Received: 25 February 2019 | Revised: 9 July 2019 | Accepted: 11 August 2019

DOI: 10.1111/ddi.12985

BIODIVERSITY RESEARCH

Diversity and Distributions WILEY

An evaluation of stringent filtering to improve species distribution models from citizen science data

Valerie A. Steen | Chris S. Elphick | Morgan W. Tingley

Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, Connecticut

Correspondence
Valerie A. Steen, Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT 06269.
Email: valerie.steen@gmail.com

Funding information
Connecticut Department of Energy and Environmental Protection

Editor: Thomas (OE) Albright

Abstract

Aim: Citizen science data are increasingly used for modelling species distributions because they offer broad spatiotemporal coverage of local observations. However, such data are often collected without experimental design or set survey methods, raising the risk that bias and noise will compromise modelled predictions. We tested the ability of species distribution models (SDMs) built from these low-structure citizen science data to match the quality of SDMs from systematically collected data and tested whether stringent data filtering improved predictions.

Location: Northeastern USA.

Methods: We evaluated models built from a rapidly growing dataset of avian occurrences reported by birders—eBird—against models built from four independent, systematically collected datasets. We developed SDMs for 96 species using both data sources and compared their predictive abilities. We also tested whether culling eBird data by applying stringent data filters on survey effort or observer expertise improved predictions.

Results: We found that SDMs built from low-structure citizen science data matched or exceeded performance of SDMs from systematically collected datasets for 12%–31% of species ($\bar{x}$ = 22%), depending on the dataset. At least one culling option produced equivalent or better performance for 40%–70% of species ($\bar{x}$ = 49%). Data culling by restricting survey effort improved predictions more than restricting by observer expertise. The optimal effort restriction differed by dataset, and for three of the datasets was further informed by species traits.

Main conclusions: Species distribution models developed using low-structure citizen science data sometimes performed as well as those from systematic data. Culling generally improved models, but results were heterogeneous, prohibiting clear recommendations for how to cull. Our results indicate that the growing availability of citizen science data holds potential for creating high-quality spatial predictions, but that time should be invested in determining how best to cull datasets and that one-size-fits-all solutions beyond basic outlier filtering may be hard to find.

KEYWORDS
citizen science, culling, data filtering, eBird, model evaluation, observer expertise, occurrence data, species distribution models, survey effort

1 | INTRODUCTION

Ecologists have growing options concerning where they source spatial data on species' occurrence and abundance. Systematically collected data are traditional ecological data products that use strict survey designs to ensure sufficient sample sizes and balanced spatial and temporal sampling (Guisan, Thuiller, & Zimmermann, 2017), while minimizing variation in the observation process. Citizen science data, by contrast, are a rapidly growing data source for ecological research as user-friendly web-based platforms for data collection proliferate (Devictor, Whittaker, & Beltrame, 2010; Lowman, D'Avanzo, & Brewer, 2009; Sullivan et al., 2014). Although some citizen science efforts are highly structured with strict protocols (e.g. Robbins, Bystrak, & Geissler, 1986), much of the recent growth involves platforms that allow untrained contributors to choose when, where and how they collect data (e.g. Hochachka et al., 2012; iNaturalist.org, 2019). The resulting data, therefore, often suffer from limitations, including spatial biases, imprecise temporal and spatial resolutions, and under- or over-reporting of species (Dickinson, Zuckerberg, & Bonter, 2010; Fitzpatrick, Preisser, Ellison, & Elkinton, 2009; Steger, Butt, & Hooten, 2017; Szabo, Vesk, Baxter, & Possingham, 2010; Tulloch, Mustin, Possingham, Szabo, & Wilson, 2013; Tulloch & Szabo, 2012; Tye, McCleery, Fletcher, Greene, & Butryn, 2017). However, because citizen scientists provide such high data volume, determining how best to use these data for ecological research will likely improve biogeographical insights, conservation decisions and conservation outcomes (Dickinson et al., 2010; La Sorte et al., 2018).

Species distribution models (SDMs) are a popular tool for using georeferenced species occurrences in relation to environmental conditions to predict occurrence across regions (Franklin, 2010). Accurate predictions of species distributions are affected by attributes of the underlying occurrence data used to build SDMs. Ideally, SDMs are built from systematically collected observational datasets, but as these data are often not available for most species and/or study regions, SDMs are commonly built with presence-only data derived from biological specimen records (Peterson et al., 2011).

Citizen science data have great potential as a cost-efficient alternative to systematic data that can provide access to more species and broader spatial coverage (Pimm et al., 2014). However, biases and noise potentially limit ability of citizen science data to be used reliably for mapped output and at scales relevant to conservation (Kremen et al., 2008; Rondinini, Wilson, Boitani, Grantham, & Possingham, 2006). A key advance in the use of citizen science data would be the ability to reliably map species distributions that meet or exceed what can be accomplished from systematic surveys.

Reducing bias and noise in citizen science datasets can be achieved by either filter-based or statistical techniques (Bird et al., 2014; Isaac, Strien, August, Zeeuw, & Roy, 2014). Filtering removes problematic observations, such as outliers, or those contributing to sampling or spatial bias (Fink et al., 2010; Bonter & Cooper, 2012; Butt, Slade, Thompson, Malhi, & Riutta, 2013; Boria, Olson, Goodman, & Anderson, 2014; Robinson et al., 2018; Tye et al., 2017). Statistical techniques fit models that address sampling bias and observation heterogeneity. For example, occupancy modelling has been successfully used to improve SDMs by correcting for imperfect detection (Higa et al., 2015; Kéry, Gardner, & Monnerat, 2010; van Strien, van Swaay, & Termaat, 2013).

Data culling, whereby stringent data filters are applied to retain only the highest quality data, is one option for improving models derived from citizen science data, but the trade-offs between reduced bias and loss of precision are poorly known. In one study, Kamp, Oppel, Heldbjerg, Nyegaard, and Donald (2016) found that culling to use only the highest quality data to estimate population trends did not overcome the effects of data loss. In comparison, Robinson, Ruiz-Gutierrez, and Fink (2018) successfully used spatial data culling to improve SDMs for a rare species.

Because species vary widely in their prevalence, aggregation patterns and the ease with which they can be identified, citizen science data quality may vary considerably (Dickinson et al., 2010; Fitzpatrick et al., 2009; Kamp et al., 2016), such that using and culling citizen science data may be more appropriate for some species than others. Taxa with smaller body sizes or lower densities often result in fewer detections in citizen science databases relative to benchmark surveys (Fitzpatrick et al., 2009; Kamp et al., 2016; Steger et al., 2017; Ward, 2014). Uncommon species may be under-reported or, conversely, over-reported when rarity increases interest (Farmer, Leonard, & Horn, 2012; Swanson, Kosmala, Lintott, & Packer, 2016). Other species may be under-reported in citizen science datasets simply because they are challenging to identify, whether requiring knowledge of vocalizations or lacking distinguishing features (Crall et al., 2011; Dickinson et al., 2010; Ratnieks et al., 2016; Shea, Peterson, Wisniewski, & Johnson, 2011; Swanson et al., 2016). Such species-specific traits which result in differential representation in citizen science databases may be indicative of whether species are best studied using systematic data (where such issues are supposedly minimized), or whether citizen science datasets can substitute.

One of the largest and fastest growing ecological citizen science datasets is eBird (Hochachka et al., 2012; Sullivan et al., 2014). eBird provides an online portal for reporting bird observations, with over 100 million sightings logged annually by hundreds of thousands of users. While the platform is flexible in how participants collect data, ancillary survey information is collected describing time, location, travelling distance, and whether all

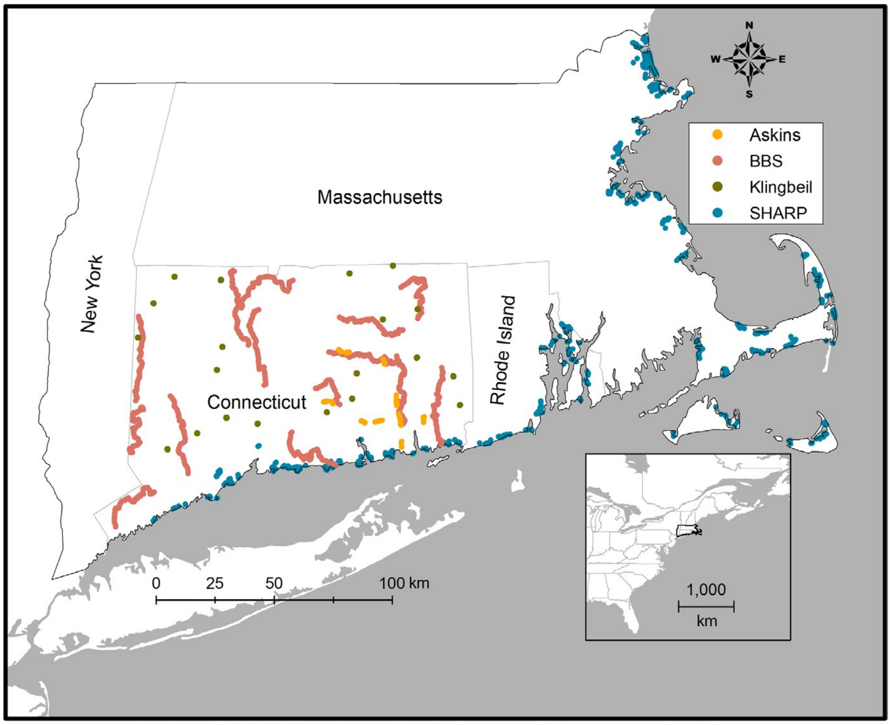


FIGURE 1 Study area in the northeastern USA showing the extent of eBird citizen science training data by the black outline in map and inset. Points show the locations of observations from the four benchmark datasets

avian species were reported. This survey information and the high density of observations in many areas make the dataset ideal for exploring whether SDMs derived from citizen science data can match those from systematic surveys, and whether data culling improves models.

Here, we challenge SDMs created from eBird data to predict occurrences independently assessed from four different 'benchmark' systematically collected datasets of species occurrence. Additionally, we investigate whether data culling of eBird based on survey effort or observer expertise—two major sources of potential bias in citizen science datasets (Kelling, Fink, et al., 2015)—improve predictions. Specifically, we ask: (a) Can low-structure citizen science data produce model predictions that match the quality of those from systematically collected data? (b) Does selective data culling improve model predictions? and (c) Do species traits explain variation in the accuracy of predictive models in ways that could guide which culling decisions to make?

2 | METHODS

2.1 | Study area

We developed models for a 43,000 km^2 region in the northeastern United States (Figure 1). Because most of our systematic data came from studies in the state of Connecticut, we used eBird data from Connecticut and surrounding states similar in climate, topography and habitat classes. This included Massachusetts to the north, Rhode Island to the east and the adjacent portion of New York west to the Hudson River. The study area is dominated by deciduous forest and developed land cover and contains some mixed conifer forest as well as spruce–fir forest. It also includes extensive coastline and small amounts of shrublands, grasslands, croplands and freshwater wetlands.

2.2 | eBird data

eBird records are submitted in checklist format listing the counts of each species encountered. Checklists include information on observation duration, distance travelled and other method-related metadata. We obtained these data by directly downloading the eBird Basic Dataset (https://ebird.org/science/download-ebird-data-products) on 6/27/2017. We restricted the data to only include (a) 'complete' checklists in which all birds observed were recorded and thus we inferred species absence when not reported on a checklist (Sullivan et al., 2014); (b) observations from 2010 to 2016—a period that overlapped the years of three of the four benchmark datasets; (c) data from 4:15 a.m. to 12:00 p.m., when many bird species are considered most detectable; (d) seasonal dates corresponding to the respective benchmark datasets (Table S1); (e) surveys that covered less than 8.1 km (Fink et al., 2010) and (f) surveys that lasted up to 300 min to increase detection data for more uncommon species.

We created three types of eBird training datasets: 'full' datasets included all eBird data that matched the criteria described above, while 'culled' and 'random' datasets were reduced subsets. To address spatial sampling bias, all datasets were spatially thinned, so that no observation was within 1 km of another, using the R package spThin v.0.1.0 (Aiello-Lammens, Boria, Radosavljevic, Vilela, & Anderson, 2015). We used R versions 3.3.3 and 3.5.2 for our various analyses. Based on results of exploratory analyses, we chose a more conservative 1-km thin over a 500-m thin and retained data from the

TABLE 1 Summaries of area under the receiver operating characteristic curve (AUC) results obtained using models built with eBird citizen science data to discriminate between occurrence and non-occurrence in four benchmark datasets that used skilled technicians and detailed data collection protocols (Askins, n = 40 species; BBS, n = 61; Klingbeil, n = 25; and SHARP, n = 47)

Species summaries	Askins			BBS			Klingbeil			SHARP		
	All	Random	Culled	All	Random	Culled	All	Random	Culled	All	Random	Culled
Mean AUC[a]	0.623	0.648	0.711	0.656	0.658	0.684	0.626	0.663	0.718	0.615	0.626	0.667
Mean AUC; benchmark data	0.731			0.694			0.748			0.652		
Proportion with AUC ≥0.7	0.275	0.175	0.575	0.328	0.328	0.393	0.200	0.320	0.560	0.149	0.106	0.277
Prop. with AUC ≥0.7; benchmark data	0.650			0.492			0.640			0.128		
Proportion of best models	0.025	0.025	0.950	0.066	0.066	0.868	0.040	0.080	0.880	0.064	0.020	0.914
Proportion meeting benchmark	0.225	0.250	0.400	0.311	0.246	0.443	0.120	0.160	0.400	0.234	0.191	0.702
Model summaries												
Mean ± *SD* AUC[b]	0.623 ± 0.095 (0.500, 0.802)	0.617 ± 0.074 (0.527, 0.816)	0.617 ± 0.092 (0.500, 0.840)	0.656 ± 0.090 (0.520, 0.830)	0.641 ± 0.085 (0.523, 0.833)	0.642 ± 0.092 (0.510, 0.830)	0.626 ± 0.120 (0.506, 0.894)	0.631 ± 0.095 (0.531, 0.879)	0.622 ± 0.100 (0.500, 0.890)	0.615 ± 0.078 (0.510, 0.798)	0.603 ± 0.064 (0.519, 0.796)	0.604 ± 0.073 (0.510, 0.790)

Note: For each comparison, we present results from models that used the full citizen science dataset, random subsets, and culled subsets. We also present cross-validated benchmark data results for mean AUC and the proportion of models with AUC ≥0.7.

[a] Averaged across species when using all data; averaged across models with highest AUC for a given species for random and culled subset models.

[b] Averaged across all models; parentheses show 0.025 and 0.975 quantiles.

full extent of the study area versus masking data to match the extent of benchmark datasets (see *Spatial Thinning* and *Masking* in Appendix S1 in Supporting Information for more details). These scales created spacing similar to that of the point-counts in our benchmark datasets.

To create the culled datasets, we first defined levels of observer expertise and survey effort for each eBird checklist. Expertise was modelled using the Poisson generalized additive mixed model described by Kelling, Johnston, et al. (2015) and Johnston, Fink, Hochachka, and Kelling (2018). This model quantifies expertise from eBird checklists by estimating the false absence reporting rate of each observer in the study region (Kelling, Johnston, et al., 2015). The model relates variation in the total number of species reported in each checklist to time of day, day of year, distance travelled, time spent observing, habitat, habitat diversity and protocol type. We extracted habitat classes and habitat diversity (using the Gini–Simpson diversity index) from the 2011 National Land Cover Database (Homer et al., 2015). Because observers are expected to vary in their ability to detect and identify different species, the model includes a random intercept for each observer as well as a random slope for each observer's effect on time spent observing. Predictions from this model for a standardized survey then provide standardized estimates of relative observer expertise along a continuous scale. Using these expertise scores, we either used all checklists ('any expertise'), those from the top two-thirds of observers ('okay'), those from the top third ('better'), or those from only the top 15% ('top').

To cull by effort, we focused on the distance travelled for a checklist. We defined four subsets using quartiles of the distribution of distances: 'short' (all checklists ≤0.805 km in distance travelled), 'medium' (>0.805, ≤1.609 km), 'lengthy' (>1.609, ≤3.219 km) and 'very lengthy' (>3.219, <8.0 km). Unlike the expertise classes, which sequentially reduce the size of the dataset, effort subsets divided the data into subsets that were equal in size. By combining the two culling criteria, we created a total of 16 'culled' datasets.

Finally, to test whether any culling effects could simply be an artefact of dataset reduction, we randomly subsetted the eBird data to create 16 'random' datasets that matched the size of the 16 'culled' datasets. We repeated each of these 16 randomizations ten times.

2.3 | Benchmark data

eBird models were evaluated separately against each of four benchmark datasets collected in different habitats in our region. While these datasets differed in the specifics of their study designs, all involved systematic surveys using point-count methods typical of many avian studies and used observers skilled in bird identification (Figure 1, Table S1). The 'Askins' dataset covered shrubland and forest edge habitats (Askins, Folsom-O'Keefe, & Hardy, 2012). 'BBS' covered primarily forest and developed habitats along roadsides as part of the North American Breeding Bird Survey (Pardieck, Ziolkowski, Lutmerding, Campbell, & Hudson, 2016), another citizen science effort, but one with a systematic survey protocol that skilled observers conduct year after year. 'Klingbeil' surveys covered forest interior sites (Klingbeil & Willig, 2015). 'SHARP' consisted of tidal marsh habitat surveys (Wiest et al., 2016).

2.4 | Environmental and observation covariates

Species distributions were modelled as functions of environmental and observation covariates (Table S2). We chose environmental covariate sets separately for each benchmark dataset given the primary habitat(s) sampled. Environmental covariates included class variables describing land cover, forest fragmentation and/or wetlands, and numerical variables describing topography, housing density and/or income. The same covariate sets were used across eBird and benchmark dataset models for a given comparison. Thus, species that appeared in multiple benchmark datasets (e.g. Red-winged Blackbird *Agelaius phoeniceus*; Table S3) had different covariates for each comparison.

All environmental covariates were modelled as the proportion (for class variables) or mean (for numerical variables) of that covariate in a 200-m radius surrounding the point-count location (systematically collected data), or the reported mid-point for the eBird checklist (citizen science data). A 200-m radius was chosen because it matched a typical detection radius for our species at a point location. Most eBird surveys are travelling surveys, and observers sometimes report the beginning or end location rather than, as suggested, the mid-point location (Munson et al., 2010). eBird surveys also, frequently, cover a larger area than the 200-m radius would encompass. However, because we were challenging noisy citizen science data to make predictions to the relatively fine-scale observations in the systematic data, we settled on a radius optimized for the systematic data, especially after finding in preliminary analyses that eBird surveys trained with covariates calculated at an 800-m or 1,500-m radius did not improve models relative to a 200-m radius (see Appendix S1 for more details, *Covariate Scale* and Table S4). To capture additional variation in the observation process for eBird records, models also included time of day, survey duration, distance covered and expertise score for the observer.

2.5 | Prediction and evaluation

We identified species with a minimum prevalence in each benchmark dataset of 0.05 for inclusion in our analyses resulting in 173 species by comparison combinations (Table S3). For each species within each benchmark dataset, we ran 177 models using training data from eBird (1 full model, 16 culled models, 160 random models). In all cases, we modelled species occurrence using classification random forests implemented in the R package randomForest v.4.6-14 (Breiman, 2001; Liaw & Wiener, 2002). We specified that a relatively large number of trees (*n* = 3,000) be used for each model. We otherwise used default settings where the number of variables tried at each split was the square root of the total number, sampling of data was done with replacement, and unstratified class sampling was used (see *Class Imbalance* in Appendix S1 for justification for not stratifying). For predictions, we used a consistent time of 7:00 a.m. Because distance and expertise vary by culled subset, and duration covaries with distance, we used the mean values from each subset as the standardized values for predictions to benchmark data.

To evaluate the eBird-derived models, we used the area under the receiver operating characteristic curve metric (AUC; Fielding & Bell, 1997) to assess whether their continuous probability predictions could discriminate between occurrence and non-occurrence (assumed from non-detection) in the benchmark data. We compared AUC results between models built with the complete eBird dataset and those from the culled and random subsets, averaging AUC values over the 10 repetitions of each random scenario.

We established benchmark model performance using the benchmark datasets, applying 10-fold cross-validation, assessing SDM performance using AUC to 10% of the data for each of 10 folds, and averaging across the 10 results. To mitigate predictive advantage owing to spatial autocorrelation among data points, we used spatial sets to group nearby observations when assigning them to testing versus training datasets within the 10-fold procedure (Roberts et al., 2017). We calculated the performance difference (ΔAUC) relative to the equivalent eBird model.

2.6 | Species traits

We summarized four species' characteristics we expected might influence citizen science data quality: prevalence, ubiquity, abundance and identifiability. Prevalence and ubiquity are measures of how common and widespread a species is, respectively. Prevalence was calculated as the frequency with which the species was reported across eBird checklists and ubiquity as the proportion of 5 × 5 km cells within the study area with records of the species (Figures S1 and S2). We used the median abundance from checklists in which a species was detected as a measure of local densities (Figure S3). To describe prevalence, ubiquity and abundance, we used only eBird checklists that met our criteria for use in model building, but further restricted the dates to 1 June–15 July to target peak breeding and further exclude migrating birds. Finally, we assigned each species a measure of how easy it is to identify, using the expected rate of reporting a species dependent on an observer's expertise, after Johnston et al. (2018). We relativized reporting rates across species by dividing that of the observer at the 97.5th expertise quantile by the observer at the 25th expertise quantile. Thus, species with identifiability values around 1 are expected to be reported by observers with modest skill at equal rates to those with high skill, whereas those with values around 0.5 would be reported half as frequently (Figure S4). To examine relationships between each of the species traits and eBird model performance relative to the benchmark performance, we plotted ΔAUC and fit a generalized additive model (GAM) to ΔAUC using a smoothing function with 3 knots (R package mgcv v.1.8-24 via the R package ggplot2 v.2_2.2.1, Wickham, 2016; Wood, 2017).

2.7 | Effort and expertise effects

We used general linear models (GLMs) and linear mixed models (LMMs) with normal error distributions to assess the effects of culling by effort and expertise on ΔAUC. For each species (*i*) by 1-4 (*j*) benchmark datasets and 1-16 (*k*) culled data subsets, the response variable, ΔAUC, was first rescaled and logit-transformed:

$$\text{logit}\,(\Delta\text{AUC})_{i,j,k} = \text{logit}\,(\text{AUC (benchmark model)}_{i,j} - \text{AUC (citizen science model)}_{i,j,k} + 0.50).$$

We rescaled by adding 0.5 to ΔAUC, as logit-transformations only take values between 0 and 1. The maximum absolute value of the difference prior to rescaling was less than 0.5.

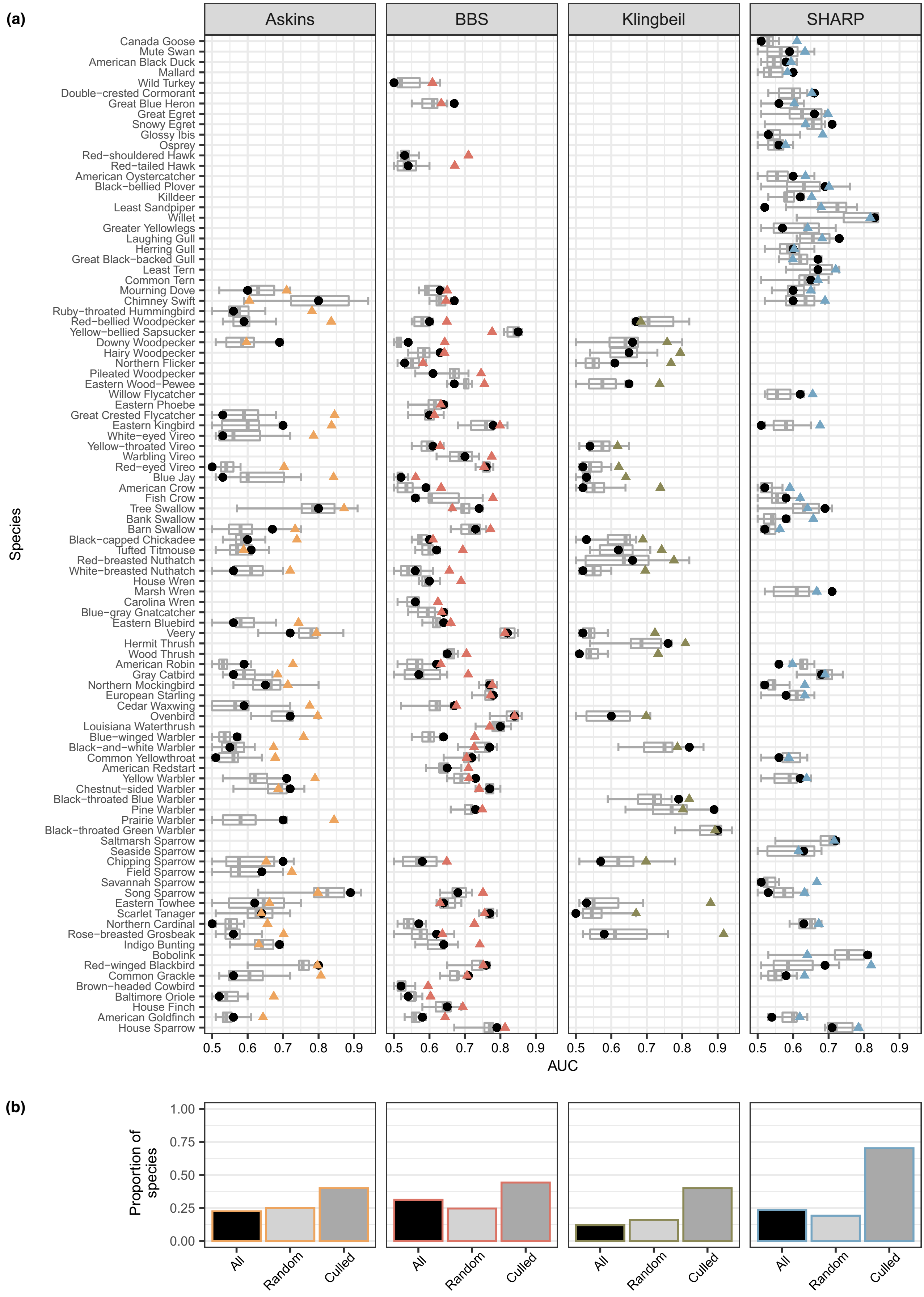
To summarize the ability of culling by effort or expertise to improve ΔAUC, we ran a set of GLMs for each species by comparison. We calculated the reduction in deviance explained by removing a given covariate from the global model (effort + expertise) and dividing that value by the deviance of the null (intercept-only) model.

We used LMMs with a random species effect in the R package lme4 v.1.1-17 (Bates et al., 2014) to assess the effects of different expertise and effort culling levels on ΔAUC. We specified non-ordinal four-level factors to model the effects of culling by the four distance (effort) levels and four expertise levels. We hypothesized that species' traits may influence the effects expertise and effort culling have on ΔAUC and included models with interactions between traits and culling covariates. We developed a set of a priori models (Table S5) and ran them for each of the four benchmark datasets. We used Akaike's information criterion (AIC) to compare models (Burnham & Anderson, 2002).

3 | RESULTS

Using the full dataset, models using eBird citizen science data performed well (AUC ≥0.7) for 15%–33% of species, depending on benchmark dataset (Table 1). Selecting the best model after data culling increased the proportion of species with good models to 28%–58%, with culled datasets generally producing better models. In contrast, random data reductions performed similarly to models using the full dataset.

FIGURE 2 (a) Area under the receiver operating characteristic curve (AUC) results quantify the ability of eBird citizen science species distribution models to predict species occurrence data for four benchmark datasets associated with different land cover types: Askins (shrubland), BBS (mostly forest and developed land), Klingbeil (interior forest) and SHARP (salt marsh). Triangles show benchmark AUC values and are based on cross-validated predictions wherein withheld data were used to evaluate models trained with the benchmark datasets. Black dots show AUC values from models that used all citizen science data that met our basic requirements. Box-whisker plots show the distribution of AUC values across 16 models that used culled subsets of the citizen science data and display the median, upper and lower quartiles, and extend to maximum and minimum values. (b) Barplots show the proportion of species for which eBird models matched or exceeded the benchmark AUC based on using all data, random subsets (not shown in 'a') or culled subsets. Random and culled subset results are based on the best model for each species


(a)
Askins
BBS
Klingbeil
SHARP
Canada Goose
Mute Swan
American Black Duck
Mallard
Wild Turkey
Double−crested Cormorant
Great Blue Heron
Great Egret
Snowy Egret
Glossy Ibis
Osprey
Red−shouldered Hawk
Red−tailed Hawk
American Oystercatcher
Black−bellied Plover
Killdeer
Least Sandpiper
Willet
Greater Yellowlegs
Laughing Gull
Herring Gull
Great Black−backed Gull
Least Tern
Common Tern
Mourning Dove
Chimney Swift
Ruby−throated Hummingbird
Red−bellied Woodpecker
Yellow−bellied Sapsucker
Downy Woodpecker
Hairy Woodpecker
Northern Flicker
Pileated Woodpecker
Eastern Wood−Pewee
Willow Flycatcher
Eastern Phoebe
Great Crested Flycatcher
Eastern Kingbird
White−eyed Vireo
Yellow−throated Vireo
Warbling Vireo
Red−eyed Vireo
Blue Jay
American Crow
Fish Crow
Tree Swallow
Bank Swallow
Barn Swallow
Black−capped Chickadee
Tufted Titmouse
Red−breasted Nuthatch
White−breasted Nuthatch
House Wren
Marsh Wren
Carolina Wren
Blue−gray Gnatcatcher
Eastern Bluebird
Veery
Hermit Thrush
Wood Thrush
American Robin
Gray Catbird
Northern Mockingbird
European Starling
Cedar Waxwing
Ovenbird
Louisiana Waterthrush
Blue−winged Warbler
Black−and−white Warbler
Common Yellowthroat
American Redstart
Yellow Warbler
Chestnut−sided Warbler
Black−throated Blue Warbler
Pine Warbler
Prairie Warbler
Black−throated Green Warbler
Saltmarsh Sparrow
Seaside Sparrow
Chipping Sparrow
Field Sparrow
Savannah Sparrow
Song Sparrow
Eastern Towhee
Scarlet Tanager
Northern Cardinal
Rose−breasted Grosbeak
Indigo Bunting
Bobolink
Red−winged Blackbird
Common Grackle
Brown−headed Cowbird
Baltimore Oriole
House Finch
American Goldfinch
House Sparrow
Species
0.5 0.6 0.7 0.8 0.9
0.5 0.6 0.7 0.8 0.9
0.5 0.6 0.7 0.8 0.9
0.5 0.6 0.7 0.8 0.9
AUC
(b)
Proportion of species
1.00
0.75
0.50
0.25
0.00
All
Random
Culled
All
Random
Culled
All
Random
Culled
All
Random
Culled

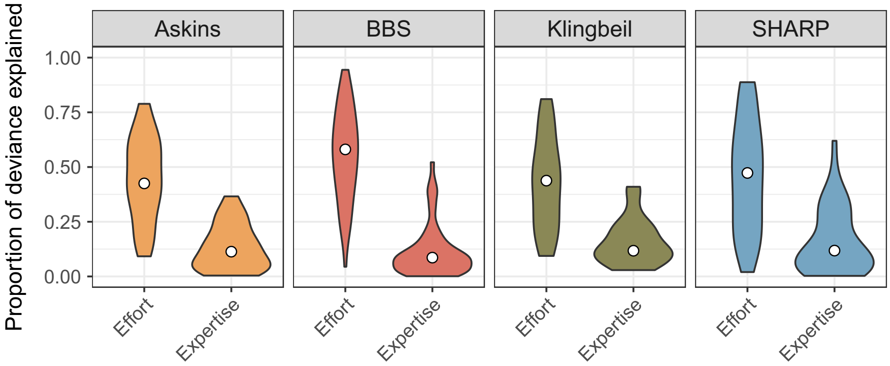


FIGURE 3 Proportion of the difference in model performance explained when culling eBird citizen science data by survey effort versus observer expertise. Performance was measured by predicting occurrence in four independent benchmark datasets ('Askins', *n* = 40 species; 'BBS', *n* = 61; 'Klingbeil', *n* = 25 species; 'SHARP', *n* = 47 species), estimating the area under the receiver operating characteristic curve (AUC) and calculating the difference (ΔAUC) compared to predictions from models derived from the benchmark data. Plots display mirrored kernel density estimates and extend from minimum to maximum values. White dots show median values

Models using the full eBird dataset were capable of matching or exceeding the benchmark performance produced directly from the benchmark datasets, but this happened for only 12%–31% of species, depending on the dataset (Figure 2, Table 1). Culled subsets reduced the number of records used by between 52% and 90% relative to the full dataset (Table S6). However, when the best culled subsets were used, the proportion meeting benchmark performance increased to 40%–70% of species. Models based on random subsets showed no improvement compared to the full dataset (16%–25% of species).

Culling eBird citizen science data by survey effort class explained more variability in performance difference than culling by expertise class across all four benchmark datasets (Figure 3). Consistent with this result, the LMMs with greatest model support all included effort as a variable but not expertise (Table 2). Although some form of culling by effort consistently improved models, there was no consistent pattern as to which survey lengths produced the best results. The best citizen science model most frequently was from 'short' distance surveys (29% of cases), but we also found many cases where the best models were from 'medium' (27%), 'lengthy' (22%) or 'very lengthy' (22%) surveys.

None of the four species traits contributed as main effects to model performance differences between the eBird citizen science models and the benchmark models (Table S5; Figure 4). Species prevalence (BBS and Klingbeil datasets) and ubiquity (SHARP dataset), however, both appeared in interaction with effort (distance class) in the top LMM models for the respective datasets (Table 2; Figure 5b-d). For the Askins dataset, no differences between distance classes were found (Figure 5a). For the BBS dataset, there

TABLE 2 Multi-model comparison showing the ability of species traits and culling (filtering) procedure to explain differences in the performance of culled eBird citizen science species distribution models and benchmark species distribution models

Benchmark dataset	Model	K	logL	AIC	ΔAIC	w
Askins	Effort	6	−112.931	237.99	0	0.54
	Effort * Ubiquity	10	−109.704	239.76	1.762	0.22
	Effort * Identifiability	10	−110.175	240.6991	2.704	0.14
BBS	Effort * Prevalence	10	640.6628	−1,261.1	0	1
	Effort * Ubiquity	10	633.4751	−1,246.72	14.375	0
Klingbeil	Effort * Prevalence	10	−71.0684	162.7025	0	0.39
	Effort	6	−75.3815	162.9768	0.2743	0.34
	Effort * Ubiquity	10	−72.5682	165.7019	2.9994	0.09
	Null	3	−80.6735	167.4076	4.7051	0.04
	Effort + Expertise	9	−74.7524	167.9663	5.2638	0.03
	Effort * Abundance	10	−73.8252	168.216	5.5135	0.02
	Effort * Identifiability	10	−73.8685	168.3025	5.6000	0.02
SHARP	Effort * Ubiquity	10	145.5602	−270.824	0	1
	Effort * Prevalence	10	139.1015	−257.906	12.917	0

Note: Candidate covariates included survey effort, observer expertise and species traits (abundance, identifiability, prevalence, and ubiquity). Columns show number of parameters (*K*), the log likelihood (logL), Akaike information criterion (AIC), the difference in AIC compared to the top model (ΔAIC) for a given dataset, and model weight (*w*). All models included a random intercept term for species. Models with ΔAIC <6 are shown unless one model had all the weight (*w* = 1) in which case, the next model is also included. Full model sets are shown in Table S5.

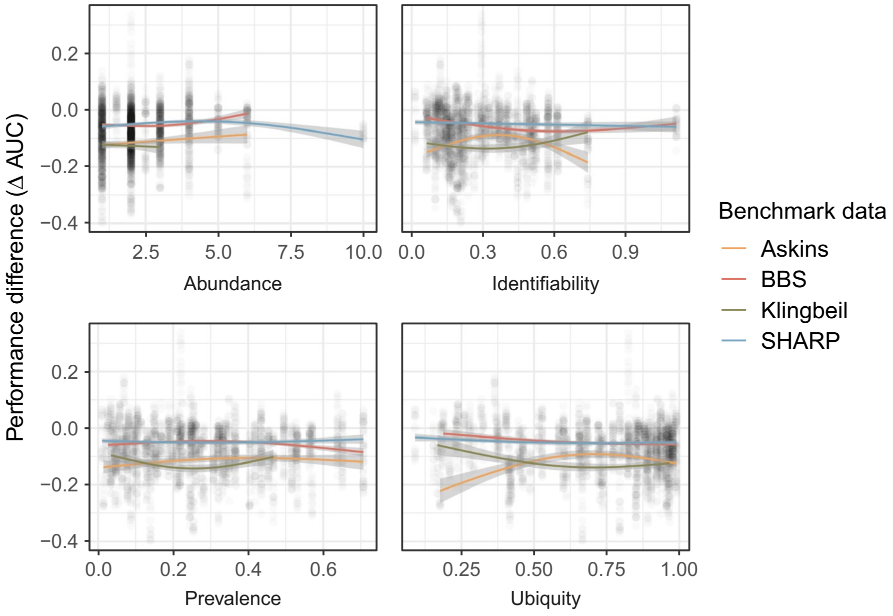


FIGURE 4 Effect of species traits on the performance of species distribution models built with culled eBird citizen science data relative to benchmark models. Performance difference was measured by predicting occurrence in four benchmark datasets ('Askins', *n* = 40 species; 'BBS', *n* = 61; 'Klingbeil', *n* = 25 species; 'SHARP', *n* = 47 species), estimating the area under the receiver operating characteristic curve (AUC) and calculating the difference (ΔAUC) compared to predictions from models derived from the benchmark data. Plots show data points, GAM fitted lines for traits as main effects and 95% confidence intervals for the four benchmark datasets

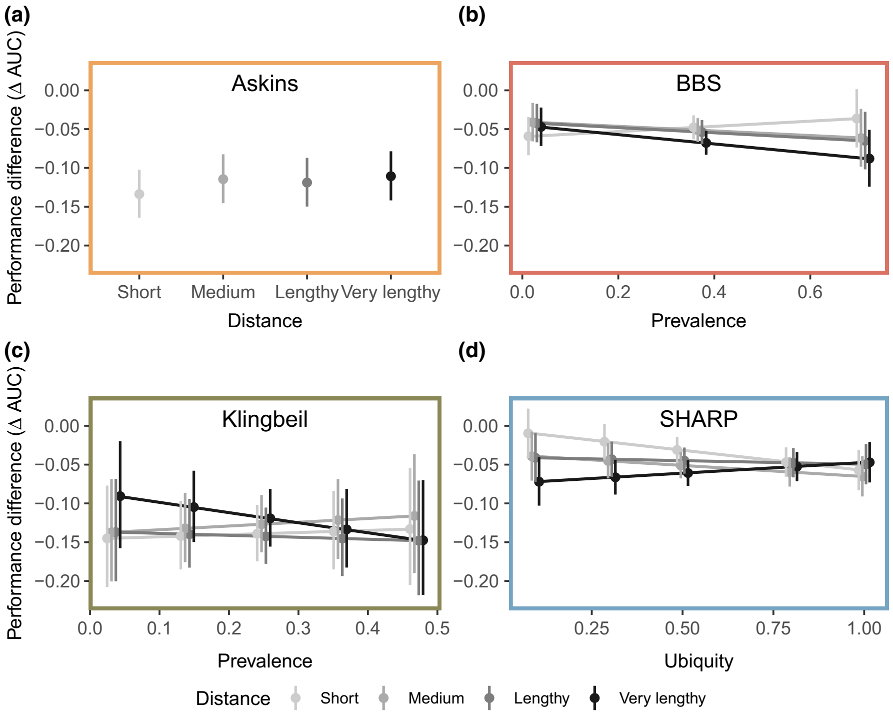


FIGURE 5 Effects of survey effort (distance travelled) on the performance of species distribution models built with culled eBird citizen science data relative to benchmark models. Performance difference was measured by predicting occurrence in four benchmark datasets ('Askins', *n* = 40 species; 'BBS', *n* = 61; 'Klingbeil', *n* = 25 species; 'SHARP', *n* = 47 species), estimating the area under the receiver operating characteristic curve (AUC) and calculating the difference (ΔAUC) compared to predictions from models derived from the benchmark data. Plots show results of best linear mixed models developed to explain ΔAUC for each comparison. For three datasets (b–d), the best model included an interaction between distance travelled and one of the species traits while in one dataset (a), the best model did not include an interaction with any species traits. Figures illustrate the line of best fit, by distance class and 95% confidence intervals (jittered horizontally in b–d)

was little difference among distance classes for species of low prevalence, but models using shorter distance checklists improved models more than those using very lengthy ones for the most prevalent species (Figure 5b). In contrast, the Klingbeil dataset indicated slightly improved performance for less prevalent species using very lengthy surveys but no difference for more prevalent species (Figure 5c). Finally, less ubiquitous species in the SHARP dataset were modelled better when using short surveys versus very lengthy ones, but this difference disappeared for ubiquitous species.

4 | DISCUSSION

Accurate maps of species distributions underlie the reliability of species conservation assessments and spatial conservation plans (Boitani et al., 2011; Kremen et al., 2008). This necessity is yet unrealized for the vast majority of species, while the need for such knowledge grows ever more critical to curtail biodiversity loss (Dirzo et al., 2014; Jetz, McPherson, & Guralnick, 2012). Citizen science datasets that provide a large volume of local observations over broad areas offer the possibility of using predictive models to develop distribution maps (Devictor et al., 2010). Such datasets, however, often lack rigorous data collection protocols and require little or no formal training of volunteers, suggesting a need to ensure that inherent biases and noise are addressed (Bird et al., 2014). We tested whether low-structure citizen science data could produce SDMs of similar quality to those derived from systematically collected data. We also tested methods for stringent data filtering to evaluate if careful data selection could improve citizen science-based SDMs.

We found that SDMs built with low-structure citizen science data can match the performance of those derived from systematic 'benchmark' surveys, but that this was only accomplished in a minority of cases. Our analysis is based on just eBird data, but this is one of the world's largest and fastest growing citizen science datasets focused on species occurrence data (Hochachka et al., 2012), and is representative of many such efforts. Although unable to match the benchmark data in a majority of cases, performance of the eBird models was still acceptable for many species and, when other data are unavailable, use of SDMs derived from citizen science data is likely to be worthwhile. Furthermore, any dataset would have an inherent disadvantage when predicting to an independent dataset versus internal cross-validation of the independent data. Therefore, our assessment of the performance of eBird models can be viewed as conservative.

While larger sample sizes have been shown to reduce bias and increase the precision and information content of citizen science datasets (Kamp et al., 2016; Munson et al., 2010), our results indicate substantial improvement by selective data reduction. Culling from a large volume of citizen science data allowed us to more than double the number of cases that met the benchmark performance (86 vs. 42 out of 173) and culled datasets almost always produced the best models (156 of 173). We primarily attribute these results to culling by survey effort (i.e. distance travelled) rather than observer expertise.

Because travel distance relates to the spatial scale of the underlying bird detections and the benchmark data were based on point surveys, we expected short distance subsets to provide the best scale match. However, the effects of distance varied by benchmark dataset and species traits indicating that spatial-scale matching for noisy citizen science data is complex. Habitat homogeneity, for example, can mitigate locational error in SDMs (Naimi, Skidmore, Groen, & Hamm, 2011). In more extensive and contiguous landcover types—for example forest and developed classes in our study area—the higher resolution obtained with shorter survey distances may not achieve the same benefits as it would for patchier habitats. For example, the improvements for shorter distance surveys in the SHARP data (Figure 5) might arise because shorter surveys are more likely to be precisely located in or close to saltmarsh habitat, which occurs only in small patches in our study region. Home range sizes—a species trait we could not test because we lacked estimates for many species—can also impact scale matching (Guisan & Thuiller, 2005).

While variation in observer expertise was related to large differences in reporting rate of individual species, this factor was not related to model improvement in culled data subsets (Figure 3; Table 2; Figure S4). This result might be surprising as mitigating false absences is central to site-level monitoring or estimating population sizes. If, however, false absences are not biased by land cover and there are sufficient true positives, occurrence predictions need not be compromised. Our expertise measure did not address false positive identifications which contribute to bird identification errors regardless of skill level (Farmer et al., 2012; Kelling, Johnston, et al., 2015). Thus, additional components of observer expertise still warrant investigation.

Our analysis used one performance metric—AUC, which measures the ability of a probabilistic model output to discriminate between presences and absences in evaluation data. We chose this measure because assessing presence–absence is central to mapping species distributions and AUC is a standard tool widely used for assessing SDM quality. Other metrics could further inform users about desired qualities of SDM performance, including metrics that assess agreement of mapped probabilities, prevalence-based metrics and calibration metrics (Fletcher & Fortin, 2018; Rödder & Engler, 2011). Many standard metrics require thresholding the SDM's probabilistic surface into a binary surface, which can be done in many ways and in a species-specific approach. To avoid over-complicating results and interpretation, we chose AUC as a single, standard metric for performance, but acknowledge that different metrics may have led to different conclusions.

Low-structure citizen science datasets have the potential to fill an important role in conservation planning by providing local observations at broad scales enabling improved knowledge of species' distributions. Although these data can produce high-quality predictions to systematically collected occurrence data, our analyses suggest that low-structure citizen science are not sufficient to replace systematic data collections in many cases. When logistical constraints make use of more rigorous methods impossible, or when formal cost-benefit analysis suggests that the benefits of increased rigour are outweighed by the lower costs of citizen science data, data culling provides a potential mechanism for improving SDM predictions from citizen science sources. Although our results show that culling can substantially improve predictions, the lack of consistent patterns across datasets or species in how best to cull data suggests that additional work is needed to explore different options and to understand how species traits and data collection methods interact to affect model performance.

ACKNOWLEDGEMENTS

Funding for this project came from Connecticut Department of Energy and Environmental Protection, and we particularly thank Min Huang. We thank eBird, Robert Askins, Brian Klingbeil, the SHARP group and the North American BBS for data; Dave Ziolkowski and Wesley Hochachka for conversations about BBS and eBird data, respectively; Alison Johnston for code and advice on estimating observer expertise; and the reviewers of this paper for comments that led to its improvement.

DATA AVAILABILITY STATEMENT

R code to run models and associated data are available at https://github.com/vavadavat/SDMfromCitSci

ORCID

Valerie A. Steen https://orcid.org/0000-0002-1417-8139
Morgan W. Tingley https://orcid.org/0000-0002-1477-2218

REFERENCES

Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A., Vilela, B., & Anderson, R. P. (2015). spThin: An R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography*, *38*(5), 541–545. https://doi.org/10.1111/ecog.01132

Askins, R. A., Folsom-O'Keefe, C. M., & Hardy, M. C. (2012). Effects of vegetation, corridor width and regional land use on early successional birds on powerline corridors. *PLoS ONE*, *7*(2), 1–10. https://doi.org/10.1371/journal.pone.0031520

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., … Grothendieck, G. (2014). *Package 'lme4'* (Vol. *12*). Vienna, Austria: R Foundation for Statistical Computing.

Bird, T. J., Bates, A. E., Lefcheck, J. S., Hill, N. A., Thomson, R. J., Edgar, G. J., … Frusher, S. (2014). Statistical solutions for error and bias in global citizen science datasets. *Biological Conservation*, *173*, 144–154. https://doi.org/10.1016/j.biocon.2013.07.037

Boitani, L., Maiorano, L., Baisero, D., Falcucci, A., Visconti, P., & Rondinini, C. (2011). What spatial data do we need to develop global mammal conservation strategies? *Philosophical Transactions of the Royal Society of London*, *366*(1578), 2623–2632. https://doi.org/10.1098/rstb.2011.0117

Bonter, D. N., & Cooper, C. B. (2012). Data validation in citizen science: A case study from project FeederWatch. *Frontiers in Ecology and the Environment*, *10*(6), 305–307. https://doi.org/10.1890/110273

Boria, R. A., Olson, L. E., Goodman, S. M., & Anderson, R. P. (2014). Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecological Modelling*, *275*, 73–77. https://doi.org/10.1016/j.ecolmodel.2013.12.012

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York: Springer Science & Business Media.

Butt, N., Slade, E., Thompson, J., Malhi, Y., & Riutta, T. (2013). Quantifying the sampling error in tree census measurements by volunteers and its effect on carbon stock estimates. *Ecological Applications*, *23*, 936–943. https://doi.org/10.1890/11-2059.1

Crall, A. W., Newman, G. J., Stohlgren, T. J., Holfelder, K. A., Graham, J., & Waller, D. M. (2011). Assessing citizen science data quality: An invasive species case study. *Conservation Letters*, *4*, 433–442. https://doi.org/10.1111/j.1755-263X.2011.00196.x

Devictor, V., Whittaker, R. J., & Beltrame, C. (2010). Beyond scarcity: Citizen science programmes as useful tools for conservation biogeography. *Diversity and Distributions*, *16*(3), 354–362. https://doi.org/10.1111/j.1472-4642.2009.00615.x

Dickinson, J. L., Zuckerberg, B., & Bonter, D. N. (2010). Citizen science as an ecological research tool: Challenges and benefits. *Annual Review of Ecology, Evolution, and Systematics*, *41*, 149–172. https://doi.org/10.1146/annurev-ecolsys-102209-144636

Dirzo, R., Young, H. S., Galetti, M., Ceballos, G., Isaac, N. J., & Collen, B. (2014). Defaunation in the anthropocene. *Science*, *345*(6195), 401–406. https://doi.org/10.1126/science.1251817

Farmer, R. G., Leonard, M. L., & Horn, A. G. (2012). Observer effects and avian-call-count survey quality: Rare-species biases and overconfidence. *The Auk*, *129*(1), 76–86.

Fielding, A. H., & Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, *24*(1), 38–49. https://doi.org/10.1017/S0376892997000088

Fink, D., Hochachka, W. M., Zuckerberg, B., Winkler, D. W., Shaby, B., Munson, M. A., … Kelling, S. (2010). Spatiotemporal exploratory models for broad-scale survey data. *Ecological Applications*, *20*(8), 2131–2147. https://doi.org/10.1890/09-1340.1

Fitzpatrick, M. C., Preisser, E. L., Ellison, A. M., & Elkinton, J. S. (2009). Observer bias and the detection of low-density populations. *Ecological Applications*, *19*(7), 1673–1679. https://doi.org/10.1890/09-0265.1

Fletcher, R., & Fortin, M. (2018). *Spatial ecology and conservation modeling*. New York: Springer.

Franklin, J. (2010). *Mapping species distributions: Spatial inference and prediction*. Cambridge: Cambridge University Press.

Guisan, A., & Thuiller, W. (2005). Predicting species distribution: Offering more than simple habitat models. *Ecology Letters*, *8*(9), 993–1009. https://doi.org/10.1111/j.1461-0248.2005.00792.x

Guisan, A., Thuiller, W., & Zimmermann, N. E. (2017). *Habitat suitability and distribution models: With applications in R*. Cambridge: Cambridge University Press.

Higa, M., Yamaura, Y., Koizumi, I., Yabuhara, Y., Senzaki, M., & Ono, S. (2015). Mapping large-scale bird distributions using occupancy models and citizen data with spatially biased sampling effort. *Diversity and Distributions*, *21*(1), 46–54. https://doi.org/10.1111/ddi.12255

Hochachka, W. M., Fink, D., Hutchinson, R. A., Sheldon, D., Wong, W., & Kelling, S. (2012). Data-intensive science applied to broad-scale citizen science. *Trends in Ecology and Evolution*, *27*(2), 130–137. https://doi.org/10.1016/j.tree.2011.11.006

Homer, C., Dewitz, J., Yang, L., Jin, S., Danielson, P., Xian, G., … Megown, K. (2015). Completion of the 2011 national land cover database for the conterminous united states–representing a decade of land cover change information. *Photogrammetric Engineering and Remote Sensing*, *81*(5), 345–354.

iNaturalist.org. (2019). *iNaturalist research-grade observations*. occurrence dataset https://doi.org/10.15468/ab3s5x accessed via GBIF.org on 2019-07-09.

Isaac, N. J., Strien, A. J., August, T. A., Zeeuw, M. P., & Roy, D. B. (2014). Statistics for citizen science: Extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution*, *5*(10), 1052–1060. https://doi.org/10.1111/2041-210X.12254

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

Johnston, A., Fink, D., Hochachka, W. M., & Kelling, S. (2018). Estimates of observer expertise improve species distributions from citizen science data. *Methods in Ecology and Evolution*, *9*(1), 88–97. https://doi.org/10.1111/2041-210X.12838

Kamp, J., Oppel, S., Heldbjerg, H., Nyegaard, T., & Donald, P. F. (2016). Unstructured citizen science data fail to detect long-term population declines of common birds in Denmark. *Diversity and Distributions*, *22*(10), 1024–1035. https://doi.org/10.1111/ddi.12463

Kelling, S., Fink, D., La Sorte, F. A., Johnston, A., Bruns, N. E., & Hochachka, W. M. (2015). Taking a 'Big data' approach to data quality in a citizen science project. *Ambio*, *44*(4), 601–611. https://doi.org/10.1007/s13280-015-0710-4

Kelling, S., Johnston, A., Hochachka, W. M., Iliff, M., Fink, D., Gerbracht, J., … Yu, J. (2015). Can observation skills of citizen scientists be estimated using species accumulation curves? *PLoS ONE*, *10*(10), e0139600. https://doi.org/10.1371/journal.pone.0139600

Kéry, M., Gardner, B., & Monnerat, C. (2010). Predicting species distributions from checklist data using site-occupancy models. *Journal of Biogeography*, *37*(10), 1851–1862. https://doi.org/10.1111/j.1365-2699.2010.02345.x

Klingbeil, B. T., & Willig, M. R. (2015). Bird biodiversity assessments in temperate forest: The value of point count versus acoustic monitoring protocols. *PeerJ*, *3*, e973. https://doi.org/10.7717/peerj.973

Kremen, C., Cameron, A., Moilanen, A., Phillips, S. J., Thomas, C. D., Beentje, H., … Zjhra, M. L. (2008). Aligning conservation priorities across taxa in madagascar with high-resolution planning tools. *Science*, *320*(5873), 222–226. https://doi.org/10.1126/science.1155193

La Sorte, F. A., Lepczyk, C. A., Burnett, J. L., Hurlbert, A. H., Tingley, M. W., & Zuckerberg, B. (2018). Opportunities and challenges for big data ornithology. *The Condor*, *120*(2), 414–426. https://doi.org/10.1650/CONDOR-17-206.1

Liaw, A., & Wiener, M. (2002). The Randomforest Package. *R News*, *2*(3), 18–22.

Lowman, M., D'Avanzo, C., & Brewer, C. (2009). A national ecological network for research and education. *Science*, *323*(5918), 1172–1173.

Munson, M. A., Caruana, R., Fink, D., Hochachka, W. M., Iliff, M., Rosenberg, K. V., … Kelling, S. (2010). A method for measuring the relative information content of data from different monitoring protocols. *Methods in Ecology and Evolution*, *1*(3), 263–273. https://doi.org/10.1111/j.2041-210X.2010.00035.x

Naimi, B., Skidmore, A. K., Groen, T. A., & Hamm, N. A. (2011). Spatial autocorrelation in predictors reduces the impact of positional uncertainty in occurrence data on species distribution modelling. *Journal of Biogeography*, *38*(8), 1497–1509. https://doi.org/10.1111/j.1365-2699.2011.02523.x

Pardieck, K., Ziolkowski, D. Jr, Lutmerding, M., Campbell, K., & Hudson, M. (2016). *North american breeding bird survey dataset 1966-2016, version 2016.0*. US Geological Survey, Patuxent Wildlife Research Center. Retrieved from www.Pwrc.Usgs.Gov/BBS/RawData. https://doi.org/10.5066/F7W0944J

Peterson, A. T., Soberón, J., Pearson, R. G., Anderson, R. P., Martínez-Meyer, E., Nakamura, M., … Araújo, M. B. (2011). *Ecological niches and geographic distributions (MPB-49)*. Princeton: Princeton University Press.

Pimm, S. L., Jenkins, C. N., Abell, R., Brooks, T. M., Gittleman, J. L., Joppa, L. N., … Sexton, J. O. (2014). The biodiversity of species and their rates of extinction, distribution, and protection. *Science*, *344*(6187), 987–998. https://doi.org/10.1126/science.1246752

Ratnieks, F. L., Schrell, F., Sheppard, R. C., Brown, E., Bristow, O. E., & Garbuzov, M. (2016). Data reliability in citizen science: Learning curve and the effects of training method, volunteer background and experience on identification accuracy of insects visiting ivy flowers. *Methods in Ecology and Evolution*, *7*, 1226–1235. https://doi.org/10.1111/2041-210X.12581

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., … Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913–929. https://doi.org/10.1111/ecog.02881

Robbins, C. S., Bystrak, D., & Geissler, P. H. (1986). *The Breeding Bird Survey: Its first fifteen years, 1965–1979* (p. 196). Washington, D.C., U.S: Fish and Wildlife Service.

Robinson, O. J., Ruiz-Gutierrez, V., & Fink, D. (2018). Correcting for bias in distribution modelling for rare species using citizen science data. *Diversity and Distributions*, *24*(4), 460–472. https://doi.org/10.1111/ddi.12698

Rödder, D., & Engler, J. (2011). Quantitative metrics of overlaps in grinnellian niches: Advances and possible drawbacks. *Global Ecology and Biogeography*, *20*(6), 915–927. https://doi.org/10.1111/j.1466-8238.2011.00659.x

Rondinini, C., Wilson, K. A., Boitani, L., Grantham, H., & Possingham, H. P. (2006). Tradeoffs of different types of species occurrence data for use in systematic conservation planning. *Ecology Letters*, *9*(10), 1136–1145. https://doi.org/10.1111/j.1461-0248.2006.00970.x

Shea, C. P., Peterson, J. T., Wisniewski, J. M., & Johnson, N. A. (2011). Misidentification of freshwater mussel species (Bivalvia: Unionidae): Contributing factors, management implications, and potential solutions. *Journal of the North American Benthological Society*, *30*, 446–458. https://doi.org/10.1899/10-073.1

Steger, C., Butt, B., & Hooten, M. B. (2017). Safari science: Assessing the reliability of citizen science data for wildlife surveys. *Journal of Applied Ecology*, *54*(6), 2053–2062. https://doi.org/10.1111/1365-2664.12921

Sullivan, B. L., Aycrigg, J. L., Barry, J. H., Bonney, R. E., Bruns, N., Cooper, C. B., … Kelling, S. (2014). The eBird enterprise: An integrated approach to development and application of citizen science. *Biological Conservation*, *169*, 31–40. https://doi.org/10.1016/j.biocon.2013.11.003

Swanson, A., Kosmala, M., Lintott, C., & Packer, C. (2016). A generalized approach for producing, quantifying, and validating citizen science data from wildlife images. *Conservation Biology*, *30*, 520–531. https://doi.org/10.1111/cobi.12695

Szabo, J. K., Vesk, P. A., Baxter, P. W., & Possingham, H. P. (2010). Regional avian species declines estimated from volunteer-collected long-term data using list length analysis. *Ecological Applications*, *20*(8), 2157–2169. https://doi.org/10.1890/09-0877.1

Tulloch, A. I., Mustin, K., Possingham, H. P., Szabo, J. K., & Wilson, K. A. (2013). To boldly go where no volunteer has gone before: Predicting volunteer activity to prioritize surveys at the landscape scale. *Diversity and Distributions*, *19*(4), 465–480. https://doi.org/10.1111/j.1472-4642.2012.00947.x

Tulloch, A. I., & Szabo, J. K. (2012). A behavioural ecology approach to understand volunteer surveying for citizen science datasets. *Emu - Austral Ornithology*, *112*(4), 313–325. https://doi.org/10.1071/MU12009

Tye, C. A., McCleery, R. A., Fletcher, R. J., Greene, D. U., & Butryn, R. S. (2017). Evaluating citizen vs. professional data for modelling distributions of a rare squirrel. *Journal of Applied Ecology*, *54*(2), 628–637.

van Strien, A. J., van Swaay, C. A., & Termaat, T. (2013). Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. *Journal of Applied Ecology*, *50*(6), 1450–1458. https://doi.org/10.1111/1365-2664.12158

Ward, D. F. (2014). Understanding sampling and taxonomic biases recorded by citizen scientists. *Journal of Insect Conservation*, *18*, 753–756. https://doi.org/10.1007/s10841-014-9676-y

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. New York: Springer.

Wiest, W. A., Correll, M. D., Olsen, B. J., Elphick, C. S., Hodgman, T. P., Curson, D. R., & Shriver, W. G. (2016). Population estimates for tidal marsh birds of high conservation concern in the northeastern USA from a design-based survey. *The Condor*, *118*(2), 274–288. https://doi.org/10.1650/CONDOR-15-30.1

Wood, S. N. (2017). *Generalized additive models: An introduction with R*. Boca Raton: CRC Press.

BIOSKETCHES

Valerie Steen is a Postdoctoral Researcher whose work focuses on improving understanding of species' populations and distributions. **Chris Elphick** is a Professor whose work focuses on conservation biology, particularly the effects of global change on birds. **Morgan Tingley** is an Assistant Professor who studies how anthropogenic drivers of change affect geographic distributions and community interactions over time.

Author contributions: All authors conceived the ideas and designed the methods. VS analysed the data and led writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Steen VA, Elphick CS, Tingley MW. An evaluation of stringent filtering to improve species distribution models from citizen science data. *Divers Distrib*. 2019;25:1857–1869. https://doi.org/10.1111/ddi.12985