

Article

Clarifying Influences of Sampling Bias (Concentration) and Locational Errors (Uncertainties) on Precision or Generality of Species Distribution Models

Brice B. Hanberry 

United States Department of Agriculture, Forest Service, Rocky Mountain Research Station,
Rapid City, SD 57702, USA; brice.hanberry@usda.gov

Abstract

Locational errors and sampling bias may produce unrepresentative species distribution models. To decompose the influence of errors, I modeled species distributions of 31 mammal species from georeferenced records and random samples from range maps, with potential sources of errors added or removed, using the random forests algorithm. Errors included the addition of (1) cities, (2) administrative centers, (3) records flagged as potential errors (e.g., outliers), and (4) urban records to range map samples; the removal of (5) flagged records and (6) urban records from georeferenced records; and the addition of (7) random points and (8) clustered points to georeferenced records. I also examined separation between thinned and unthinned (i.e., locally concentrated) records and ocean and land areas. Errors generally did not perturb species distributions, particularly if errors were located within species ranges. The greatest departure relative to unaltered models (mean niche overlap values of 0.96 out of 1) was due to the addition of administrative centers at a 13% error rate. Because locational errors overall do not occur in modern georeferenced records, outliers may provide important samples from undersampled areas. Delineating land from ocean coordinates may require a land layer at the highest available resolution and buffered to match the distance of locational uncertainty for georeferenced records. Predicted areas for species distributions increased along the spectrum of models from concentrated georeferenced records, thinned records, and random samples from range maps. Species distributions modeled with all georeferenced records will have the greatest sampling concentration (to differentiate from bias, because predictive modeling is not hypothesis testing), resulting in model locational precision, whereas species distribution models from random samples of range maps will have locational generality (rather than errors). The risk of removing samples of suitable conditions is the generation of unrepresentative models whereas the benefit of sample removal is slightly more generalized models, but which also may represent overpredictions.

Keywords: centroid; cities; location; ocean; outlier; thinning; urban



Academic Editor: Brian D. Fath

Received: 11 June 2025

Revised: 7 August 2025

Accepted: 7 August 2025

Published: 9 August 2025

Citation: Hanberry, B.B. Clarifying Influences of Sampling Bias (Concentration) and Locational Errors (Uncertainties) on Precision or Generality of Species Distribution Models. *Land* **2025**, *14*, 1620. <https://doi.org/10.3390/land14081620>

Copyright: © 2025 by the author. Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Spatial sampling at the scale to represent perfect species distributions, or other wide-ranging patterns, generally is not achievable now, resulting in varied realizations [1,2]. Nonetheless, the availability of georeferenced samples necessary for modeling numerous species distributions has progressed due to opportunistic sampling by community naturalists [3]. Opportunistic sampling results in sampled areas, yet at the same time, the

preferential sampling of accessible locations, such as near urban areas, is disproportionate relative to other locations [4]. Differential sampling is considered to be sampling bias due to overrepresentation of available yet suitable environmental conditions, with not all conditions being equally and randomly sampled [5]. However, conditions absolutely must be sampled for representative species distribution models, and sampled areas with correct yet uneven sampling from suitable locations is an advancement compared to unsampled areas [6,7].

Removing adjacent samples to reduce concentrated georeferenced records in clusters has not been demonstrated to be beneficial, due to information loss about conditions suitable for species [2,8–10]. To be clear, localized thinning to reduce clustered records is different than treating regional or continental differentials in sampling intensity, as occurs with high sampling intensity in Europe for most taxa relative to the rest of the world. Otherwise, samples from suitable conditions can supply information about missing environmental gradients in unsampled locations. For example, concentrated samples near cities or roads may differ in temperature relative to other locations and urban or road samples also can represent a wide gradient of climate conditions [11,12]. The applied species distribution modeling algorithm (e.g., random forests) can extend predictions from conditions in well-sampled (e.g., urban) areas to undersampled (e.g., rural) areas because modeling algorithms perform predictive modeling rather than statistical testing [13]. Therefore, sampling for predictive modeling can be more opportunistic than the assumption of simple random sampling with balanced representation necessary for statistical tests. If entire areas of species distributions are missing from samples, removing or reducing (i.e., thinning) samples from suitable locations to reduce sampling bias will not fix and balance lack of sampling. For small samples, removing samples will result in loss of information necessary to differentiate suitable from unsuitable areas [8–10]. That is, the precaution may cause the outcome that is meant to be prevented.

Similarly, the potential for erroneous locations of records has given rise to a standard practice of removing records to avoid issues, typically without any evidence for errors, rather than trying to avoid loss of samples ('coordinate cleaning'; [2,14–18]). Relatively few coordinates are flagged as potential issues by programs (typically <5% flagged records in the Global Biodiversity Information Facility database; GBIF; [15,19,20]). Of these flagged records, the most frequent errors encompass duplicated or missing coordinates, fossils or unknown types of observations, and older records [19,21], which unquestionably should be addressed as part of workflows for processing data, to meet objectives. Records collected before the year 1993 and development of georeferencing technology may be prone to locational errors if records are georeferenced based on imprecise descriptions of collection sites [22]. Locational errors may arise from mistakes in data entry of georeferenced coordinates, including placement at biodiversity institutions and administrative unit centers (i.e., centers of countries, cities, counties; [15]). The coordinate cleaning programs flag urban locations, because these locations may identify cultivated individuals or may be locations that previously were wildlands [15,16]. Conversely, flagged outliers may represent the missing, unsampled species distributions [18]. Ocean coordinates also are among the most frequently flagged records [19,21]. However, coordinates from georeferenced records for species are probable to be more accurate than a 1:110,000,000 land polygon used to flag ocean coordinates ([23]; default of 110 million scale documented in Zizka et al. [15]). Map scales are ratios of map distances to the corresponding distances on the ground, such that even the 10 million map scale printed on paper may be equivalent to 3 km resolution, albeit with finer resolutions along coastlines [24]. After basic processing of data, true locational errors are probably rare for modern georeferenced records within specified locational uncertainty distances (e.g., 1 km), particularly if from automated georeferencing within

current data collection platforms, and not likely to affect species distribution models, given sufficient sample sizes (i.e., at least 300 to 400 records; [8,18,25,26]).

Sampling bias and coordinate errors have been attributed to be important issues for species distribution models that need to be addressed by removing samples [2,17]. However, without any particular foundation, reduction in sample sizes as a preventative measure may result in lack of complete species distribution models (omission errors) or incorrect models outside of distributions (commission errors); therefore, data cleaning beyond basic corrections for unique coordinates of observations during specific years may obscure the samples [27]. For this study, I weighed the influence on species distribution models of sampling bias and various errors commonly identified by coordinate cleaning packages, by modeling 31 mammal species, with georeferenced records of at least 1000 samples (i.e., relatively stable due to enough information from samples; before thinning although potential errors and bias were added to thinned records) that provided realized sampling and also 1000 randomly generated samples from range maps, as objective measures of unbiased sampling with imperfect locations [28]. Relatedly, I contrasted differences between models from thinned and unthinned (i.e., with sampling bias due to concentrated species observations) georeferenced records. The georeferenced records, of which 94% of records were sourced from iNaturalist [3] in the GBIF database [20], were a real, not virtual, source of potential errors. Basic processing limited observations to unique terrestrial coordinates with tagged uncertainty of less than 1 km, and few records occurred before even 2015 (i.e., <5%). These are impeccable georeferenced records, not museum records. All the randomly generated samples from range maps were sampled without bias yet locationally uncertain because no species occurrences in fact were sampled. Without empirical observations, these pseudopresence samples are generally correct representations of the extent of occurrence but without any locational precision of the area of occupancy. I performed the following assessments, holding all else in modeling equal, to examine how locational errors and sampling bias affect species distribution models (Figure 1).

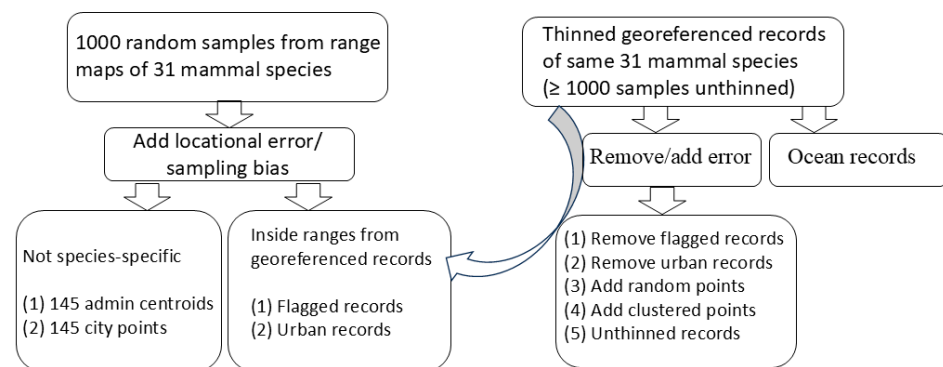


Figure 1. Workflow for generating or removing potential locational errors and sampling biases. Randomly generated presence samples from range maps were sampled without concentration (bias) but with locational uncertainty (error) because no species were observed at locations. Nonetheless, range maps provide objective measures of suitable climate conditions for species, as designated by expert knowledge. Conversely, georeferenced records are precisely located but with potential, realized errors, particularly sampling concentration (bias) in urban locations.

(1) Incorporation of obviously incorrect locations: I created centroid points from cities and administrative centers [23], randomly selected 145 points, and then added the same erroneous points to 1000 range map samples for each of 31 species. This was a 13% error rate, which was at least double the <5% flagged records [15,19], with points that were not related to any species ranges. I added this type of error only to the range map samples

because the range map samples do not contain concern about potential errors embedded in georeferenced records.

(2) Inclusion and exclusion of flagged records: For about 650 georeferenced records out of 126,000 georeferenced records flagged as potential errors (0.2% error rate of a few city and administrative centroids and primarily biodiversity institutions and outliers), I added flagged records to range map samples from each respective species, for an overall 2% error for 31,000 range map samples, varying by species. Conversely, I removed the flagged errors from the 126,000 georeferenced records.

(3) Inclusion and exclusion of sampling bias: I applied the urban samples from georeferenced records to random samples of range maps and also removed the urban samples from georeferenced records, artificially splitting species into a country subspecies. The sampling bias rate was high, with more urban samples (39,800 records) than randomly generated points from range maps (31,000). I also modeled urban points to demonstrate the extent of climate conditions contained within urban areas.

(4) Addition of adjacent random and clustered samples: To create comprehensive species-specific locational errors and sampling bias of uneven sampling to georeferenced records, I added random points (20% mean error rate, varying by species) and clustered points (40% mean error rate) to georeferenced records up to 35 km distances from the records, which is the maximum distance uncertainty for almost all georeferenced records (e.g., 99%, depending on percentage of records from iNaturalist [3]). These types of error are relatively comparable to the addition of urban samples from the georeferenced records to range maps. While these added points may seem like distortions, in fact adding random samples near to georeferenced records is conceptually similar to synthetic samples generated by varying nearest neighbors approaches, to increase sample sizes and balance classes [29].

(5) Differentiation of coordinates on land and at sea: I used two different scales of land area to demonstrate errors associated with this type of coordinate removal, and provided flexible solutions.

Samples from range maps, unthinned georeferenced records (with unique coordinates), and thinned georeferenced records, were modeled using consistent methods aside from eight additions or removals of errors to range maps or thinned georeferenced records.

2. Materials and Methods

2.1. Overview of Species Distribution Modeling

For unbiased species distribution models, the input information needs to be as complete as possible to allow the modeling algorithm to separate suitable conditions for presence classes from conditions for absence classes ([6]; Figure 2, step 4). To achieve enough information, the following conditions need to be met: use of large sample sizes relative to range areas (at least 300 to 400 records, or perhaps at least 100 records with expert knowledge of the species ranges; [8]; Figure 2, step 1), samples not truncated to small administrative boundaries relative to the size of the species distributions ([6]; Figure 2, steps 1 and 2), and retention of critical predictor information of conditions (e.g., annual mean temperature that spatiotemporally anchors species distributions; [30]; Figure 2, step 3). For simple binary classification, accuracies should meet a minimum mean value of about 0.9, varying by the scale of the metric, to indicate successful differentiation of classes (see evaluation section below; Figure 2, step 5). In addition to supplying as much information as available to the modeling algorithm, the task of the researcher is to evaluate the models, and not simply through accuracy metrics, which are only as good as the input samples (i.e., poor models can have perfect accuracy metrics; [31]). Evaluations include visual inspections of predicted areas of presence, with comparison to expert range maps or other

expert knowledge, and additional metrics, such as areas of predicted presence and niche overlap (see evaluation section below; Figure 2, step 6b).

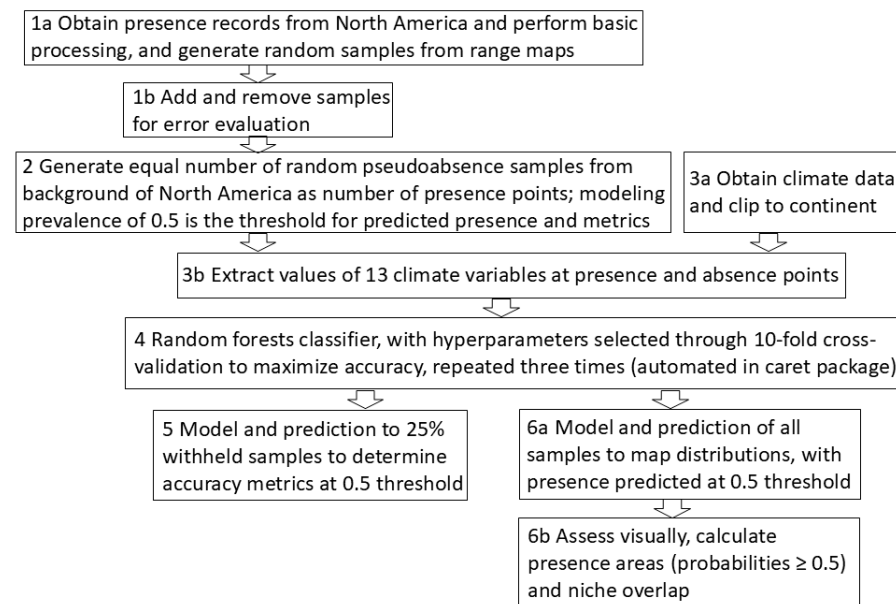


Figure 2. Overview of modeling steps.

2.2. Samples

Terrestrial mammal species were selected if they had 1000 records with unique coordinates limited to North America, excluding species with known range extirpations and small ranges and most small nocturnal mammals (*Chiroptera*, *Muridae*, to ensure intersection between records and ranges but in fact these taxa have a median percentage of 92.5% intersection between georeferenced records and range maps; B. Hanberry, USDA Forest Service, unpublished data). Range maps for 31 mammal species were from IUCN ([28]; World Geodetic System WGS 1984 geographic coordinate system), with 1000 randomly generated samples within ranges for presence samples. For georeferenced records of the same 31 mammal species, I downloaded North American records during years 1990 to 2020 with coordinate uncertainty less than 1 km ([20]; GBIF occurrence download <https://doi.org/10.15468/dl.2wq3r3> (accessed on 27 July 2024); World Geodetic System WGS 1984 geographic coordinate system). Basic processing eliminated duplicated information, selected collection year, removed fossils and unknown record types, and separated terrestrial from ocean locations. These species had ≥ 1000 unthinned samples, after removal of duplicate coordinates by species and only records georeferenced in the North American continent that matched with climate data. Unthinned sample sizes for 31 species ranged from 1061 to 31,115. Sample sizes for the species thinned to one sample per climate cell (resolution of 30 s; see climate below) ranged from 621 to 20,395 (dismo package; [32]), which were used for the assessments of error and bias. For pseudoabsences, an equal number of background samples as records or range samples, matching the presence samples with or without errors, were generated from random samples of the North American continent.

2.3. Potential Error and Bias

I added deliberate locational errors, which were manifestly wrong in having no regard for any species ranges (Figure 1). For locational errors that were wrong, I generated centroid points from cities and administrative centers [23], randomly selected 145 points, and then added the points to the range map samples. This was a 13% error rate, greater than intrinsic

locational error rates identified by coordinate cleaning (<5% flagged records; [15,19]), with erroneous points that were not related to the species samples.

I incorporated and removed potential locational errors flagged by a coordinate cleaning package [15]. I added the 650 total coordinates flagged as potential errors (city and administrative centroids, biodiversity institutions, and outliers) in the georeferenced records to each respective species in the range map samples, for an overall 2% error, varying by species. Conversely, I removed the flagged errors from the georeferenced records.

For more realistic (than the cities and administrative center errors) yet more complete (than the few flagged errors) addition of errors and sampling bias to all samples, I also added random or clustered points within 35 km of georeferenced records. The random points averaged a 20% error rate that varied by species, with a mean distance of 1.5 km and distances ranging from 1 to 35 km (applying create random points function in ArcGIS Pro 3.3.4, ESRI, Redlands, California). The clustered points averaged a 40% error rate, with a mean distance of 2.2 km and distances ranging from 1 to 35 km (applying create random points function followed by near function in ArcGIS Pro 3.3.4, ESRI, Redlands, CA, USA). The distance of 35 km was selected because filtering distance uncertainties of georeferenced records to about 35 km will contain nearly all records (e.g., 99%, depending on percentage of records from iNaturalist [3]).

Equally, for sampling bias, I applied the urban samples (flagged from 1:50 million scale or medium scale; [15,23]) from georeferenced records to randomly sampled range maps and removed the urban samples from georeferenced records. The sampling bias rate overall was 56%, with more urban samples (39,800 coordinates) than randomly generated points from range maps (31,000). The number of urban points varied by species, with up to 8000 samples. I also modeled 1000 randomly sampled urban points (from 1:50 million scale or medium scale; [23]) to demonstrate the climate conditions contained within the small extent of urban areas (<5% of terrestrial land, depending on definition; [33]). Additionally, I modeled the unthinned, or concentrated, georeferenced records.

Lastly, I addressed the determination of samples located in the ocean. I applied the 110 and 10 million scales of land layers (World Geodetic System WGS 1984 geographic coordinate system; [23]) to demonstrate error in separating land from ocean. Several flexible solutions were suggested to reduce error from this type of coordinate removal.

2.4. Climate Values for Samples

Climate values, during 1981–2010, were extracted at each of the georeferenced coordinates and samples from ranges with an equal number of randomly generated background coordinates for pseudoabsences throughout the North American land extent (and see the results section related to defining the North American land extent). Pseudoabsence samples therefore were untruncated and contained the full continental range of climate. Debaised and downscaled climatologies quantify temperature and precipitation at annual, seasonal, and monthly intervals (resolution 30 arc seconds; CHELSA climate; [34,35]). Climate variables were 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. I clipped 13 bioclimatic variables to match the North American extent (analysis of the North American extent below).

2.5. Modeling Framework

For modeling of the untruncated samples for each species, I applied the random forests classifier, an ensemble nonlinear model, with hyperparameters determined through 10-fold cross-validation, repeated three times (hyperparameters automatically selected by the caret package; [36,37]). I modeled one time with 75% random split of the samples to determine how well the model predicted the withheld 25% random split 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, along a continuous range of climate suitability. I modeled a second time with all of the samples to map predictions to North America during 1981–2010. All methods and variables (13) were held constant except for the source of samples. Correlated climate variables are not an issue for predictions but for standard error of estimates in data models [30].

2.6. Evaluations

Accuracy metrics measure how well the modeling algorithm classified the provided samples, not how representative models are of species distributions. Models, particularly from small samples sizes, can have perfect accuracy and yet not resemble the intended species distributions due to unrepresentative presence and absence samples. Furthermore, different metrics have received various critiques [38]. Therefore, I selected three straightforward accuracy metrics of balanced accuracy (mean of sensitivity and specificity; scaled from 0 to 1 but similar to the true skill statistic, which is scaled from -1 to 1), sensitivity or true positive rate, and specificity or true negative rate. These accuracy metrics simply indicated whether the modeling algorithm was successful, with values about 0.90 to be accurate for binary classification with climate predictors, depending on scaling of the accuracy metric (e.g., greater minimum values for AUC).

To evaluate robustness and representativeness of models, I visually assessed model predictions, calculated predicted presence areas (probabilities ≥ 0.50 , based on modeling prevalence of the presence and absence classes) and also niche overlap (i.e., geographic distance or similarity index) between predictions from species distribution models with added and removed errors and the baseline species distribution models from thinned georeferenced records and range maps. For niche overlap, or difference in predicted probabilities between two models, I selected the Hellinger distance similarity index scaled from 0 to 1 to match the range of predicted probabilities ([39]; dismo package; [32]). Values of zero indicate geographic separation and values of one indicate overlap. While rules are arbitrary, niche overlap values of at least 0.9 typically represent relatively similar appearance of species distributions and values less than 0.8 start to indicate noticeable divergence in predictions. To focus overlap on areas where species were predicted, I used the rescaled Hellinger distance for areas where one of the compared models predicted values ≥ 0.50 (i.e., masking areas where both models predicted values < 0.5).

3. Results

3.1. Niche Overlap

Overall, addition and removal of potential errors caused little difference in species distribution models (Figure 3). Regardless of added or removed errors, all species distribution models for each error type based on georeferenced records had mean niche overlap values of 1.0, with minimum values of 0.98 for a few species. Mean niche overlap values of 1.0 occurred for comparisons between model predictions for addition of random and clustered points near georeferenced records, removal of urban and flagged records, and also unthinned records relative to model predictions from thinned records without any added or removed errors.

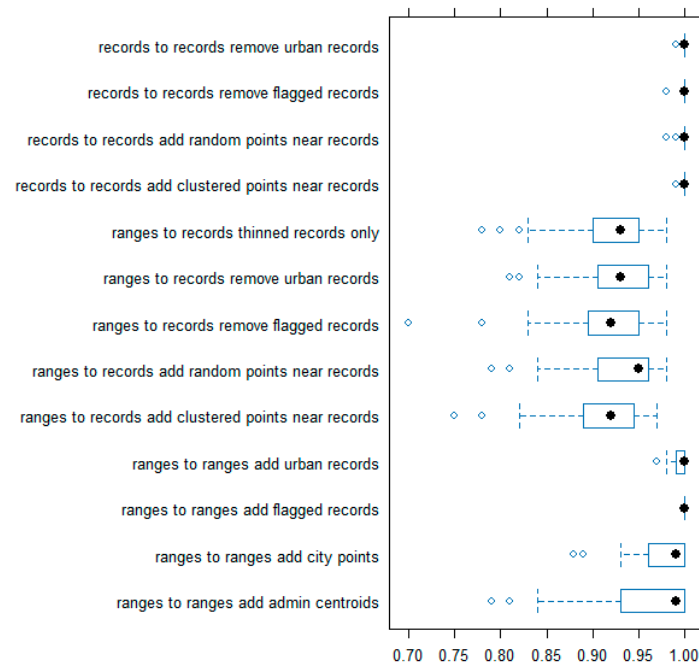


Figure 3. Niche overlap comparisons to predictions of species distributions (predicted probabilities ≥ 0.5) from either georeferenced records or range maps, for eight different additions or removals of potential locational and sampling bias errors (box plots with the box of first quartile to the third quartile, or interquartile range, with low and high whisker values, or dotted lines, no further than $1.5 \times$ interquartile range, and outlying points). Georeferenced records and random samples from range maps were both robust to perturbation, and urban sampling bias was not an issue. All comparisons for species distribution models from georeferenced records, for each error type, had mean niche overlap values of 1.0, with minimum values of 0.98 for a few species. Relative to models from range maps, removal of flagged records slightly increased divergence in niche overlap values, for which outliers provided critical information from unsampled areas. Addition of unreasonable errors (at almost 15% error rate) from unrelated city and administrative centroid points to range maps resulted in the errors that overall produced the greatest mean departure of 0.03 and 0.04 (out of 1.0), values which were relative to range maps with no added errors.

For comparisons of predictions from unaltered georeferenced records to predictions from unaltered range maps, as an objective measure, the baseline comparison between predictions from thinned georeferenced records to predictions from range maps was 0.916. That is, species distributions from georeferenced records were similar but not the same as species distributions from range maps. Models differed due to data sources of georeferenced records and range maps, with observed records that were not incorporated into range maps developed by experts and expert knowledge about locations that were not observed in georeferenced records. Regardless of added or removed errors, all comparisons by error type averaged 0.90 to 0.92. The most divergent niche overlap value for a species was 0.70 (from a baseline of 0.78 for this species, for the comparison between predictions from thinned georeferenced records to predictions from range maps). This was due to removal of the flagged records of *Marmota monax*, a wide-ranging species. Outliers were flagged as potential problems but for this species, the outliers provided information from otherwise unsampled areas. Likewise, removal of flagged records resulted in loss of the northern distribution for *Otospermophilus beecheyi*, reducing niche overlap with models from range map samples from 0.97 to 0.93 (Figure 4A,B).

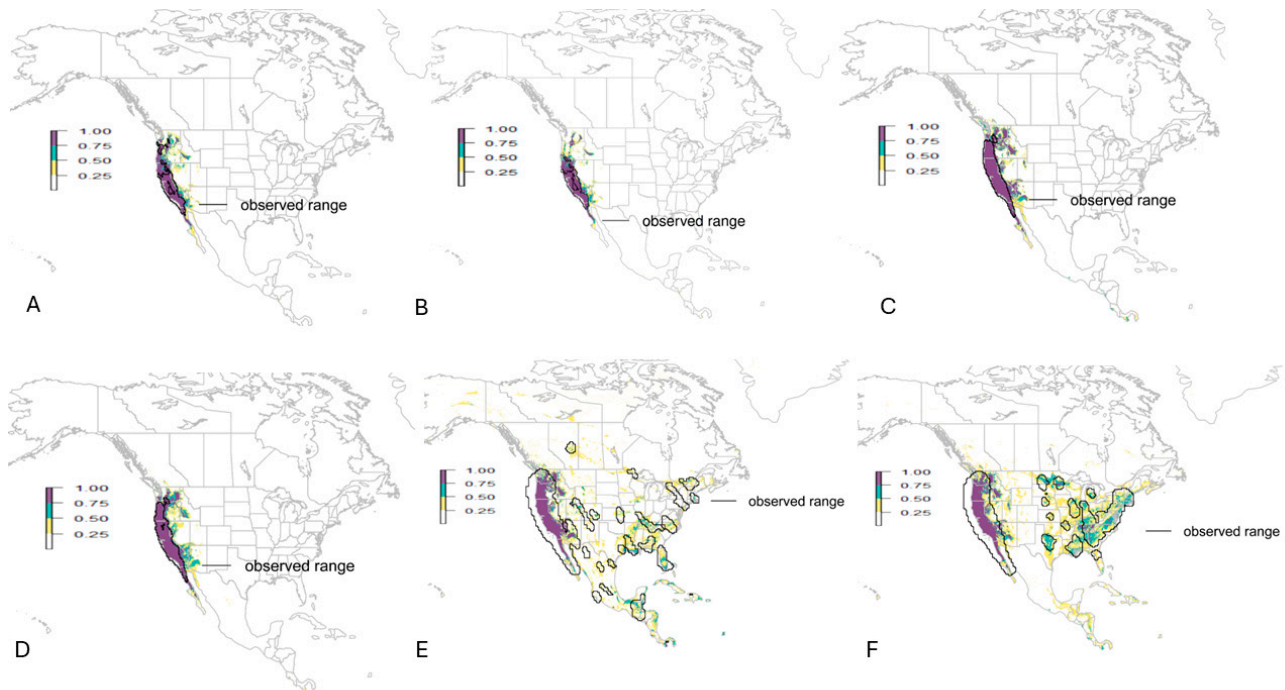


Figure 4. Removal of flagged records included outliers of otherwise unsampled areas, resulting in loss of predicted areas of presence for species distribution models from georeferenced records. For *Otospermophilus beecheyi*, the species distribution model (predicted probabilities ≥ 0.5) from all thinned georeferenced records (A) compared to loss of the northern distribution predicted after removal of flagged records, reducing niche overlap with models of range maps from 0.97 to 0.93 (B). Addition of 145 city and administrative points (a 13% error rate), unrelated to any georeferenced records, had the greatest influence on species distributions from random samples of range maps. Errors outside of range map samples for *Otospermophilus beecheyi* resulted in the least niche overlap comparisons to species distribution models from random samples of range maps without added errors (C). Addition of about 1000 urban samples (about 50% error rate) from georeferenced records (D) slightly reduced niche overlap to 0.99, but addition of city points (E) and administrative centroid points (F) reduced overlap to 0.88 and 0.79, respectively. The species range was maintained without distortion (purple; high climate suitability), but commission error increased (green; moderate climate suitability and an indicator of uncertainty relative to predicted probabilities ≥ 0.75), with additional expression of uncertainty (yellow; low climate suitability and not classified as present). Observed range (black outline) indicates location of species presence samples.

For comparisons of predictions for range maps, addition of flagrant errors (at almost a 15% error rate) from unrelated city and administrative centroid points were the errors that overall produced the greatest mean departure relative to range maps with no added errors. Comparisons averaged 0.96 and 0.97 for addition of administrative centroid and city points, respectively. For these two added types of error to samples from range maps, differences were small for most species, but addition of administrative centroid and city points to *Otospermophilus beecheyi* resulted in the least niche overlap (0.79 and 0.88, respectively) between species distributions from range maps (Figure 4). Nonetheless, the real species range was maintained, with some far-outlying added commission error, which the modeling algorithm recognized with uncertainty (predicted probabilities less than 0.75) despite the sample locations. Additions within ranges, that is, the urban and flagged records, were unimportant, with average niche overlap values of 0.99 and 1.00 for addition of urban and flagged records, respectively, with the least values by species of 0.97 and 1.00, respectively.

3.2. Areas, Including Urban

Predicted presence areas of species distribution models can range from very precise representation of exact locations of species observations, with sampling bias from clustered samples to very generalized representations of species ranges, without sampling bias because species were not sampled (i.e., ranges were sampled with random points; Figure 5). All predicted species distributions from georeferenced records were more precisely located representations of observed species, reflecting narrower climate amplitudes in areas of occupancy, than species distributions from completely sampled range maps, reflecting wider climate amplitudes in extents of occurrence. This resulted in less predicted presence area for georeferenced records than range maps. Equally, within predicted species distributions from georeferenced records, models from unthinned georeferenced records were the most precisely located because the concentrated samples provide a weight in locations, allowing the modeling algorithm to favor those locations, unlike the greater generality and predicted area of models from thinned georeferenced records. Predicted probabilities from unthinned records had less uncertainty (predicted values less than 0.75 and greater than 0) of 30% relative to 34% for thinned models.

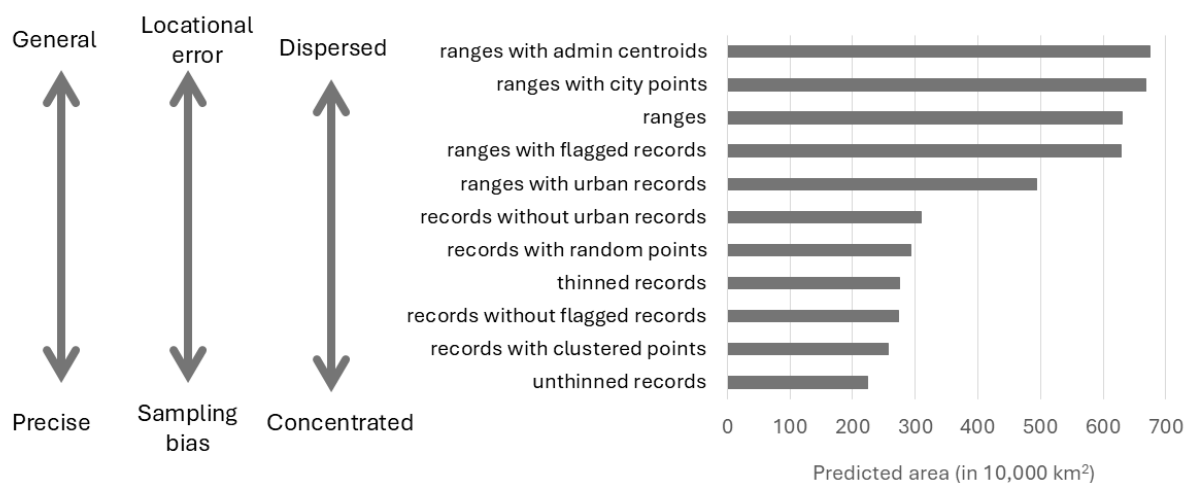


Figure 5. Area of predictions for species distributions (predicted probabilities ≥ 0.5) from unthinned georeferenced records, thinned georeferenced records, range maps, and eight different additions or removals of potential locational and sampling bias errors. As area of predicted species distributions increased, generality, locational errors, and dispersion increased and precision, sampling bias, and concentration decreased.

While urban areas cover small extents, about 164,980 km² in North America, they may be disproportionately sampled (32% of these georeferenced records). However, after modeling 1000 samples of urban areas, predictions generated 4,156,735 km² of comparable climate conditions (Figure 6). This area is greater than areas of most species distributions.

3.3. Accuracies

Mean sensitivity values (from 25% withheld samples) ranged from 0.89 (ranges with city points) to 0.97 (ranges, ranges with flagged records, unthinned records, records with clustered records). Mean specificity values ranged from 0.91 to 0.95. Mean balanced accuracies (from 25% withheld samples) overall were high, ranging from low values of 0.90 and 0.91 for the flagrant errors of added administrative center and city points, to 0.93 to 0.96 for the remaining models with different samples. Balanced accuracies were slightly greater for models from range maps than georeferenced records, but models from range maps were enhanced by complete separation between presence and absence classes (i.e., within or outside of ranges), allowing generation of stronger classification of samples. Balanced

accuracies reflected model precision within models for georeferenced records, increasing slightly with concentrated, clustered samples.

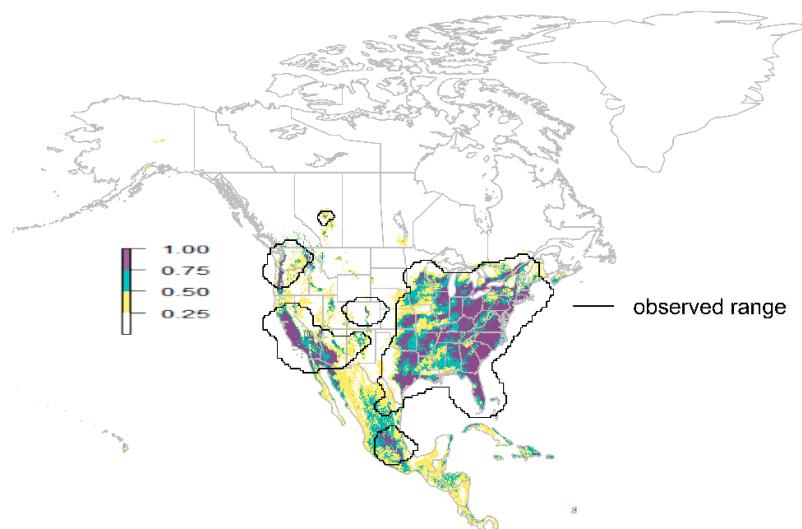


Figure 6. Area with similar climate conditions as urban area samples. Observed range (black outline) indicates location of urban samples.

3.4. Ocean Analysis

After downloading species with unique coordinates and thinning to one georeferenced record per cell in a global climate layer, 131,066 samples remained. Of these samples, 3607 samples (2.8%) and 2555 samples (1.9%) were in the ocean based on the 110 and 10 million scales, respectively, of Natural Earth [23]. Only 871 samples (0.7%) did not intersect with either land layer, and yet appeared to be at least within 1 km (within the distance uncertainty of the georeferenced records) to the land layers. Indeed, adding a 1 km buffer, or the acceptable distance uncertainty of georeferenced records, retained all but 150 samples (0.1%). The 150 samples also appeared to be correctly located. Therefore, one solution to reduce the number of removed samples was to merge the two scales of land layers and add a buffer that matches the uncertainty distance of the georeferenced records.

If desired, for terrestrial species, which do not necessarily inhabit permanent aquatic cover as part of occupied distributions, this also gives an opportunity to erase large lakes [23], and associated records, by performing this clip before buffering. With large lakes erased, 753 samples were removed, although these sample locations were likely to be correct within 1 km, but the lake layer did not capture all the island locations where species could occur. Therefore, another option was to simply erase the very largest lakes (scaleranks of 0 and 1 in Natural Earth [23]), resulting in removal of 620 water samples.

A smoother possibility was applying only a 1:10 m map scale of an administrative layer, which may have additional information of interest, such as continents [23]. Then, if desired, erasing the largest lakes and buffering by 2 km resulted in 284 samples (0.2%) not included within the North American continent. Small islands can be excluded, to speed up processing and reduce processing extent area.

4. Discussion

4.1. Muddying the Waters by Cleaning Coordinates

Potential biased sampling and locational errors are avoided by routine removal of samples. However, based on decomposing effects of errors added to samples from range maps and errors removed or added to georeferenced records, sampling and locations of geo-

referenced records did not produce biased species distribution models. Added errors that occurred within ranges were compatible with the other samples. Even for the poorest, most divergent model, arising from almost a 15% error rate added outside of the species range, the modeling algorithm of random forests could recognize disruptive errors that were added in locations beyond reason for the range, and the modeling algorithm incorporated those administrative centroid samples with uncertainty (predicted probabilities less than 0.75) into models. The georeferenced records did not contain enough errors to influence models, as apparent from models with flagged and urban records transferred to range map samples or from models with flagged and urban records compared to models without the records. Species distribution models proved extremely robust to error, particularly to added points within ranges. Errors and sampling bias do not matter if the climate amplitude of the species is sampled in the erroneous or biased points [7].

The small amount of locational errors contained in georeferenced records will not influence the outcomes, given enough samples to establish suitable conditions. If errors are inside ranges and compatible with sampled climate conditions, they become samples that help the modeling algorithm differentiate suitable conditions from unsuitable conditions. If errors are outliers to the climate samples, the modeling algorithm will be able to use them to fill in ranges. If errors are not compatible with other samples, they will be discounted because the modeling algorithm maximizes classification accuracy. Even flagrant errors added without regard to species ranges did not perturb primary representations of species ranges. In this study, definite errors were almost 15% of samples from cities or administrative centers without any relation to georeferenced records and most likely beyond error rates in any untampered samples. For the poorest model of a species (0.79 niche overlap between the model with added locational error and model with no error), the modeling algorithm identified with high model certainty (predicted probabilities ≥ 0.75) that the range was limited to the Pacific Coast of the United States and communicated uncertainty about addition of samples scattered throughout the eastern United States. That is, the modeling algorithm was able to differentiate the corrupted samples because they were not compatible with the bulk of the samples, although the model predictions of presence were degraded by predictions outside of the range. Effects of errors depend on the species distributions, error rates relative to sample sizes, and how different the added errors are from the species samples.

Rather than discounting outlying samples, the modeling algorithm may determine that outlying samples are compatible with the other georeferenced records, enhancing the species distribution model [18]. Generally, the addition or removal of the few georeferenced records flagged as potential errors (about 650 records, or 0.2% of all records) made no difference to species distribution models. Nonetheless, removal of outlying points deteriorated models of a few species, resulting in reduced predicted presence within ranges. Particularly for wide-ranging species, outliers may represent critical information, not errors [18]. Because unsampled areas are an issue for generating representative species distribution models, sampled outliers may be the most important records of the potential errors to keep in terms of species representation. Outliers may help increase the width of climate amplitudes similar to synthetic samples generated by nearest neighbors [29].

In contrast to georeferenced records that did not contain enough errors to influence models, models were more probable to be biased when samples were removed, due to loss of information about where species occurred. Indeed, species distribution models became less representative by removing georeferenced records of outlying records in undersampled areas. Similarly, predicted presence area for one of the modeled species from thinned records was greater than (114%) the range area of the species. Although the range map may be too small and captured only 94% of georeferenced records, the unthinned records

avoided this error of overprediction and predicted presence area that was 84% of the range of the species. Even with more than 400 georeferenced records, thinned records may not provide enough information to allow the modeling algorithm to differentiate the climate of the presence records from the climate of the background samples. The unsampling procedures to remove samples decreased information content of input samples, resulting in unclear information delivered to the modeling algorithm. Deliberately withholding information by cleaning and other unsampling approaches can create modeling uncertainty, overpredictions, and underpredictions. Representative models can best be achieved by greater number of records that increase information about climate conditions for presence and absence classes, even if the records contain sampling bias and locational errors without cleaning [18,26,27,40].

Recognition of the trade-off in information contained by retaining samples and information lost by routine coordinate cleaning and record thinning is important. Given enough input information, the choice to thin samples or not to thin samples and clean coordinates or not clean coordinates resulted in model variation between precision and generality. If removal of samples does not affect representativeness of models, as indicated by visual inspection of predicted distribution compared to range maps, niche overlap, and area predicted as present relative to range maps, then exclusion of samples will result in slightly more general models than models from all samples. This is a small benefit relative to the risk of commission error outside of species ranges due to thinning of samples or omission error due to removal of samples from undersampled areas. Sufficient information content that includes potential errors produces representative models, with a probability of model improvement from greater sample sizes, because locational uncertainty and sampling bias are unimportant relative to sample size [18,25–27,40–42].

Thinning samples and cleaning coordinates will not correct lack of information from unsampled areas and may reduce information needed by the modeling algorithm [8]. Sampling bias may be reasonable to reduce if there is a surfeit of information provided by tens of thousands of samples of slightly different climate conditions to the modeling algorithm, and even necessary if this concentration of information overwhelms the information content from outlying samples, causing the modeling algorithm to discount these locations (specifically for continental sampling gradients between Europe and Asia for Eurasian species). Locational errors present in old records that are georeferenced by imprecise descriptions after collection may be helpful to identify and evaluate. Otherwise without evidence, reducing information through routine filtering and coordinate cleaning may be useless steps at best and detrimental to models if information is lost necessary to differentiate climate of presence samples from climate of the background (i.e., resulting in overpredictions outside of ranges) or more completely sample the distributions (i.e., resulting in underpredictions due to truncation of sampled conditions, including removal of outlying samples; [18,26,27,40]). To be clear, missing information about the climate of species distributions, rather than sampling bias and locational errors, results in poor, unrepresentative models [6]. Therefore, removing samples with the intent to avoid potential errors and sampling bias can result in information loss and biased models, with added commission or omission error.

4.2. Locational Uncertainties and Concentrated Samples Rather than Error and Bias

In terms of sampling bias, the addition of urban and clustered samples changed the specificity of the species distribution models. Urban areas represent a wide range of climate, and modeling algorithms can use information from urban records to fill in unknown areas (i.e., models of climate from urban samples predicted ‘presence’ areas, or areas with similar climate greater by a factor of 25 than urban extents; Figure 6; [7]). Urban samples

were representative of sampling bias, and remained extremely similar in comparison to unperturbed species distributions from range maps and georeferenced records. Despite spatial concentration, urban records captured meaningful ecological gradients. Nonetheless, due to concentration of samples, addition of urban records from georeferenced records to range map samples pushed species distribution models closer in similarity to the specificity of species distribution models from georeferenced records. Likewise, removal of urban records from georeferenced records pushed species distribution models closer in similarity to the generality of species distribution models from range maps. Random (20% error rate) and clustered (40% error rate) points added near georeferenced records had the same effects of diffuse and concentrated samples. The concentrated, clustered points pushed the predictions closer on the spectrum to the models from unthinned georeferenced records; likewise, the random points pushed the predictions closer on the spectrum to the models from thinned georeferenced records.

Assuming sufficient input information about climate within species ranges, thinned records will offer slightly different representations and characteristics than unthinned records, a trade-off between precision and generality quantified directly by area of predicted presence, which is a concrete measurement that demonstrates precision or generality in area of predictions. Greatest precision for model predictions required use of unthinned georeferenced records, for which the concentrated samples corroborated validity and the predictions were precise to sample locations, without generalizing to greater areas. Treating sampling bias involves removal of the concentrated samples by thinning samples, and results in more uncertain, generalized predictions, which are potentially erroneous by including commission error. Nevertheless, species distribution models from georeferenced records were extremely similar, specifically as judged by a mean niche overlap value of 1.0, with no niche overlap value less than 0.99 by species. This overlap in models is likely greater than model divergence generated by application of other modeling options (e.g., modeling algorithms) or the varied realizations of range maps developed by different expert opinions.

Characterizations of precision, sampling bias, and concentration on one end of a gradient and generality, locational errors, and dispersion at the other end of a gradient help to interpret the influences of potential locational errors and sampling bias. Species distribution models from concentrated georeferenced records are more locationally correct than species distribution models from thinned georeferenced records. The concentrated georeferenced records corroborated locational correctness of the georeferenced records and also limited generalizing model predictions to additional areas, resulting in decreased areas of potential locational error of predictions, relative to thinned georeferenced records. Referring to localized, clustered samples and sample concentration, rather than sampling bias, differentiates that predictive modeling is not hypothesis testing. Concentrated georeferenced records do have the greatest sampling bias, in the sense that correctly sampled locations have greater weight than lesser-known locations [8]. Sampling bias in this regard is reduced by thinning georeferenced records and consequent model generalization to greater areas, but at the risk of known commission error, sometimes with potential overpredictions from species distribution models that exceed range areas. Unthinned georeferenced records create species distribution models characterized by precision, fit to samples of species occurrences, sampling bias or concentrated distributions relative to species distribution models from thinned records, which are characterized in terms of generality, reduced fit to samples, locational error or uncertainty, and dispersed distributions.

Addressing locational errors, which may result in commission error and model generalization, and sampling bias, which may result in omission error and model concentration, are conflicting objectives. All of the random samples from range maps were locational

errors because the randomly generated samples have no evidence of correct observation, beyond generally being located within species ranges. Despite 100% locational errors, no sampling bias occurred as samples were completely random, and species distribution models were representative of species ranges, as developed by experts. In contrast, georeferenced records, particularly with a coordinate uncertainty of less than 1 km, are as locationally correct as possible, but with uneven sampling and unsampled areas. Given enough samples, georeferenced records allow generation of alternative representations of species ranges, the areas of occupancy, which are more precisely located and concentrated than general, continuously dispersed range maps.

4.3. Removal of Non-Terrestrial Points as Part of the Coordinate Cleaning Workflow

Intentional coordinate cleaning to simply remove duplicated coordinates and differentiate land areas is straightforward to add to workflows (Figure 7). The major concerns, aside from coordinates in urban areas, are duplicated information, collection year, basis of record (i.e., observations, preserved specimens, fossils, unknown), and separation of land from water [19]. Processing data to remove duplicated coordinates by species and fossils and unknown records is basic. Collection years likewise can be specified easily. Older records from museums probably contain errors but these require careful examination to understand the specific issues rather than automatic cleaning. Similarly taxonomic resolution also entails attention to preserve and reconcile at least primary synonyms, due to deprecation or unresolved and conflicting taxonomies, rather than simply removing samples of species with different scientific names in favor of one accepted name.

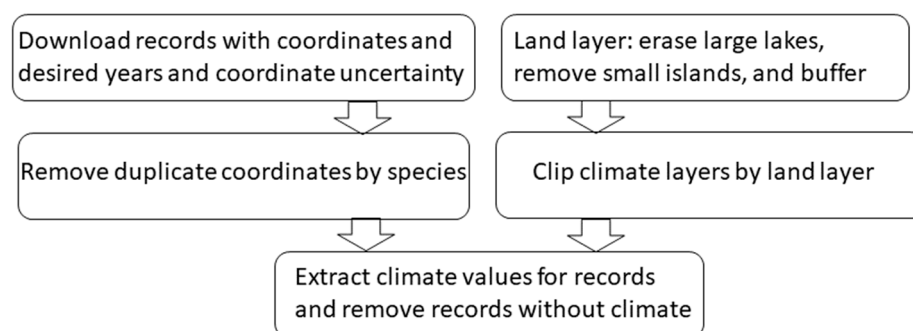


Figure 7. Simple workflow for cleaning coordinates.

While terrestrial species with samples located in oceans, or water in general, seems to be the most straightforward concern related to coordinate cleaning to resolve, in actuality, this issue is complicated by the lack of resolution between water and land. The georeferenced records had a coordinate uncertainty less than 1 km, which is less than the coordinate uncertainty or resolution of map layers at 1:10,000,000 or 1:110,000,000 scales [24]. Overall, then, the locations of the georeferenced coordinates are likely to be more accurate than the resolution of the land layers. In this study, adding a 2 km buffer to a 1:10 m land layer was able to salvage most georeferenced records (99.8%), as opposed to losing 2% to 3% of all records with a 1:10 m land layer or 1:110 m land layer, respectively, without buffering. While 2% to 3% of all records may not be that critical, these records will be proportionally greater in species with ranges along coastlines.

Indeed, even if exclusion of aquatic records for terrestrial species is not a particular interest, as a practical matter, if the study extent, such as the North American continent, contains a boundary between land and ocean, then the coastline will act as the border to retain or remove records. Delineation of coastlines based on resolution will affect the number of records kept in the study extent. Addition of a buffer is necessary to avoid

unintentional sample exclusion along coastlines, and the buffer could match the maximum distance uncertainty of the georeferenced records (i.e., about 35 km).

It may be desirable to also remove georeferenced records (and predictions) located in inland water, but perhaps limited to relatively stable large lakes. With large lakes removed, lakes and rivers are 5% of the North American land surface [43]. Lakes are not monolithic, and often have islands or inlets with trees, including in large lakes [44]. Rivers can be extremely dynamic, with oxbow formation and other processes. Many rivers are not correctly delineated in land covers, particularly coarse ones (e.g., [45]). Applying a coarse scale map to account for <5% of land area actually would contain unknown errors and be misleading in that it would appear that the water area is removed when in fact water remained and land was removed. The error may be magnified over time to the future predictions, if inland water levels changed with shifting precipitation and increased evapotranspiration under warmer temperatures, whereas sea level rose. If agricultural intensification occurs in Canada, water sources and hydrological networks may be modified through tile drainage to move water away from crops and terrestrial land, changing water locations [46]. Making an adjustment to identify inland water beyond large lakes may be only an appearance of an adjustment, due to spatiotemporal dynamics. Moreover, unless resolution is very fine (e.g., 10 m), then multiple cover classes, including intermixed land and water, may be contained within one pixel. Assignment of one class for heterogeneous classes is a limitation of coarse resolution data sources, along with accuracy of classes [47].

5. Conclusions

Species distribution models using georeferenced records or random samples from range maps are both robust to perturbation by concentrated samples and locational uncertainties, given enough samples for the modeling algorithm to differentiate suitable and unsuitable conditions. Locational uncertainties inside species ranges are compatible with sampled climate conditions. If locational uncertainties are outliers in relation to the climate samples, the modeling algorithm will be able to use them to fill in the climate amplitude of species ranges and differentiate suitable conditions from unsuitable conditions. If the flagged errors are errors that are not compatible with other samples, they will be discarded or discounted by the modeling algorithm. Procedures to remove samples decrease information content of input samples, resulting in unclear information being delivered to the modeling algorithm, and resultant omission and commission errors. The risks of poor models due to unclear samples of suitable conditions after sample removal outweigh the benefits of limited generalization of models with thinned georeferenced records. Rather than removing samples unnecessarily to avoid potential errors, the major prerequisite for species distribution models is that researchers deliver clear information, by not cleaning samples beyond necessary, to allow the modeling algorithm to separate presence from absence classes.

In modeling, precision, sampling bias, and concentration characterize one end of the gradient and generality, locational errors, and dispersion characterize the other end of the gradient. To clarify, models with precise representation from concentrated samples (rather than sampling bias), with limited locational and commission error, represent one side of the modeling spectrum. Model generalization from thinned samples or samples from range maps, with limited sampling bias and omission error yet locational uncertainties (rather than errors), represents the other side of the spectrum. Randomly generated points from range maps are an extension of the concept of thinning. In fact, all of the random samples from within ranges may be locationally incorrect at fine resolutions but have no sampling bias and models are general representations of species ranges. Different realizations occur when depicting any landscape patterns, varying from models of comprehensive

range coverage (generality, with locational/commission error) or models of the recorded observations (species area of occupancy, with omission error). Species distribution models require complete information available from samples, despite data cleaning and thinning becoming common procedures in the modeling of species distributions, which may result in different inferences that shape ecological interpretations and conservation outcomes.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author. Due to deliberate generation of poor models, the models were not shared at a public repository.

Acknowledgments: I thank GBIF, iNaturalist, IUCN, and contributors to the range maps and georeferenced records. I acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, and I thank the climate modeling groups for producing and making available model outputs. Likewise, I thank the Global Biodiversity Information Facility and contributors. This research was supported by the USDA Forest Service, Rocky Mountain Research Station, internal funding from the Infrastructure Investment and Jobs Act of 2021. 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.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Hanberry, B.B. Imposing consistent global definitions of urban populations with gridded population density models: Irreconcilable differences at the national scale. *Landsc. Urban Plan.* **2022**, *226*, 104493. [CrossRef]
2. Moudrý, V.; Bazzichetto, M.; Remelgado, R.; Devillers, R.; Lenoir, J.; Mateo, R.G.; Lembrechts, J.J.; Sillero, N.; Lecours, V.; Cord, A.F.; et al. Optimising occurrence data in species distribution models: Sample size, positional uncertainty, and sampling bias matter. *Ecography* **2024**, *12*, e07294. [CrossRef]
3. iNaturalist. A Community for Naturalists iNaturalist. Available online: <https://www.inaturalist.org/> (accessed on 24 November 2024).
4. Zizka, A.; Antonelli, A.; Silvestro, D. Sampbias, a method for quantifying geographic sampling biases in species distribution data. *Ecography* **2021**, *44*, 25–32. [CrossRef]
5. Inman, R.; Franklin, J.; Esque, T.; Nussear, K. Comparing sample bias correction methods for species distribution modeling using virtual species. *Ecosphere* **2021**, *12*, e03422. [CrossRef]
6. Thuiller, W.; Brotons, L.; Araújo, M.B.; Lavorel, S. Effects of restricting environmental range of data to project current and future species distributions. *Ecography* **2004**, *27*, 165–172. [CrossRef]
7. McCarthy, K.; Fletcher, R.J., Jr.; Rota, C.T.; Hutto, R.L. Predicting species distributions from samples collected along roadsides. *Conserv. Biol.* **2012**, *26*, 68–77. [CrossRef]
8. Gábor, L.; Moudrý, V.; Barták, V.; Lecours, V. How do species and data characteristics affect species distribution models and when to use environmental filtering? *Int. J. Geogr. Inf. Sci.* **2020**, *34*, 1567–1584. [CrossRef]
9. Ten Caten, C.; Dallas, T. Thinning occurrence points does not improve species distribution model performance. *Ecosphere* **2023**, *14*, e4703. [CrossRef]
10. Lamboley, Q.; Fourcade, Y. No optimal spatial filtering distance for mitigating sampling bias in ecological niche models. *J. Biogeogr.* **2024**, *51*, 1783–1794. [CrossRef]
11. Kadmon, R.; Farber, O.; Danin, A. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecol. Appl.* **2004**, *14*, 401–413. [CrossRef]
12. Hanberry, B.B. Global population densities, climate change, and the maximum monthly temperature threshold as a potential tipping point for high urban densities. *Ecol. Indic.* **2022**, *135*, 108512. [CrossRef]
13. Breiman, L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Stat. Sci.* **2001**, *16*, 199–231. [CrossRef]
14. Robertson, M.; Visser, V.; Hui, C. Biogeo: An R package for assessing and improving data quality of occurrence record datasets. *Ecography* **2016**, *39*, 394–401. [CrossRef]
15. Zizka, A.; Silvestro, D.; Andermann, T.; Azevedo, J.; Duarte Ritter, C.; Edler, D.; Farooq, H.; Herdean, A.; Ariza, M.; Scharn, R.; et al. CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. *Methods Ecol. Evol.* **2019**, *10*, 744–751. [CrossRef]

16. Ribeiro, B.R.; Velazco, S.J.E.; Guidoni-Martins, K.; Tessarolo, G.; Jardim, L.; Bachman, S.; Loyola, R. bdc: A toolkit for standardizing, integrating and cleaning biodiversity data. *Methods Ecol. Evol.* **2022**, *13*, 1421–1428. [CrossRef]
17. Gábor, L.; Jetz, W.; Zarzo-Arias, A.; Winner, K.; Yanco, S.; Pinkert, S.; Marsh, C.J.; Rogan, M.S.; Mäkinen, J.; Rocchini, D.; et al. Species distribution models affected by positional uncertainty in species occurrences can still be ecologically interpretable. *Ecography* **2023**, *2023*, e06358. [CrossRef]
18. Smith, A.B.; Murphy, S.J.; Henderson, D.; Erickson, K.D. Including imprecisely georeferenced specimens improves accuracy of species distribution models and estimates of niche breadth. *Glob. Ecol. Biogeogr.* **2023**, *32*, 342–355. [CrossRef]
19. 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.; et al. No one-size-fits-all solution to clean GBIF. *PeerJ* **2020**, *8*, e9916. [CrossRef]
20. Global Biodiversity Information Facility [GBIF]. Free and Open Access to Biodiversity Data. 2024. Available online: www.gbif.org (accessed on 2 August 2024).
21. Führling-Potschkat, P.; Kreft, H.; Ickert-Bond, S.M. Influence of different data cleaning solutions of point-occurrence records on downstream macroecological diversity models. *Ecol. Evol.* **2022**, *12*, e9168. [CrossRef] [PubMed]
22. Feeley, K.J.; Silman, M.R. 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.* **2010**, *37*, 733–740. [CrossRef]
23. Natural Earth. 2024. Available online: <https://www.naturalearthdata.com/downloads/> (accessed on 24 November 2024).
24. GIS StackExchange. What Does 1:10m Mean Related to Map Resolution? 2017. Available online: <https://gis.stackexchange.com/questions/21391/what-does-110m-mean-related-to-map-resolution> (accessed on 24 November 2024).
25. Graham, C.H.; Elith, J.; Hijmans, R.J.; Guisan, A.; Townsend Peterson, A.; Loiselle, B.A. and NCEAS Predicting Species Distributions Working Group. The influence of spatial errors in species occurrence data used in distribution models. *J. Appl. Ecol.* **2008**, *45*, 239–247. [CrossRef]
26. Mitchell, J.; Monk, J.; Laurenson, L. Sensitivity of fine-scale species distribution models to locational uncertainty in occurrence data across multiple sample sizes. *Methods Ecol. Evol.* **2017**, *8*, 12–21. [CrossRef]
27. Gaul, W.; Sadykova, D.; White, H.J.; Leon-Sanchez, L.; Caplat, P.; Emmerson, M.C.; Yearsley, J.M. Data quantity is more important than its spatial bias for predictive species distribution modelling. *PeerJ* **2020**, *8*, e10411. [CrossRef]
28. IUCN. Spatial data download. Available online: <http://www.iucnredlist.org/resources/spatial-data-download> (accessed on 10 May 2024).
29. Menardi, G.; Torelli, N. Training and assessing classification rules with imbalanced data. *Data Min. Knowl. Discov.* **2014**, *28*, 92–122. [CrossRef]
30. Hanberry, B.B. Practical guide for retaining correlated climate variables and unthinned samples in species distribution modeling, using random forests. *Ecol. Inform.* **2024**, *79*, 102406. [CrossRef]
31. Moudry, V. Modelling species distributions with simulated virtual species. *J. Biogeogr.* **2015**, *42*, 1365. [CrossRef]
32. Hijmans, R.J.; Phillips, S.; Leathwick, J.; Elith, J. Dismo Package for R. 2011. Available online: <https://cran.r-project.org/package=dismo> (accessed on 20 August 2024).
33. Hanberry, B.B. Urban land expansion and decreased urban sprawl at global, national, and city scales during 2000 to 2020. *Ecosyst. Health Sustain.* **2023**, *9*, 0074. [CrossRef]
34. Karger, D.N.; Conrad, O.; Böhner, J.; Kawohl, T.; Kreft, H.; Soria-Auza, R.W.; Zimmermann, N.E.; Linder, H.; Kessler, M. Climatologies at high resolution for the Earth land surface areas. *Sci. Data* **2017**, *4*, 170122. [CrossRef] [PubMed]
35. Karger, D.N.; Conrad, O.; Böhner, J.; Kawohl, T.; Kreft, H.; Soria-Auza, R.W.; Zimmermann, N.E.; Linder, H.; Kessler, M. Data from: Climatologies at high resolution for the earth's land surface areas. *EnviDat* **2018**. [CrossRef]
36. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **2008**, *28*, 1–26. [CrossRef]
37. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2024.
38. Lobo, J.M.; Jiménez-Valverde, A.; Real, R. AUC: A misleading measure of the performance of predictive distribution models. *Glob. Ecol. Biogeogr.* **2008**, *17*, 145–151. [CrossRef]
39. Warren, D.L.; Glor, R.E.; Turelli, M. Environmental niche equivalency versus conservatism: Quantitative approaches to niche evolution. *Evolution* **2008**, *62*, 2868–2883. [CrossRef] [PubMed]
40. Steen, V.A.; Tingley, M.W.; Paton, W.; Elphick, C.S. Spatial thinning and class balancing: Key choices lead to variation in the performance of species distribution models with citizen science data. *Methods Ecol. Evol.* **2021**, *12*, 216–226. [CrossRef]
41. Osborne, E.; Leitao, J. Effects of species and habitat positional errors on the performance and interpretation of species distribution models. *Divers. Distrib.* **2009**, *15*, 671–681. [CrossRef]
42. Soultan, A.; Safi, K. The interplay of various sources of noise on reliability of species distribution models hinges on ecological specialisation. *PLoS ONE* **2017**, *12*, e0187906. [CrossRef] [PubMed]
43. Commission for Environmental Cooperation 2005. Land Cover, 2005 (MODIS, 250m). Available online: <http://www.cec.org/north-american-environmental-atlas/land-cover-2005-modis-250m/> (accessed on 24 November 2024).

44. NASA, Island in a Lake on an Island in a Lake on an Island. 2024. Available online: <https://earthobservatory.nasa.gov/images/85342/island-in-a-lake-on-an-island-in-a-lake-on-an-island>. (accessed on 24 November 2024).
45. Hanberry, B.B.; Hanberry, P. Illustrating land cover change associated with erosion management of the Little Blue River, Kansas, USA. *River* **2023**, *2*, 421–432. [[CrossRef](#)]
46. Cristine, L.E.; Kerr, J.T. Habitat loss, climate change, and emerging conservation challenges in Canada. *Can. J. Zool.* **2011**, *89*, 435–451. [[CrossRef](#)]
47. Pouliot, D.; Latifovic, R.; Zabcic, N.; Guindon, L.; Olthof, I. Development and assessment of a 250 m spatial resolution MODIS annual land cover time series (2000–2011) for the forest region of Canada derived from change-based updating. *Remote Sens. Environ.* **2014**, *140*, 731–743. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.