Received: 15 January 2021 | Accepted: 6 April 2021

DOI: 10.1111/2041-210X.13629

RESEARCH ARTICLE

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

bRacatus: A method to estimate the accuracy and biogeographical status of georeferenced biological data

Eduardo Arlé[1,2] | Alexander Zizka[1,3] | Petr Keil[1,4,5] | Marten Winter[1] | Franz Essl[6] | Tiffany Knight[1,7,8] | Patrick Weigelt[9] | Marina Jiménez-Muñoz[1] | Carsten Meyer[1,2,10]

[1]German Centre for Integrative Biodiversity Research (iDiv), Halle-Jena-Leipzig, Leipzig, Germany; [2]Faculty of Biosciences, Pharmacy and Psychology, University of Leipzig, Leipzig, Germany; [3]Naturalis Biodiversity Center, Leiden, The Netherlands; [4]Institute of Computer Science, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany; [5]Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Praha, Czech Republic; [6]Division of Conservation, Vegetation and Landscape Ecology, Department of Botany and Biodiversity Research, University Vienna, Vienna, Austria; [7]Institute of Biology, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany; [8]Department of Community Ecology, Helmholtz Centre for Environmental Research - UFZ, Halle (Saale), Germany; [9]Biodiversity, Macroecology & Biogeography, University of Goettingen, Göttingen, Germany and [10]Institute of Geosciences and Geography, Martin Luther University Halle-Wittenberg, Halle, Germany

Correspondence
Eduardo Arlé
Email: eduardo.arle@idiv.de

Carsten Meyer
Email: carsten.meyer@idiv.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: FZT-118; Volkswagen Foundation, Grant/Award Number: A118199; Universitaet Leipzig

Handling Editor: Nick Isaac

Abstract

1. Georeferenced biological data of species distributions, abundances or traits are critical for ecological and evolutionary research. However, the accuracy (true vs. false records) and biogeographical status (native vs. alien) of individual georeferenced records are often unclear, which limits their use in species distribution modelling, analyses of biodiversity change and other applications.
2. Here, we introduced bRacatus, a new method and R package to estimate a given georeferenced record's probability of being true or false and of corresponding to a native or an alien occurrence. Our framework avoided artificial thresholds of data filtering and instead implemented a probabilistic framework which allowed propagating uncertainties in subsequent analyses. We trained and tested our method on 400 terrestrial species of amphibians, birds, terrestrial mammals and vascular plants from four continents.
3. bRacatus showed good predictive power (mean AUC higher than 0.9; mean RMSE lower than 0.3) for both the accuracy and biogeographical status. Model performance was similar among continents, range sizes and taxa not used in the training. Tests were robust using either range maps or regional checklists of differing levels of data completeness as reference regions.
4. bRacatus was implemented as a user-friendly R package that enabled researchers to assess the accuracy and biogeographical status of species occurrences, population abundances, community composition or any other type of georeferenced biodiversity records. We proposed this method as a routine step in addressing the inherent uncertainty of point observations to promote more accurate ecological inference and predictions.

KEYWORDS
checklists, data quality, data uncertainty, GBIF, invasion biology, range maps, vegetation plot data

1 | INTRODUCTION

One of the fundamental questions in biological sciences is how species are distributed in space and what drives this distribution (Hortal et al., 2012). Understanding geographical patterns in species occurrences, abundances or traits is critical to understanding the current state and trends in biodiversity and guiding biodiversity conservation decision-making (Jetz et al., 2019). Thanks to the mobilisation of records from biological collections, scientific literature and citizen science (Chandler et al., 2017), georeferenced data compilations have rapidly increased in the past two decades. Several databases and networks provide coordinate sets representing point and vegetation-plot records of occurrence, abundance, traits, movement and other biological phenomena assigned to particular species (e.g. GBIF, www.gbif.org; SpeciesLink, splink.cria.org.br; OBIS, obis.org; sPlot, Bruelheide et al., 2019; BioTIME, Dornelas et al., 2018; Movebank, Kranstauber et al., 2011; Open Traits Network, Gallagher et al., 2020; Atlas of Living Australia, www.ala.org.au).

Available georeferenced biological data are limited by data inaccuracy and unclear biogeographical status (Meyer et al., 2016; Troudet et al., 2017). Data inaccuracy commonly occurs because of errors in georeferencing processes (Murphy et al., 2004) and species misidentifications (Scott & Hallam, 2002). These data inaccuracies result in wrong geographical information in species records, and are common problems in GBIF and other biological databases (Maldonado et al., 2015; Zizka et al., 2019). If not addressed, such inaccuracies can distort results of analyses such as species distribution modelling (SDM; Sporbert et al., 2019) or richness assessments (Walther & Moore, 2005) and conservation assessments (Panter et al., 2020; Zizka et al., 2021).

Understanding biogeographical status—whether a record is a native or alien occurrence—is important information for many research questions. For example, information on biogeographical status is necessary for biological invasion science. Knowing their distribution is vital for a dynamic conservation management and policy decisions, given that alien species numbers are increasing world-wide (Seebens et al., 2017) and are one of the main drivers of biodiversity change (Essl et al., 2018; Pagad et al., 2018). Information about biogeographical status is also necessary for many questions focused on native species. For instance, research focused on native biodiversity will overestimate the ranges and environmental niches of species if alien records are erroneously included in analyses (Meyer et al., 2016). Despite its importance, information on the biogeographical status of biological records is largely unavailable. For example, the *establishmentMeans* field in GBIF, meant to provide such information, is filled for only circa 1% of the *c*. 1.4 billion GBIF-facilitated records (accessed via GBIF.org on 01 April 2020; Table S1).

Researchers often clean datasets manually, which requires taxon-specific expertise. This task is time-consuming (Pérez et al., 2015), unfeasible for large datasets and highly subjective and prone to human error (Zizka et al., 2019). Despite the urgent need to address the aforementioned data limitations efficiently, reliable methods and tools to validate the accuracy and to estimate the biogeographical status of records are still lacking. The available alternatives to manual data management rely on threshold-based methods, such as outlier detection or gazetteer-based flagging (e.g. Robertson et al., 2016; Maitner et al., 2018; Zizka et al., 2019, 2021), and miss many important sources of inaccuracies. Moreover, such approaches lead to binary classification of records, keeping remaining uncertainty invisible to subsequent analyses.

Overcoming these limitations can be achieved by harnessing the complementary strengths and weaknesses of different data types. In contrast to occurrence records, species checklists and expert-based range maps provide geographically coarser but usually more complete global coverage of species' ranges. These reference regions are fairly reliable at informing that a species occurs somewhere within those imprecise areas (Hurlbert & Jetz, 2007; König et al., 2019). Moreover, while occurrence records usually lack information on biogeographical status, range maps and checklists often provide this information (Weigelt et al., 2020). The taxonomic coverage of databases providing range maps and checklists is expanding (e.g. IUCN, www.iucn.org; GloNAF, van Kleunen et al., 2019; GIFT, Weigelt et al., 2020; GRIIS, Pagad et al., 2018; GABI, Guénard et al., 2017). Combining and integrating information derived from different data types, such as checklists and range maps with point occurrences and community data, has great potential to improve the knowledge on species biogeography and can help overcome data limitations (Isaac et al., 2020; Jetz et al., 2012; Keil & Chase, 2019; König et al., 2019), but has not yet been developed for detecting and quantifying uncertainties in the accuracy and biogeographical status of biological records.

Here, we address this gap by providing a novel methodological framework (bRacatus) for reliably estimating the geographical and taxonomic accuracy and the biogeographical status of biological point records. 'bracatus' is a Latin word used as an epithet for foreign or barbarian. bRacatus validates georeferenced species records by considering their position relative to reference regions (e.g. derived from range maps or checklists, Figure 1a) in a probabilistic framework. bRacatus is implemented as an R package, including functionalities for data downloaded from GBIF, but also allowing users to provide their own data. By estimating records' probabilities of being accurate and of being native (Figure 1b), bRacatus avoids subjective thresholds of data filtering and instead allows propagating uncertainties in subsequent analyses.

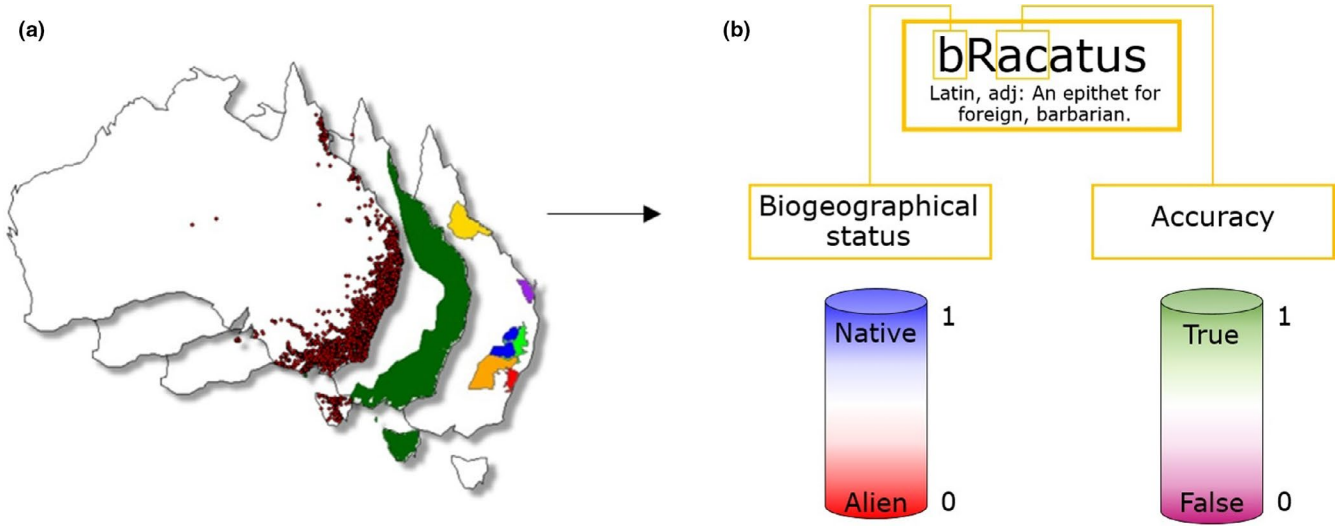

FIGURE 1 Schematic overview showing (a) exemplary types of georeferenced biological data (here: occurrence records) and reference data (here: range maps and checklists) that can be used together by the bRacatus method to probabilistically estimate the accuracy and biogeographical status of the georeferenced records and (b) output values as continuous likelihood measures

2 | WHO CAN BENEFIT FROM BRACATUS?

Scientists working with any type of spatially detailed biological data of uncertain accuracy or biogeographical status can benefit from bRacatus. Our implementation focuses on fine-grain data such as point records of species occurrence, abundance, traits or movement. bRacatus provides support for researchers or resource managers to prepare data for spatial applications, including SDMs, biodiversity change assessments, analysis of community structure, conservation planning and invasive alien species management. bRacatus specifically allows users to automatically assess which records are geographically reliable, and which ones are likely to represent native or alien occurrences. The statistics emerging from bRacatus are record specific; hence, the method can be applied to both single records and larger datasets. bRacatus provides probabilistic values of both accuracy and biogeographical status ranging from 0 to 1 as outputs, allowing users to carry uncertainty to subsequent analyses, for instance, by weighting the contribution of individual records in SDMs (Fletcher et al., 2019) or by using the uncertainty to determine prior probabilities in Bayesian approaches (Winkler, 1967). Alternatively, users can define probability thresholds for record exclusion or discrimination between native and alien occurrences.

3 | METHOD

bRacatus uses binomial GLMs to estimate the accuracy (*acc*) and biogeographical status (*bgs*) of biological records based on their geographical position relative to trusted reference regions, that is, all regions known to form part of the respective species' native or alien distribution. The theoretical foundation of these models is geographical distance decay of similarity (Tobler, 1970). Even without biological assumptions, the expectation is that species' records collected closer to their respective known native or alien ranges are more likely to be accurate and have the same biogeographical status. This is reflected in models considering the distance-decaying signals sent from all cells within reference regions to each grid cell in which a georeferenced record may be located. To estimate the default model parameters implemented in the associated bRacatus R package, we trained and validated the GLMs for predictive performance with a combination of real and simulated species occurrence information. In the following sections, we provide an overview of the data preparation for building these models, and a detailed account of the models' construction, evaluation and validation steps.

3.1 | Method development using empirical and simulated data

We developed bRacatus with occurrence data for 400 species, representing amphibians, birds, terrestrial mammals and vascular plants with 100 species each. We selected the species based on the following three criteria to ensure broad generality and applicability for the models: (a) availability of species range maps for both native and alien ranges (the latter applied for 148 out of the 400 species known to occur outside their native ranges), (b) availability of ≥5 unique GBIF records per species and (c) the species' representation of 14 different terrestrial biomes (Olson et al., 2001), four continents and four range size classes (Table S2). All calculations were performed in R (R Core Team, 2019). We chose 0.5°-grid cell resolution (corresponding to ~25 x 25 km at the equator) as the minimal spatial grain for distinguishing accuracy and biogeographical status, both because we deemed native versus alien status distinctions over shorter distances biologically dubious for most taxa, and to enable fast computation times using the bRacatus R package even for large datasets.

3.1.1 | Georeferenced biological records

We obtained species point-occurrence records from www.gbif.org (GBIF) (GBIF Occurrence Download, 2020). Afterwards, to avoid carrying spatial sampling bias into the models (Anderson, 2012), we thinned the points to a maximum of one record per 0.5°-grid cell (Figure 2a). Subsequently, we classified the points by accuracy for further model validation. Specifically, we classified a species'

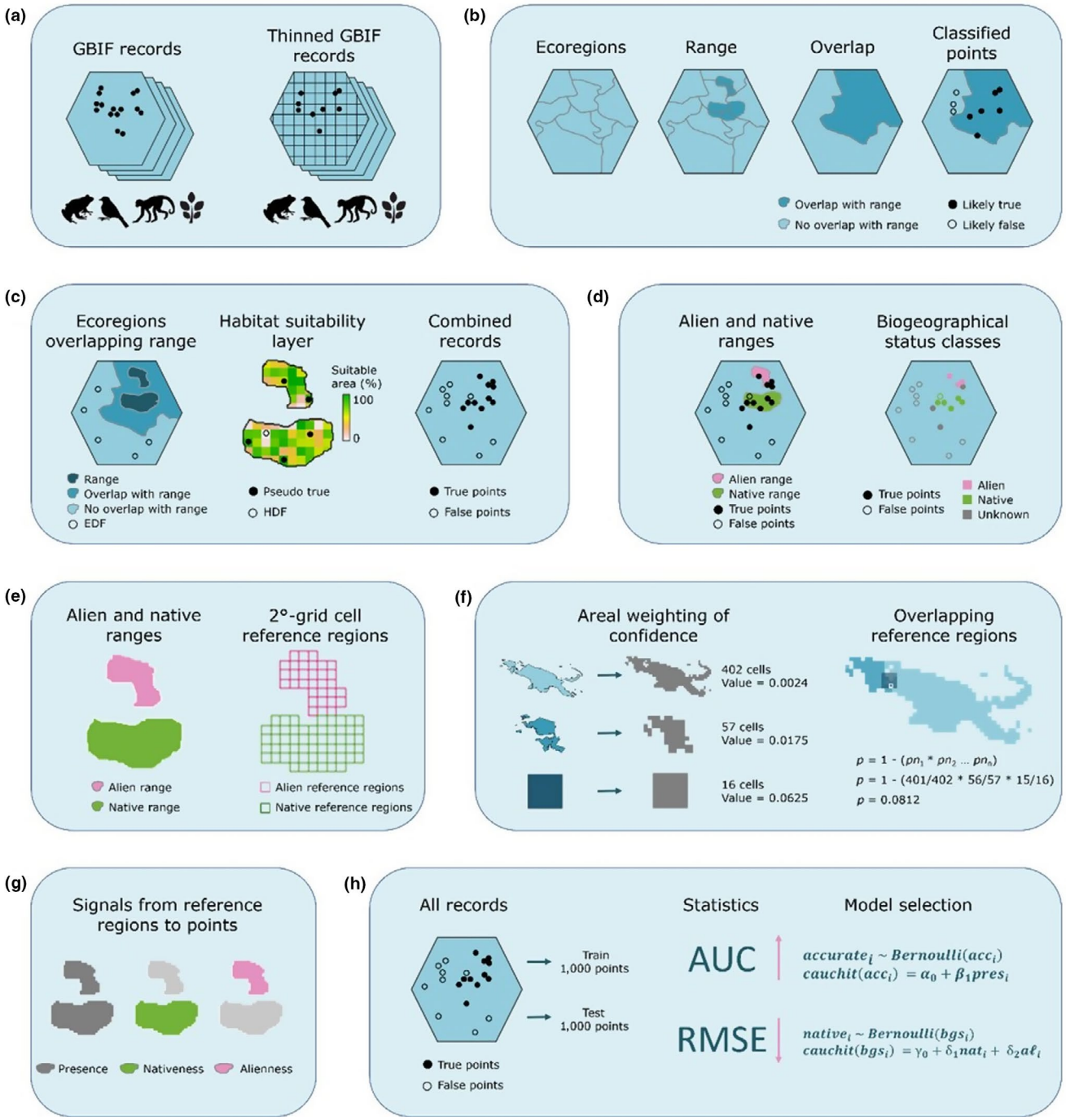

FIGURE 2 Steps in developing the bRacatus methodology a–h: (a) Selection of occurrence data for 400 species distributed across four taxa. (b) Classification of point-occurrence records as likely true or likely false according to intersection with ecoregions overlapping the species range map. (c) Simulation of additional records as easy-to-detect-false records (EDF) distributed in non-range-overlapping ecoregions, pseudo-true records (PT) and hard-to-detect-false records (HDF) within species ranges in pixels with higher and lower habitat suitability respectively. (d) Classification of point records as native, alien or unknown according to species range maps. (e) Transformation of range maps into collections of 2°-grid-cell reference regions. (f) Reference regions gridded to a 0.5°-resolution, a priori confidence of occurrence assessed to each cell and calculation of joint a priori confidence for cells part of multiple overlapping reference regions (see Supporting Information 3 for details). (g) Signals of presence, nativeness and alienness sent from reference regions to each occurrence record. (h) Model selection by AUC and RMSE values

records falling within the terrestrial ecoregions (Olson et al., 2001) overlapping their respective ranges as 'likely true', and as 'likely false' when falling outside those limits (Figure 2b). In addition to these range-validated GBIF data, we simulated three categories of records ('easy-to-detect false' (EDF), 'hard-to-detect false' (HDF) and 'pseudo-true' (PT) occurrences), to compensate for remaining sampling bias while mimicking common data errors. To simulate likely locations of EDF and HDF records, we considered species-specific habitat suitability, considering species' expert-based habitat preferences and elevational limits (Figure 2c; see Supporting Information 1 for details).

In order to evaluate the performance of the biogeographical models, we classified the combined GBIF and simulated records according to their presumed biogeographical status, that is, those falling within their 1°-buffered native ranges as 'likely native', those within their 1°-buffered alien ranges as 'likely alien' and those falling outside of both buffers as 'unknown status' (Figure 2d).

The selection and simulation of point-occurrence records resulted in 377,796 unique point occurrences for all species combined (Supporting Information 1, Table S2). Note that the records' binary classifications of both accuracy and biogeographical status according to the above protocol had the sole aim of assessing model performance, and are independent of the general bRacatus method.

3.1.2 | Reference regions

The definition of species reference regions may be based on expert-drawn range maps or on regional checklists. We derived the reference regions used for training and validating our models from range map data for birds (BirdLife International, 2019), amphibians, terrestrial mammals and vascular plants (IUCN, 2019). Since range maps are not available for most taxonomic groups, less precise and potentially more incomplete regional checklists are often the only option for reference regions. Therefore, we additionally validated our models with such checklist-based reference regions. To control the checklist regions' degrees of imprecision and incompleteness, we simulated regional checklists of different realistic levels of geographical precision and completeness, by overlaying the range maps with checklist-region boundaries stored in the GIFT database (a comprehensive resource of regional vascular plant species distributions based on checklists and floras; Weigelt et al., 2020; see Section 3.2.2).

We estimated the model parameters from records' positions relative to all available reference regions, with each region sending an independent distance-decaying signal to all records. The bRacatus method accommodates for the large heterogeneity in sizes and shapes of reference regions and the spatial grain (resolution) at which species occupancy can be reliably inferred from these data types, by distributing this signal over the region's entire area.

Specifically, range maps only delimit the outer range boundaries within which species are expected to be present (Jetz et al., 2012), but do not indicate which precise areas are occupied. However, it has been shown that they can estimate species occupancies at coarse grains of circa 2° (Hurlbert & Jetz, 2007). Hence, each 2°-grid cell overlaying a range map is considered an independent reference region that sends its own signal (Figure 2e). Unlike in expert-based range maps, the sizes and shapes of the politically defined sampling units of regional checklists are not indicative of the extents of occurrence of the listed species, but merely confirm that those species were recorded at least once somewhere within those regions. Without further information, a priori confidence that a species was recorded in any particular subregion within those regional boundaries (or, in the case of range maps, within a 2°-grid cell) is thus inversely proportional to the share of the larger region represented by the subregion.

bRacatus represents location uncertainty for three categories: presence (*pres*), nativeness (*nat*) and alienness (*al*). To do this, it calculates an area weighting of confidence for each uncertainty category by rasterising reference regions to 0.5°-resolution and assigning each raster cell overlapping the region a value calculated by

$$apc_h = \frac{1}{n_h}, \tag{1}$$

where apc_h is the a priori confidence that the species in category $h \in \{pres, nat, al\}$ has been detected in each cell and n_h is the number of cells covering the region of that category (Figure 2f).

All reference regions used for model training during our testing and default parametrisation of bRacatus were 2°-grid cells artificially derived from range maps (Figure 2e), having exactly the same number of cells and no overlap. For each species, we generated three raster layers from the checklists, carrying the information on presence, nativeness and alienness. Due to this area weighting of confidence, the bRacatus framework can probabilistically validate fine-scale biological records without needing to assume that reference regions can indicate species occupancy at fine scales. It does assume, however, that the *broad-scale* evidence on species' native or alien presences provided by range maps and checklists is credible (for tests of the method's sensitivity to violating this assumption, see Supporting Information 3).

3.1.3 | Signals sent from reference regions to points

Each record receives distance-decaying signals from three raster layers, presence, nativeness and alienness respectively (Figure 2g). To calculate the strength of the signals reaching each record, bRacatus first uses pre-computed pairwise geographical distances between all 0.5°-cells globally ($d_{i,j}$). It then normalises $d_{i,j}$ to obtain a proximity index ($td_{i,j}$) between all pairs of cells as

$$td_{i,j} = 1 - \frac{d_{i,j}}{\max(d)}, \tag{2}$$

where $\max(d)$ is the maximum value of all $d_{i,j}$. While developing bRacatus, we identified the distance-decay function that would lead to the

best predictive power by comparing 13 alternative exponential decays given by

$$dd_{i,j} = td_{i,j}^{2^m}, \quad (3)$$

where $dd_{i,j}$ represents the decayed proximity indices and m ranges from 0 to 12. bRacatus calculates the index for all species by computing the signals sent from all cells within every independent reference region to each individual record according to the formula

$$Vpt_{h,i} = \frac{\sum_j apc_{h,j} dd_{i,j}}{\max(Vpt)}, \quad (4)$$

where $Vpt_{h,i}$ is the value assigned to a record that will further be used in the model, $apc_{h,j}$ is the a priori confidence of each cell within the species range, $dd_{i,j}$ is the distance-decayed proximity index between the cell i under consideration and the cell j sending the signal and $\max(Vpt)$ is the maximum value obtained in the signal calculation. The denominator ranges from 0 to 1, making these values comparable among species. The rationale behind this transformation is that the individual record receiving the highest signal has the highest probability of being accurate, or the highest probability of representing the correct biogeographical status among all the species' records. During bRacatus development, we repeated this process for the aforementioned 12 exponential decay functions and a linear decay function ($m = 0$), thus producing 13 alternative versions per records of each index: *pres*, *nat* and *al*.

3.2 | Model construction, evaluation and validation steps

We developed our models based on binomial GLMs for both the accuracy and the biogeographical status analyses (Supporting Information 2). Accuracy models use the *pres* index as the only predictor. Biogeographical status models use both *nat* and *al* indices as predictors, with species having only native range reference regions receiving an *al* score of 0 for all points. The model output is continuous probabilities, ranging from 'most-likely false' (0) to 'most-likely true' (1) for the accuracy analysis and from 'most-likely alien' (0) to 'most-likely native' (1) for the biogeographical status analysis (Figure 1b).

We conducted in-sample and out-of-sample predictive tests to verify the models' performance and ensure their broad applicability, testing for potential biases in model performance towards certain taxa, range sizes or continents, and whether our models can be extrapolated to other taxa, range size bins and continents than those that were used for model training (Supporting Information 2). The model selection relied on two metrics—the area under the receiver operating characteristic curve (AUC) and the root mean squared error (RMSE; Figure 2h). The AUC ranges from 0 to 1 and informs about the model's ability to separate classes in a prediction (Swets, 1988). For the biogeographical status analyses, we applied a variation of AUC calculation, the multiclass receiver operating characteristic (ROC), which allows analysing multiclass data (Wandishin & Mullen, 2009). The RMSE indicates how close the predictions are to the actual values (Chai & Draxler, 2014). High AUC and low RMSE values indicate better performing models. To evaluate the models considering both the metrics simultaneously, we calculated the Euclidean distance from the AUC and RMSE obtained in each model to the ideal values (1 and 0 respectively) of these metrics (Draisma et al., 2014).

3.2.1 | Model selection

Initially, we tested 260,000 models (see Supporting Information 2 for details) to identify the best-performing distance-decay function, not considering other parameters. Subsequently, we combined three variations of the signal calculations (*Vpt*), four link functions and two other covariates (average distance to other occurrence points and background sampling effort), resulting in 460,000 different models. We trained and tested all models with all combinations among the aforementioned variables. We deliberately avoided variables based on biological grounded relationships, such as environmental distances, to ensure greatest-possible applicability in downstream analyses without risks of circularities (see Supporting Information 2 for further details).

The three variations of *Vpt* aimed to ensure that records within reference regions' boundaries are assessed with higher values than those in neighbouring areas. Thus, we performed an analysis on how the models would perform with stronger signals sent from the very cell where a point is located, computing two extra versions of the indices by multiplying the signal sent from the cell where each point is located by 10 (sig10) and by 100 (sig100). For the accuracy analysis, only points assessed as 'likely true' underwent signal variation, as the HDF points seeded within the range would magnify the noise they represent. We trained all models with different link functions: logit, probit, cauchit and cloglog. The following equations depict the models using a cauchit link function, which generally performed best:

$$\text{accurate}_i \sim \text{Bernoulli}(acc_i), \quad (5 - \text{accuracy})$$

$$\text{cauchit}(acc_i) = \alpha_0 + \beta_1 pres_i;$$

$$\text{native}_i \sim \text{Bernoulli}(bgs_i), \quad (6 - \text{biogeographical status})$$

$$\text{cauchit}(bgs_i) = \gamma_0 + \delta_1 nat_i + \delta_2 al_i.$$

For each record $i, i \in 1{:}r$, where r is the number of records; acc_i is estimated accuracy of each record; α_0 is the intercept of Equation 5; β_1 is the slope associated with the covariate $pres_i$, which represents the presence index of each record; bgs_i is estimated biogeographical status of each record; γ_0 is the intercept of Equation 6; δ_1 is the slope associated with the covariate nat_i, which represents the nativeness index of

each record; and δ_2 is the slope associated with the covariate *al*, which represents the alienness index of each point.

We further included two other covariates in the models: (a) average distance to the closest five occurrence points, to account for the extent to which records are geographical outliers; and (b) density of records in the same taxonomic order of the focus species, to represent background sampling intensity (Supporting Information 3).

3.2.2 | Sensitivity tests with reference regions based on regional checklists

To simulate realistic data and evaluate our models' performance when working with checklists instead of range maps, we used regions' shapefiles derived from the GIFT database. We created artificial checklists for each species, gridded and combined them to account for possible overlaps, according to

$$p = 1 - (pn_1 * pn_2 \ldots pn_k), \quad (7)$$

where p is the final a priori confidence in the cell and pn_i is the confidence of no occurrence informed by each checklist, i, represented in the cell (Figure 2f). We then applied the accuracy and the biogeographical models using the reference regions derived from the simulated checklists and calculated the evaluation metrics (Figure 2h). Checklist data are not necessarily complete and differ in geographical precision. Thus, we ran additional sensitivity tests to evaluate the models' performance under different levels of checklist data incompleteness and different region sizes (Supporting Information 3).

4 | RESULTS

4.1 | Model construction, evaluation and validation steps

We found a strong difference in the performance of the 13 different distance-decay functions. Models using predictors calculated with $m = 5$ in Equation 3 performed best in 31.2% of the tests for the accuracy and in 70.2% of the tests for the biogeographical status analysis (measured by AUC and RMSE - Table S3).

The tests including other variables indicated that the signal variation sig10 performed best in terms of AUC and RMSE. A cauchit link function consistently yielded the best models. Including the two covariates representing the extent to which records are geographical outliers and background sampling intensity in the accuracy models did not show consistent improvements, so we excluded both from the final algorithm (Supporting Information 2).

Our tests showed high predictive power both for models trained and tested with the whole dataset, and when testing for different taxa, range size bins and continents, and both for in-sample and for out-of-sample tests. The biogeographical status analysis produced overall better results (mean AUC ≈ 0.98, mean RMSE ≈ 0.15) than the accuracy analysis (mean AUC ≈ 0.91, mean RMSE ≈ 0.32; Figure 3). Our tests also point to marginal difference in model performance for different taxa, continents or species range size quartiles (Figure S3).

4.2 | Tests with checklists

The tests performed with the simulated checklist data indicated that the bRacatus method had good predictive power even when no range maps but only checklists were available as sources of reference regions. As with range maps, the biogeographical status analysis produced overall better results (mean AUC ≈ 0.96, mean RMSE ≈ 0.18) than the accuracy analysis (mean AUC ≈ 0.90, mean RMSE ≈ 0.39). More incomplete checklist collections and data composed solely or mostly by vast regions, such as large countries or subcontinents, decreased the accuracy of bRacatus-based estimations of record accuracy and biogeographical status (see Supporting Information 3, Figures S1 and S2). Overall, these sensitivity tests indicate that our models perform satisfactorily even when provided with reference regions containing reasonably large geographical uncertainties.

5 | DISCUSSION

Our results show that bRacatus is a reliable method for automatically validating georeferenced biological datasets. We demonstrated that the taxonomic and geographical accuracy (true vs. false) and biogeographical status (alien vs. native) of biological field records can be reliably predicted by using coarse-grain distribution data as geographical reference. The variation in model performance among taxa, range size classes and continents was marginal, indicating a high degree of generality and transferability of the presented methodological framework (Figure 3). Our models were robust to using checklists instead of range maps as reference regions. Analyses considering information gaps in the checklists showed that average-sized regions corresponding roughly to country level (100,000 to 1,000,000 km^2) still produced satisfactory results (Supporting Information 3). These results broaden the applicability of bRacatus. Users can manually provide any checklist data, or, for plants, benefit from a function in the bRacatus R package (giftRegions) that automatically accesses and inputs species-level checklists available via the GIFT database (Weigelt et al., 2020). As analogous data-lookup services may eventually become available for other taxonomic groups (e.g. based on IUCN range maps), we intend to include new functionalities to retrieve the corresponding reference regions in the future versions of the package.

bRacatus's accuracy and biogeographical status estimations are less reliable if the reference regions are themselves highly inaccurate or extremely imprecise or incomplete, for instance, because they were derived from checklists for (sub)continental regions or with large gaps in coverage. In such cases, bRacatus outputs can be corrected with further specialist curation, which is facilitated by the graphic visualisation provided in the accompanying R package.

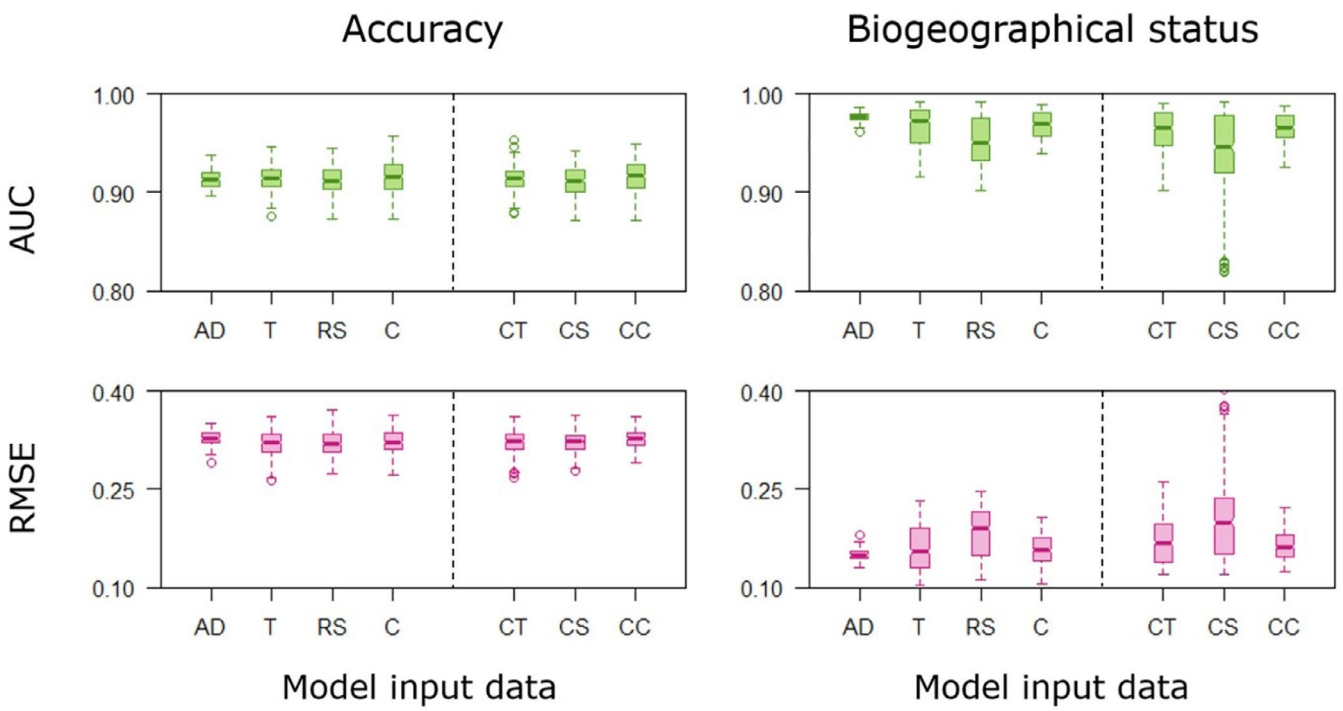

FIGURE 3 Box plots showing the models' performance in estimating the accuracy and the biogeographical status of individual records under different in-sample (AD, T, RS, C) and out-of-sample (CT, CS, CC) tests. AD = All data, both the train and the test data come from the complete data pool; T = Taxa, train and test data from the same taxon; RS = Range size, train and test data from the same range size class; C = Continent, train and test data from the same continent; CT = Cross-taxa, train data from one taxon and test data from all other taxa; CS = Cross-range size, train data from one range size class and test data from all other range size classes; CC = Cross-continent, train data from one continent and test data from all other continents. Lower and upper box boundaries represent the 1st and 3rd quartiles, respectively, line inside box represents the median, lower and upper error lines represent the 10th and 90th percentiles respectively. Circles represent data falling outside the 10th and 90th percentiles

Small-scale geolocation errors that would still be geographically plausible, such as rounded coordinates or small derivations from real localities, will probably not be detected by the method. This could potentially lead to erroneous interpretations of the environmental context in which the species have been recorded, particularly in regions of high environmentally heterogeneity over short distances such as along tropical mountain slopes. Similarly, taxonomic misidentifications between sympatric species may lead to erroneous assessments of either species' records as accurate due to the spatial proximity between both species' reference regions. Similarly, bRacatus outputs may not reliably discriminate biogeographical status in some cases where alien and native ranges are geographically very close or nested, requiring further specialist curation (Figure 4).

By providing estimates ranging from 0 to 1, bRacatus avoids arbitrary filtering thresholds and instead allows propagating continuous uncertainties in subsequent analyses. For example, the uncertainties of individual records could be picked up by SDMs or other methods by using weights or by probabilistically sampling from alternative record interpretations. By enabling the use of all available data while explicitly accounting for individual records' uncertainties, the bRacatus framework helps to address the common trade-off in ecological studies between data coverage and data uncertainty (Meyer et al., 2016). Such effective use of all available information is arguably an imperative for sound ecological inference and applications in the majority of the most biodiverse, tropical regions, which tend to be particularly data scarce (Meyer et al., 2015).

Although currently designed for assess spatially detailed biological records (e.g. point data), the current methodology could be easily adapted for coarser data. For example, the accuracy model could potentially be used to validate small-scale species inventories based on better curated data, and the biogeographical status model could be used to estimate whether a region of unknown status represents part of the species' alien or the native range. Such applications could further contribute to the cross-information and mutual quality enhancement of diverse data types such as checklists, protected-area inventories or transect data.

Further developments of the bRacatus R package may add possibilities to include additional information to further improve the accuracy and/or biogeographical status estimations. For example, species-specific dispersal-related traits could be incorporated in the models, as well as alternative distance matrices reflecting environmental similarities, economic trade links or natural dispersal barriers among regions. Such extensions could serve more specific biogeographical applications such as analyses of niche shifts between native and alien ranges or invasion risk assessments. However, we caution that such extensions will also impose trade-offs due to

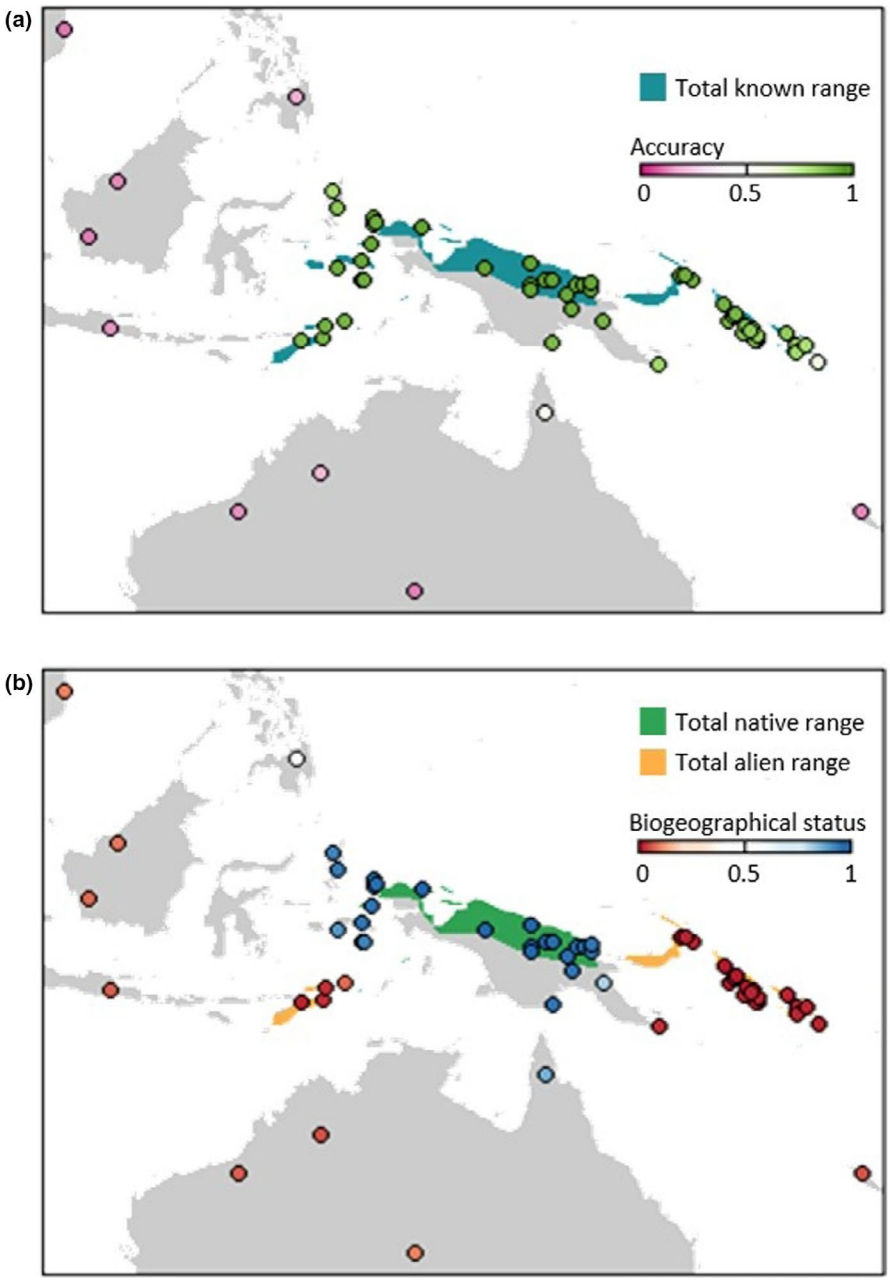

FIGURE 4 Example of typical bRacatus output for point-occurrence records for the species *Phalanger orientalis*. (a) Estimated accuracy values ranging from most-likely false (0) to most-likely true (1) occurrences, produced with function *plotAccuracy*. (b) Estimated biogeographical status values ranging from most-likely alien (0) to most-likely native (1) occurrences, produced with function *plotBiogeo*. Shaded regions indicate species total (blue), native (green) and alien (orange) range maps

potentially increased risks of circular reasoning. A key strength of the presented, simpler implementation of the bRacatus method is that it yields robust estimations of accuracy and biogeographical status by solely relying on the highly general theory of geographical distance decay of similarity (Tobler, 1970) without requiring that any additional ecological assumptions be 'built into' the records' assessments.

ACKNOWLEDGEMENTS

E.A., F.E., T.K. C.M. and M.W. acknowledge funding for this work through iDiv's Flexpool mechanism (FZT-118, DFG). C.M. acknowledges funding from the Volkswagen Foundation through a Freigeist Fellowship (A118199). A.Z. is thankful for funding by iDiv via the German Research Foundation (FZT-118, DGF), specifically through sDiv, the Synthesis Centre of iDiv. All the authors acknowledge the contribution of Holger Kreft to this work by facilitating our access to the GIFT database, and Ruben Remelgado for the technical support to produce our models. The authors have no conflict of interest. They also acknowledge the valuable comments and suggestions made by the anonymous reviewers of MEE that helped us to improve this manuscript. Open Access funding enabled and organized by Projekt DEAL.

AUTHORS' CONTRIBUTIONS

E.A., F.E., P.K., T.K., C.M., P.W. and M.W. designed this study; E.A., P.K. and M.J.-M. designed and implemented the algorithms; E.A. and A.Z. built the R package; E.A. wrote the manuscript, with major contributions from T.K. and C.M. All the authors contributed extensively in two rounds of revisions, read and approved the final version of the manuscript.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13629.

DATA AVAILABILITY STATEMENT

The code of bRacatus is open source and is available on Zenodo under a GPL (>=2) licence (https://zenodo.org/record/4698910#.YHs9OjxXb0, https://doi.org/10.5281/zenodo.4698910) (Arlé, 2021). The package is available on CRAN (https://cran.r-project.org/web/packages/bRacatus/index.html).

ORCID

Eduardo Arlé https://orcid.org/0000-0003-4776-6161
Alexander Zizka https://orcid.org/0000-0002-1680-9192
Petr Keil https://orcid.org/0000-0003-3017-1858
Marten Winter https://orcid.org/0000-0002-9593-7300
Franz Essl https://orcid.org/0000-0001-8253-2112
Tiffany Knight https://orcid.org/0000-0003-0318-1567
Patrick Weigelt https://orcid.org/0000-0002-2485-3708
Marina Jiménez-Muñoz https://orcid.org/0000-0002-4543-2929
Carsten Meyer https://orcid.org/0000-0003-3927-5856

REFERENCES

Anderson, R. P. (2012). Harnessing the world's biodiversity data: Promise and peril in ecological niche modeling of species distributions. *Annals of the New York Academy of Sciences*, *1260*(1), 66–80. https://doi.org/10.1111/j.1749-6632.2011.06440.x

Arlé, E. (2021). Data from: EduardoArle/bRacatus_code: bRacatus code (Version v.1.0.0). *Zenodo*, http://doi.org/10.5281/zenodo.4698910

BirdLife. (2019). *BirdLife International*. Retrieved from https://www.birdlife.org/datazone/info/spcdownload

Bruelheide, H., Dengler, J., Jiménez-Alfaro, B., Purschke, O., Hennekens, S. M., Chytrý, M., Pillar, V. D., Jansen, F., Kattge, J., Sandel, B., Aubin, I., Biurrun, I., Field, R., Haider, S., Jandt, U., Lenoir, J., Peet, R. K., Peyre, G., Sabatini, F. M., … Zverev, A. (2019). sPlot – A new tool for global vegetation analyses. *Journal of Vegetation Science*, *30*(2), 161–186. https://doi.org/10.1111/jvs.12710

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Chandler, M., See, L., Copas, K., Bonde, A. M. Z., López, B. C., Danielsen, F., Legind, J. K., Masinde, S., Miller-Rushing, A. J., Newman, G., Rosemartin, A., & Turak, E. (2017). Contribution of citizen science towards international biodiversity monitoring. *Biological Conservation*, *213*, 280–294. https://doi.org/10.1016/j.biocon.2016.09.004

Dornelas, M., Antão, L. H., Moyes, F., Bates, A. E., Magurran, A. E., Adam, D., Akhmetzhanova, A. A., Appeltans, W., Arcos, J. M., Arnold, H., Ayyappan, N., Badihi, G., Baird, A. H., Barbosa, M., Barreto, T. E., Bässler, C., Bellgrove, A., Belmaker, J., Benedetti-Cecchi, L., … Zettler, M. L. (2018). BioTIME: A database of biodiversity time series for the Anthropocene. *Global Ecology and Biogeography*, *27*(7), 760–786. https://doi.org/10.1111/geb.12729

Draisma, J., Horobeț, E., Ottaviani, G., Sturmfels, B., & Thomas, R. (2014). The euclidean distance degree. *Proceedings of the 2014 Symposium on Symbolic-Numeric Computation, SNC 2014*. https://doi.org/10.1145/2631948.2631951

Essl, F., Bacher, S., Genovesi, P., Hulme, P. E., Jeschke, J. M., Katsanevakis, S., Kowarik, I., Kühn, I., Pyšek, P., Rabitsch, W., Schindler, S., van Kleunen, M., Vilà, M., Wilson, J. R. U., & Richardson, D. M. (2018). Which taxa are alien? Criteria, applications, and uncertainties. *BioScience*, *68*(7), 496–509. https://doi.org/10.1093/biosci/biy057

Fletcher, R. J., Hefley, T. J., Robertson, E. P., Zuckerberg, B., McCleery, R. A., & Dorazio, R. M. (2019). A practical guide for combining data to model species distributions. *Ecology*, *100*(6), e02710. https://doi.org/10.1002/ecy.2710

Gallagher, R. V., Falster, D. S., Maitner, B. S., Salguero-Gómez, R., Vandvik, V., Pearse, W. D., Schneider, F. D., Kattge, J., Poelen, J. H., Madin, J. S., Ankenbrand, M. J., Penone, C., Feng, X., Adams, V. M., Alroy, J., Andrew, S. C., Balk, M. A., Bland, L. M., Boyle, B. L., … Enquist, B. J. (2020). Open Science principles for accelerating trait-based science across the Tree of Life. *Nature Ecology & Evolution*, *4*(3), 294–303. https://doi.org/10.1038/s41559-020-1109-6

GBIF Occurrence Download. (01 April 2020). https://doi.org/10.15468/dl.3pa1nh

Guénard, B., Weiser, M., Gómez, K., Narula, N., & Economo, E. (2017). The Global Ant Biodiversity Informatics (GABI) database: Synthesizing data on the geographic distribution of ant species (Hymenoptera: Formicidae). *Myrmecological News*, *24*, 83–89.

Hortal, J., Lobo, J., & Jimenez-Valverde, A. (2012). Basic questions in biogeography and the (lack of) simplicity of species distributions: Putting species distribution models in the right place. *Natureza & Conservação Revista Brasileira de Conservação da Natureza*, *10*, 108–118. https://doi.org/10.4322/natcon.2012.029

Hurlbert, A. H., & Jetz, W. (2007). Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(33), 13384–13389. https://doi.org/10.1073/pnas.0704469104

Isaac, N. J. B., Jarzyna, M. A., Keil, P., Dambly, L. I., Boersch-Supan, P. H., Browning, E., Freeman, S. N., Golding, N., Guillera-Arroita, G., Henrys, P. A., Jarvis, S., Lahoz-Monfort, J., Pangel, J., Pescott, O. L., Schmucki, R., Simmonds, E. G., & O'Hara, R. B. (2020). Data integration for large-scale models of species distributions. *Trends in Ecology & Evolution*, *35*(1), 56–67. https://doi.org/10.1016/j.tree.2019.08.006

IUCN. (2019). *The IUCN Red List of threatened species (version 2019–3)*. Retrieved from http://www.iucnredlist.org

Jetz, W., McGeoch, M. A., Guralnick, R., Ferrier, S., Beck, J., Costello, M. J., Fernandez, M., Geller, G. N., Keil, P., Merow, C., Meyer, C., Muller-Karger, F. E., Pereira, H. M., Regan, E. C., Schmeller, D. S., & Turak, E. (2019). Essential biodiversity variables for mapping and monitoring species populations. *Nature Ecology & Evolution*, *3*(4), 539–551. https://doi.org/10.1038/s41559-019-0826-1

Jetz, W., McPherson, J. M., & Guralnick, R. P. (2012). Integrating biodiversity distribution knowledge: Toward a global map of life. *Trends in Ecology & Evolution*, *27*(3), 151–159. https://doi.org/10.1016/j.tree.2011.09.007

Keil, P., & Chase, J. (2019). Global patterns and drivers of tree diversity integrated across a continuum of spatial grains. *Nature Ecology & Evolution*, *3*(3), 390–399. https://doi.org/10.1038/s41559-019-0799-0

König, 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*(3), 1–16. https://doi.org/10.1371/journal.pbio.3000183

Kranstauber, B., Cameron, A., Weinzerl, R., Fountain, T., Tilak, S., Wikelski, M., & Kays, R. (2011). The Movebank data model for animal tracking. *Environmental Modelling & Software*, *26*(6), 834–835. https://doi.org/10.1016/j.envsoft.2010.12.005

Maitner, B. S., Boyle, B., Casler, N., Condit, R., Donoghue-II, J., Durán, S. M., Guaderrama, D., Hinchliff, C. E., Jørgensen, P. M., Kraft, N. J. B., McGill, B., Merow, C., Morueta-Holme, N., Peet, R. K., Sandel, B.,

Schildhauer, M., Smith, S. A., Svenning, J. C., Thiers, B., … Enquist, B. J. (2018). The BIEN R package: A tool to access the Botanical Information and Ecology Network (BIEN) database. *Methods in Ecology and Evolution*, *9*(2), 373–379. https://doi.org/10.1111/2041-210X.12861

Maldonado, C., Molina, C. I., Zizka, A., Persson, C., Taylor, C. M., Albán, J., Chilquillo, E., Rønsted, N., & Antonelli, A. (2015). Species diversity and distribution in the era of Big Data. *Global Ecology and Biogeography*, *24*, 973–984. https://doi.org/10.1111/geb.12326

Meyer, C., Kreft, H., Guralnick, R., & Jetz, W. (2015). Global priorities for an effective information basis of biodiversity distributions. *Nature Communications*, *6*(1), 1–8. https://doi.org/10.1038/ncomms9221

Meyer, C., Weigelt, P., & Kreft, H. (2016). Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters*, *19*(8), 992–1006. https://doi.org/10.1111/ele.12624

Murphy, P. C., Guralnick, R. P., Glaubitz, R., Neufeld, D., & Ryan, J. A. (2004). Georeferencing of museum collections: A review of problems and automated tools, and the methodology developed by the Mountain and Plains Spatio-Temporal Database-Informatics Initiative (Mapstedi). *Zenodo*, https://doi.org/10.5281/zenodo.59792

Olson, D. M., Dinerstein, E., Wikramanayake, E. D., Burgess, N. D., Powell, G. V. N., Underwood, E. C., D'amico, J. A., Itoua, I., Strand, H. E., Morrison, J. C., Loucks, C. J., Allnutt, T. F., Ricketts, T. H., Kura, Y., Lamoreux, J. F., Wettengel, W. W., Hedao, P., & Kassem, K. R. (2001). Terrestrial ecoregions of the world: A new map of life on earth: A new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. *BioScience*, *51*(11), 933–938. https://doi.org/10.1641/0006-3568(2001)051[0933:TEOTWA[2.0.CO;2

Pagad, S., Genovesi, P., Carnevali, L., Schigel, D., & McGeoch, M. A. (2018). Introducing the global register of introduced and invasive species. *Scientific Data*, *5*, 170–202. https://doi.org/10.1038/sdata.2017.202

Panter, C. T., Clegg, R. L., Moat, J., Bachman, S. P., Klitgård, B. B., & White, R. L. (2020). To clean or not to clean: Cleaning open-source data improves extinction risk assessments for threatened plant species. *Conservation Science and Practice*, *2*, e311. https://doi.org/10.1111/csp2.311

Pérez, J., Iturbide, E., Olivares, V., Hidalgo, M., Martínez, A., & Almanza, N. (2015). A data preparation methodology in data mining applied to mortality population databases. *Journal of Medical Systems*, *39*(11), 152. https://doi.org/10.1007/s10916-015-0312-5

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

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

Scott, W. A., & Hallam, C. J. (2002). Assessing species misidentification rates through quality assurance of vegetation monitoring. *Plant Ecology*, *165*(1), 101–115. https://doi.org/10.1023/A:1021441331839

Seebens, H., Blackburn, T. M., Dyer, E. E., Genovesi, P., Hulme, P. E., Jeschke, J. M., Pagad, S., Pyšek, P., Winter, M., Arianoutsou, M., Bacher, S., Blasius, B., Brundu, G., Capinha, C., Celesti-Grapow, L., Dawson, W., Dullinger, S., Fuentes, N., Jäger, H., … Essl, F. (2017). No saturation in the accumulation of alien species worldwide. *Nature Communications*, *8*, 14435. https://doi.org/10.1038/ncomms14435

Sporbert, M., Bruelheide, H., Seidler, G., Keil, P., Jandt, U., Austrheim, G., Biurrun, I., Campos, J. A., Čarni, A., Chytrý, M., Csiky, J., De Bie, E., Dengler, J., Golub, V., Grytnes, J.-A., Indreica, A., Jansen, F., Jiroušek, M., Lenoir, J., … Welk, E. (2019). Assessing sampling coverage of species distribution in biodiversity databases. *Journal of Vegetation Science*, *30*(4), 620–632. https://doi.org/10.1111/jvs.12763

Swets, J. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*(4857), 1285–1293. https://doi.org/10.1126/science.3287615

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, *46*, 234–240. https://doi.org/10.2307/143141

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

van Kleunen, M., Pyšek, P., Dawson, W., Essl, F., Kreft, H., Pergl, J., Weigelt, P., Stein, A., Dullinger, S., König, C., Lenzner, B., Maurel, N., Moser, D., Seebens, H., Kartesz, J., Nishino, M., Aleksanyan, A., Ansong, M., Antonova, L. A., … Winter, M. (2019). The Global Naturalized Alien Flora (GloNAF) database. *Ecology*, *100*(1), e02542. https://doi.org/10.1002/ecy.2542

Walther, B. A., & Moore, J. L. (2005). The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography*, *28*(6), 815–829. https://doi.org/10.1111/j.2005.0906-7590.04112.x

Wandishin, M. S., & Mullen, S. J. (2009). Multiclass ROC analysis. *Weather and Forecasting*, *24*(2), 530–547. https://doi.org/10.1175/2008WAF2222119.1

Weigelt, P., König, C., & Kreft, H. (2020). GIFT – A Global Inventory of Floras and Traits for macroecology and biogeography. *Journal of Biogeography*, *47*, 16–43. https://doi.org/10.1111/jbi.13623

Winkler, R. L. (1967). The assessment of prior distributions in Bayesian analysis. *Journal of the American Statistical Association*, *62*(319), 776–800. https://doi.org/10.1080/01621459.1967.10500894

Zizka, A., Antonelli, A., & Silvestro, D. (2021). sampbias, a method for quantifying geographic sampling biases in species distribution data. *Ecography*, *44*, 25–32. https://doi.org/10.1111/ecog.05102

Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Ritter, C. D., Edler, D., Farooq, H., Herdean, A., Ariza, M., Scharn, R., Svantesson, S., Wengström, N., Zizka, V., & Antonelli, A. (2019). CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. *Methods in Ecology and Evolution*, *10*(5), 744–751. https://doi.org/10.1111/2041-210X.13152

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Arlé E, Zizka A, Keil P, et al. bRacatus: A method to estimate the accuracy and biogeographical status of georeferenced biological data. *Methods Ecol Evol*. 2021;12:1609–1619. https://doi.org/10.1111/2041-210X.13629