Acta Oecologica 129 (2025) 104137

Contents lists available at ScienceDirect

Acta Oecologica

journal homepage: www.elsevier.com/locate/actoec

Validating ranges, georeferenced records, and species distribution models, for vertebrate and plant species globally

Brice B. Hanberry

USDA Forest Service, Rocky Mountain Research Station, Rapid City, SD, 57702, USA

ARTICLE INFO

Keywords:
Area
Commission
Error
IUCN
Sample sizes
MDD
Omission

ABSTRACT

Species range maps and georeferenced records of species occurrences supply complementary information about where species occur. I applied range maps and georeferenced records to examine intersection between range maps and the most certain and less certain locations of georeferenced records (i.e., coordinates can have an associated field of coordinate uncertainty in meters), intersection between an alternative range map product and georeferenced records, and areas of range maps relative to areas predicted by species distribution models from georeferenced records or range maps, with varying sample sizes. For all combined 3853 vertebrate and plant species, 25 % of species had at least 96.8 % intersection between georeferenced records of distance uncertainties <1 km and range maps, 50 % of species had at least 91.7 % intersection, and 75 % of species had at least 81.0 % intersection. The 25 % of species range maps with less than 80 % intersection may need reevaluation for mislocation, with omission and commission error. Similar results occurred for intersection with ranges of georeferenced records for distance uncertainties greater than 1 km or unknown, relative to uncertainties <1 km, and the Mammal Diversity Database (MDD) range maps, relative to International Union for Conservation of Nature (IUCN) range maps. Species distribution models based on small sample sizes overpredicted total area of occurrence, with about half of the predicted area outside of range maps. Range maps and georeferenced records triangulated, regardless of distance uncertainties. Range maps can help validate species distribution models or indicate issues of area overprediction and poor representation due to small samples, whereas georeferenced records detected potential anomalies in range maps.

1. Introduction

Ranges delineate the geographic distributions of species. The International Union for Conservation of Nature (IUCN, 2024) has produced the most comprehensive set of accessible range maps for taxa including vertebrates and plants. Various groups have produced competing versions of range maps, including the Mammal Diversity Database (e.g., Marsh et al., 2022). Regardless of specific accuracy, range maps are generally accepted coarse realizations of species distributions (Mathias et al., 2004; Hurlbert and White, 2005; McPherson and Jetz, 2007; Jetz et al., 2008; Ficetola et al., 2014). Objective reality is captured through approximate representation (Denzin, 2012).

Given relative temporal correspondence (e.g., extant ranges and recent species observations), range maps and georeferenced records of species occurrences offer complementary information about species distributions, with range maps containing commission error, due to monolithic delineation, whereas georeferenced records contain omission error (Hurlbert and Jetz, 2007; Ficetola et al., 2014). Range maps typically are more comprehensive than sampling coverage by observed species occurrences, but range maps may include locations where species do not occur or rarely occur (Gaston and Fuller, 2009; Graham and Hijmans, 2006; McPherson and Jetz, 2007; Hawkins et al., 2008; Ficetola et al., 2014). Georeferenced records represent precise yet uneven sampling of species because continents are undersampled, with unsampled areas (Graham and Hijmans, 2006; McPherson and Jetz, 2007; Hughes et al., 2024; Daru and Rodriguez, 2023; Garcia-Rosello et al., 2023). Some research has supported that both data types appear equally suitable for use, without relying on both, albeit with recognition that complete coverage may require both types for wide-ranging and poorly sampled species (Alhajeri and Fourcade, 2019; Arenas-Castro et al., 2022).

Georeferenced records are likely to be correctly located, due to fine resolution of point coordinates (Graham and Hijmans, 2006; McPherson and Jetz, 2007). Nonetheless, georeferenced records are considered to have locational errors, to the extent a standard practice is to remove samples that potentially may have errors (Zizka et al., 2020).

E-mail address: brice.hanberry@usda.gov.

https://doi.org/10.1016/j.actao.2025.104137
Received 8 August 2025; Received in revised form 6 November 2025; Accepted 9 November 2025
Available online 15 November 2025

Undeniably, museum records georeferenced years after collection based on imprecise locations are not expected to be accurate (Feeley and Silman, 2010), and records need to be filtered to the time frame of interest, given species extirpations and range changes over time (Faurby and Araújo, 2018; WWF et al., 2018). However, no other mapped product overall has greater locational accuracy relative to georeferenced coordinates, particularly georeferenced coordinates with an associated distance uncertainty. That is, 1:110 million land layers are less accurate (Chapman and Wieczorek, 2020), whereas range maps may have distance uncertainties of 200 km (Hurlbert and Jetz, 2007). A distance uncertainty <1 km means that coordinates for the records are offset by no more than 1 km. While a maximum offset distance of 1 km may appear to be distant, mammals and birds are mobile and may move away from human observers, which also may be in motion. Furthermore, distance uncertainties are the maximum and may be increased to match the standards of the reporting institution.

Species distribution models use georeferenced records, predictor variables, and a modeling algorithm to fill in geographical space with environmental conditions similar to occurrence points, albeit limited by available samples of conditions (Barry and Elith, 2006; Thuiller and Münkemüller, 2010). Certainly, most species distribution models have not made use of range maps, and associated expert input, as references to confirm representativeness of species models (Merow et al., 2017). Indeed, Herkt et al. (2017) concluded that IUCN range maps underestimated the complete geographical ranges of bat species in Africa relative to species distribution models, for which they used small sample sizes (mean of 60, median of 21). They recommended that IUCN range maps should not be used where species are not sampled or undersampled. However, due to small sample sizes, species distribution models can be unrepresentative of species ranges, particularly by overpredicting areas of occurrence (Mitchell et al., 2017; Gábor et al., 2020). Likewise, other modeling decisions such as truncated samples and choices in threshold between area classified as presence and absence can affect predicted areas in species distribution models. That is, it may be more probable that poorly sampled areas will result in small samples and species distribution models that contain commission (and omission) error rather than that range maps contain omission error (Mitchell et al., 2017; Gábor et al., 2020).

Given that currently available range maps and georeferenced records seemingly are derived from different data streams and processes, at least with divergence over time, co-occurrences in location can help confirm where species occur, while also validating the accuracies of range maps and georeferenced records by research triangulation of different data sources (Ficetola et al., 2014; Merow et al., 2017; Rotenberry and Balasubramaniam, 2020; Bans-Akutey and Tiimub, 2021). Alhajeri and Fourcade (2019) determined that more than half of 1315 rodent species had more than 75 % of their georeferenced records within their respective IUCN range maps. Ficetola et al. (2014) verified that 89 % of georeferenced records intersected amphibian range maps, and 95 % of georeferenced coordinates were within 31 km of range boundaries, but with reduced intersection in South America and Asia. Georeferenced records that occur within range maps corroborate the accuracy of range maps and georeferenced records that occur within range maps are not errors, in that species occurrences correctly are within designated ranges (Ficetola et al., 2014; Alhajeri and Fourcade, 2019). Alignment of different data sources is known as research triangulation, to confirm credibility and reliability of data, while also applying the strengths of one dataset to identify conflicting data in another dataset (Bans-Akutey and Tiimub, 2021).

Range maps and georeferenced records offer many different applications, including by extension, species distribution models. Expanding on previous work to validate the value of range maps and also locational certainty of georeferenced records, my first objective was to evaluate the percentages of georeferenced records with location distance uncertainty <1 km that intersected IUCN range maps for 732 mammal species, 367 anuran species, 528 reptile species, 829 plant species, and 1397 bird species (Supplementary Table 1). To help substantiate the relative accuracy of georeferenced records with greater distance uncertainties or unknown distance uncertainty, my second objective was to determine the percentages of range intersection for georeferenced records with different distance uncertainties (i.e., a field that may be recorded along with coordinates in georeferenced records; Chapman and Wieczorek, 2020; Marcer et al., 2022; iNaturalist, 2025), varying by taxa. My third objective was to compare range maps from a different source, for record intersection with range maps from IUCN relative to the Mammal Diversity Database (MDD; Marsh et al., 2022). Rather than using small sample sizes, which may be unrepresentative of species distributions plus contribute a limited range of percentage values for intersection if sample sizes are less than 100, species were filtered to have at least 400 unique records for all combined distance uncertainties. Relatedly, my last objective was to demonstrate that small sample sizes are one of many modeling choices that may generate species distributions models that predict greater areas than range maps, and also have wide variation in accuracy, reducing robustness of accuracy metrics. Georeferenced records typically do not sample conditions of entire ranges, but to account for artifacts of modeling, the appropriate comparison for areas of species distribution models from georeferenced records is areas of species distribution models from ranges, with application of consistent modeling methods.

2. Methods

2.1. Species range maps and georeferenced records

Range maps provided the candidate list of species. Because small ranges likely are well-described, I selected extant ranges with areas ≥10,000 square kilometers. Terrestrial mammal, anuran, squamate reptile, and plant species ranges were from IUCN (2024), for available species. Range maps for bird species, resident-only and terrestrial, were from BirdLife International and Handbook of the Birds of the World (2023), which coordinated formats with IUCN. Ranges were filtered to extant ranges. As an additional comparison to IUCN maps, range maps for mammal species from the most abundant orders (Artiodactyla, Carnivora, Chiroptera, Diprotodontia, Eulipotyphla or Soricomorpha, Lagomorpha, and Rodentia) were obtained from Mammal Diversity Database (MDD; Marsh et al., 2022), and I compared species with the same scientific names in both databases to help ensure that species were considered the same entities.

To have robust comparisons of intersections for a range of distance uncertainties, only species with at least 400 georeferenced records were retained. Observations occurred during 1990–2024 with unique coordinates not in oceans (Global Biodiversity Information Facility [GBIF] 2024; https://doi.org/10.15468/dl.ewt6xj; https://doi.org/10.15468/dl.9kafm2; https://doi.org/10.15468/dl.u8exef; https://doi.org/10.15468/dl.w2krf2; https://doi.org/10.15468/dl.nzcwed; https://doi.org/10.15468/dl.k6k2cg; https://doi.org/10.15468/dl.cqsusc; https://doi.org/10.15468/dl.neehbu; https://doi.org/10.15468/dl.nr9yek; https://doi.org/10.15468/dl.6bkr63; https://doi.org/10.15468/dl.92s77y; https://doi.org/10.15468/dl.jhu36x; https://doi.org/10.15468/dl.6fehdk). Associated with coordinates of georeferenced records, location distance uncertainties were unknown distance uncertainties (i.e., not entered) combined with distance uncertainties <1 km for birds, <10 km for plants, and <36 km for mammals, anurans, and reptiles. For bird species, I did exclude from analysis more than 100 bird species that had at least 20,000 records with distance uncertainties of < 1 km, or totaling over 4.5 million records, given that records with unknown distance certainties could be an order of magnitude greater. This resulted in 732 mammal species, 367 anuran species, 528 reptile species, 829 plant species, and 1397 bird species. For less than 5 % of species, GBIF identified one species, providing both an accepted species name and synonym, for two or (three) species in IUCN. I merged ranges of the species, according to IUCN, to combine subspecies into species,

according to GBIF. If GBIF and IUCN both identified species as separate species, despite only one source separating the species geographically, then I did not try to correct these more complex issues (e.g., *Cebus capucinus* and *Cebus imitator*). Reconciling taxonomies from different classification systems is a challenge, given different accepted species, synonyms, and changes over time (Franz and Sterner, 2018).

2.2. Objectives 1 to 3: intersection comparisons

To determine accuracy of range maps, I calculated percentage intersection of georeferenced records of <1 km distance uncertainty, by species, with range maps (intersections calculated in terra package; Hijmans, 2024; R Core Team, 2024). I also compared percentage intersection with ranges of georeferenced records of <1 km distance uncertainty to percentage range intersection with georeferenced 1 km–6 km, 6 km–36 km, and unknown distance uncertainties. For mammal species, I explored intersection of range maps for 464 shared mammal species (i.e., same scientific names) from IUCN and MDD (Marsh et al., 2022).

2.3. Objective 4: species distribution areas and accuracies for small and sufficient sample sizes

To generate species distribution models, and then calculate total areas and areas within ranges, I used both small samples and large samples from georeferenced records and randomly generated points from IUCN range maps of African bat species (see modeling methods below). For georeferenced records, only four species limited to Africa in range met very minimum sufficient sample sizes of around 300 (Gábor et al., 2020), ranging from 297 to 484 records: *Epomophorus gambianus*, *Epomophorus wahlbergi*, *Lavia frons*, and *Taphozous mauritianus*. I modeled these species with the minimum sufficient sample sizes and then modeled small samples, both for 10 sets of 25 random samples from the georeferenced records and for 10 sets of samples by year, with annual samples ranging from 11 to 69 (mean = 23) during years 2014–2023. If sample sizes were less than 11 or greater than 100, I excluded those years and added years 2013 or 2012. Annual slices represent real ways that small samples may be less representative than the whole; likewise, random small samples can be more biased than larger samples (Barry and Elith, 2006). For IUCN ranges of these four species, species distribution models were from 400 spatially random samples of the IUCN range, and 10 sets of 25 random samples from the 400 samples. To increase the number of species, I compared the IUCN range areas of 156 bat species in Africa with areas predicted by species distribution models from 400 spatially random samples of the IUCN ranges and 25 samples from the 400 random samples.

Random forests is one of the most commonly applied algorithms for species distribution models that is highly accurate (Fernández-Delgado et al., 2014; Santini et al., 2021). For modeling of each species, I applied the random forests classifier, an ensemble (i.e., 500 classification trees as a robust number) nonlinear model, with the model configuration hyperparameter (mtry, the number of randomly selected predictors considered at each split in a tree) determined through 10-fold cross-validation, repeated three times (automated in the caret package; Kuhn, 2008; R Core Team, 2024).

Model classification was based on an equal number of presence and absence samples (i.e., modeling prevalence of 0.5). An equal number of pseudoabsence samples as presence samples were generated randomly from Africa as the background range of values. Modeling prevalence was the statistical basis to set the threshold between classification as present or absent at 0.5, and as another benefit allows comparability among models. That is, the modeling prevalence already has been applied by the modeling algorithm to determine the binary classification of presence and absence. One standard practice is to match the modeling prevalence, or ratio of presence samples to absence samples, to the classification threshold because when sample classes are balanced, the modeling algorithm is forced to also balance predictions between classes, rather than favoring predictions of one class over another (Hanberry and He, 2013).

The model ran one time (with 500 trees) using all the samples to map predictions for area calculations and another time (with 500 trees) with 25 % withheld samples for accuracy metric calculations of accuracy, sensitivity, and specificity. Accuracy metrics measure how well the modeling algorithm classified the provided samples, not how representative models are of species distributions. Accuracy metrics are similar, because they are based on variations of the confusion matrix (true positives or sensitivity, true negatives or specificity, false positives, and false negatives), although some accuracy metrics have been promoted as better than others or critiqued as worse than others (e.g., Lobo et al., 2008).

Climate predictors are proximal variables that directly describe environmental conditions of species occurrences and have direct effects on species (Elith and Leathwick, 2009; Mod et al., 2016). Climate overall is the most influential variable for modeling species distributions, and due to resultant model accuracy, climate variables are the most commonly applied predictors for species distribution models, and may be the only types of predictors selected for modeling (Mod et al., 2016; Titeux et al., 2016; Fourcade et al., 2018). Modeling algorithms favor the relatively predictable grids of climate variables, or other continuously distributed variables, over variable predictor variables (Hanberry, 2023a). Model predictors were 13 climate variables during 1981–2010: mean annual temperature, maximum temperature of the hottest month, minimum temperature of the coldest month, mean temperature of the coldest and hottest three months, mean temperature of the driest and wettest three months, annual precipitation, precipitation during the driest month and driest three months, precipitation during the wettest month, and precipitation during the coldest and warmest three months (resolution 30 arc seconds; CHELSA climate; Karger et al., 2017, 2018). Correlated climate variables are not an issue for predictions but a problem for standard error of estimates in data models (Hanberry, 2024). All methods and variables were held constant except for the source and number of samples.

3. Results

3.1. Objective 1: percentage intersection with ranges of georeferenced records for <1 km distance uncertainty

Regarding percentage intersection with ranges of georeferenced records for <1 km distance uncertainty, for 732 mammal species the 75th percentile was 96.4 % intersection (i.e., 25 % of species had at least 96.4 % of georeferenced records within the respective IUCN range maps), the 50th percentile was 92.0 % intersection, and the 25th percentile was 82.0 % intersection (Table 1). For anuran and reptile species, percentages were slightly greater than for mammals (anurans 25th percentile was 84.2, 50th percentile was 94.0, 75th percentile was 97.7; reptiles 25th percentile was 83.7, 50th percentile was 92.5, 75th percentile was 97.0). For bird species, percentages were slightly less than for mammals (25th percentile was 81.1, 50th percentile was 90.9, 75th percentile was 96.0). For plant species, the 75th percentile (97.3) and the 50th percentile (91.3) were similar to other taxa, but the 25th percentile was the least value, at 74.8 %. For all 3853 species combined, the 75th percentile was 96.8 % intersection, the 50th percentile was 91.7 % intersection, and the 25th percentile was 81.0 % intersection (Fig. 1).

3.2. Objective 2: percentage intersection with ranges of georeferenced records for ≥1 km or unknown distance uncertainty

For all combined species, percentage intersection with ranges of georeferenced records for ≥1 km or unknown distance uncertainty was only slightly less than percentage intersection with ranges of georeferenced records for <1 km distance uncertainty (see Table 1 for results by taxa). The 75th percentile was 96.1 % intersection with ranges of

Table 1
Taxa, distance uncertainties (NA = unknown, not entered with georeferenced records) and percentiles of species (75th percentile = 75 % of species below the value) with percentage intersection of georeferenced records with ranges (NA = not calculated due to the magnitude of records that had no distance uncertainty entry).

Taxa	Distance	25th percentile	50th percentile	75th percentile
Anurans	Distance 0–1 km	84.2	94.0	97.7
	Distance 1–6 km	82.5	93.2	97.6
	Distance 6–36 km	81.0	92.2	97.0
	Distance NA	81.2	92.5	97.5
	Distance total	81.0	93.1	97.3
birds	Distance 0–1 km	81.1	90.9	96.0
	Distance 1–6 km	NA	NA	NA
	Distance 6–36 km	NA	NA	NA
	Distance NA	81.0	89.8	94.9
	Distance total	81.1	90.0	95.0
mammals	Distance 0–1 km	82.0	92.0	96.4
	Distance 1–6 km	80.7	92.2	97.0
	Distance 6–36 km	79.2	91.0	96.1
	Distance NA	80.2	91.9	96.9
	Distance total	81.2	91.8	96.1
plants	Distance 0–1 km	74.8	91.3	97.3
	Distance 1–6 km	72.7	91.0	96.8
	Distance 6–36 km (10 km)	68.7	89.6	96.3
	Distance NA	68.5	89.4	96.4
	Distance total	74.6	90.7	96.9
reptiles	Distance 0–1 km	83.7	92.5	97.0
	Distance 1–6 km	81.5	91.5	96.5
	Distance 6–36 km	83.4	91.6	96.1
	Distance NA	83.2	91.9	96.7
	Distance total	83.5	92.4	96.4

georeferenced records of unknown distance uncertainty (relative to 96.8 % range intersection for georeferenced records of <1 km), the 50th percentile was 90.7 % intersection (relative to 91.7 % intersection), and the 25th percentile was 79.5 % (relative to 81.0 % intersection). For all distances combined, the 75th percentile was 96.1 % intersection with ranges, the 50th percentile was 91.1 % intersection, and the 25th percentile was 80.5 %.

3.3. Objective 3: percentage intersection of IUCN and Mammal Diversity Database (MDD) ranges with georeferenced records

For 464 mammal species, intersection percentage of georeferenced records with range maps from IUCN relative to MDD were similar (Fig. 2). However, the percentage intersection for the first quartile was less for MDD range maps than IUCN range maps. For MDD range maps, the 75th percentile was 97.4 % intersection for georeferenced records of <1 km distance uncertainty relative to 96.3 % intersection with IUCN range maps, the 50th percentile was 91.8 % intersection for georeferenced records of <1 km distance uncertainty relative to 92.0 % intersection with IUCN range maps, and the 25th percentile was 78.0 % intersection for georeferenced records of <1 km distance uncertainty relative to 82.1 % intersection with IUCN range maps. Either as a benefit or detriment, the percentage intersection for the first quartile of species was about 74 % for the other distance uncertainty categories with MDD range maps. Examples of North American species below the first quartile of georeferenced record and range intersections for both IUCN and MDD range maps include *Sciurus carolinensis*, a wide-ranging and expanding species, *Urocitellus elegans*, which has a specifically shaped distribution, and *Bison bison,* which is contained within fences and precisely located, by and large (although *Bison bison* was excluded from the intersection comparison due to the scientific name of *Bos bison* in MDD; Fig. 3).

3.4. Objective 4: species distribution areas and accuracies for small and sufficient sample sizes

Species distribution models of four bat species from 25 random samples generated greater areas of predicted presence than species distribution models from larger samples (Fig. 4). Specifically, the mean predicted area for four species of 10 species distribution models (i.e., 40 models) based on 25 random samples from georeferenced records (mean accuracy = 0.83, sensitivity = 0.85, specificity = 0.81) was greater by a

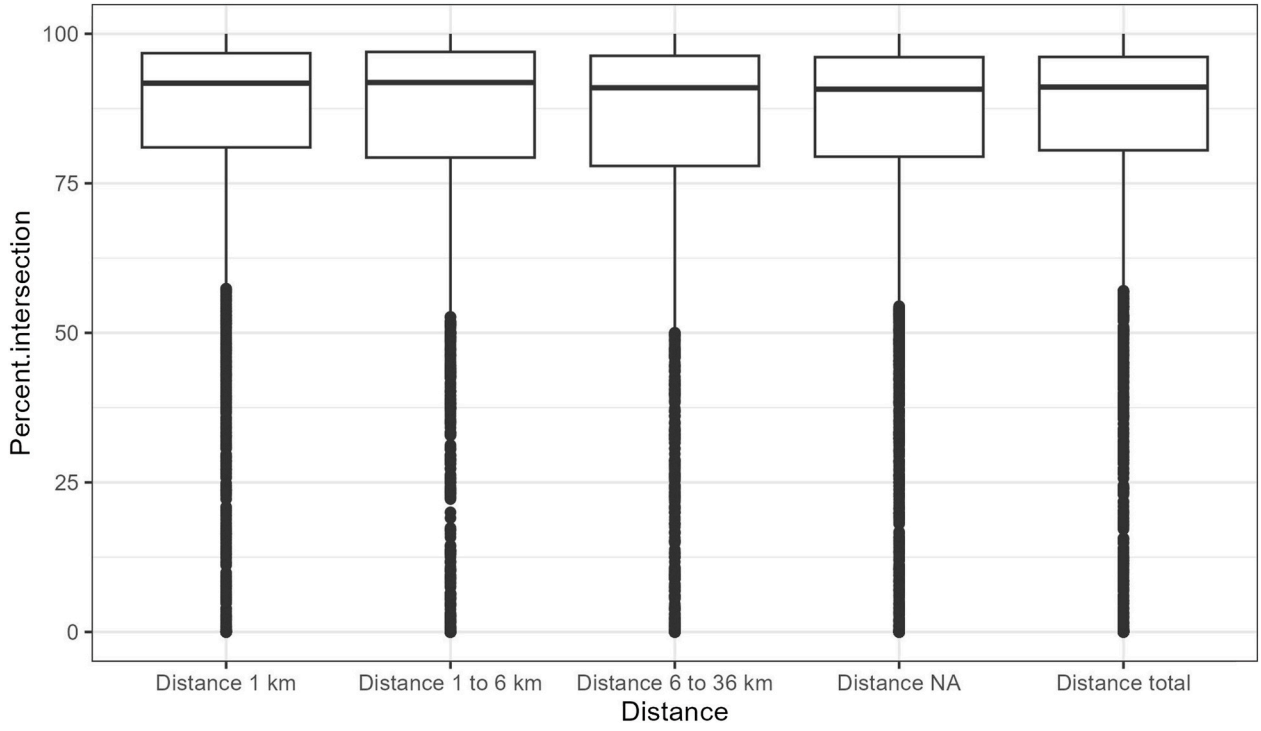

Fig. 1. Percentage intersection with International Union for Conservation of Nature (IUCN) range maps of georeferenced records at distance uncertainties of <1 km, 1 – < 6 km, 6 – < 36 km, unknown (NA) distance uncertainties, and all (total) distance uncertainties for all combined species.

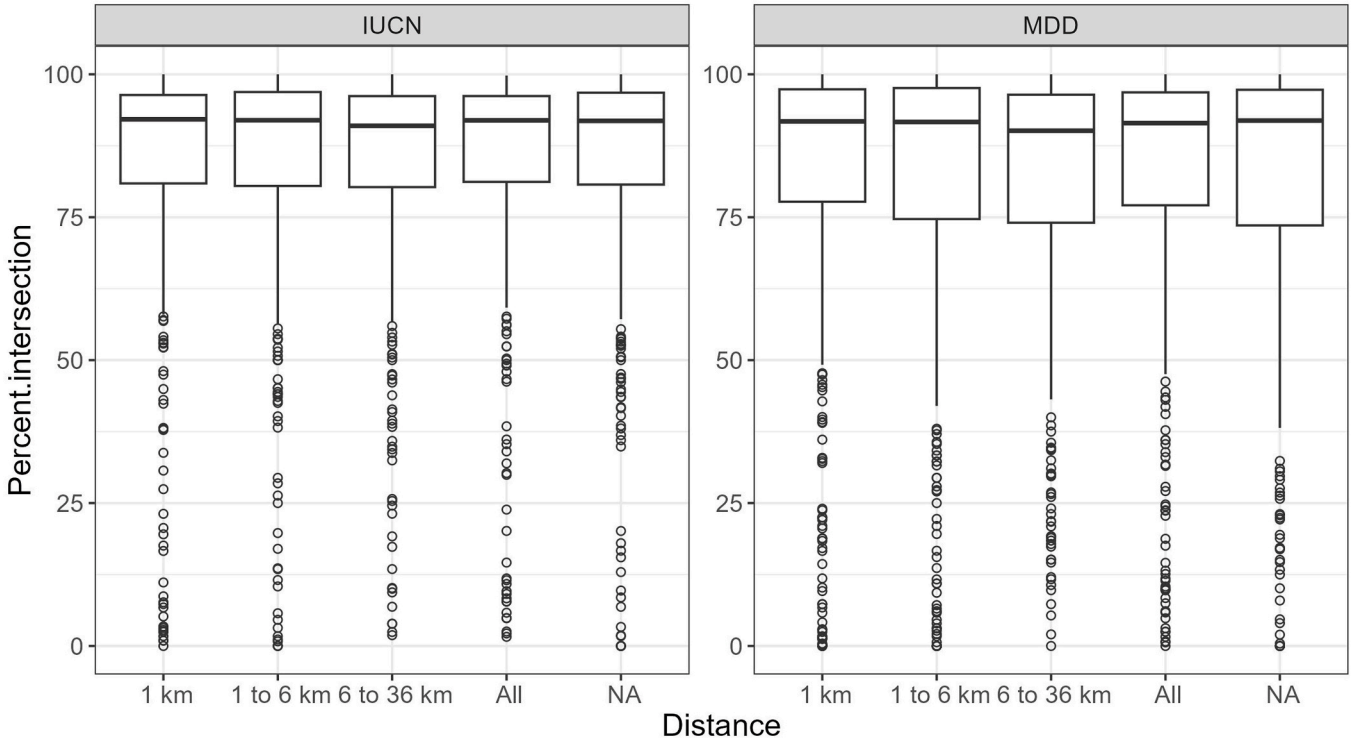

Fig. 2. Percentage intersection with International Union for Conservation of Nature (IUCN) range maps and Mammal Diversity Database (MDD) range maps of georeferenced records at distance uncertainties of <1 km, 1 – < 6 km, 6 – < 36 km, unknown (NA) distance uncertainties, and all (total) distance uncertainties for 464 mammal species.

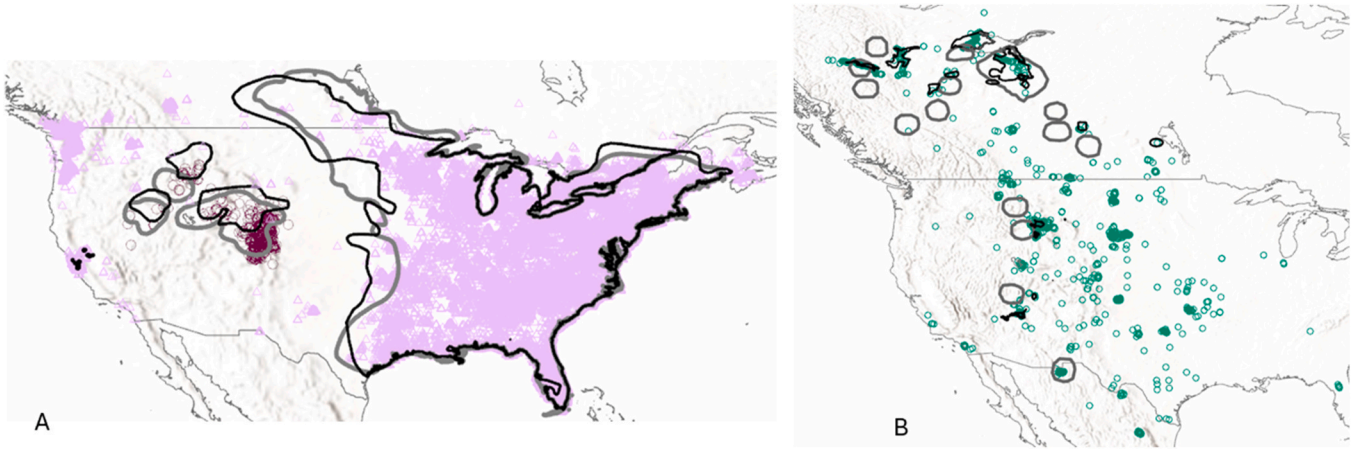

Fig. 3. Three examples of intersection of georeferenced records (light purple circles of *Sciurus carolinensis* and dark red circles of *Urocitellus elegans* in panel A; green circles of *Bison bison* in panel B) and ranges (black outline of IUCN ranges and thick gray outline of MMD ranges) in North America, centered on the United States. Intersection of about 50 % of georeferenced records with IUCN and MDD range maps of wide-ranging and expanding *Sciurus carolinensis*; intersection of about 48 % of georeferenced records with IUCN range maps and intersection of 12 % of georeferenced records with MDD range maps of the specifically-shaped range of *Urocitellus elegans*; and intersection of about 44 % of georeferenced records with IUCN range maps and intersection of 3.4 % of georeferenced records with MDD range maps of the precisely located range of *Bison bison*.

factor of 2.6 than the mean area for four species distribution from all records (mean accuracy = 0.91, sensitivity = 0.92, specificity = 0.91). The mean predicted area for four species of 10 species distribution models (i.e., 40 models) based on about 23 annual samples from georeferenced records (mean accuracy = 0.81, sensitivity = 0.87, specificity = 0.74) was greater by a factor of 2 than the mean area for four species distribution models from all records. Because the 25 samples of 400 random samples were perfectly sampled from range maps, the mean area for four species of 10 species distribution models (i.e., 40 models; mean accuracy = 0.85, sensitivity = 0.88, specificity = 0.82) based on 25 samples of range maps was only 113 % of the mean area for four species distribution models from 400 random samples of range maps (mean accuracy = 0.91, sensitivity = 0.93, specificity = 0.88).

Overall, areas of IUCN ranges and predicted areas of species distribution models from ranges were greater than predicted areas of species distribution models from georeferenced records (Figs. 4 and 5). For the four bat species, species distribution models from minimum sufficient sample sizes of at least 300 georeferenced records were 39 % of the range areas, with 67 % of predicted presence areas within the ranges. Species distribution models from 25 random samples of georeferenced

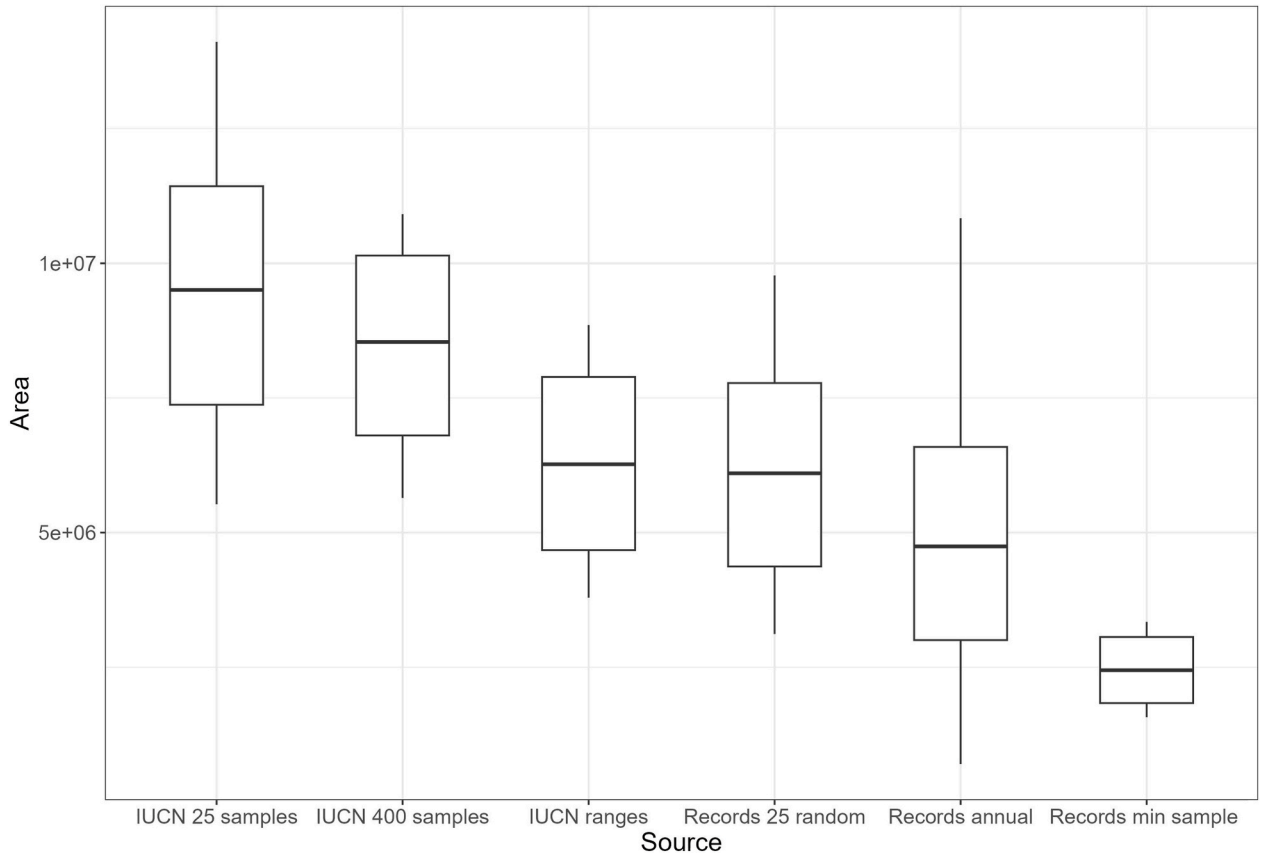

Fig. 4. For four bat species in Africa, species distribution models from 25 random samples of IUCN range maps generated greater areas of predicted presence than species distribution models from larger samples, 400 random samples of IUCN range maps, which were greater in area than IUCN range maps. Species distribution models from about 25 random samples of georeferenced records generated similar predicted presence areas to IUCN range maps whereas very minimum sufficient sample sizes of 300–500 georeferenced records generated smaller predicted presence areas than IUCN range maps.

records were of equal sizes to ranges on average, with 49 % of predicted presence areas within the ranges. Of the 40 species distribution models from 25 random samples of georeferenced records, 18 of 40 (45 %) models predicted areas that exceeded range maps. Similarly, species distribution models from 400 random samples of range maps overpredicted range areas by 137 %, with 68 % of predicted presence areas within the ranges, whereas the models based on 25 range samples overpredicted range by 156 %, with 57 % of predicted presence areas within the ranges. For 156 bat species in Africa with areas predicted by species distribution models from 400 spatially random samples of the IUCN ranges (mean accuracy = 0.95, sensitivity = 0.97, specificity = 0.93) and 25 samples of the 400 samples (mean accuracy = 0.90, sensitivity = 0.93, specificity = 0.87), areas from species distribution models were greater than range areas (162 % greater than range areas and 242 % greater than range areas, respectively; Fig. 5). In addition to overprediction of areas, species distribution models from small sample sizes may not be representative of species ranges, due to the amount of area predicted outside of species ranges (Fig. 6).

Small samples also can result in variation in accuracy, from poor to perfect classification, for similar representations of species (Fig. 7). For one species, accuracy ranged from 0.33 to 1.0 for approximately 25 annual samples from georeferenced records and from 0.58 to 1.0 for 25 random samples from georeferenced records. As a comparison of accuracy for sample sizes of 25 and 400 for a similar number of species (i.e., 156), albeit with perfectly random sampling unlike georeferenced records, accuracy for sample sizes of 25 was 0.90, with a range of 0.58 and standard deviation of 0.11 and accuracy for sample sizes of 400 was 0.95, with a range of 0.15 and standard deviation of 0.04 (Fig. 8).

4. Discussion

4.1. Georeferenced records and range maps

Georeferenced records and range maps are complementary and congruent sources of information about species ranges (Alhajeri and Fourcade, 2019). Georeferenced records supply confirmed points of occupancy and range maps contribute expert knowledge about potential extents of occurrence that may not yet be confirmed by readily available, compiled georeferenced records of species occurrences (Alhajeri and Fourcade, 2019). Key findings for this study were that georeferenced records, with great certainty of location, validated accuracy and location of most (≥75 %) range maps, with limited to mildly moderate omission error, for 3853 mammal, anuran, reptile, bird, and plant species. Greater omission error occurred for about 25 % of species ranges that had less than 81 % intersection with georeferenced records. In general, commission error of range maps cannot be determined with georeferenced records, because the absence of georeferenced records may simply reflect lack of sampling, and hence the valuable contribution of range maps (Hughes et al., 2024; Daru and Rodriguez, 2023; Garcia-Rosello et al., 2023). Nonetheless, low intersection with georeferenced records for specific range shapes, with positional offset, indicate potential mislocation and commission error in range maps (Fig. 3). Range maps verified that georeferenced records with uncertainty in location still are representative of species ranges. Despite different sources, range maps and georeferenced records on the whole aligned, corroborating the credibility and validity of each source through research triangulation (Bans-Akutey and Tiimub, 2021). The strengths of one data source can isolate the weaknesses of the other data source, by indicating range

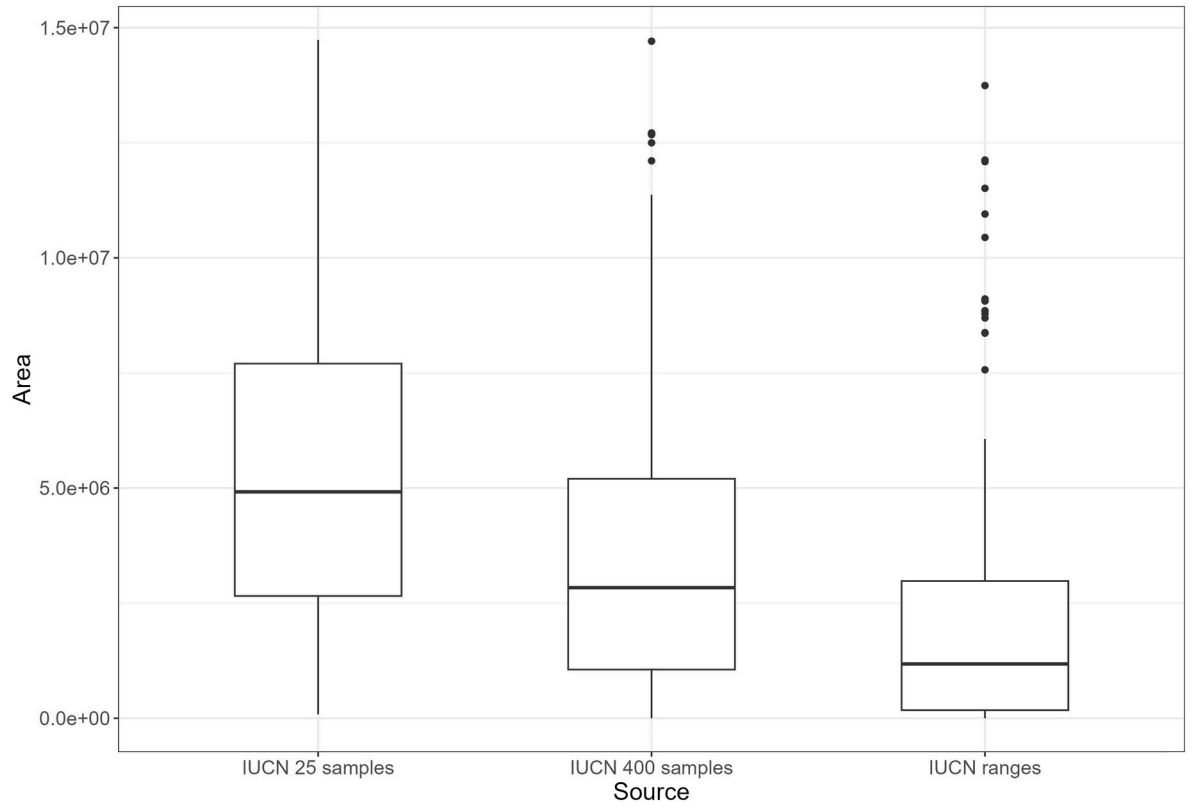

Fig. 5. For 156 bat species in Africa, species distribution models from 25 random samples of IUCN range maps predicted greater area of presence than species distribution models from 400 random samples and areas of IUCN range maps.

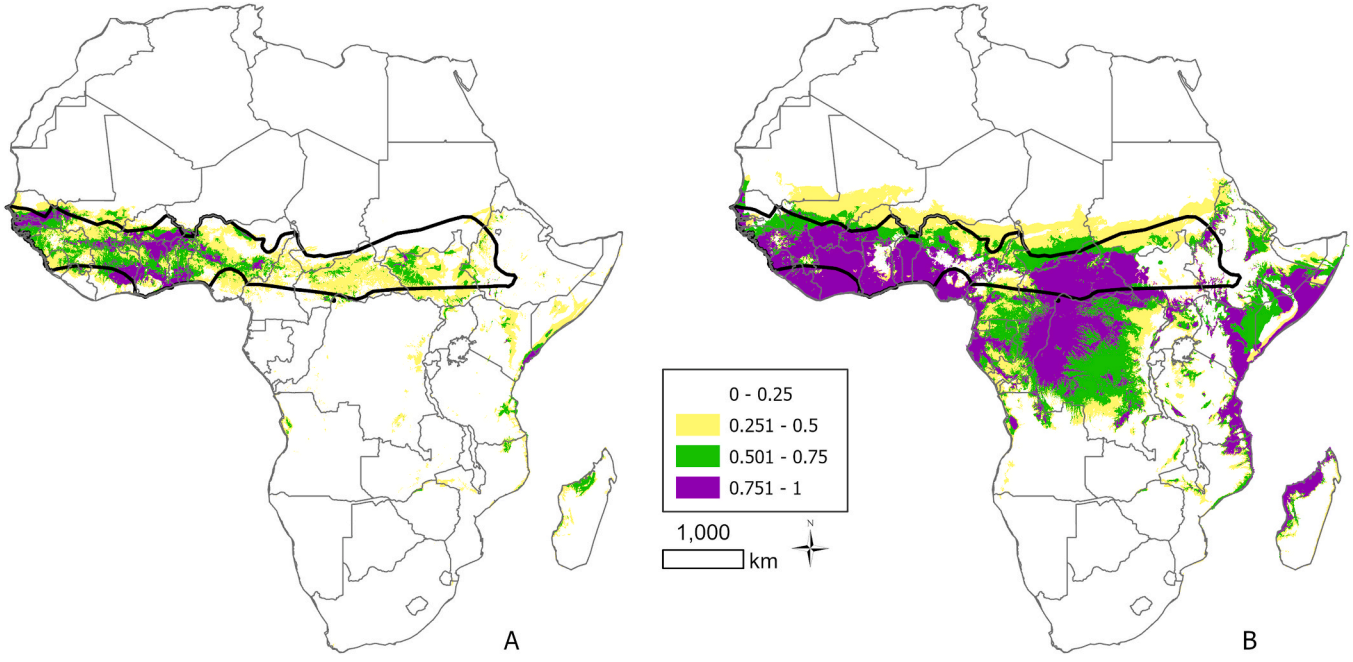

Fig. 6. Species distribution models from almost 500 samples of georeferenced records (green and purple colors represent predicted probabilities of presence; A) and 25 samples of georeferenced records (green and purple colors; B) of *Epomophorus gambianus*, with IUCN range map for comparison (black outline). Species distribution models from small samples are more probable to overpredict areas due to commission errors than species distribution models from minimum sufficient sample sizes of about 300–500 samples, which are required to achieve a higher probability of representative models.

maps that likely have omission error and species distribution models with commission error (Bans-Akutey and Tiimub, 2021). Rather than a call for caution (Herkt et al., 2017), these results demonstrated the value of range maps to support incompletely sampled georeferenced records and associated species distribution models with omission error.

Intersection between georeferenced records with very accurate and

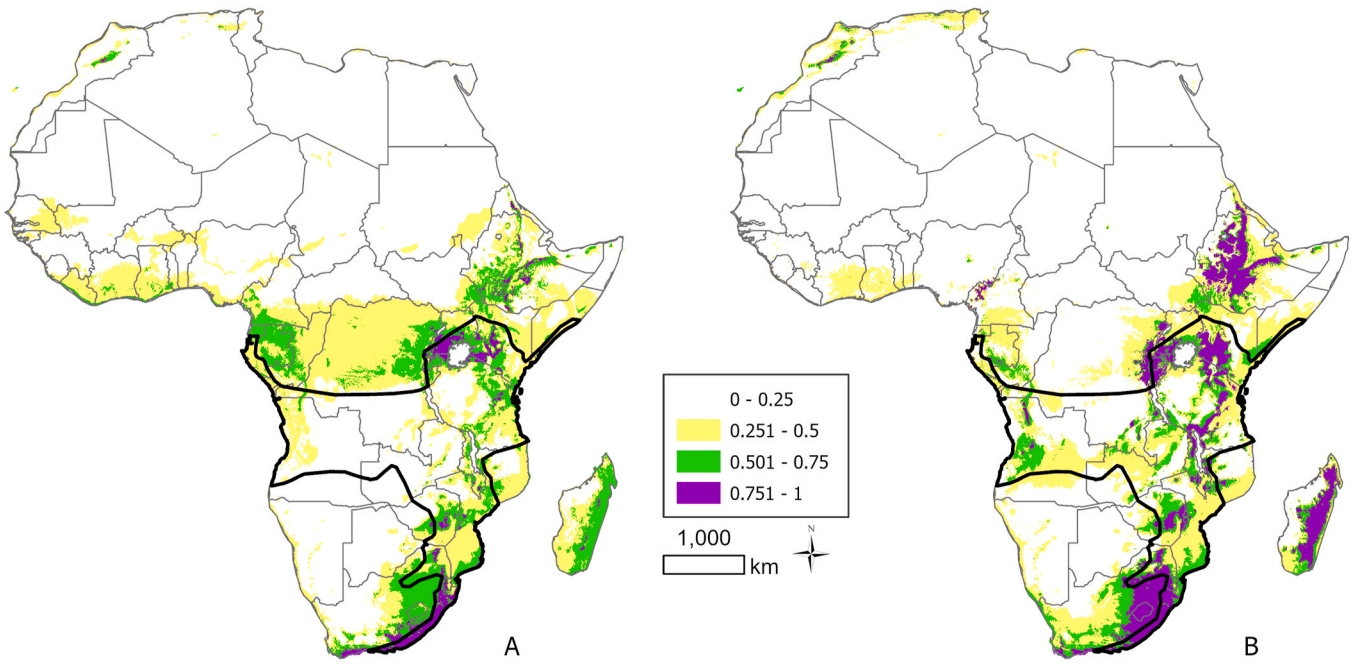

Fig. 7. Species distribution models from about 25 samples of georeferenced records during 2019 (green and purple colors represent predicted probabilities of presence; A) and about 25 samples of georeferenced records during 2014 (green and purple colors; B) of *Epomophorus wahlbergi*, with IUCN range map for comparison (black outline). These small samples did not overpredict presence, but with similar representations of the species, had widely varying balanced accuracies of 0.67 (A) and 1.0 (B).

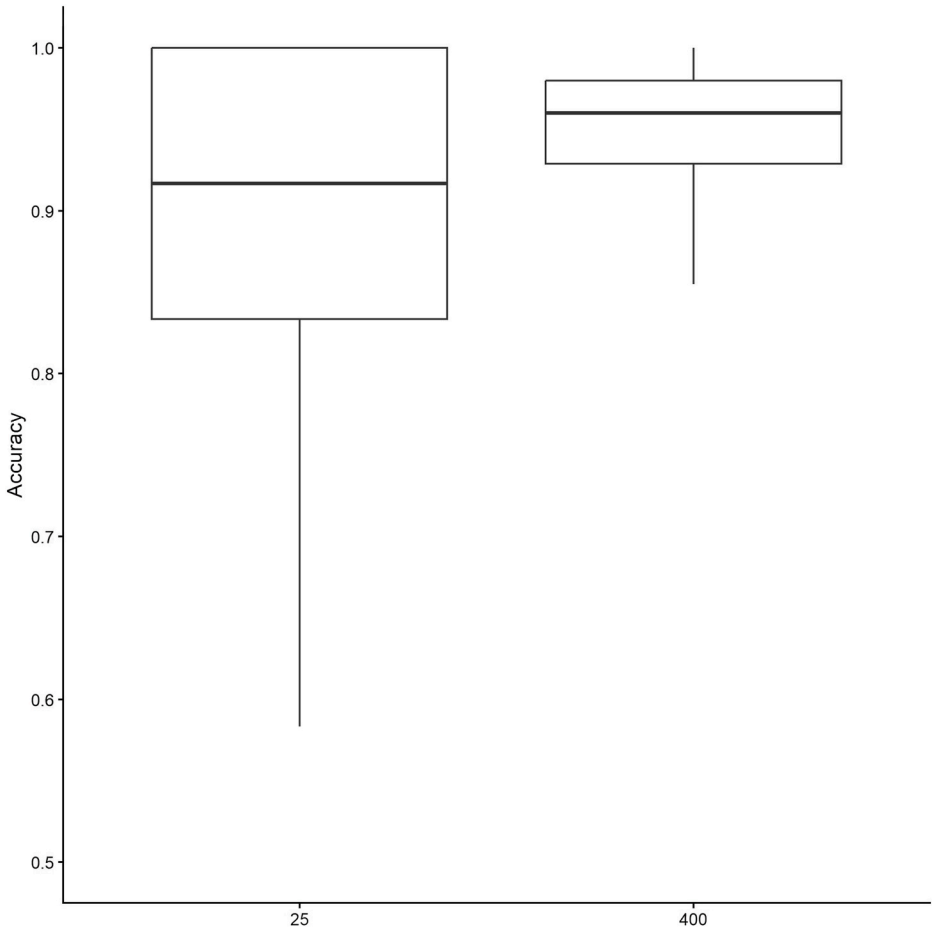

Fig. 8. For 156 bat species in Africa, species distribution models from 25 random samples of IUCN range maps had a wide range of accuracy relative to species distribution models from 400 random samples of IUCN range maps.

certain location (distance uncertainties <1 km) and range maps indicated most (≥75 %) range maps of species are located accurately, with some omission error. According to this study, 75 % of 3853 mammal, anuran, reptile, bird, and plant species had at least 81.0 % intersection between georeferenced records of great locational certainty, or fine resolution, and general, coarse range maps. Moreover, 50 % of species had at least 91.7 % intersection between records and range maps, which is an improvement upon results reported by Alhajeri and Fourcade (2019) of more than half of 1315 rodent species with more than 75 % intersection between georeferenced records and range maps. Differences in results are due most likely to thresholds for inclusion in analysis (e.g., I filtered species to ones with robust sample sizes of at least 400 records), although taxa and species (rodents relative to mammal, anuran, reptile, bird, and plant species) may play a role. Ficetola et al. (2014) used georeferenced records from the GBIF database when half of records were from museums, with uncertain georeferenced coordinates. They found 73.8 % of records intersected with species ranges, and when coordinates were filtered to records during or after year 1990, with distance uncertainty of less than 1 km, 89 % of records intersected with species' ranges. Rather than locational errors based on distance uncertainty, filtering by year to remove museum records may be the source of improvement, based on findings from this study. Observations recorded before accessible global positioning systems may be georeferenced incorrectly based on vague site descriptions (Chapman and Wieczorek, 2020; Marcer et al., 2022).

Errors may occur in georeferenced records, but intersections suggested that georeferenced records with greater uncertainty in location (distance uncertainties >1 km or unknown) are not likely to be substantially more erroneous in location than georeferenced records with less uncertainty in location. While ranges are at coarse resolutions, percentage intersection with ranges of georeferenced records for distance uncertainties that were greater than 1 km or unknown was essentially equivalent to percentage intersection with ranges of georeferenced records for distance uncertainties <1 km. Therefore, at least for some applications, georeferenced records with distance uncertainties >1 km or unknown may provide additional samples. For bird species, this may allow use of the preponderance of records from eBird (2024; Cornell Lab of Ornithology), which do not contain a tag about distance uncertainty. To be clear, one option to select for locations in eBird (2024) is nearby public hotspots, which may be less accurate than actual locations. For another data sharing platform, iNaturalist (2020), locations can be obscured to a random point within a 0.2 by 0.2-degree area, for protection of the species or landowner. Nonetheless, establishment of species occupancies is about probabilities. That is, the more records, and the more records that occur in proximity, the more probable that species ranges are being demarcated and that species occur in the vicinity. Georeferenced records with unknown distance uncertainties and distance uncertainties greater than 1 km may supply valuable information about species ranges.

The approximately 80 %–90 % intersection of georeferenced records with ranges for about 25 % of species, above the first quartile, did reveal mildly moderate omission error, by not capturing 10 %–20 % of species occurrences. To increase necessary coverage for minor omissions may simply entail widening the range, up to 100–200 km suggested by Jetz et al. (2008), or 31 km by Ficetola et al. (2014), while attempting to be conservative in terms of commission error. However, adding wider buffers to ranges will not correct disjunct species ranges and will expand commission error all the way around if the omitted areas are specifically located.

Range maps for the 25 % of species below the first quartile of intersections, or less than 80 % intersection with georeferenced records, may need reevaluation (Supplementary Table 1; Fig. 3). For example, most of the populations of large mammal species may be contained within wildlands refugia, such as mountainous terrain or designated wildlife reserves. American bison (i.e., plains bison or *Bison bison bison* largely distributed in the central grasslands of North America and wood bison or *Bison bison athabascae* in Canada and Alaska) are fenced primarily within protected areas (over 90 % of records), although historically American bison were a wide-ranging species (Hanberry, 2023b). The IUCN extant range captured 44 % of georeferenced records, whereas the Mammalian Diversity Database range (for *Bos bison*) captured less than 4 % of records. Range maps failed to delineate a charismatic megafauna with locations on public lands, which are well-documented with greater than 10,000 georeferenced records (GBIF, 2024). Range maps for bison had omission and likely commission error, due to mislocation, evident with basic knowledge of the species. Intersection of range maps with georeferenced records supplies an efficient method to detect potential anomalies in range maps.

For the variation encompassed by thousands of species ranges, omission and commission error may fit a variety of patterns. Indeed, IUCN guidelines for range map development specify a precautionary approach to err on the side of omission error to prevent commission error (IUCN Standards and Petitions Committee, 2024). Ranges may be mislocated, with some refinement in shape necessary to fit better to landscapes, such as patchy distributions that follow mountainous terrain or nature preserves (Fig. 3). Alternatively, rather than considering ranges to encompass maximum areas of suitability, instead ranges may contain areas of greatest density and areas outside may comprise less suitable conditions and dispersed individuals, perhaps even sink or transient populations following metapopulation dynamics (Ficetola et al., 2014). Ranges also may represent an earlier interval and lack representation of recent expansions (Fig. 3). Of course, the objective for range mapping may be to represent native ranges before recent expansion, but after historical extirpation.

The Mammalian Diversity Database source of mammal species range maps (Marsh et al., 2022) overall had similar intersection percentages with the georeferenced records as IUCN range maps, indicating equivalent performance, albeit varying by species. The percentage intersection of the first quartile was less for MDD range maps than IUCN range maps. Larger range maps will increase the chance of intersections, but the IUCN range maps were approximately the same in mean area (i.e., 102 %) as MDD range maps. Unreconciled taxonomy was another factor that may influence intersections, although not affecting most species. In any event, alternative sources of range maps may not be superior to IUCN range maps.

4.2. Species distribution models and range maps

Rather than considering IUCN ranges to be too small relative to predicted presence areas of species distribution models (Herkt et al., 2017), overpredicted areas of presence are artifacts of species distribution models. Most species distribution models use small samples (<50 for 62 % of 250 randomly selected publications; Santini et al., 2021), despite reduced probability of creating fully representative ranges and equally representative species distribution models with a modest sample size, a long-standing issue (Barry and Elith, 2006). Small samples indicate lack of sampled conditions, resulting in a greater probability of being biased. The probability of samples being representative of species ranges decreases with sample size, and below 100 samples is particularly unreliable (Hanberry et al., 2012; Mitchell et al., 2017; Gábor et al., 2020). Species distribution models from small samples (25) of range maps had greater mean area than species distribution models from minimum sufficient samples (400) of range maps, which had greater mean area than range maps. It is not reasonable to expect 25 samples to describe conditions of entire species distributions, and accuracy metrics from species distribution models may not be robust, due to wide variation in small samples. High accuracy metrics need to be tempered with input data limitations (Moudrý, 2015). Modeling with small samples is an unreliable approach due to commission error resulting from sample sizes that are too small for the modeling algorithm to differentiate conditions between presence and absence samples (Fig. 5; Mitchell et al., 2017; Gábor et al., 2020).

Given sufficient samples, typically at minimum 300 to 400 (Gábor et al., 2020), georeferenced records are probable to be relatively characteristic of species ranges, with some omission error because field sampling does not cover all conditions. While species distribution models from small samples of georeferenced records predicted the same mean area as range maps, species distribution models from minimum sufficient samples of georeferenced records predicted smaller areas than range maps because sampled conditions are narrower than conditions outlined by range maps. Greater sample sizes of 300–500 georeferenced records did not overpredict presence relative to range maps, unlike about half of species distribution models from small samples of georeferenced records. With minimum sufficient samples for species distribution models, the smaller areas of predicted presence for species distribution models relative to range maps represent omission error due to incompletely sampled ranges and also that species are not likely to occur completely throughout ranges (Jetz et al., 2008). As number of samples increases from the minimum sufficient size, incorporating a greater range of sampled conditions, the predicted area of presence increases, but more consistently representing ranges than when information about conditions from samples is deficient.

Instead of species distribution models indicating issues with range maps, the range maps indicate underlying issues with species distribution models. Species distribution models based on small samples, of limited conditions, generally overpredicted species distributions relative to range maps. In addition to the greater areas, about 50 % of predicted presence area in species distribution models based on small samples was outside of species ranges, indicating commission error that was not as present for predicted area in species distribution models from minimum sufficient sample sizes of at least 300 georeferenced records, which predicted 33 % of presence area outside of species ranges, and for smaller areas. Even 400 samples, if not representative of underlying species distributions, can produce unrepresentative models that overpredict areas of presence (Mitchell et al., 2017; Gábor et al., 2020). While it is not impossible for small samples to be completely representative of ranges, the smaller the samples, the less probable that species distribution models will be representative and more probable that omission and commission errors have occurred (Barry and Elith, 2006; Mitchell et al., 2017; Gábor et al., 2020).

5. Conclusions

There is much information to be gained from range maps, georeferenced records, and triangulation between the two data sources. Georeferenced records validated range maps, while indicating range maps that likely need reevaluation for at minimum omission error. Range maps validated the overall accuracy of georeferenced records, even records with unknown distance uncertainty. Range maps likewise can corroborate species distribution models, and identify erroneous models. In the future, range maps likely will be updated more frequently with readily available georeferenced records, creating greater correspondence between the two sources of data about species ranges.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

I thank the Global Biodiversity Information Facility, International Union for Conservation of Nature, and contributors to the range maps and georeferenced observations. This research was supported by the U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station, and the U.S. Department of Agriculture, Forest Service, with funding from the Infrastructure Investment and Jobs Act of 2021. The findings and conclusions in this publication are those of the author and should not be construed to represent any official USDA or U.S. Government determination or policy.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.actao.2025.104137.

Data availability

The data and code are available at Zenodo https://doi.org/10.5281/zenodo.15851509.

References

Alhajeri, B.H., Fourcade, Y., 2019. High correlation between species-level environmental data estimates extracted from IUCN expert range maps and from GBIF occurrence data. J. Biogeogr. 46 (7), 1329–1341.
Arenas-Castro, S., Regos, A., Martins, I., Honrado, J., Alonso, J., 2022. Effects of input data sources on species distribution model predictions across species with different distributional ranges. J. Biogeogr. 49 (7), 1299–1312.
Bans-Akutey, A., Tiimub, B.M., 2021. Triangulation in research. Academia Letters 2 (3392), 1–7.
Barry, S., Elith, J., 2006. Error and uncertainty in habitat models. J. Appl. Ecol. 43 (3), 413–423.
BirdLife International and Handbook of the Birds of the World, 2023. Bird species distribution maps of the world. Version 2023.1. http://datazone.birdlife.org/species/requestdis.
Chapman, A.D., Wieczorek, J.R., 2020. Georeferencing Best Practices. GBIF Secretariat, Copenhagen. https://doi.org/10.15468/doc-gg7h-s853. (Accessed 25 September 2025).
Denzin, N.K., 2012. Triangulation 2.0. J. Mix. Methods Res. 6 (2), 80–88.
Daru, B.H., Rodriguez, J., 2023. Mass production of unvouchered records fails to represent global biodiversity patterns. Nat. Ecol. Evol. 7 (6), 816–831.
eBird, 2024. Choosing and Managing Locations in eBird. Available at https://support.ebird.org/en/support/solutions/articles/48000850891-choosing-and-managing-locations-in-ebird. (Accessed 16 February 2025).
Elith, J., Leathwick, J.R., 2009. Species distribution models: ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics 40 (1), 677–697.
Faurby, S., Araújo, M.B., 2018. Anthropogenic range contractions bias species climate change forecasts. Nat. Clim. Change 8, 252–256.
Feeley, K.J., Silman, M.R., 2010. Modelling the responses of Andean and Amazonian plant species to climate change: the effects of georeferencing errors and the importance of data filtering. J. Biogeogr. 37 (4), 733–740.
Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., 2014. Do we need hundreds of classifiers to solve real world classification problems? J. Mach. Learn. Res. 15 (1), 3133–3181.
Ficetola, G.F., Rondinini, C., Bonardi, A., Katariya, V., Padoa-Schioppa, E., Angulo, A., 2014. An evaluation of the robustness of global amphibian range maps. J. Biogeogr. 41 (2), 211–221.
Fourcade, Y., Besnard, A.G., Secondi, J., 2018. Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. Global Ecol. Biogeogr. 27, 245–256.
Franz, N.M., Sterner, B.W., 2018. To Increase Trust, Change the Social Design Behind Aggregated Biodiversity Data. Database, bax100.
Gábor, L., Moudrý, V., Barták, V., Lecours, V., 2020. How do species and data characteristics affect species distribution models and when to use environmental filtering? Int. J. Geogr. Inf. Sci. 34 (8), 1567–1584.
Garcia-Rosello, E., Gonzalez-Dacosta, J., Guisande, C., Lobo, J.M., 2023. GBIF falls short of providing a representative picture of the global distribution of insects. Syst. Entomol. 48 (4), 489–497.
Gaston, K.J., Fuller, R.A., 2009. The sizes of species' geographic ranges. J. Appl. Ecol. 46 (1), 1–9.
Global Biodiversity Information Facility [GBIF], 2024. Free and open access to biodiversity data. Available at: www.gbif.org. (Accessed 2 January 2025).
Graham, C.H., Hijmans, R.J., 2006. A comparison of methods for mapping species ranges and species richness. Global Ecol. Biogeogr. 15 (6), 578–587.
Hanberry, B.B., He, H.S., Dey, D.C., 2012. Sample sizes and model comparison metrics for species distribution models. Ecol. Model. 227, 29–33.
Hanberry, B.B., He, H.S., 2013. Prevalence, statistical thresholds, and accuracy assessment for species distribution models. Web Ecol. 13 (1), 13–19.
Hanberry, B.B., 2023a. Effectiveness of land use and disturbance measures, compared to climate, soil, and topography, for modeling relative abundances of tree species in the eastern United States. Ecol. Inform. 75, 102110.
Hanberry, B.B., 2023b. A solution for perfect bioclimate envelopes that are imperfect for extirpated species. Environ. Res.: Ecology 2 (2), 025005.
Hanberry, B.B., 2024. Practical guide for retaining correlated climate variables and unthinned samples in species distribution modeling, using random forests. Ecol. Inform. 79, 102406.

Hawkins, B.A., Rueda, M., Rodríguez, M.Á., 2008. What do range maps and surveys tell us about diversity patterns? Folia Geobot. 43, 345–355.
Herkt, K.M.B., Skidmore, A.K., Fahr, J., 2017. Macroecological conclusions based on IUCN expert maps: a call for caution. Global Ecol. Biogeogr. 26 (8), 930–941.
Hijmans, R., 2024. Terra: spatial data analysis. R package version 1, 7–81. Available online at: https://github.com/rspatial/terra.
Hughes, A.C., Dorey, J.B., Bossert, S., Qiao, H., Orr, M.C., 2024. Big data, big problems? How to circumvent problems in biodiversity mapping and ensure meaningful results. Ecography 2024 (8), e07115.
Hurlbert, A.H., White, E.P., 2005. Disparity between range map-and survey-based analyses of species richness: patterns, processes and implications. Ecol. Lett. 8 (3), 319–327.
Hurlbert, A.H., Jetz, W., 2007. Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. Proc. Natl. Acad. Sci. 104 (33), 13384–13389.
iNaturalist, 2020. To Obscure, make private or leave your observation Public/Open? https://www.inaturalist.org/posts/34066-to-obscure-make-private-or-leave-your-observation-public-open. (Accessed 16 February 2025).
iNaturalist, 2025. What is geoprivacy? What does it mean for an observation to be obscured? https://www.inaturalist.org/pages/archived+help#geoprivacy. (Accessed 16 February 2025).
International Union for Conservation of Nature [IUCN], 2024. The IUCN red list of threatened species. Spatial Data Download. https://www.iucnredlist.org/resources/spatial-data-download. (Accessed 25 March 2025).
IUCN Standards and Petitions Committee, 2024. Guidelines for using the IUCN red list categories and criteria. Version 16. Prepared by the standards and petitions Committee. Downloadable from. https://www.iucnredlist.org/documents/RedListGuidelines.pdf. (Accessed 27 March 2025).
Jetz, W., Sekercioglu, C.H., Watson, J.E., 2008. Ecological correlates and conservation implications of overestimating species geographic ranges. Conserv. Biol. 22 (1), 110–119.
Karger, D.N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R.W., Zimmermann, N.E., Linder, H., Kessler, M., 2017. Climatologies at high resolution for the Earth land surface areas. Sci. Data 4, 170122.
Karger, D.N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R.W., Zimmermann, N.E., Linder, H., Kessler, M., 2018. Data from: climatologies at high resolution for the earth's land surface areas. EnviDat. https://doi.org/10.16904/envidat.228.v2.1.
Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Software 28, 1–26.
Lobo, J.M., Jiménez-Valverde, A., Real, R., 2008. AUC: a misleading measure of the performance of predictive distribution models. Global Ecol. Biogeogr. 17 (2), 145–151.
Marcer, A., Chapman, A.D., Wieczorek, J.R., Xavier Picó, F., Uribe, F., Waller, J., Ariño, A.H., 2022. Uncertainty matters: ascertaining where specimens in natural history collections come from and its implications for predicting species distributions. Ecography 2022 (9), e06025.
Marsh, C.J., Sica, Y.V., Burgin, C.J., Dorman, W.A., Anderson, R.C., del Toro Mijares, I., Vigneron, J.G., Barve, V., Dombrowik, V.L., Duong, M., Guralnick, R., et al., 2022. Expert range maps of global mammal distributions harmonised to three taxonomic authorities. J. Biogeogr. 49 (5), 979–992.
Mathias, P.V., Mendonça, C.V., Rangel, T.F., Diniz-Filho, J.A., 2004. Sensitivity of macroecological patterns of South American parrots to differences in data sources. Global Ecol. Biogeogr. 13 (3), 193–198.
McPherson, J.M., Jetz, W., 2007. Type and spatial structure of distribution data and the perceived determinants of geographical gradients in ecology: the species richness of African birds. Global Ecol. Biogeogr. 16 (5), 657–667.
Merow, C., Wilson, A.M., Jetz, W., 2017. Integrating occurrence data and expert maps for improved species range predictions. Global Ecol. Biogeogr. 26 (2), 243–258.
Mitchell, J., Monk, J., Laurenson, L., 2017. Sensitivity of fine-scale species distribution models to locational uncertainty in occurrence data across multiple sample sizes. Methods Ecol. Evol. 8 (1), 12–21.
Mod, H.K., Scherrer, D., Luoto, M., Guisan, A., 2016. What we use is not what we know: environmental predictors in plant distribution models. J. Veg. Sci. 27, 1308–1322.
Moudrý, V., 2015. Modelling species distributions with simulated virtual species. J. Biogeogr. 42 (8).
R Core Team, 2024. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Rotenberry, J.T., Balasubramaniam, P., 2020. Connecting species' geographical distributions to environmental variables: range maps versus observed points of occurrence. Ecography 43 (6), 897–913.
Santini, L., Benítez-López, A., Maiorano, L., Čengić, M., Huijbregts, M.A., 2021. Assessing the reliability of species distribution projections in climate change research. Divers. Distrib. 27 (6), 1035–1050.
Thuiller, W., Münkemüller, T., 2010. Habitat suitability modeling. In: Moeller, A.P., Fielder, W., Berthold, P. (Eds.), Effects of Climate Change on Birds. Oxford University Press, pp. 77–85.
Titeux, N., Henle, K., Mihoub, J.B., Regos, A., Geijzendorffer, I.R., Cramer, W., Verburg, P.H., Brotons, L., 2016. Biodiversity scenarios neglect future land-use changes. Glob. Change Biol. 22, 2505–2515.
WWF, 2018. In: Grooten, M., Almond, R.E.A. (Eds.), Living Planet Report - 2018: Aiming Higher. WWF, Gland, Switzerland.
Zizka, A., Carvalho, F.A., Calvente, A., Baez-Lizarazo, M.R., Cabral, A., Coelho, J.F.R., Colli-Silva, M., Fantinati, M.R., Fernandes, M.F., Ferreira-Araújo, T., Moreira, F.G.L., 2020. No one-size-fits-all solution to clean GBIF. PeerJ 8, e9916.