.F. (2015) Bridging the gap in marine and terrestrial studies. *Ecosphere* , **6** , 1–4.

Newmark, W.D., Jenkins, C.N., Pimm, S.L., McNeally, P.B. & Halley, J.M. (2017) Targeted habitat restoration can reduce extinction rates in fragmented forests. *Proceedings of the National Academy of Sciences* , **114** , 9635–9640.

Oksanen, J., Simpson, G.L., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., Solymos, P., Stevens, M.H.H., Szoecs, E., Wagner, H., Barbour, M., Bedward, M., Bolker, B., Borcard, D., Carvalho, G., Chirico, M., De Caceres, M., Durand, S., Antoniazi Evangelista, H.B., FitzJohn, R., Friendly, M., Furneaux, B., Hannigan, G., Hill, M.O., Lahti, L., McGlinn, D., Ouellette, M.-H., Ribeiro Cunha, E., Smith, T., Stier, A., Ter Braak, C.J.F. & Weedon, J.S. (2024) vegan: Community Ecology Package.

Pebesma, E. & Bivand, R. (2023) *Spatial Data Science: With Applications in R* , 1st ed. Chapman and Hall/CRC.

Pilowsky, J.A., Colwell, R.K., Rahbek, C. & Fordham, D.A. (2022) Process-explicit models reveal the structure and dynamics of biodiversity patterns. *Science Advances* , **8** , eabj2271.

Porch, N., Smith, T.R., & Greig, K. (2020) Five new Pycnomerus Erichson (Coleoptera: Zopheridae: Pycnomerini) from Raivavae, French Polynesia.*Zootaxa* , **4718** (2), 239-250. doi: 10.11646/zootaxa.4718.2.5. PMID: 32230018.

R Core Team (2024) R: A language and environment for statistical computing.

Raffaelli, D., Solan, M. & Webb, T.J. (2005) Do marine and terrestrial ecologists do it differently? *Marine Ecology Progress Series* ,**304** , 283–289.

Ramage, T. (2017) Checklist of the terrestrial and freshwater arthropods of French Polynesia (Chelicerata; Myriapoda; Crustacea; Hexapoda).*Zoosystema* , **39** , 213.

Ramirez, F., Sbragaglia, V., Soacha, K., Coll, M. & Piera, J. (2022) Challenges for Marine Ecological Assessments: Completeness of Findable, Accessible, Interoperable, and Reusable Biodiversity Data in European Seas. *Frontiers in Marine Science* , **8** , 802235.

Reddy, S. & Davalos, L.M. (2003) Geographical sampling bias and its implications for conservation priorities in in Africa. *Journal of Biogeography* , **30** , 1719–1727.

Reisser, C., Lo, C., Schikorski, D., Sham Koua, M., Planes, S. & Ky, C.-L. (2019) Strong genetic isolation of the black-lipped pearl oyster (Pinctada margaritifera) in the Marquesas archipelago (French Polynesia).

Scientific Reports , **9** , 11420.

Richling, I. & Bouchet, P. (2013) Extinct even before scientific recognition: a remarkable radiation of helicinid snails (Helicinidae) on the Gambier Islands, French Polynesia. *Biodiversity and Conservation* , **22** , 2433–2468.

Rocchini, D., Tordoni, E., Marchetto, E., Marcantonio, M., Barbosa, A.M., Bazzichetto, M., Beierkuhnlein, C., Castelnuovo, E., Gatti, R.C., Chiarucci, A., Chieffallo, L., Da Re, D., Di Musciano, M., Foody, G.M., Gabor, L., Garzon-Lopez, C.X., Guisan, A., Hattab, T., Hortal, J., Kunin, W.E., Jordan, F., Lenoir, J., Mirri, S., Moudry, V., Naimi, B., Nowosad, J., Sabatini, F.M., Schweiger, A.H., Šímová, P., Tessarolo, G., Zannini, P. & Malavasi, M. (2023) A quixotic view of spatial bias in modelling the distribution of species and their diversity. *npj Biodiversity* , **2** , 10.

Rocha-Ortega, M., Rodriguez, P. & Córdoba-Aguilar, A. (2021) Geographical, temporal and taxonomic biases in insect GBIF data on biodiversity and extinction. *Ecological Entomology* , **46** , 718–728.

Russell, J.C. & Kueffer, C. (2019) Island Biodiversity in the Anthropocene. *Annual Review of Environment and Resources* ,**44** , 31–60.

Salvat, B. (2009) Dominant benthic mollusks in closed atolls, French Polynesia. *Galaxea, Journal of Coral Reef Studies* , **11** , 197–206.

Salvat, B. & Tröndlé, J. (2017) Biogéographie des mollusques marins de Polynésie française. *Revue d'Écologie (La Terre et La Vie)* ,**72** , 215–257.

Sánchez-Fernández, D., Fox, R., Dennis, R.L.H. & Lobo, J.M. (2021) How complete are insect inventories? An assessment of the british butterfly database highlighting the influence of dynamic distribution shifts on sampling completeness. *Biodiversity and Conservation* ,**30** , 889–902.

Schiesari, L., Grillitsch, B. & Grillitsch, H. (2007) Biogeographic Biases in Research and Their Consequences for Linking Amphibian Declines to Pollution. *Conservation Biology* , **21** , 465–471.

Shirey, V., Belitz, M.W., Barve, V. & Guralnick, R. (2021) A complete inventory of North American butterfly occurrence data: narrowing data gaps, but increasing bias. *Ecography* , **44** , 537–547.

Simberloff, D. (2000) Extinction-proneness of island species-causes and management implications. *Raffles Bulletin of Zoology* ,**48** , 1–9.

Singh, J.S. (2002) The biodiversity crisis: A multifaceted review.*Current Science* , **82** , 638–647.

Siu, G., Bacchet, P., Bernardi, G., Brooks, A.J., Carlot, J., Causse, R., Claudet, J., Clua, E., Delrieu-Trottin, E., Espiau, B., Harmelin-Vivien, M., Keith, P., Lecchini, D., Madi Moussa, R., Parravicini, V., Planes, S., Ponsonnet, C., Randall, J.E., Sasal, P., Taquet, M., Williams, J.T. & Galzin, R. (2017) Shore fishes of French Polynesia. *Cybium* , **41** , 245–278.

Soberón, J., Jiménez, R., Golubov, J. & Koleff, P. (2007) Assessing completeness of biodiversity databases at different spatial scales.*Ecography* , **30** , 152–160.

Soberón, J. & Peterson, T. (2004) Biodiversity informatics: managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* ,**359** , 689–698.

Soulé, M.E. (1985) What is Conservation Biology?: A new synthetic discipline addresses the dynamics and problems of perturbed species, communities, and ecosystems. *BioScience* , **35** , 727–734.

Stephenson, P., Brooks, T.M., Butchart, S.H., Fegraus, E., Geller, G.N., Hoft, R., Hutton, J., Kingston, N., Long, B. & McRae, L. (2017) Priorities for big biodiversity data. *Frontiers in Ecology and the Environment* , **15** , 124–125.

Stokes, D.L. (2007) Things We Like: Human Preferences among Similar Organisms and Implications for Conservation. *Human Ecology* ,**35** , 361–369.

Takashina, N. & Kusumoto, B. (2023) A perspective on biodiversity data and applications for spatio-temporally robust spatial planning for area-based conservation. *Discover Sustainability* , **4** , 1.

Thibault, J.-C. & Cibois, A. (2017) *Birds of eastern Polynesia: a biogeographic atlas* , Lynx, Barcelona, Spain.

Troia, M.J. & McManamay, R.A. (2016) Filling in the GAPS : evaluating completeness and coverage of open-access biodiversity databases in the United States. *Ecology and Evolution* ,**6** , 4654–4669.

Trondle, J. & Boutet, M. (2009) Inventory of Marine Molluscs of French Polynesia. *Atoll Research Bulletin* , 1–87.

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

Underwood, E., Taylor, K. & Tucker, G. (2018) The use of biodiversity data in spatial planning and impact assessment in Europe. *Research Ideas and Outcomes* , **4** , e28045.

Vieira, C., De Clerck, O., De Ramon N'Yeurt, A., D'hondt, S., Millet, L., Kim, M.S., Payri, C. & Zubia, M. (2023) Diversity, systematics and biogeography of French Polynesian *Lobophora* (Dictyotales, Phaeophyceae). *European Journal of Phycology* , **58** , 226–253.

Vieira, C., Steen, F., D'hondt, S., Bafort, Q., Tyberghein, L., Fernandez-Garcia, C., Wysor, B., Tronholm, A., Mattio, L., Payri, C., Kawai, H., Saunders, G., Leliaert, F., Verbruggen, H. & De Clerck, O. (2021) Global biogeography and diversification of a group of brown seaweeds (Phaeophyceae) driven by clade-specific evolutionary processes.*Journal of Biogeography* , **48** , 703–715.

Warren, B.H., Simberloff, D., Ricklefs, R.E., Aguilee, R., Condamine, F.L., Gravel, D., Morlon, H., Mouquet, N., Rosindell, J., Casquet, J., Conti, E., Cornuault, J., Fernandez-Palacios, J.M., Hengl, T., Norder, S.J., Rijsdijk, K.F., Sanmartin, I., Strasberg, D., Triantis, K.A., Valente, L.M., Whittaker, R.J., Gillespie, R.G., Emerson, B.C. & Thebaud, C. (2015) Islands as model systems in ecology and evolution: prospects fifty years after MacArthur-Wilson. *Ecology Letters* ,**18** , 200–217.

Whittaker, R.J., Fernandez-Palacios, J.M., Matthews, T.J., Borregaard, M.K. & Triantis, K.A. (2017) Island biogeography: Taking the long view of nature's laboratories. *Science* , **357** , eaam8326.

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Doring, M., Giovanni, R., Robertson, T. & Vieglais, D. (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLOS ONE* ,**7** , e29715.

Williams, J.T., Delrieu-Trottin, E. & Planes, S. (2012) A new species of Indo-Pacific fish, Canthigaster criobe, with comments on other Canthigaster (Tetraodontiformes: Tetraodontidae) at the Gambier Archipelago. *Zootaxa* , **3523** .

Wuest, R.O., Zimmermann, N.E., Zurell, D., Alexander, J.M., Fritz, S.A., Hof, C., Kreft, H., Normand, S., Cabral, J.S., Szekely, E., Thuiller, W., Wikelski, M. & Karger, D.N. (2020) Macroecology in the age of Big Data – Where to go from here? *Journal of Biogeography* ,**47** , 1–12.

Zimmerman, G., Gargominy, O. & Fontaine, B. (2009) *Quatre especes nouvelles d'Endodontidae (Mollusca, Pulmonata) eteints de Rurutu (Iles Australes, Polynesie francaise) . Zoosystema* , **31** (4) 791–805.

Zizka, A., Antonelli, A. & Silvestro, D. (2021) *sampbias* , a method for quantifying geographic sampling biases in species distribution data. *Ecography* , **44** , 25–32.

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-Araujo, 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. & Antonelli, A. (2020) No one-size-fits-all solution to clean GBIF. *PeerJ* , **8** , e9916.

Figures

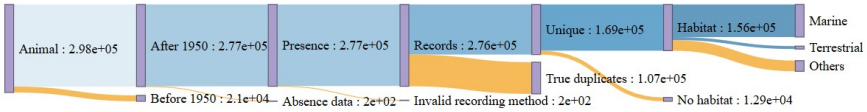

Figure 1. Sankey diagram illustrating the data filtering and quality-control steps. To obtain the final marine, terrestrial and mixed-habitat animal dataset, we removed: records earlier than 1950, absence data, occurrences based on invalid recording methods, true duplicates, and records without habitat information. Data from habitats that are not exclusively marine or terrestrial were excluded from subsequent analyses.

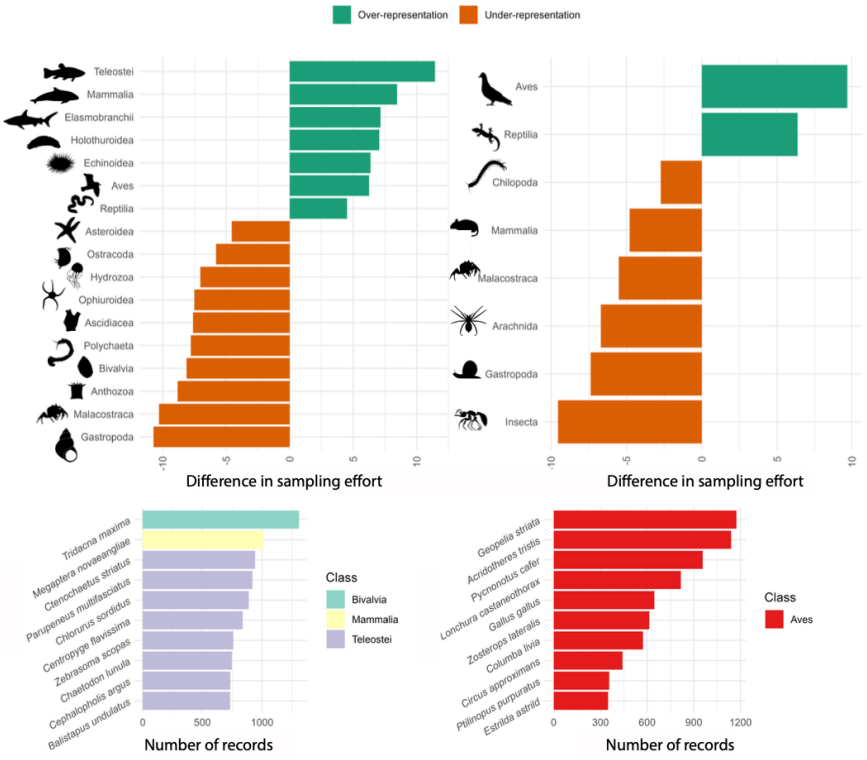

Figure 2. Taxonomic bias assessment. (A) Major class representation in sampling effort (i.e., observed - ideal). Over- and under-representation of each class are illustrated by the green and orange bars respectively. An inverse hyperbolic-sine transformation was used for the x-axis. (B) Number of records for the top 10 most-sampled marine (left) and terrestrial (right) species.

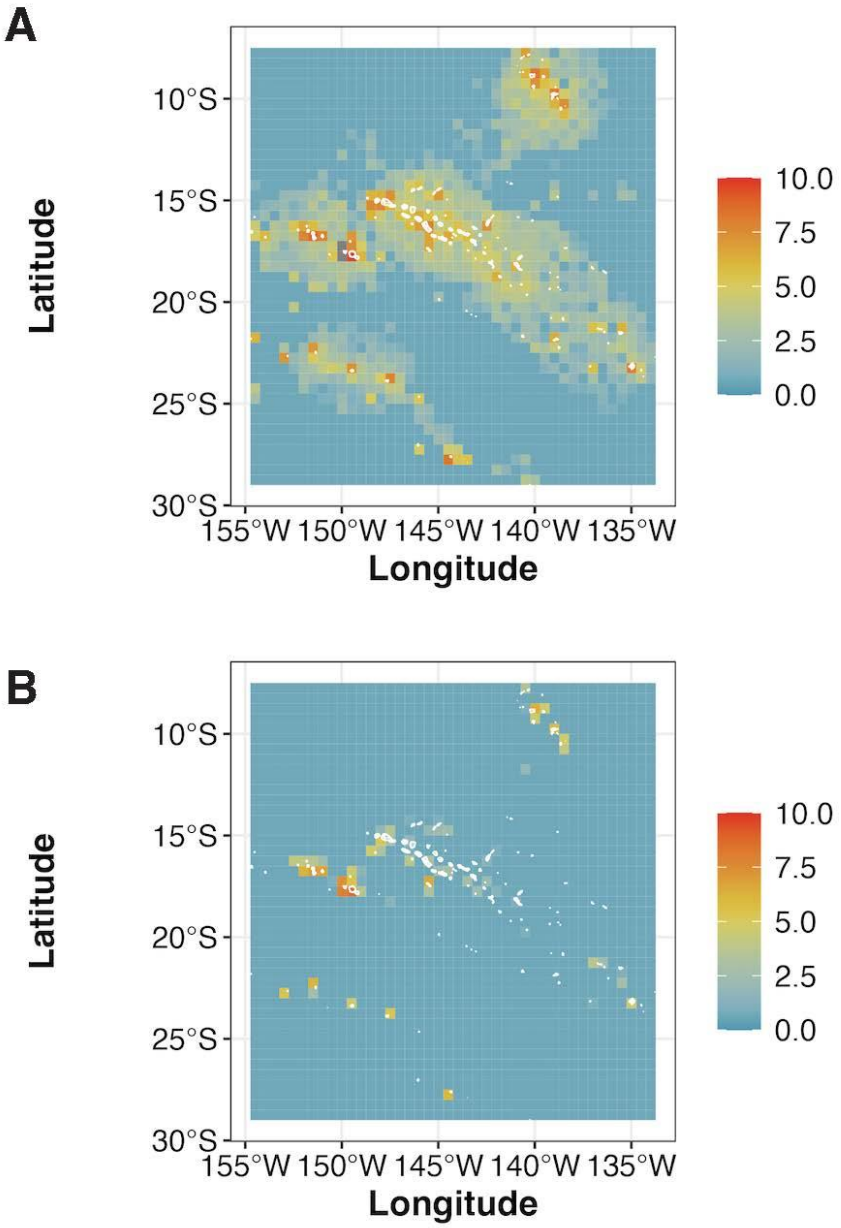

Figure 3. Number of GBIF database records (log-10 transformed) across French Polynesia for marine (**A**) and terrestrial (**B**) species. Blue and red colours indicate a low and high number of occurrences (proxy of sampling rate) respectively, for a 0.05-degree resolution.

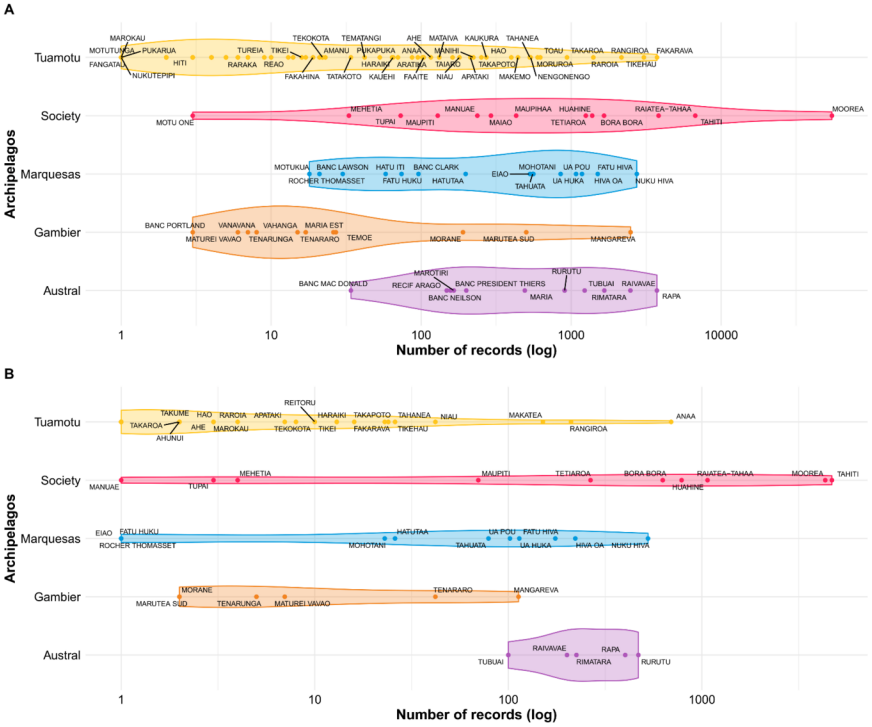

Figure 4 . Number of marine (**A**) and terrestrial (**B**) records per island and archipelago. Islands without records are excluded. The x-axis is presented on a logarithmic scale.

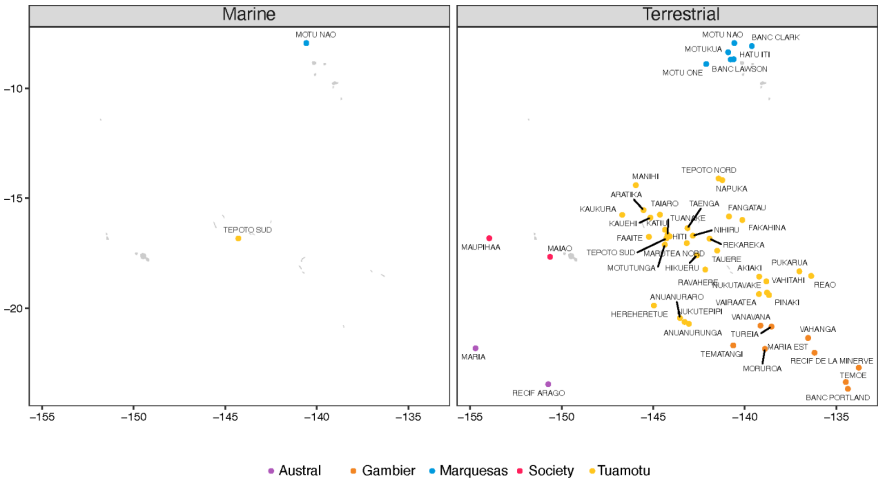

Figure 5 . Map of French Polynesian “neglected” islands (i.e., without species records) in marine (left panel) and terrestrial (right panel) datasets. The islands are color-coded according to their respective archipelagos.

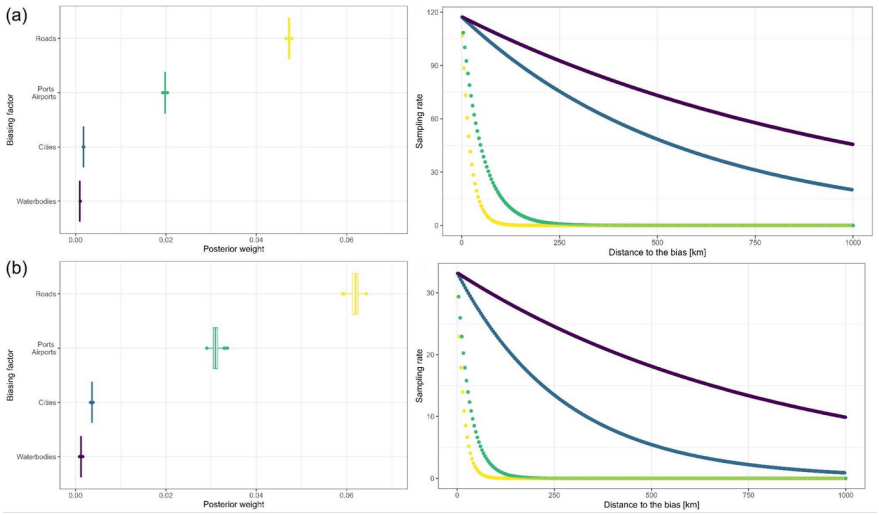

Figure 6. Accessibility bias. Results from the *calculate bias* function (*Sampbias* package) which estimates marine (a) and terrestrial (b) expected occurrences based on accessibility, illustrating the impact of various infrastructure types on species sampling efforts. **(Left)** The posterior weight shows the relative importance of anthropic accessibility factors: -roads, ports/airports, cities, and waterbodies—on sampling efforts in both ecosystems. **(Right)** The sampling rate, i.e., the expected number of occurrences, as a function of distance from accessibility factors is depicted for marine (a) and terrestrial (b) environments.

Table 1 . Archipelago-scale Michaelis-Menten model output parameters based on a records approach (N records as sampling units). Analyses are per archipelago and for each ecosystem. Parameters provided are the number of observed species (*Sobs*), the maximum number of species estimated (*Sest*), the number of cells/records required to capture 50% of the maximum number of species estimated (K), and the inventory completeness in percentage (C).

		Records approach				
Ecosystem	Archipelago	*Sobs*	*Sobs*	K	C	N
Marine	Austral	1,881	2615.52	4036.62	75.59	11,208
	Gambier	788	1023.53	1324.89	76.99	4,012
	Marquesas	1,929	2744.47	4294.65	71.42	9,636
	Society	4,160	5174.48	20207.04	80.41	70800
	Tuamotu	1,602	2027.48	5042.25	79.01	16234
Terrestrial	Austral	283	389.51	561.55	72.65	1,398
	Gambier	49	76.49	98.18	64.06	172
	Marquesas	275	571.69	1402.90	48.10	1,271
	Society	613	728.29	2684.62	84.04	11,868
	Tuamotu	102	120.71	301.91	84.50	1,267

Table 2 . Island-scale Michaelis-Menten model output parameters based on a records approach (N records as sampling units) for the marine ecosystem. Parameters provided are the number of observed species (*Sobs*), the maximum number of species estimated (*Sest*), the number of records required to capture 50% of the maximum number of species estimated (K), and the inventory completeness in percentage (C).

Island	*Sobs*	*Sest*	*K*	*C*	N
AHE	103	379.44	784.20	27.14	116

Island	*Sobs*	*Sest*	*K*	*C*	N
AMANU	17	27.50	155.50	61.81	23
ANAA	112	146.67	429.35	76.36	87
APATAKI	175	483.86	571.29	36.17	213
BANC CLARK	84	270.50	484.44	31.05	96
BANC LAWSON	32	53.13	139.61	60.23	30
BANC MAC DONALD	29	62.29	97.23	46.55	34
BANC NEILSON	85	162.55	161.77	52.29	148
BANC PRESIDENT THIERS	122	250.41	261.07	48.72	200
BORA BORA	635	995.59	2098.04	63.78	1657
EIAO	349	1007.21	1078.27	34.65	532
FAAITE	116	1065.12	1192.85	10.89	104
FAKAHINA	12	16.83	57.23	71.30	19
FAKARAVA	489	587.76	1075.76	83.20	3739
FANGATAUFA	53	126.73	179.44	41.82	70
FATU HIVA	1243	4601.19	6291.63	27.01	1504
FATU HUKU	73	191.57	418.02	38.11	74
HAO	207	514.68	949.77	40.22	399
HARAIKI	69	216.66	296.69	31.85	53
HATU ITI	51	145.37	178.43	35.08	58
HATUTAA	206	761.01	950.68	27.07	198
HEREHERETUE	17	72.91	76.28	23.32	13
HIVA OA	660	1244.07	1676.62	53.05	1181
HUAHINE	849	1591.33	2462.25	53.35	1381
KAUEHI	39	66.16	134.54	58.95	65
KAUKURA	169	361.51	495.58	46.75	272
MAIAO	154	335.14	781.80	45.95	292
MAKATEA	131	221.45	495.28	59.15	95
MAKEMO	199	332.88	421.34	59.78	442
MANGAREVA	734	1048.09	1410.89	70.03	2490
MANIHI	142	328.86	455.88	43.18	223
MANUAE	149	365.42	341.95	40.77	237
MARIA	219	395.93	407.31	55.31	491
MARIA EST	29	91.69	192.61	31.63	26
MAROTIRI	109	209.67	230.64	51.99	165
MARUTEA SUD	180	260.99	303.74	68.97	502
MATAIVA	107	282.83	442.22	37.83	132
MAUPIHAA	264	575.39	776.64	45.88	430
MAUPITI	157	400.49	612.72	39.20	129
MEHETIA	45	64.77	132.15	69.47	33
MOHOTANI	242	385.00	494.80	62.86	560
MOOREA	2759	3279.32	11526.00	84.13	54761
MORANE	168	451.15	595.76	37.24	190
MORUROA	370	890.29	945.25	41.56	598
MOTU ONE (Mar)	135	572.28	713.19	23.59	140
MOTUKUA	19	32.05	33.94	59.28	18
NENGONENGO	176	243.46	269.42	72.29	537
NIAU	155	312.98	460.46	49.52	180
NUKU HIVA	1103	1780.24	2657.89	61.96	2741
PUKAPUKA	70	270.77	253.29	25.85	64
RAIATEA-TAHAA	2598	7575.24	11475.92	34.30	3829

Island	*Sobs*	*Sest*	*K*	*C*	N
RAIVAVAE	634	1013.92	2085.85	62.53	2483
RANGIROA	898	1608.88	2861.20	55.82	2166
RAPA	1027	1514.15	2297.56	67.83	3731
RAROIA	417	592.38	703.65	70.39	1406
REAO	17	37.44	60.20	45.41	14
RECIF ARAGO	103	238.31	379.73	43.22	157
REITORU	45	103.56	307.88	43.45	17
RIMATARA	1158	3642.02	4230.51	31.80	1664
ROCHER THOMASSET	32	61.25	299.27	52.24	21
RURUTU	543	943.70	1341.85	57.54	905
TAHANEA	218	304.68	327.19	71.55	530
TAHITI	1337	1926.10	3205.79	69.40	6722
TAHUATA	410	971.02	1363.12	42.22	554
TAIARO	83	151.51	254.60	54.78	162
TAKAPOTO	176	393.65	474.64	44.71	252
TAKAROA	246	312.78	310.40	78.65	938
TATAKOTO	37	90.91	217.39	40.70	34
TEKOKOTA	45	175.27	362.37	25.67	22
TEMATANGI	40	93.96	164.33	42.57	42
TEMOE	37	90.72	148.68	40.79	27
TENARARO	41	47.76	85.83	85.85	17
TEPOTO NORD	36	50.37	58.03	71.47	21
TETIAROA	445	607.43	1223.13	73.26	1253
TIKEHAU	637	898.12	1529.56	70.93	3053
TIKEI	42	131.02	454.24	32.06	16
TOAU	222	321.87	347.30	68.97	623
TUBUAI	500	792.99	1222.56	63.05	1230
TUPAI	83	271.05	828.04	30.62	73
TUREIA	14	20.29	43.38	69.00	10
UA HUKA	503	1052.07	1509.22	47.81	851
UA POU	613	1213.43	1925.82	50.52	1078
VAHANGA	19	42.51	124.54	44.70	15

Table 3 . Island-scale Michaelis-Menten model output parameters based on a records approach (N records as sampling units) for the terrestrial ecosystem. Parameters provided are the number of observed species (*Sobs*), the maximum number of species estimated (*Vm*), the number of records required to capture 50% of the maximum number of species estimated (*K*), and the inventory completeness in percentage (*C*).

Island	*Sobs*	*Sest*	*K*	*C*	N
ANAA	29	29.78	26.48	97.39	694
UA HUKA	15	16.74	18.12	89.59	114
TENARARO	5	5.61	10.13	89.09	42
HATUTAA	4	4.55	5.79	87.84	26
TAHUATA	13	15.02	12.98	86.53	79
RAIVAVAE	32	37.81	44.44	84.64	201
RIMATARA	50	62.47	59.03	80.04	225
MOHOTANI	6	7.69	7.01	78.06	23
TIKEHAU	7	9.15	9.87	76.50	26
RURUTU	113	151.07	174.10	74.80	470

Island	*Sobs*	*Sest*	*K*	*C*	N
TAKAPOTO	5	6.69	5.77	74.70	16
BORA BORA	54	74.67	275.37	72.32	628
MOOREA	527	746.89	1907.54	70.56	4341
MAKATEA	39	57.69	75.58	67.60	151
UA POU	30	46.17	62.60	64.97	102
TAHITI	393	621.56	2899.46	63.23	4699
TETIAROA	72	115.44	170.00	62.37	266
RANGIROA	62	99.61	131.25	62.24	211
TUBUAI	37	59.94	65.72	61.73	100
RAIATEA-TAHAA	268	441.26	728.54	60.74	1070
MANGAREVA	47	80.13	79.56	58.65	113
HUAHINE	239	409.14	579.88	58.42	786
NIAU	17	31.32	37.64	54.28	42
TAHANEA	9	17.19	23.96	52.35	24
RAPA	182	363.61	406.82	50.05	402
HARAIKI	8	18.29	16.43	43.74	13
NUKU HIVA	133	322.48	779.53	41.24	527
FATU HIVA	90	255.38	326.85	35.24	175
MAUPITI	45	133.07	138.71	33.82	70
HIVA OA	118	362.32	469.02	32.57	222
FAKARAVA	17	67.82	69.42	25.07	23
TIKEI	12	140.60	139.25	8.54	13

Supplementary Material

Table S1. Bias posterior weights (mean ± SD) showing the relative importance of anthropic accessibility factors (roads, ports/airports, cities, and waterbodies) on sampling efforts in both marine and terrestrial ecosystems.

	Marine	Terrestrial
Cities	0.00176 ± 0.00002	0.00361 ± 0.00008
Roads	0.04725 ± 0.00023	0.06189 ± 0.00088
Waterbodies	0.00094 ± 0.00002	0.00121 ± 0.00011
Airports	0.01982 ± 0.00017	0.03098 ± 0.00068

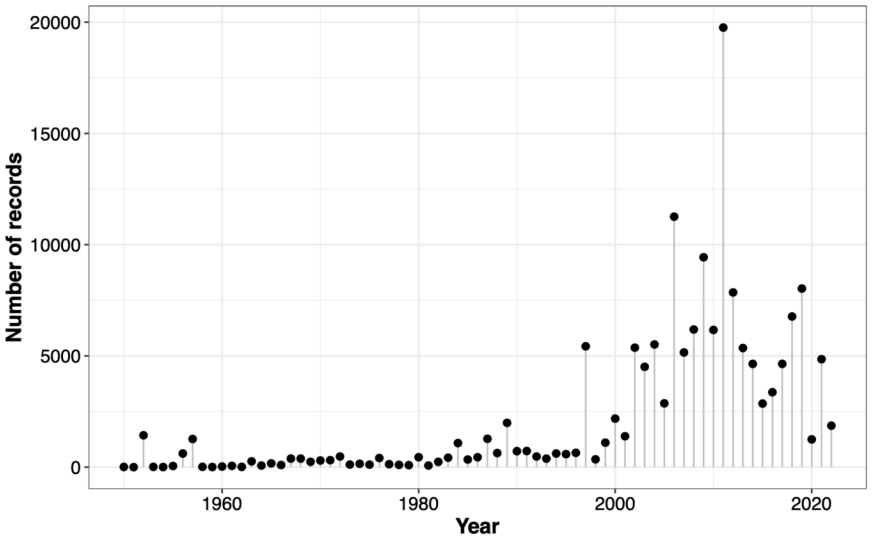

Figure S1. Number of GBIF (Global Biodiversity Information Facility) records for fauna in French Polynesia between 1950 and 2023.

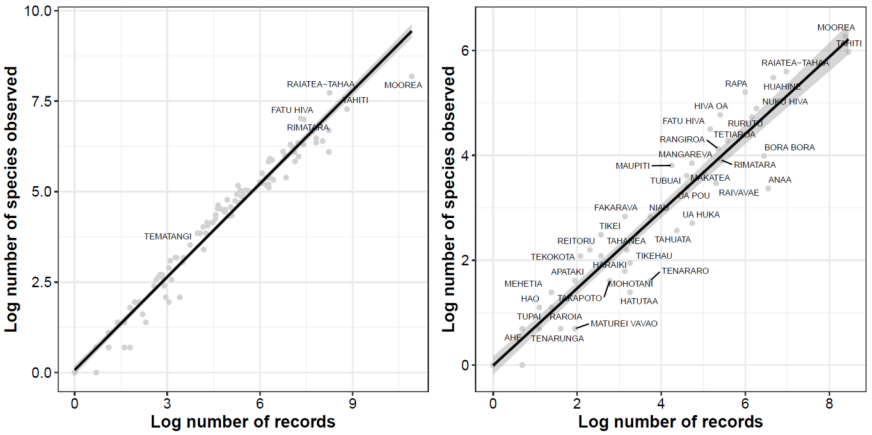

Figure S2. Record bias. Correlation between the number of species per island and records across French Polynesia for marine (left panel) and terrestrial (right panel) habitats.

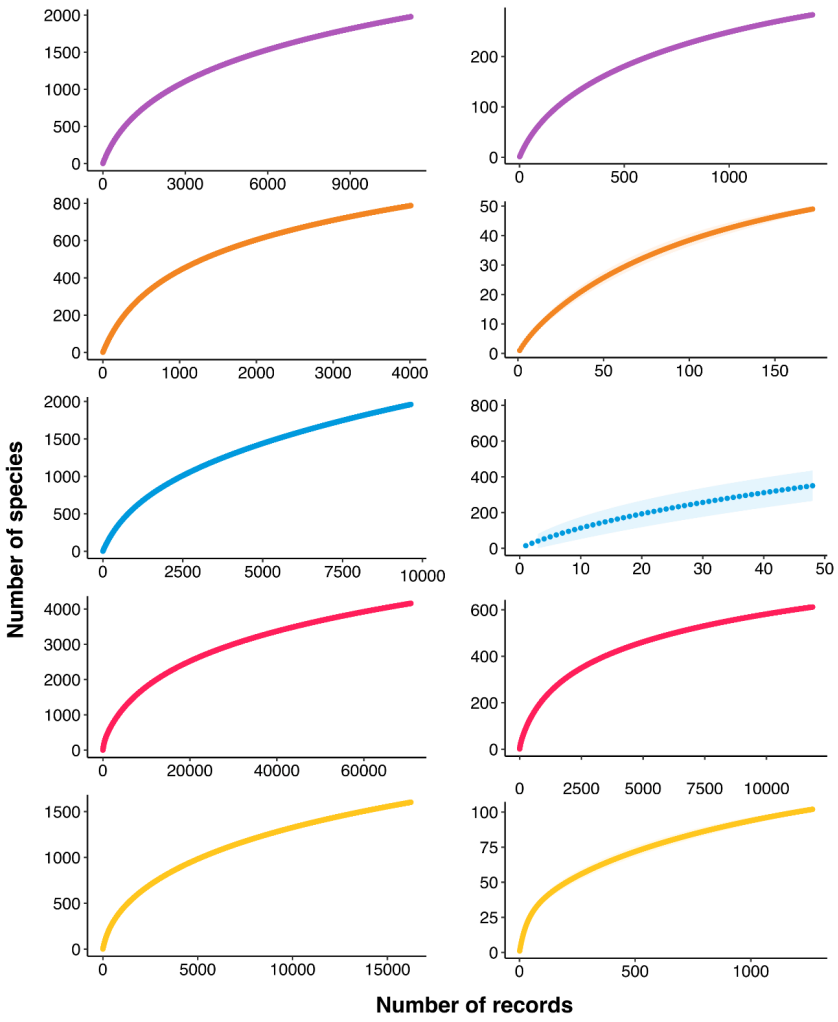

Figure S3 . Species Accumulation Curves produced from resampling. Modelled relationships between the number of species and the umber records for marine (left panels) and terrestrial (right panels) habitats. The lines represent Michaelis-Menten model fits, each archipelago is represented by a specific colour (purple: Austral, orange: Gambier, blue: Marquesas, pink: Society, yellow: Tuamotu) and the shaded zone illustrate 95% confidence intervals. The half-saturation constant (K), representing the area required to capture 50% of the total number of expected species, is shown with vertical dashed lines.

Figure S4. Poorly-documented French Polynesian islands for both marine (left) and terrestrial (right) ecosystems. The number of records for these islands is represented by variable point sizes (from 1 to 10 records). Islands are colour-coded according to their respective archipelagos.