



OPEN ACCESS

EDITED BY
Emilio Berti,
German Centre for Integrative
Biodiversity Research (iDiv), Germany

REVIEWED BY
Lifei Wang,
University of Toronto, Canada
Felipe Nunes,
Center for Territorial Intelligence, Brazil

*CORRESPONDENCE
Brice B. Hanberry
✉ brice.hanberry@usda.gov

RECEIVED 16 March 2026
REVISED 11 May 2026
ACCEPTED 12 May 2026
PUBLISHED 03 June 2026

CITATION

Hanberry BB (2026) Beyond thinning:
equalized sampling based on global
sampling intensity gradients for species
distribution models.
Front. Ecol. Evol. 14:1831501.
doi: 10.3389/fevo.2026.1831501

COPYRIGHT

© 2026 Hanberry. This is an open-access
article distributed under the terms of the
[Creative Commons Attribution License](#)
(CC BY). The use, distribution or
reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication
in this journal is cited, in accordance
with accepted academic practice. No
use, distribution or reproduction is
permitted which does not comply with
these terms.

Beyond thinning: equalized sampling based on global sampling intensity gradients for species distribution models

Brice B. Hanberry *

USDA Forest Service, Rocky Mountain Research Station, Rapid City, SD, United States

Spatial thinning of samples to remove clustered georeferenced records at local scales has become standard practice for species distribution models. Alternatively, I demonstrated the sampling issue of unequal sampling at global scales, and provided and tested a solution by equalizing samples. First, I used 17,200 species to characterize intensity gradients, with kernel density grids of georeferenced records, for 12 taxa. Secondly, I developed a method to even sampling intensity, by dividing kernel density grids into 10 quantiles of sampling intensity classes by taxa, and for each species, I determined the midpoint value of sample number in sampling intensity classes and removed records that exceeded the midpoint number. Thirdly, I compared species distribution models from equalized, unthinned, and thinned records, including locational distance uncertainty of records, to models from range maps. Only bird species deviated from greatest sampling intensity in Europe. Equalization reduced 42% of 140 million records to 15% of 50 million records in the greatest intensity sampling class. Models from equalized sampling maximized both the intersection and area predicted as present with models from range maps, with limited overprediction relative to models from thinned records. No sample processing approach generated models with the coverage of wide-ranging species. Models from any distance uncertainties were similar, but inclusion of records with distance uncertainties of 5 km resulted in even greater concentration of records in Europe and predicted area concentrated in Europe, for non-equalized records. Models from equalized sampling were able to avoid greater concentration of samples and omission error, due to recruiting records from a range of distance uncertainties in less sampled areas to increase predicted areas of presence, relative to unthinned samples, and reduce commission error, relative to thinned samples.

KEYWORDS

bias, distance uncertainty, filtering, intensity, thinning, virtual species

1 Introduction

Sampling intensity of species occurrence records is unequal at global scales (Meyer et al., 2016; Hughes et al., 2021; Daru and Rodriguez, 2023; Garcia-Rosello et al., 2023). Concentrated sampling efforts routinely occur in Europe, and typically for most taxa, in urban centers of the United States and western Australia (Meyer et al., 2016; Hughes et al., 2021; Daru and Rodriguez, 2023; Garcia-Rosello et al., 2023). However, accounting for

differential sampling intensity is an issue for any research that applies species records, such as species richness grids and species distribution models, because the distributions reflect sampling effort (Hughes et al., 2021). That is, modeling algorithms will model the underlying sampling gradients for Eurasian species, rather than the underlying species distributions, and will predict that Eurasian species are concentrated in Europe, which are the correct models based on the input samples (Fourcade et al., 2013).

To date, correction for sampling bias has been more focused on localized than global scales, with concerns about clustered and urban samples resulting from sampling at accessible locations rather than continental gradients in sampling intensity (Aiello-Lammens et al., 2015; Zizka et al., 2020). Spatial thinning or filtering of georeferenced records has become standard practice, by imposing a minimum distance between georeferenced records, often limited to one record per grid cell (< 1 km) or within 100 km (Lamboley and Fourcade, 2024). Distance remains a researcher choice because no specific distance has been identified to improve models (Lamboley and Fourcade, 2024). In fact, this sampling approach has not been proven to improve models (Gábor et al., 2020; Ten Caten and Dallas, 2023; Lamboley and Fourcade, 2024; Moudry et al., 2024). Another general approach entails clustering absence samples near georeferenced records, which preserves presence samples (Barber et al., 2022; Pili et al., 2025). However, targeted pseudoabsences, or generating absence samples where georeferenced records are most concentrated, enforce overlap in predictor variable values between presence and absence samples, preventing differentiation of presence and absence samples (Chefaoui and Lobo, 2008). The modeling algorithm in consequence will underpredict presence, not only where species definitively are present, but throughout predictions of species distributions, and overpredict presence outside of presence samples, creating unreliable, inverse models (Chefaoui and Lobo, 2008).

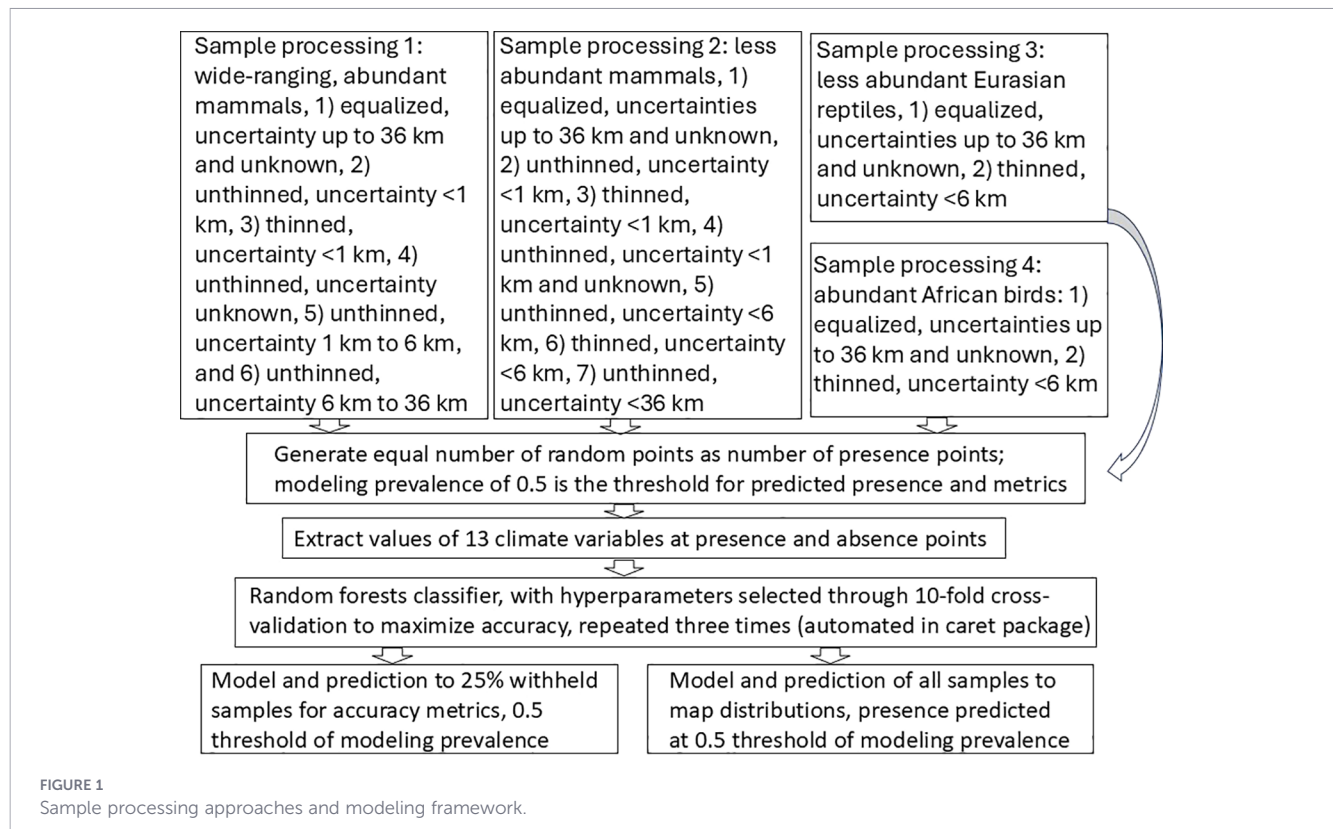
Relatedly, because locational inaccuracy creates obvious potential for error, samples often are filtered to remove samples with distance uncertainties that exceed acceptable distance uncertainties. While there is not any particular consensus about the maximum tolerable uncertainty, the recommendation is to match the spatial resolution of the predictor variables with presence samples (Moudry et al., 2023; Smith et al., 2023). A general match with spatial resolution of climate data of 30 second resolution would mean excluding georeferenced records with unknown distance uncertainties and distance uncertainties greater than 1 km, to prevent spatial mismatch in resolutions (Moudry et al., 2023).

Thinning of clustered samples and filtering based on locational distance uncertainties both reduce sample sizes. Thinning samples clustered at local scales will result in more generalized species distribution models, which may be beneficial (Hanberry, 2025a). Nevertheless, proactive removal of samples to prevent potential errors may increase the risk of outcomes that sample removal was intended to avoid (Hanberry, 2025a). Typically, small sample sizes are a clear way to increase commission errors and probability of unrepresentative models (Gábor et al., 2020; Lamboley and Fourcade, 2024; Hanberry, 2025a).

Removal of adjacent, local samples is not a remedy for uneven continental sampling intensities, particularly for species distributions

that extend from Europe, most of which is highly sampled, across Asia, which has sampling deficits (Hughes et al., 2021; Daru and Rodriguez, 2023; Garcia-Rosello et al., 2023). The concentration differential between Europe and Asia can be a real issue for species with Eurasian ranges, resulting in species distribution models that also are concentrated in Europe, missing the distributions in Asia (Fourcade et al., 2013). Removing clustered samples does not reduce concentration where thousands or tens of thousands of records are spread across Europe and results in loss of the opportunistic and accessible tens or hundreds of records from rarely sampled locations. The sampling concentration imbalance cannot be rectified by thinning of records with the same specified distance, without adjusting for different sampling intensities. An indirect approach would be to specify different sample removal distances by continent (Aiello-Lammens et al., 2015).

Instead, a direct method is to adjust georeferenced records for uneven sampling concentration based on stratification of global sampling intensity, by taxa. My objective was to evaluate models from unthinned and thinned records and equalized records, using global sampling intensity gradients, and including reported locational distance uncertainty of records, compared to models from complete samples of range maps, with an ensemble species distribution modeling framework. To accomplish the overall goal of solving the problem to equalize sampling intensity in georeferenced records, I first isolated sampling intensities for mammals, anurans, scaled reptiles, and (resident) birds; plants; and invertebrate species of odonates, butterflies and moths, flies, bees, wasps, beetles, and spiders. Secondly, I presented a method to address the sampling issue of uneven sampling gradients. Thirdly, I evaluated the method by applying four datasets of species distribution models, each dataset varying in range area and number of samples or taxa and continents: wide-ranging abundantly recorded mammal species located in a variety of continents, mammal species with more limited number of records and any continents, reptile species in Eurasia, and bird species in Africa (Figure 1). Sample processing approaches included different combinations of equalized, unthinned, and thinned records and distance uncertainties of records. While equalizing reduces samples where sampling is more abundant, equalizing also allows ordering of samples by distance certainty and use of samples with greater distance uncertainty where sampling intensity classes do not meet midpoint sampling intensities. To help alleviate the issue of sample sizes, generally only species with a minimum of 400 unique terrestrial georeferenced records, before sample removal, were included to generate robust species distribution models, particularly after sample removal (Gábor et al., 2020), although this may amplify sampling gradients. Selected species for comparisons had at least 70% intersection of georeferenced records with species ranges developed by experts (BirdLife International (BLI), 2023; International Union for Conservation of Nature (IUCN), 2025). Because objective reality is based on approximations, the representativeness of species distribution models from sample processing approaches was judged by percentage intersection in predicted area and percentage predicted area relative to species distribution models from range map samples (i.e., intersection and ratio indices; BLI, 2023; IUCN, 2025). In general, ranges encompass greater areas



than sampled areas (Hurlbert and White, 2005; Jetz et al., 2008); therefore, predicted areas from ranges should exceed predicted areas from occurrence records, unless overprediction of presence areas ensues due to information deficits in the georeferenced records about species distributions. Although the objective was to focus on real sampling of species and not to compare modeling algorithms, I provided an example of this method applied to a virtual species and with the Maximum Entropy (Maxent) model and extreme gradient boosting as alternative modeling algorithms (Supplementary File 1).

2 Methods

2.1 Data

Taxa included terrestrial mammals, anurans, scaled reptiles, and (resident) birds; plants; and invertebrate species of odonates, butterflies and moths, bees, wasps, beetles, flies, and spiders. The Global Biodiversity Information Facility (GBIF, 2025) is a rapidly increasing centralized data repository of georeferenced records of species observations, currently with about two billion species occurrence records from 60,000 datasets. Georeferenced records occurred during 1990 to 2024, and into the first quarter of 2025, with unique terrestrial coordinates (GBIF, 2025; <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.nzcwcd>; <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.uaskaz>; <https://doi.org/10.15468/dl.jhu36x>; <https://doi.org/10.15468/dl.6fehdk>; <https://doi.org/10.15468/dl.f4xa7r>; <https://doi.org/10.15468/dl.mnzzxe>; <https://www.gbif.org/occurrence/download/0009004-250325103851331>; <https://www.gbif.org/occurrence/download/0009319-250325103851331>; <https://www.gbif.org/occurrence/download/0009304-250325103851331>; <https://www.gbif.org/occurrence/download/0009004-250325103851331>; <https://www.gbif.org/occurrence/download/0009338-250325103851331>). Location distance uncertainties were < 36 km or unknown for most taxa, except unknown distance uncertainties alone or combined with < 1 km for bird species (due to the preponderance of records with unknown uncertainties for birds; Hughes et al., 2021) and < 10 km and unknown distance uncertainties for plants (immobile). Most vertebrate and plant species had at least 400 unique georeferenced records and range areas of at least 10,000 square kilometers, but I also added 220 vertebrate and vascular plant species to represent smaller range areas or less than 400 records. Odonates, butterflies and moths, bees, wasps, beetles, flies, and spiders were filtered to at least 400 unique terrestrial georeferenced records. In total the following characterized the intensity gradients: 8,748,181 records of 905 mammal species, 3,194,352 records of 529 reptile species, 2,295,736 records of 373 anuran species, 42,187,155 records of 1505 bird species, 13,971,078 records of 873 plant species, 3,266,841 records of 225 odonate species, 44,945,531 records of 5700 butterflies and moth species, 4,032,263 records of 1613 fly species, 3,239,902 records of 582 bee species, 1,532,043 records of 586 wasp species, 9,027,018 records of 3371 beetle species, and 2,707,565 records of 963 spider species.

Location distance uncertainties were < 36 km or unknown for most taxa, except unknown distance uncertainties alone or combined with < 1 km for bird species (due to the preponderance of records with unknown uncertainties for birds; Hughes et al., 2021) and < 10 km and unknown distance uncertainties for plants (immobile). Most vertebrate and plant species had at least 400 unique georeferenced records and range areas of at least 10,000 square kilometers, but I also added 220 vertebrate and vascular plant species to represent smaller range areas or less than 400 records. Odonates, butterflies and moths, bees, wasps, beetles, flies, and spiders were filtered to at least 400 unique terrestrial georeferenced records. In total the following characterized the intensity gradients: 8,748,181 records of 905 mammal species, 3,194,352 records of 529 reptile species, 2,295,736 records of 373 anuran species, 42,187,155 records of 1505 bird species, 13,971,078 records of 873 plant species, 3,266,841 records of 225 odonate species, 44,945,531 records of 5700 butterflies and moth species, 4,032,263 records of 1613 fly species, 3,239,902 records of 582 bee species, 1,532,043 records of 586 wasp species, 9,027,018 records of 3371 beetle species, and 2,707,565 records of 963 spider species.

2.2 Sampling intensity and adjustment

By taxa, I determined sampling intensity with the kernel density tool in ArcGIS Pro (Environmental Systems Research Institute (ESRI), Redlands, CA, USA; [ESRI, 2025](#)). This kernel function (i.e., smoothing distributions) calculates a grid of magnitude per unit area from point features. A similar function is available in the free and open source [QGIS \(2025\)](#). For specifications in the kernel density tool, I selected a cell size of 0.5 degrees (around 55 km length at the equator), which was 1) approximately the default value, 2) a slightly finer resolution than the 1 degree resolution for macroecological species richness studies ([Hurlbert and Jetz, 2007](#); [Ballesteros-Mejia et al., 2017](#)), 3) within a typical distance range for spatial thinning ([Lambole and Fourcade, 2024](#)), and more likely to contain samples than the minimum potential resolution of climate data (30 seconds). The planar option with no projection of the point features processed data efficiently. I reclassified the kernel density grid for each taxa into 10 quantiles of sampling intensity classes, although the least class essentially was unsampled areas.

For equalizing samples, I determined the midpoint value of sample number for each species in sampling intensity classes by taxa. Not all species were sampled in all sampling intensity classes. For each species, the midpoint value was the maximum value of either the 1) mean of the samples in each sampling intensity class, 2) median of the samples in each sampling intensity class, or 3) maximum of the samples relative to number of sampling intensity classes, but excluding the greatest and least sampling intensity classes (the least sampling intensity class consistently was unsampled). I did not make any adjustment (i.e., kept the maximum number of samples) for species that had a range of sampling intensity classes of three or less or only sampled in the greatest sampling intensity class; species with at least 90% of records in one class; and species that had a maximum value of 100 georeferenced records in a sampling intensity class. Likewise, for species with an abundance of records, I set the maximum value at 3500. After determining the maximum midpoint value by species, I then reduced samples for any sampling intensity classes with more than that number of samples.

Within a sampling intensity class, I sorted records by least distance uncertainty, with unknown uncertainty set at 1999 meters; years, favoring in descending order from 2010s, 2000s, and the combined 1990s and 2020s; and then randomly generated numbers. This was to isolate the highest quality records, in terms of locational accuracy, and with a preference to reduce spread in time of recorded observations for modeling but also preferring any observations to no observations. In addition, for species that still had more than 30,000 records after equalizing classes (out of a maximum of 31,500 samples, or 3500 for nine classes with samples), I randomly selected 30,000 records.

After equalizing the sampling intensity classes, by reducing samples for each class to midpoint values for every species, I reiterated the kernel density function. I reclassified the second kernel density grid with the same classes as the first kernel density grid, with all samples. I compared the number of samples in each class before and after sample reduction.

2.3 Species distribution models with different sample processing

First, I tested species distribution models of different samples from sampling approaches and distance uncertainties, for 21 abundant, wide-ranging mammal species, testing coverage of species distribution models relative to species ranges. Species had at least 70% of georeferenced records that intersected with range maps and at least 1000 records for each model, regardless of species, distance uncertainties, and sampling approach. I compared species distribution models from georeferenced records that were 1) equalized, with a range of distance uncertainties up to 36 km and unknown distance uncertainties, 2) unthinned and distance uncertainty less than 1 km, 3) thinned and distance uncertainty less than 1 km, 4) unthinned and distance uncertainty unknown only, 5) unthinned and distance uncertainty 1 km to 6 km, and 6) unthinned and distance uncertainty 6 km to 36 km relative to 7) 1000 random samples of range maps.

The next comparison was of 21 less abundant mammal species, with at least 400 but less than 1000 records in the unthinned 1 km distance uncertainty category, with at least 70% of georeferenced records intersected with range maps. This illustrated the effects of increasing sample sizes by including additional distance categories and also potential issues related to smaller sample sizes. I compared species distribution models from georeferenced records that were 1) equalized, with a range of distance uncertainties up to 36 km and unknown distance uncertainties, 2) unthinned and distance uncertainty less than 1 km, 3) thinned and distance uncertainty less than 1 km, 4) unthinned and distance uncertainty less than 1 km and unknown, 5) unthinned and distance uncertainty less than 6 km, 6) thinned and distance uncertainty less than 6 km, 7) unthinned and distance uncertainty less than 36 km relative to 8) 1000 random samples of range maps for the species.

The third and fourth comparisons for reptile species in Eurasia and bird species in Africa were of species distribution models from equalized sampling and thinned records of distance uncertainties less than 6 km and unknown distance uncertainties, relative to 1000 random samples of range maps for the species. Three reptile species had less than 400 records for thinned samples, out of 33 reptile species with ranges in Europe, Asia, or Eurasia. Only one bird species had less than 1000 records for thinned samples and three bird species had less than 1000 records for equalized samples, out of 65 bird species in Africa. At least 70% of georeferenced records intersected with range maps.

Pseudoabsence samples were generated for the continent or continents of the species distributions in georeferenced records. For the mammal species from diverse continents, I limited *Myotis emarginatus* to Eurasia for comparison because georeferenced records did not include northern Africa. If a continent or two continents were a particular focus, then I limited comparisons to those extents (i.e., African birds, Eurasian reptiles).

Predictor variables were 13 climate variables during 1981–2010 clipped to continents (resolution 30 arc seconds; Climatologies at High resolution for the Earth's Land Surface Areas (CHELSA) climate dataset; [Karger et al., 2017, 2018](#)). The resolution of 30 arc seconds (< 1 km) was tested against the reported distance

uncertainties of the georeferenced records. Predictors were proximal temperature and precipitation variables at month, quarterly, and annual intervals (untransformed values; [Elith and Leathwick, 2009](#)): 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. Correlated climate variables are not an issue for predictions for linear or nonlinear models ([Hanberry, 2024](#)).

For modeling of each species, I applied the random forests classifier developed for predictive modeling, to differentiate between two or more classes ([Breiman, 2001](#)). Random forests is a standard non-linear ensemble classifier that is highly accurate ([Fernández-Delgado et al., 2014](#)) and among the most commonly applied classifiers for species distribution modeling ([Santini et al., 2021](#)). While not necessary given general robustness of any value for number of climate variables, the one minor hyperparameter of number variables at each split was established through 10-fold cross-validation, repeated three times (automated procedure in the caret package; [Kuhn, 2008](#); [Probst et al., 2019](#); [R Core Team, 2025](#)). Models were based on an equal number of presence and absence samples (i.e., modeling prevalence of 0.5), which was the modeling basis of the 0.5 threshold between classification as present or absent, which as another benefit allows comparability among models ([Hanberry and He, 2013](#)). All methods and variables were held constant except for variation in samples, based on filtering by distance uncertainties, thinning, and equalizing.

I modeled one time with all samples for predictions and one time with 75% of the samples to determine how well the model predicted the withheld 25% of samples. The modeling prevalence, or ratio of presence and absence points, was 0.5, which is the basis for establishing a threshold for presence at 0.5 predicted probability. To establish that the modeling algorithm was able to accomplish the

technical task of differentiating presence and absence classes for modeling, I selected three accuracy metrics of balanced accuracy (mean of sensitivity and specificity), sensitivity or true positive rate, and specificity or true negative rate.

Comparisons of representativeness were to species distribution models from range maps based on predicted areas of presence (probabilities ≥ 0.50 , based on modeling prevalence of the presence and absence classes; areas calculated in terra R package; [Hijmans, 2024](#)). To detect overprediction, comparing predicted areas from species distributions to predicted areas from species ranges rather than predicted areas from species distributions directly to range maps will hold comparisons as constant as possible (*ceteris paribus*), including modeling artefacts. One comparison index was percentage area of intersection between species distribution models from the sample processing approaches and species distribution models from range maps, relative to the area of species distribution models from the sample processing approaches, referred to as the intersection index, an isolation of commission error. Another index was percentage predicted area of species distribution models (i.e., total area, not the intersection) from the sample processing approaches relative to predicted area of species distribution models from range maps, referred to as the ratio index, to indicate both omission error by underfilling and commission error. Ratios equal to or greater than 1 generally counted as overprediction, particularly combined with commission error registered in the intersection index of area that did not intersect.

3 Results

3.1 Sampling intensity gradients: before and after equalizing

For sampling intensity patterns, only bird species of the 12 taxa deviated from greatest sampling intensity in Europe ([Figures 2, 3](#)).

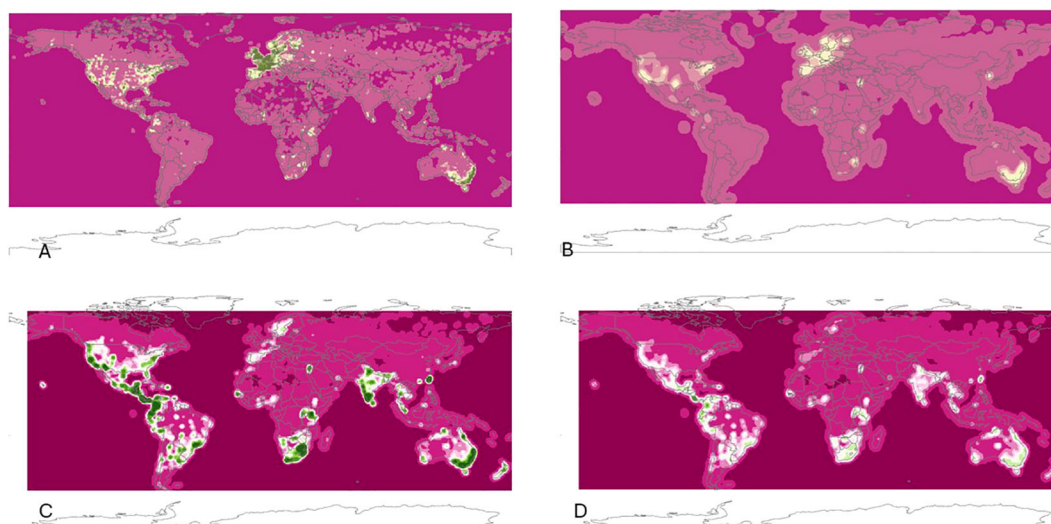


FIGURE 2

Mammal species before (A) and after (B) equalized sampling, which displays the typical concentration of sampling intensity in Europe, with the exception of bird species before (C) and after (D) equalized sampling. (Green designates greatest sampling intensity, white and yellow designate moderate sampling intensity, and pinks designate least sampling intensity).

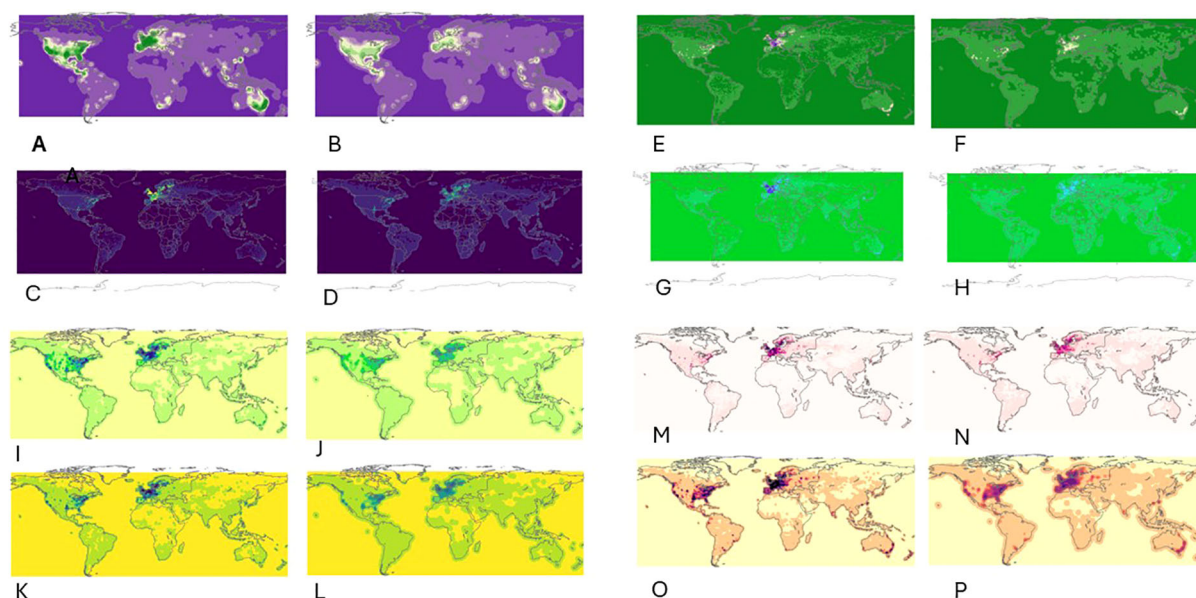


FIGURE 3

Reptile species before (A) and after (B) equalized sampling, which was very similar to anuran sampling intensity (green designates greatest sampling intensity, yellow designates moderate sampling intensity, and purple designates least sampling intensity). Butterfly and moth species before (C) and after (D) equalized sampling, which was very similar to fly sampling intensity (yellow designates greatest sampling intensity, blue designates moderate sampling intensity, and purple designates least sampling intensity). Plant species before (E) and after (F) equalized sampling (purple designates greatest sampling intensity, yellow designates moderate sampling intensity, and green designates least sampling intensity). Odonate species before (G) and after (H) equalized sampling (dark blue designates greatest sampling intensity, light blue designates moderate sampling intensity, and green designates least sampling intensity). Bee species before (I) and after (J) equalized sampling and wasp species before (K) and after (L) equalized sampling (dark blue designates greatest sampling intensity, light blue designates moderate sampling intensity, and yellow designates least sampling intensity). Beetle species before (M) and after (N) equalized sampling and spider species before (O) and after (P) equalized sampling (dark colors designate greatest sampling intensity and light colors designate least sampling intensity).

Sampling of bird species had concentrated locations in every continent but Europe. For the 11 other taxa, sampling was concentrated in Europe with variations based on the amount of area covering the United States, southwestern Australia, and sampling hotspots in smaller areas. Reptiles and anurans had the largest areas of great sampling intensities.

Before equalizing, about 42% of almost 140 million records were in the greatest intensity sampling class, ranging by taxa from 37% for plants to 59% for anurans. After equalizing, which was done uniquely within taxa, 15% of more than 50 million records were in the greatest intensity sampling class, ranging by taxa from 11% for plants to 31% for anurans. Sampling gradients were reduced, not eliminated, because this would require reducing samples to the least intensity sampling class, excluding the non-sampled class.

3.2 Species distribution models

The first comparison, for 21 mammal species that were abundant and wide-ranging (mean predicted range area of 13.4 million km² from range maps), allowed comparison of species distribution models for range size coverage with sample sizes unlikely to have overprediction issues for six sample processing approaches (minimum sample size of 1000 records): equalized sampling, with a range of distance uncertainties up to 36 km and unknown distance uncertainties; unthinned samples with different distance uncertainties of less than 1 km, 1 km to 6 km, 6 km to 36 km, and unknown; and thinned samples, distance uncertainty less than 1 km. Mean

accuracies for species distribution models for all six sampling approaches from georeferenced records and species distribution models from randomly sampled range maps ranged from 0.95 (IUCN ranges) to 0.97 (unthinned and distance uncertainty 1 km to 6 km), with sensitivity values of 0.96 to 0.98 and specificity values of 0.94 to 0.96. Models from all of the six sample processing approaches generated species distribution models that had similar values for the two area indices relative to species distribution models from range maps, despite distance uncertainties. For the intersection index, a mean of 91% (unthinned and distance uncertainty 6 km to 36 km) to 93% (unthinned and distance uncertainty less than 1 km) of the predicted presence area from sample processing approach models intersected the predicted areas of models from range maps. However, for the ratio index, models from sample processing approaches were unable to achieve the coverage of wide-ranging species, predicting a mean presence area of 17% (unthinned and distance uncertainty less than 1 km) to 25% (equalized, with a range of distance uncertainties up to 36 km and unknown distance uncertainties; and unthinned and distance uncertainty unknown) of the predicted areas from models of range maps.

The next comparison was of 21 less abundant mammal species, with at least 400 but less than 1000 records at the 1 km distance uncertainty category (i.e., potential overprediction due to sample size), and which had a mean predicted range area of 5.4 million km² from range maps. Mean accuracies for species distribution models for all seven sampling approaches from georeferenced records and

species distribution models from randomly sampled range maps ranged from 0.94 (6 km thinned) to 0.96 (unthinned and distance uncertainty less than 1 km and unknown distance uncertainties; unthinned and distance uncertainty less than 6 km), with sensitivity values of 0.96 to 0.97 and specificity values of 0.92 to 0.94. Some separation occurred in model predictions, depending on approach (Figure 4). For the intersection index, species distribution models from the thinned records with distance uncertainty of less than 1 km had a mean of 80% predicted presence area that intersected with predicted areas of models from range maps, indicating potential commission error due to 20% of area outside of the range maps. Conversely, species distribution models from equalized records and from unthinned records with distance uncertainty less than 1 km and unknown distance uncertainties had a mean of 87% predicted presence area that matched the predicted areas of range maps. For the ratio index, the species distribution models from the thinned records with distance uncertainty of less than 1 km also predicted the greatest mean presence area, at 69% of the predicted areas from models of range maps, albeit with 20% of area outside of the range maps, according to the intersection index. However, models for three species out of 21 species (15% of species) exceeded the predicted areas of range maps, with one model at 150% of the predicted area of the range map (Figure 4). In contrast, species distribution models from the equalized records predicted 58% of the predicted areas of range maps, with one model that exceeded (139%) the predicted area of range map and one model of the same predicted area (100%).

Visualization allowed illustration of the sampling issue, for the example of *Myotis emarginatus*, which has a range that extends across the steep sampling gradient from heavily sampled Europe into lightly sampled Asia (Figure 5). Equalized sampling was able to avoid concentration of samples by continent. Conversely, thinning to within one climate raster cell was not able to prevent

concentration when even more samples were added from Europe, specifically with distance uncertainty of 5 km. Models from equalized sampling did not predict as much of the range area, inside or outside of the range, as models from unthinned and thinned samples with distance uncertainties less than 1 km.

The third comparison was for species distribution models of 33 reptile species in Eurasia, generally with at least 400 samples (i.e., potential overprediction due to sample size), which had a mean predicted area of 4.3 million km² from range maps. The two sampling approaches were equalized samples and thinned georeferenced records with unknown distance uncertainties and distance uncertainties up to 6 km. Mean accuracies ranged from 0.97 to 0.98 with sensitivity values of 0.98 and specificity values of 0.96 to 0.97. For the intersection index, mean predicted presence area was 85% and 89% for species distribution models from thinned and equalized samples, respectively, that intersected the predicted areas of models from range maps. For the ratio index, the species distribution models from the thinned records also predicted the greatest mean presence area, at 82% of the predicted areas of models from range maps. However, models from thinned records for six species (18% of species) met or exceeded the predicted areas of models from range maps, with one model at 372% of the predicted area of the range map. Species distribution models from the equalized records predicted 68% of the predicted areas of range maps, with three models that exceeded (maximum by 139%) the predicted area of range maps.

The fourth comparison was for 51 bird species in Africa, generally with at least 1000 samples, which had mean predicted area of 5.7 million km² from range maps. The two sampling approaches were thinned georeferenced records with unknown distance uncertainties and distance uncertainties up to 6 km and equalized samples. Mean accuracies ranged from 0.93 to 0.96, with sensitivity values of 0.96 to 0.97 and specificity values of 0.91 to 0.95.

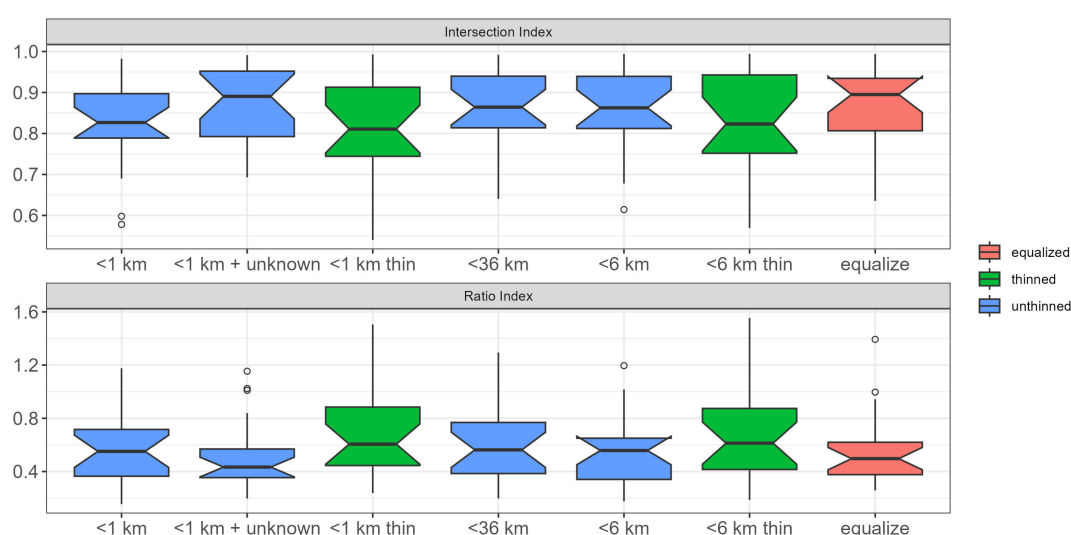


FIGURE 4

Models from equalized sampling were able to maximize both the intersection with predicted area from random samples of IUCN ranges, while also maximizing area predicted as present without overpredicting (i.e., ratio greater than 1), which is the issue with thinned samples. Boxplots: the horizontal line in the box is the median, the lower/upper hinges of the box boundaries are the 25th and 75th percentiles, whisker extremes represent minimum and maximum values (within 1.5 of the interquartile range), and small circles represent outliers.

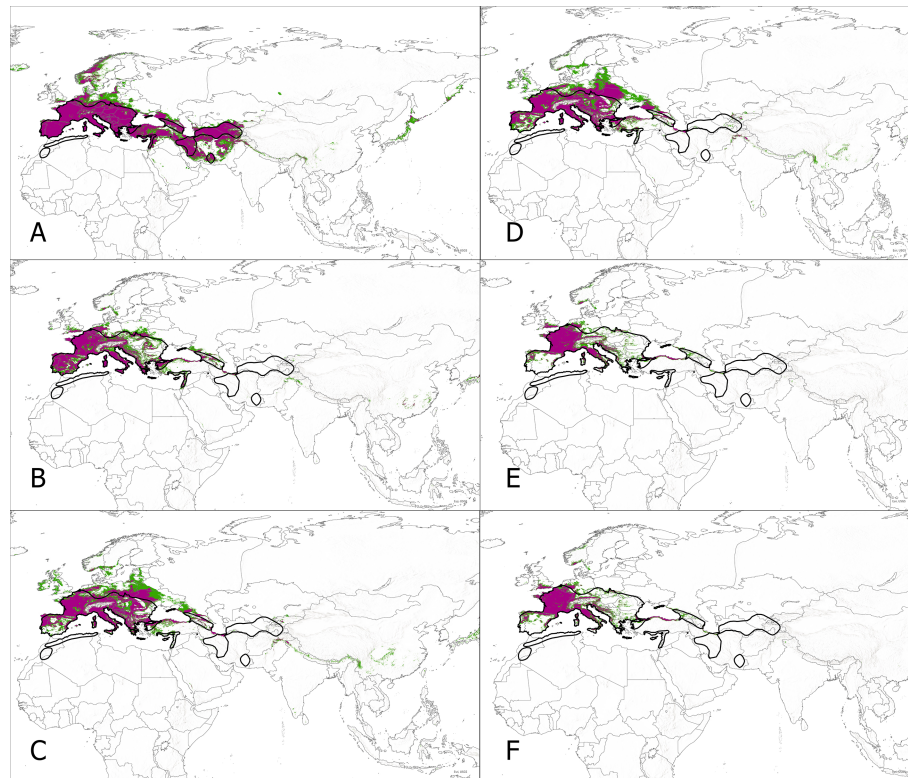


FIGURE 5

Illustration of the sampling issue, for *Myotis emarginatus*, which has a range that extends across the steep sampling gradient from heavily sampled Europe into lightly sampled Asia [(A); predicted presence area from species distribution model from random samples of the IUCN range map]. None of the sampling approaches were successful in achieving range coverage (species distribution models with 18% to 53% of predicted presence area from the species distributions of the range map, with 81% to 99% intersection with the area from the species distributions of the range map. Equalized samples (B), unthinned samples and distance uncertainty less than 1 km (C), and thinned samples and distance uncertainty less than 1 km (D) avoided the concentrated models of thinned and distance uncertainty less than 6 km (E), and unthinned and distance uncertainty less than 36 km (F), when even more samples, with distance uncertainty of 5 km specifically, were included from Europe. Thinning does not prevent concentration by continent, unlike equalizing, which included a range of distance uncertainties up to 36 km and unknown distance uncertainties that were even by sampling intensity.

For the intersection index, mean predicted presence area was 86% for species distribution models from both thinned and equalized samples that intersected the predicted areas of models from range maps. For the ratio index, the species distribution models from the thinned and equalized samples predicted 41% and 43%, respectively, of the predicted areas of models from range maps. Only one species distribution models from the thinned samples exceeded (106%) the predicted area of range maps. No species distribution models from equalized samples exceeded the predicted area of range maps.

4 Discussion

Differential sampling is considered to be sampling bias because of overrepresentation of available yet suitable environmental conditions, with not all conditions being equally and randomly sampled (Inman et al., 2021). Thinning samples that are clustered locally, based on a minimum distance between occurrence records, is a method developed to reduce effects of concentrated samples (Lambole and

Fourcade, 2024). However, thinning is not clearly necessary or always an improvement for species distribution models (Gábor et al., 2020; Ten Caten and Dallas, 2023; Lambole and Fourcade, 2024; Moudry et al., 2024). A greater issue is that thinning will result in overprediction of species presence and added commission error in models, particularly after reducing already small samples (Hanberry, 2025a). Thinning based on one specified distance to reduce localized concentrations of clustered records is different than adjusting continental differentials in sampling intensity, as occurs with high sampling intensity for most taxa in Europe relative to the rest of the world (Fourcade et al., 2013; Meyer et al., 2016; Hughes et al., 2021; Daru and Rodriguez, 2023; Garcia-Rosello et al., 2023). Here, after determining sampling intensity grids, from kernel densities, for mammal, reptile, anuran, bird, plant, odonate, butterfly and moth, fly, bee, wasp, beetle, and spider taxa, only bird species out of the 12 taxa deviated from greatest sampling intensity in Europe. Application of the equalized sampling method, based on the maximum midpoint value of sampling intensity classes, for species distribution models demonstrated that equalized samples were able to avoid concentration of samples in Europe, while employing georeferenced records from a range of distance uncertainties. This allowed greater range

coverage than sampling approaches of concentrated unthinned records and limited georeferenced records based on distance uncertainties. Equalization also produced less overprediction of presence areas, relative to models from range maps, than models from thinned samples. Moreover, no sample processing approaches could replicate areas of wide-ranging species. To be clear, sampled areas with uneven sampling effort are suitable for predictive modeling, as opposed to testing hypotheses and deriving estimates for inferential statistical analysis, and sampled areas are not a problem relative to unsampled areas for modeling (Thuiller et al., 2004; McCarthy et al., 2012).

For all but bird taxa, sampling intensity was concentrated in Europe with variations based on sampling coverage in the United States, southwestern Australia, and sampling hotspots in smaller areas. The patterns for the 11 taxa were similar to the pattern in Hughes et al. (2021), who displayed 742 million records of 374,900 species, and also appeared similar to Garcia-Rosello et al. (2023), who illustrated 22 million records of 292,273 insect species. The overall concentration in Europe generally corresponded with Daru and Rodriguez (2023), who used 1.9 billion occurrence records of terrestrial plants, butterflies, amphibians, birds, reptiles and mammals. For bird species, sampling intensity varied from characteristic patterns resulting from sampling of species records, perhaps displaying underlying species richness patterns instead of underlying sampling patterns.

The steep sampling intensity gradient between Europe and Asia can be a large issue for species with Eurasian ranges (Fourcade et al., 2013). The sample concentration differential can result in species distribution models that also are concentrated in Europe, missing the distribution in Asia (Figure 5; Fourcade et al., 2013). At this time, the concentrated model outcome particularly arises from including georeferenced records with 5 km distance uncertainty, which recruits more records from European institutions that report a 5 km distance uncertainty as a standard (as apparent in the datasets and models), intensifying the gradient. Although thinning does produce more generalized models, not concentrated in modeling the samples like models from unthinned records, thinning did not correct for the greater sampling concentration in Europe when georeferenced records with distance uncertainties of less than 6 km were included. Perhaps the thinning distance could be matched to the distance uncertainty of the coordinates, but at the cost of necessary records outside of Europe if the same distance is applied. There is no one specified distance for thinning that accomplishes the intended purpose of evening samples at all distances, from localized clustering to continental concentrations (Lambole and Fourcade, 2024). Instead, different distances would need to be applied based on sampling intensities, but that distance is unknown. Essentially, thinning needs to be modified based on sampling intensities, as applied here based on sampling intensities to equalize samples.

In the future, perhaps more spatially balanced sampling will develop, if sampling by institutions is fully saturated in Europe. Conversely, contributions from community naturalists of georeferenced records, with little distance uncertainty, likely will continue at ever increasing rates as during the past decade (apparent in datasets), amplifying the current patterns of sampling. This

may enforce the concentrated model outcome to any locational distance uncertainty and perhaps in locations beyond Eurasia. Depending on taxa, irregular sampling hotspots occur at different locations, consistently in urban United States and southwestern Australia and in smaller countries or nature preserves, such as for mammals. In the future, concentrated models may become more probable with any sample processing approach that does not account for sampling intensity gradients, similar to the one presented here for equilibrating sampling intensities based on reducing samples from the most sampled areas and engaging samples from the least sampled areas, despite distance uncertainties greater than 1 km or unknown. Before equalizing, about 42% of almost 140 million records were in the greatest intensity sampling class, ranging by taxa from 37% for plants to 59% for anurans. After equalizing, which was done uniquely within taxa, 15% of more than 50 million records were in the greatest intensity sampling class, ranging by taxa from 11% for plants to 31% for anurans. Equalized sampling by design evaded concentrated sampling.

Although the equalization method maximized predicted areas of presence with limited overpredictions, none of the sampling approaches were successful in achieving range coverage, for wide-ranging species and in undersampled areas, such as parts of Asia. For wide-ranging species, species distribution models from sample processing approaches contained only 17% (unthinned and distance uncertainty less than 1 km) to 25% (equalized, with a range of distance uncertainties up to 36 km and unknown distance uncertainties; also unthinned and distance uncertainty unknown) of the predicted areas from models of range maps. While omission error occurred, commission error was minimal, with most of the predicted areas of presence from models of different sampling processing approaches contained with predicted areas models from range maps. Extensive omission error reinforces that sampling of conditions must occur for generation of representative species distribution models (McCarthy et al., 2012). No panacea exists to fix lack of georeferenced records from unsampled areas.

Previous research has found that species typically occupy about 40–65% of delineated ranges (Hurlbert and White, 2005; Jetz et al., 2008). For less wide-ranging species, species distributions from georeferenced records predicted better coverage of mammals and reptiles (55% to 75%, depending on the variety of sample processing approaches) than bird species in Africa (41% or 43% for two sample processing approaches), relative to predicted areas from range maps. Certainly, species do not occur at all locations and all times throughout their ranges (Hurlbert and White, 2005; Jetz et al., 2008). Within ranges, resources, disturbance, land use, and competition are some of the factors that influence areas of occupancy (Early and Sax, 2014). Moreover, georeferenced records characteristically have sampling deficits that limit achievement of modeling the full climate space of the species. Location (i.e., highly sampled Europe) likely has the greatest influence on sampling coverage. However, a variety of factors may influence sampling coverage, such as abundance and detectability of species (Yates et al., 2018; Wang and Jackson, 2023). Nonetheless, ranges provide a general outline of suitable climate for species, and with greater percentage of range

coverage comes greater commission error, particularly evident for models from thinned samples.

Modeling demonstrated overprediction issues for thinning, evaded by unthinned and equalized sampling approaches. I focused on robust sample sizes (commonly at least 400 records) to isolate sampling intensity effects, rather than to illustrate instances of small samples when thinning causes issues of overgeneralization and commission error due to information loss. Most publications of species distribution models use small samples of less than 50 observations (62% of 250 reviewed papers; Santini et al., 2021). One issue with small sample sizes and lack of information about conditions is overprediction of presence areas. Despite the attempt to avoid this issue, thinned samples generated models with greater presence areas that had less intersection with models from range maps, that is, commission error. Models from thinned samples had greater coverage than models from other sample processing approaches, but models from thinned samples were more prone to predict presence areas to a greater extent than predicted presence areas from models from range maps. The range maps and models from range maps were for large ranges, overall, and they encompassed at least 70% of georeferenced records. That is, ranges were generally correct as corroborated by georeferenced records through research triangulation (Denzin, 2012). Additionally, comparisons were species distribution models to species distribution models, so that models from range samples also can represent suitable climate beyond range area (Figure 5). It is unlikely for georeferenced records to cover a greater range of conditions than random samples from range maps (Hurlbert and White, 2005; Jetz et al., 2008). Even if species have expanded beyond range maps, and georeferenced records sample beyond the ranges, thinned samples were not accessing information about locations that were unavailable to unthinned samples or equalized samples.

Unthinned and equalized samples were less likely to overpredict area as much as predictions from thinned samples, which had information loss from sample reduction. Sample removal for equalized sampling was proportional to sampling intensity, minimizing cost of information loss, and also allowed for recruitment of records with greater distance uncertainties where information was lacking. Equalized sampling did require extra effort that may not be worthwhile for irregular sampling intensity gradients in most continents, at least that exist at this time. If equalized samples are not applied for species distribution modeling, the next best option appears to be modeling with unthinned records and distance uncertainties less than 1 km and unknown distance uncertainties, to maximize both the intersection and total area predicted as present with models from range maps, with limited overprediction and avoidance of concentrated models. Alternatively, unthinned records and distance uncertainties less than 6 km and unknown distance is another option, if avoiding species, aside from birds, with Eurasian ranges.

In addition to providing contrasts among equalized, unthinned, and thinned samples, I also incorporated the influence of locational distance uncertainty, in particular to recruit information available in records with unknown distance uncertainties and greater than 1 km. Georeferenced records with unknown distance uncertainties

and a range of distance uncertainties are available to help increase area sampled to attain more comprehensive models. Due to potential locational uncertainties and scale/resolution mismatches, records may be discarded (Smith et al., 2023). However, species distribution models from different distance uncertainty classes were similar. For the most part, reported distance uncertainties may reflect institutional standards because some institutions appeared to only have georeferenced records with distance uncertainties of 5 km, which then are unlikely to all be mis-located. That is, the records likely are more accurate than the maximum possible distance uncertainty, with no reason to exclude the records based on distance uncertainty. However, the distance uncertainties at about 30 km do have positional offsets built in by iNaturalist (2020) to protect specific locations. For recruitment of older records, spatial observations recorded before accessible satellite navigation may be georeferenced incorrectly based on vague site descriptions (Zizka et al., 2020). For bird species, inclusion of records with unknown distance uncertainties will permit accessing the preponderance of bird records (eBird, 2024; Cornell Lab of Ornithology), which do not contain a tag about distance uncertainty. Overall, it may be a mistake to exclude information from records with greater distance uncertainties because the distances uncertainties are not the actual error.

Moreover, modeling algorithms are robust to some locational error because they maximize accuracy (Hanberry, 2025b). Locations that are erroneous will not contribute to overall accuracy and will be discounted by the modeling algorithm. The more records that occur in proximity, the more probable that species ranges are being demarcated and that species occur in the vicinity. However, this may cause greater concentration of sampling intensities in well-sampled continents, even more, as demonstrated by models that include distance uncertainties less than 6 km, whether unthinned or unthinned (Figure 5). Equalized sampling can take advantage of a range of distance uncertainties without concentration, which resulted in greater coverage. With equalized sampling, whether or not distance uncertainties will concentrate sampling gradients is not an issue.

The method for equalizing samples has potential for improvement and refinement, to retain as many samples as possible by equalizing the least. Sampling gradients were reduced, not eliminated, because this would require reducing samples to the least intensity sampling class, excluding the non-sampled class. I developed a method to even unequal number of samples in sampling intensity classes, with kernel density grids of samples divided into 10 quantiles of sampling intensity classes, with the least class representing unsampled areas. For equalizing samples, I determined the midpoint value of sample number for each species in sampling intensity classes by taxa, as the maximum of either 1) the mean of the samples in each sampling intensity class, 2) the median of the samples in each sampling intensity class, or 3) maximum of the samples relative to number of sampling intensity classes, excluding the greatest and least sampling intensity classes. I did not adjust samples for species that had a range of sampling intensity classes of three or less or only sampled in the greatest sampling intensity class and species with at least 90% of records in one class. By varying these options, including perhaps another approach than kernel

densities (Daru and Rodriguez, 2023; Garcia-Rosello et al., 2023), perhaps a more conservative approach can be developed.

Any removal of records has a cost of information loss, leading to commission errors and unrepresentative models (Smith et al., 2023). Rather than no information, I preferred georeferenced records up to 36 km, a threshold that retains almost all records. That is, for equalization, records were sorted by distance uncertainties and longer distances were only recruited if necessary to meet the midpoint sampling threshold. If number of samples in the sampling class exceeded the midpoint threshold, longer distances were removed. For some objectives, such as for specialist species with narrow environmental tolerances, this margin of error may not be suitable and no information would be preferable. Also, sampling was not spatial because no particular benefit appears to be derived from localized unclustered samples. However, if this is a concern, spatial thinning of unequalized samples can be nested within the equalizing of sampling intensities, with the number of retained samples specified.

5 Conclusions

These analyses highlighted the sampling intensity gradient issue in Eurasia, reiterated overprediction and commission error consequences of using thinned samples, developed a method of equalized samples that balanced commission and omission error while evading concentrated models by design, and recommended use of unknown and increased distance uncertainties. In general, models from a range of distance uncertainties were similar. Indeed, locational distance uncertainties appeared to reflect institutional standards, with no reason to exclude the records based on distance uncertainty, which most likely had greater accuracy than the indicated maximum potential distance uncertainty. However, inclusion of records with distance uncertainties of 5 km resulted in even greater concentration of samples in Europe and predicted area concentrated in Europe for species distribution models. Models from thinned records were liable to overpredict presence area, for about twice the number of species as models from equalized sampling when sample size was below 1000 records. Equalized sampling was able to avoid greater concentration of samples, while employing records from a range of distance uncertainties. Therefore, models from equalized sampling maximized both the intersection with predicted presence area from random samples of range maps and area predicted as present without overpredicting as much as models from thinned samples or underpredicting as much as unthinned samples. If equalized sampling is not applied, unthinned samples with unknown and distance uncertainties less than 1 km may be the next best option or unthinned samples with unknown and distance uncertainties less than 6 km if avoiding species with Eurasian ranges, although sampling gradients may

intensity over time in other locations. Future research can continue to test equalization more comprehensively, although reducing the sampling gradient was effective for other modeling algorithms for a virtual species (Supplementary File 1). No approach to sample processing could correct for unsampled areas.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

Author contributions

BH: Validation, Funding acquisition, Project administration, Formal analysis, Writing – review & editing, Supervision, Software, Resources, Methodology, Writing – original draft, Data curation, Conceptualization, Investigation, Visualization.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This research was supported by the USDA Forest Service, Rocky Mountain Research Station, internal funding from the Infrastructure Investment and Jobs Act of 2021.

Acknowledgments

I thank the Global Biodiversity Information Facility, International Union for Conservation of Nature, BirdLife International, and contributors to the range maps and georeferenced observations.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author BH declared that they were an editorial board member of *Frontiers*, at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be

made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

The findings and conclusions in this publication are those of the authors and should not be construed to represent any official USDA or U.S. Government determination or policy.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fevo.2026.1831501/full#supplementary-material>

References

- Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A., Vilela, B., and Anderson, R. P. (2015). Sphthin: an R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography* 38, 541–545. doi: 10.1111/ecog.01132
- Ballesteros-Mejia, L., Kitching, I. J., Jetz, W., and Beck, J. (2017). Putting insects on the map: Near-global variation in sphingid moth richness along spatial and environmental gradients. *Ecography* 40, 698–708. doi: 10.1111/ecog.02438
- Barber, R. A., Ball, S. G., Morris, R. K. A., and Gilbert, F. (2022). Target-group backgrounds prove effective at correcting sampling bias in Maxent models. *Divers. Distrib.* 28, 128–141. doi: 10.1111/ddi.13442
- BirdLife International and Handbook of the Birds of the World [BLI] (2023). Bird species distribution maps of the world. Version 2023.1. Available online at: <http://datazone.birdlife.org/species/requestdis> (Accessed April 3, 2023).
- Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Stat. Sci.* 16, 199–231. doi: 10.1214/ss/1009213726
- Chefaoui, R. M., and Lobo, J. M. (2008). Assessing the effects of pseudo-absences on predictive distribution model performance. *Ecol. Modell.* 210, 478–486. doi: 10.1016/j.ecolmodel.2007.08.010
- Daru, B. H., and Rodriguez, J. (2023). Mass production of unvouchered records fails to represent global biodiversity patterns. *Nat. Ecol. Evol.* 7, 816–831. doi: 10.1038/s41559-023-02047-3
- Denzin, N. K. (2012). Triangulation 2.0. *J. Mixed. Methods Res.* 6, 80–88. doi: 10.1002/9781405165518.wbeost050.pub2
- Early, R., and Sax, D. F. (2014). Climatic niche shifts between species' native and naturalized ranges raise concern for ecological forecasts during invasions and climate change. *Global Ecol. Biogeogr.* 23, 1356–1365. doi: 10.1111/geb.12208
- eBird (2024). Choosing and Managing Locations in eBird. Available online at: <https://support.ebird.org/en/support/solutions/articles/48000850891-choosing-and-managing-locations-in-ebird> (Accessed February 16, 2025).
- Elith, J., and Leathwick, J. R. (2009). Species distribution models: ecological explanation and prediction across space and time. *Annu. Rev. Ecol. Syst.* 40, 677–697. doi: 10.1146/annurev.ecolsys.110308.120159
- ESRI (2025). Kernel density (Spatial analyst). Available online at: <https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-analyst/kernel-density.htm> (Accessed May 2, 2025).
- Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.* 15, 3133–3181.
- Fourcade, Y., Engler, J. O., Besnard, A. G., Rödder, D., and Secondi, J. (2013). Confronting expert-based and modelled distributions for species with uncertain conservation status: a case study from the corncrake (*Crex crex*). *Biol. Conserv.* 167, 161–171. doi: 10.1016/j.biocon.2013.08.009
- Gábor, L., Moudrý, V., Barták, V., and 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, 1567–1584. doi: 10.1080/13658816.2019.1615070
- García-Rosello, E., Gonzalez-Dacosta, J., Guisande, C., and Lobo, J. M. (2023). Gbif falls short of providing a representative picture of the global distribution of insects. *Syst. Entomol.* 48, 489–497. doi: 10.1111/syen.12589
- Global Biodiversity Information Facility [GBIF] (2025). Free and open access to biodiversity data. Available online at: www.gbif.org (Accessed April 2, 2025).
- Hanberry, B. B. (2024). Practical guide for retaining correlated climate variables and unthinned samples in species distribution modeling, using random forests. *Ecol. Inf.* 79, 102406. doi: 10.1016/j.ecoinf.2023.102406
- Hanberry, B. B. (2025a). Potential expanded pollinator distributions in North America under future climate. *Ecol. Solutions. Evid.* 6, e70025. doi: 10.1002/2688-8319.70025
- Hanberry, B. B. (2025b). Clarifying influences of sampling bias (concentration) and locational errors (uncertainties) on precision or generality of species distribution models. *Land* 14, 1620. doi: 10.3390/land14081620
- Hanberry, B. B., and He, H. S. (2013). Prevalence, statistical thresholds, and accuracy assessment for species distribution models. *Web Ecol.* 13, 13–19. doi: 10.5194/we-13-13-2013
- Hijmans, R. (2024). Terra: Spatial Data Analysis. R package version 1.7-81. Available online at: <https://github.com/rspsatial/terra> (Accessed October 31, 2024).
- Hughes, A. C., Orr, M. C., Ma, K., Costello, M. J., Waller, J., Provoost, P., et al. (2021). Sampling biases shape our view of the natural world. *Ecography* 44, 1259–1269. doi: 10.22541/au.159419540.02222358
- Hurlbert, A. H., and Jetz, W. (2007). Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. *Proc. Natl. Acad. Sci.* 104, 13384–13389. doi: 10.1073/pnas.0704469104
- Hurlbert, A. H., and White, E. P. (2005). Disparity between range map-and survey-based analyses of species richness: patterns, processes and implications. *Ecol. Lett.* 8, 319–327. doi: 10.1111/j.1461-0248.2005.00726.x
- iNaturalist (2020). To Obscure, make Private or leave your observation Public/Open? Available online at: <https://www.inaturalist.org/posts/34066-to-obscure-make-private-or-leave-your-observation-public-open> (Accessed February 16, 2025).
- Inman, R., Franklin, J., Esque, T., and Nussear, K. (2021). Comparing sample bias correction methods for species distribution modeling using virtual species. *Ecosphere* 12, e03422. doi: 10.1002/ecs2.3422
- International Union for Conservation of Nature [IUCN] (2025). The IUCN Red List of Threatened Species. Spatial data download. Available online at: <https://www.iucnredlist.org/resources/spatial-data-download> (Accessed March 25, 2025).
- Jetz, W., Sekercioglu, C. H., and Watson, J. E. (2008). Ecological correlates and conservation implications of overestimating species geographic ranges. *Conserv. Biol.* 22, 110–119. doi: 10.1111/j.1523-1739.2007.00847.x
- Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., et al. (2017). Climatologies at high resolution for the Earth land surface areas. *Sci. Data* 4, 170122. doi: 10.1038/sdata.2017.122

- Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., et al. (2018). Data from: Climatologies at high resolution for the earth's land surface areas. *EnviDat*. doi: 10.16904/enviDat.228.v2.1
- Kuhn, M. (2008). Building predictive models in R using the caret package. *J. Stat. Software* 28, 1–26. doi: 10.18637/jss.v028.i05
- Lambole, Q., and Fourcade, Y. (2024). No optimal spatial filtering distance for mitigating sampling bias in ecological niche models. *J. Biogeogr.* 51, 1783–1794. doi: 10.1111/jbi.14854
- McCarthy, K., Fletcher Jr, R. J., Rota, C. T., and Hutto, R. L. (2012). Predicting species distributions from samples collected along roadsides. *Conservation Biology* 26(1), 68–77
- Meyer, C., Weigelt, P., and Kreft, H. (2016). Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecol. Lett.* 19, 992–1006. doi: 10.7287/peerj.preprints.1326v1
- Moudry, V., Bazzichetto, M., Remelgado, R., Devillers, R., Lenoir, J., Mateo, R. G., et al. (2024). Optimising occurrence data in species distribution models: sample size, positional uncertainty, and sampling bias matter. *Ecography*, e07294. doi: 10.1111/ecog.07294
- Moudry, V., Keil, P., Gábor, L., Lecours, V., Zarzo-Arias, A., Barták, V., et al. (2023). Scale mismatches between predictor and response variables in species distribution modelling: A review of practices for appropriate grain selection. *Prog. Phys. Geograp.: Earth Environ.* 47, 467–482. doi: 10.1177/03091333231156362
- Pili, A., Leroy, B., and Zurell, D. (2025). Correcting environmental sampling bias improves transferability of species distribution models. *Ecography* 2025, e08002. doi: 10.1002/ecog.08002
- Probst, P., Wright, M. N., and Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley. Interdiscip. Rev.: Data Min. Knowledge. Discov.* 9, e1301. doi: 10.1002/widm.1301
- QGIS.org (2025). QGIS geographic information system. Open source geospatial foundation project. Available online at: <http://qgis.org> (Accessed May 2, 2025).
- R Core Team (2025). *R: A language and environment for statistical computing* (Vienna, Austria: R Foundation for Statistical Computing).
- Santini, L., Benítez-López, A., Maiorano, L., Čengić, M., and Huijbregts, M. A. (2021). Assessing the reliability of species distribution projections in climate change research. *Divers. Distrib.* 27, 1035–1050. doi: 10.1111/ddi.13252
- Smith, A. B., Murphy, S. J., Henderson, D., and Erickson, K. D. (2023). Including imprecisely georeferenced specimens improves accuracy of species distribution models and estimates of niche breadth. *Global Ecol. Biogeogr.* 32, 342–355. doi: 10.1111/geb.13628
- Ten Caten, C., and Dallas, T. (2023). Thinning occurrence points does not improve species distribution model performance. *Ecosphere* 14, e4703. doi: 10.1002/ecs2.4703
- Thuiller, W., Brotons, L., Araújo, M. B., and Lavorel, S. (2004). Effects of restricting environmental range of data to project current and future species distributions. *Ecography* 27, 165–172. doi: 10.1111/j.0906-7590.2004.03673.x
- Wang, L., and Jackson, D. A. (2023). Effects of sample size, data quality, and species response in environmental space on modeling species distributions. *Landscape Ecol.* 38, 4009–4031. doi: 10.1007/s10980-023-01771-2
- Yates, K. L., Bouchet, P. J., Caley, M. J., Mengersen, K., Randin, C. F., Parnell, S., et al. (2018). Outstanding challenges in the transferability of ecological models. *Trends Ecol. Evol.* 33, 790–802. doi: 10.1016/j.tree.2018.08.001
- Zizka, A., Carvalho, F. A., Calvente, A., Baez-Lizarazo, M. R., Cabral, A., Coelho, J. F. R., et al. (2020). No one-size-fits-all solution to clean GBIF. *PeerJ* 8, e9916. doi: 10.7717/peerj.9916