Received: 26 October 2020 | Revised: 4 March 2021 | Accepted: 5 March 2021

DOI: 10.1111/ddi.13271

METHOD

Diversity and Distributions WILEY

Analytical guidelines to increase the value of community science data: An example using eBird data to estimate species distributions

Alison Johnston | Wesley M. Hochachka | Matthew E. Strimas-Mackey | Viviana Ruiz Gutierrez | Orin J. Robinson | Eliot T. Miller | Tom Auer | Steve T. Kelling | Daniel Fink

Cornell Lab of Ornithology, Cornell University, Ithaca, NY, USA

Correspondence
Alison Johnston, Cornell Lab of Ornithology, Cornell University, 159 Sapsucker Woods Road, Ithaca, NY 14850, USA.
Email: aj327@cornell.edu

Funding information
Wolf Creek Foundation; National Science Foundation, Grant/Award Number: CCF-1522054, CNS-1059284, DBI-1939187 and ICER-1927646; National Aeronautics and Space Administration, Grant/Award Number: NNH12ZDA001N-ECOF; Leon Levy Foundation

Editor: Yoan Fourcade

Abstract

Aim: Ecological data collected by the general public are valuable for addressing a wide range of ecological research and conservation planning, and there has been a rapid increase in the scope and volume of data available. However, data from eBird or other large-scale projects with volunteer observers typically present several challenges that can impede robust ecological inferences. These challenges include spatial bias, variation in effort and species reporting bias.

Innovation: We use the example of estimating species distributions with data from eBird, a community science or citizen science (CS) project. We estimate two widely used metrics of species distributions: encounter rate and occupancy probability. For each metric, we critically assess the impact of data processing steps that either degrade or refine the data used in the analyses. CS data density varies widely across the globe, so we also test whether differences in model performance are robust to sample size.

Main conclusions: Model performance improved when data processing and analytical methods addressed the challenges arising from CS data; however, the degree of improvement varied with species and data density. The largest gains we observed in model performance were achieved with 1) the use of complete checklists (where observers report all the species they detect and identify, allowing non-detections to be inferred) and 2) the use of covariates describing variation in effort and detectability for each checklist. Occupancy models were more robust to a lack of complete checklists. Improvements in model performance with data refinement were more evident with larger sample sizes. In general, we found that the value of each refinement varied by situation and we encourage researchers to assess the benefits in other scenarios. These approaches will enable researchers to more effectively harness the vast ecological knowledge that exists within CS data for conservation and basic research.

KEYWORDS
citizen science, community science, detectability, eBird, encounter rate, occupancy model, species distribution model

1 | INTRODUCTION

Community science or citizen science (CS) data are increasingly making important contributions to applied ecological research and conservation planning. One of the most common forms of CS data is the recording of species observations by members of the public. These observations are being collected for a diverse array of taxa, including butterflies (Howard et al., 2010), sharks (Vianna et al., 2014), lichen (Casanovas et al., 2014), bats (Newson et al., 2015) and birds (Sauer et al., 2017). The number of these CS projects has been growing exponentially, but they vary widely in complexity, data collection flexibility and participation (Pocock et al., 2017; Wiggins & Crowston, 2011). Projects occur on a spectrum from those with a predefined sampling structure that resembles more traditional survey designs, to those that are unstructured and collect observations opportunistically. Projects with study designs and defined protocols generally produce data that are more informative for a particular objective, but are often limited to a specific time frame and region and have fewer participants. This can lead to a trade-off between the quality and quantity of data collected by CS projects (Bird et al., 2014; Pacifici et al., 2017). Semi-structured CS projects have unstructured data collection, but critically also collect data on the observation process, which can be used to retrospectively account for many sources of noise introduced by data collection (Altwegg & Nichols, 2019; Kelling et al., 2019). With the increasing popularity in the use and application of CS data, we describe and evaluate steps for data processing and analysis that maximize the value of semi-structured CS data (Sullivan et al., 2014).

Data consisting of species observations from volunteers present three general challenges that are not as prevalent in conventional scientific data. Firstly, the locations selected by participants to collect data are usually strongly spatially biased. For example, participants may preferentially visit locations that are close to where they live (Dennis & Thomas, 2000; Mair & Ruete, 2016), are more accessible (Botts et al., 2011; Kadmon et al., 2004), contain high species diversity (Hijmans et al., 2000; Tulloch et al., 2013) or are within protected areas (Tulloch et al., 2013). Secondly, the observation process is heterogeneous, with large variation in effort, time of day, observers and weather, all of which can affect the detectability of species (Ellis & Taylor, 2018; Hochachka et al., 2021; Oliveira et al., 2018). Thirdly, participants often have preferences for certain species, which may lead to preferential recording of some species over others (Troudet et al., 2017; Tulloch & Szabo, 2012). Nonetheless, CS data can fill critical gaps in our knowledge of the biodiversity of many parts of the world, and the growing scale and scope of CS data will likely increase our understanding of global biodiversity into the future. Therefore, it is imperative to define approaches that can maximize the value of the increasing volumes of CS species observations.

Imperfect detection in the observation process means that not every individual is detected by an observer, and consequently, some species are falsely absent from the data. The three CS data challenges listed above each result in false absences in the species recorded. The spatial bias has the strongest impact, since an absence of observers in an area results in no species being recorded. The other two challenges affect whether a species is recorded, conditional on an observer visiting a location where a species is present. Some facets of observer effort affect whether a species is *available* for detection—e.g. whether it is a time of day when the species is present in that place and behaves in a way that makes it detectable (Diefenbach et al., 2007; Hochachka et al., 2009). Other facets of effort affect whether an observer detects and identifies an available species, for example the duration and distance travelled while observing (Fuller & Langslow, 1984), or the skills and equipment associated with a particular observer (Kelling et al., 2015).

False negatives due to imperfect detection are ubiquitous in ecological data and require careful data analysis for robust inference. There are two main approaches for addressing the challenges of false negatives: 1) imposing a more structured protocol onto the dataset after collection via data filtering (Kamp et al., 2016) and 2) using an analytical framework that accommodate the false negatives, such as including covariates in a model to account for the variation in the causes of false negatives (Miller et al., 2019). In this paper, we advocate combining both of these approaches to increase the reliability of inferences made using CS observations.

We describe analytical approaches for using semi-structured CS data, using the example of estimating species distributions from data collected by the eBird CS project (Sullivan et al., 2014). We evaluate the efficacy of using two critical aspects of these CS data that facilitate robust ecological inference. Firstly, data submitted to eBird are structured as “checklists,” where each checklist is a list of the numbers of individuals of each bird species recorded during a period of bird-watching. The majority of these checklists record every individual bird the observer detected and identified, so we can infer when a species was not detected. Secondly, eBird is a semi-structured CS project, which means most eBird checklists have associated metadata describing the “effort” or observation process (Kelling et al., 2019), which allow us to model variation in the probability of detection. While our examples focus on the use of eBird data for estimating species occurrence, our results are applicable to similar CS datasets tackling similar ecological questions, and these results can also help inform the design of future CS surveys.

2 | METHODS

We explored the impact of various analytical practices when using CS data to estimate species distributions. We used different modelling approaches to estimate 1) encounter rate with Maxent and random forest models and 2) occupancy rate with an occupancy model. Species encounters arise as a compound process requiring both the species to occur at a site and to be detected at that site. Encounter rate is defined as the average rate at which observers encounter the species, so it reflects the product of occurrence and detectability. It

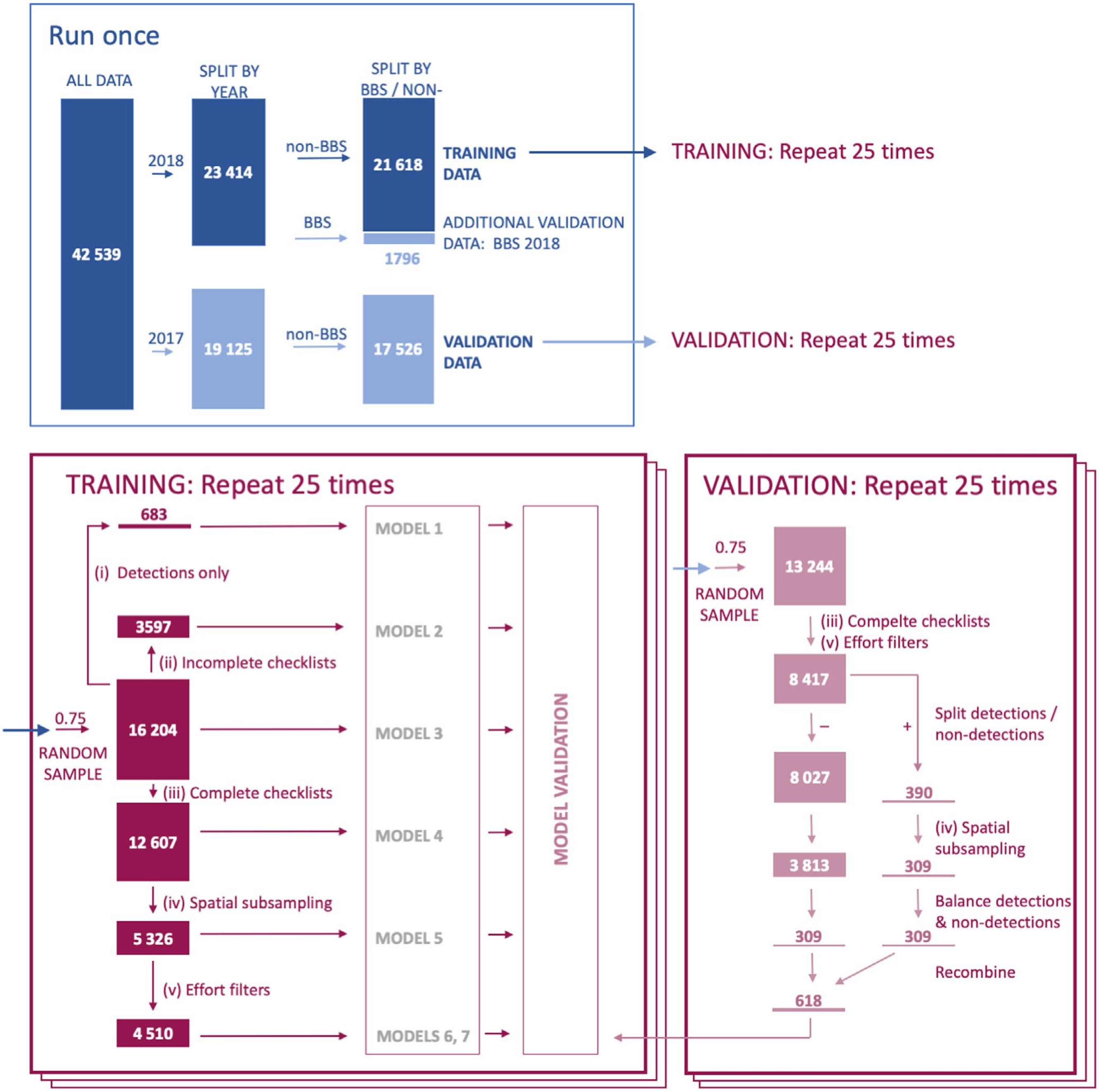


FIGURE 1 Schematic diagram of the flow of data into each of the 7 model types for the encounter rate model. The sizes of the boxes and the numbers inside them are the number of checklists. The blue processes occur once, and the pink processes occur 25 times, once for each model run. The numbers shown will therefore vary slightly each time within the pink box. The dark colours represent training data and the pale colours validation data. Arrows represent data processing steps or projection of the same data forward to the next stage

can also be considered to describe the "apparent distribution" of the species: the distribution of where observers encounter, detect, identify and record the species. Occupancy is defined as the probability that a species is present in a given location, with the model structure separating occurrence and detectability. For the random forest and occupancy models, we use detection/non-detection data as the response variable, while Maxent uses only detection (or "presence-only") data and combines these with pseudo-absences. All analyses were conducted with R (R Core Team, 2018).

2.1 | eBird data selection

We used data from the eBird Basic Dataset (EBD), which is global in extent and updated monthly (www.ebird.org/science/download-ebird-data-products). The most current version of the EBD can be freely accessed via an online data portal and processed with the *auk* R package (Strimas-Mackey et al., 2017). eBird has a robust review process, focussed on ensuring correct locations and species identification, that is conducted before data enter the

EBD. This review process removes unlikely false positives from the data, that is species records without adequate evidence of the identification, for locations and times of year that they are not expected to occur. This process does not remove false positives that are plausible observations based on species distributions and phenology. We provide further details on this review process and other aspects of eBird data in Appendix S1. Our data are from the EBD version released in May 2019. To model distribution in the breeding season, we used checklists from 15 May to 30 June. We used a geographically restricted subset of data, from Bird Conservation Region 27 "Southeastern coastal plain" (BCR27), a biogeographically distinct region that covers and includes parts of the states: Mississippi, Alabama, Florida, Georgia, North Carolina and South Carolina (NABCI, 2000).

For our primary case study, we focussed on wood thrush *Hylocichla mustelina* in the breeding season. Wood thrush is a relatively common nesting passerine across much of eastern North America, that is easily detected by its distinctive song. We also present some supplementary results from modelling the distribution of chuck-will's-widow (*Antrostomus carolinensis*), in order to illustrate how decisions about data analysis may have different impacts in different species. Chuck-will's-widows are camouflaged and nocturnally active, when their loud and distinctive vocalizations make them highly detectable. Their different daily activity patterns and habitats provide a good contrast between these two example species.

2.2 | eBird data processing—training data

We split the eBird data into a dataset to train (or fit) the models and semi-independent datasets to validate (or test) the models (Figure 1 and Figure S1). To train the models, we used eBird data from 2018. We used a hierarchy of data processing steps on the training data, applying these sequentially to create a set of differently processed datasets. These data processing steps were designed to highlight or address the challenges with CS data outlined in the introduction. We applied these datasets to each of the two model types to estimate both species encounter rate and occupancy (Table 1).

Two data processing steps were designed to demonstrate the differences in model estimates when the dataset does not contain key information. These two steps both degraded the eBird data to produce datasets that mimic common CS data structures. The first data degrading step, i) select only detections (Table 1), produced a dataset of "presence-only" information. This structure of data is common with CS projects that do not collect lists of species. The data degrading step, ii) select only "incomplete" checklists (Table 1), produced a dataset of checklists for which observers explicitly indicated that not all species were recorded. In this subset, non-detections cannot be separated from species bias when observers decide not to record a particular species that they have detected and identified. The models with these data (models 1 and 2) highlight the impact of using similar data to estimate species occupancy or encounter rates.

Three data processing steps were designed to demonstrate the impact of refining the eBird data and show the relative value of smaller, but more selective datasets, compared to larger and less refined datasets. These refinements were additively imposed on the raw data, so each cumulatively refined the data further. The data refinement steps were iii) select only "complete" checklists, to provide data with non-detections; iv) spatially subsample the data, to reduce the influence of spatial bias; and v) select checklists within standard range of effort, to reduce the influence of checklists with unusual effort (Table 1). Using non-detections allows the model to

TABLE 1 Descriptions of the elements in models 1–7 that include different data processing treatments. Model 3 uses all the raw data with no processing. Models 1–2 use data degraded in different ways by processes (i) and (ii). Models 4–6 use data refined in different ways by processes (iii), (iv) and (v). Model 7 uses the same data as model 6, but additionally includes effort variables as covariates

			Model						
Data processing treatment			**1**	**2**	**3**	**4**	**5**	**6**	**7**
Degrade	i)	Select detections only ("presence-only")	✓						
	ii)	Select incomplete checklists only		✓					
Refine	iii)	Select complete checklists only				✓	✓	✓	✓
	iv)	Spatial subsampling					✓	✓	✓
	v)	Effort filters						✓	✓
	vi)	Effort covariates							✓
Model structures									
Encounter rate model		Model type	Maxent	Random forest					
		No. of land cover covariates	16	16					
		No. of effort covariates	0	0					5
Occupancy model		Occupancy model		Single-season occupancy model					
		No. of land cover covariates		4					
		No. of effort covariates		0					5

have knowledge of where effort was expended, but the species was not recorded. Data processing step iii) ensures that all the inferred "non-detections" are actually non-detections, by only including complete checklists where observers report all the species they could detect and identify. This addresses the challenge of species reporting bias. Step iv) spatial subsampling reduces the over-influence of well-surveyed locations in the analysis. This addresses the challenge of spatial bias. Step v) reduces the range of checklist effort, creating a more consistent and standardized set of checklists for analysis. This addresses the challenge of variable effort. Methodological details for how we performed each of these data processing steps are given in Appendix S2.

2.3 | eBird data processing—validation data

We created two validation datasets, chosen because of their different forms of independence from the 2018 training data. In general, the form of validation data should be tailored to a specific intent (Valavi et al., 2018).

Our main validation set was temporally independent, using eBird data from 2017. We split data with species detections from data with species non-detections and spatially subsampled each set of data to reduce the influence of spatial bias. We then randomly subsampled the non-detections so there were equal numbers of detections and non-detections. Recombining these two datasets gave us a balanced validation set (with equal detections and non-detections) and with reduced spatial bias (Figure 1). The reduced spatial bias ensured that the data represent the study region more evenly, and the balance of detections and non-detections was designed to test the ability of the model to discriminate between areas of species presence and absence. As 2017 is a different year, it would not provide good validation for species that change their distribution substantially from year-to-year, such as irruptive species, but we have no reason to expect such inter-annual variability for our example species.

Our second validation dataset was designed to compare estimates from eBird data with estimates from data collected with a standardized and pre-designed survey. We used data from the 2018 North American Breeding Bird Survey (BBS) that were also submitted to eBird (Figure 1). We used the BBS data submitted to eBird to enable us to use data with precise location information for each stop on the 25-mile BBS routes. We extracted BBS data from eBird by identifying sets of at least 40 × 3-min point counts conducted on the same day, by the same observer, at locations that were spatially and temporally separated according to expectations for BBS stops. We also removed these same data from the training data. See Appendix S2 for more details of both validation datasets.

Using the model fitted with the training data, we estimated counts on checklists in both the validation datasets. We compared the estimated occurrence rates to the actual occurrence, enabling us to understand the quality of the models to predict to different datasets. See Appendix S2 for more details of the validation procedures.

2.4 | eBird data processing—occupancy models

Preparing data for the occupancy models required some additional data processing. There are many decisions required when using CS data for occupancy models and we describe these in greater detail in Appendix S1 with only a brief overview here. We defined a "site" as a location (defined by latitude and longitude) with at least two visits during 15 May-30 June 2018. Where there were more than 10 visits to a single site, we randomly selected 10 of the visits.

For the occupancy models, we created a third validation set. We wanted to validate the estimates of occupancy, while limiting the effects of detectability. We used the models to estimate occupancy and detectability at all sites. We calculated the cumulative estimated detectability across all visits at a single site by using the formula: $p_i = 1 - \prod_{t=1}^{10} \left(1 - p_{it}\right)$ where p_{it} is the estimated detectability at site i and visit t and p_i is the cumulative detectability at site i across all visits. We used the cumulative detectability to select only sites with high detectability ($p_i \geq 0.90$) and determined whether the focal species was recorded on any visit. Using these validation data, we compared the estimated occupancy to the observed occurrence at each site. Using only the sites with high detectability ensured that we were getting close to comparing our estimates of occupancy with true species occurrence.

2.5 | Environmental data

As environmental covariates, we used land cover data derived from the MODIS product MCD12Q1 v006 (Friedl & Sulla-Menashe, 2015). We estimated the land cover associated with each checklist as the proportion of each land cover category in a 2.5 km × 2.5 km square surrounding the checklist location in the year the observations were made. We included the proportions of each of 16 land cover types in the UMD LC_Type2 classification of MODIS MCD12Q1 v006 classification (Friedl & Sulla-Menashe, 2015). See Appendix S2 for a list of the land cover types.

2.6 | Effort covariates

We used effort covariates that describe heterogeneity in observer effort that we expect to be associated with differences in detectability. eBird checklists contain information on the following effort covariates: start time of birding activity, duration of birding activity, whether observers were travelling or stationary, distance travelled and the number of observers. For occupancy models, we also included the square of "start time of birding activity," to enable quadratic relationships with time of day. Each of these covariates describes variation in effort that will impact detectability. We expect that all of these will usually be important descriptors of heterogeneity in effort, but the effect of these on detectability is likely to vary by species, region and season. Not all eBird checklists contain each of these variables, but all complete checklists contain each of these; by filtering to only complete checklists in step iii), we ensure that each of the checklists in the training and validation data all contains the effort variables.

2.7 | Estimating species encounter rate

We estimated the *encounter rate* of the two species on eBird checklists in relation to the environmental covariates for each of the seven treatments of the data (Table 1). We fitted models to 25 versions of data, from each of the seven treatments of the data. For each of the 25 versions, we randomly selected 0.75 of the training and validation datasets before applying the relevant data processing treatments (Figure 1). The response was the detection/non-detection of each species, and the environmental covariates were 16 land cover covariates described in Appendix S2 (Friedl & Sulla-Menashe, 2015). Model 1 used presence-only records of the species on a checklist, fitted with a Maxent model through the R package *maxnet* (Phillips, 2016). Models 2–7 fitted a random forest with a response of detection/non-detection records on checklists, followed by calibration with a generalized additive model (GAM). The random forest models were fitted with the R package *ranger* (Wright & Ziegler, 2017) and the calibration GAMs within R package *scam* (Pya, 2013). For further details of the model fitting, see Appendix S2 and the code in supporting information A3.

We used the validation datasets to validate the estimates either from the Maxent model or from the combination of the random forest and the calibration GAM. We used a range of performance metrics to compare the estimates to the observations: sensitivity, specificity, true skill statistic (TSS), area under the curve (AUC), kappa and mean squared error (MSE, also named Brier score). To quantify the benefit or detriment of the seven data refining or degrading steps, we calculated the differences in performance metrics between each of the 7 models and model 3. We selected model 3 as the "baseline" because it used no data degrading or refinement. We examined the distribution of these differences across the 25 different runs of the model sets.

Randomly selecting one of the twenty-five iterations of fitting the set of seven models, we mapped the estimated encounter rates across the whole region of the BCR27. We produced a dataset with the land cover for each 2.5 km × 2.5 km grid cell across the entire region and we set effort variables to be constant across the region. The predictions were the hypothetical encounter rate of an average eBird participant conducting a 1 hr, 1 km complete checklist on 15 June 2018 at the optimal time of day for species detection. We estimated encounter rate for this standardized checklist in each grid cell in BCR27, using each of the seven models.

2.8 | Estimating species occupancy

To assess the effects of these data processing steps in an alternative modelling framework, we applied single-species occupancy models to estimate occupancy and detectability. We modelled occupancy probability as a function of MODIS land cover (Friedl & Sulla-Menashe, 2015). However rather than using all 16 land cover variables as above, we selected four categories considered *a priori* to have the most ecological relevance for wood thrush (deciduous broadleaf forest, mixed forest, croplands and urban) and chuck-will's-widow (evergreen needleleaf, deciduous broadleaf, mixed forest, urban). For modelling detectability, we used five effort covariates described above and the square of start time of birding activity. We used the R package *unmarked* to fit single-season occupancy models (Fiske & Chandler, 2011). We could not run an occupancy model with the detection only data (model 1) above, but we ran these occupancy models using six different combinations of data processing that matched encounter rate models 2–7 (Table 1). The data degrading and refinement steps took place before we prepared the data for occupancy models. For further details of the data processing and model fitting, see Appendix S2. Given the more stringent data processing for occupancy models, there was less value in repeating this analysis several times as the datasets would be relatively similar; therefore, we did not repeat this analysis 25 times.

We validated the estimates from the occupancy model using the occupancy validation dataset described above. As above, we also mapped the occupancy rate across the whole region by predicting to the whole of BCR27.

2.9 | Varying sample size

Our study area has a relatively high density of eBird data, but other regions and other CS projects often have fewer data. Therefore, we wanted to assess whether the results we found would be similar with smaller datasets. We estimated wood thrush encounter rate using only models 3 and 7 for a range of sample sizes. As above, for each model pair (model 3 and model 7) we randomly selected 0.75 of both the training and validation datasets. We then further subsampled these new datasets to varying proportions of the new total: 0.1, 0.3, 0.5, 0.7 or 0.9. We ran this set of 10 analyses (five sample sizes, two models) 25 times. For each run, we compared the difference in predictive performance metrics (as described above) between model 7 and model 3.

3 | RESULTS

3.1 | Estimating species encounter rate

Both wood thrush and chuck-will's-widow results show model 7 had the highest estimates of encounter rate (Figure 2, Figures S4, S10 and S11) and the best model performance (Figure 3, Figures S2, S13 and S14). Model performance was consistently the best with model 7, across both validation datasets and most of the performance metrics (Figure 3, Figures S2, S13 and S14). Thus, the combination of all data processing steps resulted in the best model, and using complete checklists produced the biggest improvement for wood thrush, while adding covariates produced the biggest improvement for chuck-will's-widow (compare models 2 and 3).

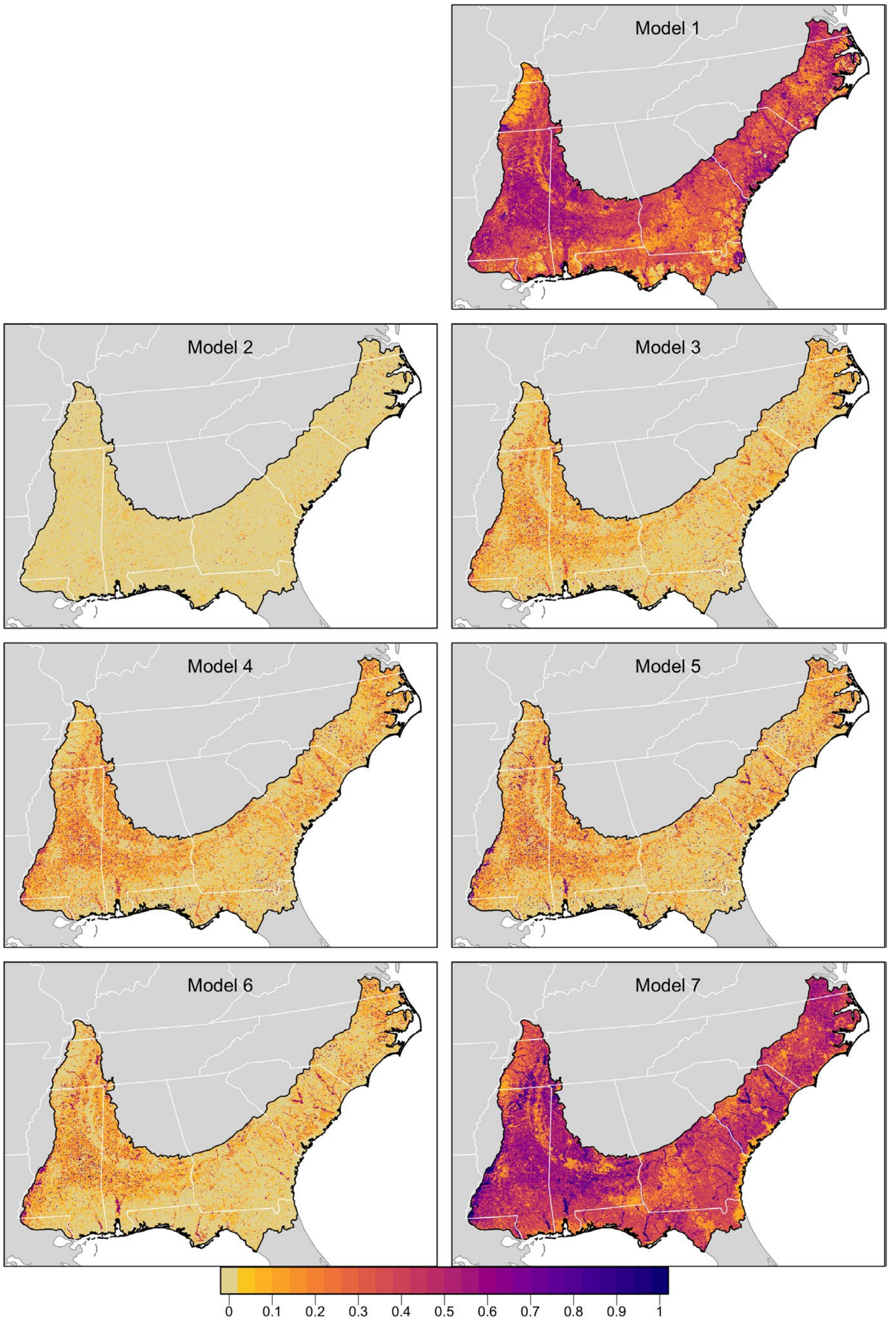


FIGURE 2 Estimated wood thrush encounter rate across the BCR27 region for models 1–7. Estimated encounter rate is the expected proportion of standardized checklists that would record wood thrush. These hypothetical standardized checklists are conducted by an average eBirder, travelling 1 km over 1 hr, at the optimal time of day for detecting wood thrush

With wood thrush data, both models 1 and 2 had substantially worse model performance than other models, which was evident with both of the validation datasets (Figure 3 and Figure S2). The estimates of encounter rate from models 1 and 2 were poorly correlated with those from model 7 (Figure S3), although there are some broad similarities in spatial patterns (Figure 2). These results demonstrate that for wood thrush using presence-only or casual observations (not part of a complete checklist) is likely to result in poorer ecological inference. Models 3–6 all displayed similar model performance (Figure 3 and Figure S2), similar absolute encounter rate (Figure S4) and similar correlations with the predictions from model 7 (Figure S3). As a contrast, chuck-will's-widow showed the greatest gains in model performance with the addition of effort variables as covariates and with the use of non-detections in model 2. There was smaller improvement for the other model refinement steps (Figures S13 and S14). Overall, due to the strong effect of time of day on the estimated encounter rate, most estimates had a poor correlation with those from model 7 (Figure S12). All these results suggest that the largest gains in model performance may vary with characteristics of the data, which we expect to vary by species, season and region.

3.2 | Estimating species occupancy

Across models, the estimates of occupancy for both wood thrush and chuck-will's-widow were less variable (within species) than those of encounter rate. The six occupancy models showed relatively consistent spatial patterns (Figure 4 and Figure S15) and high correlation

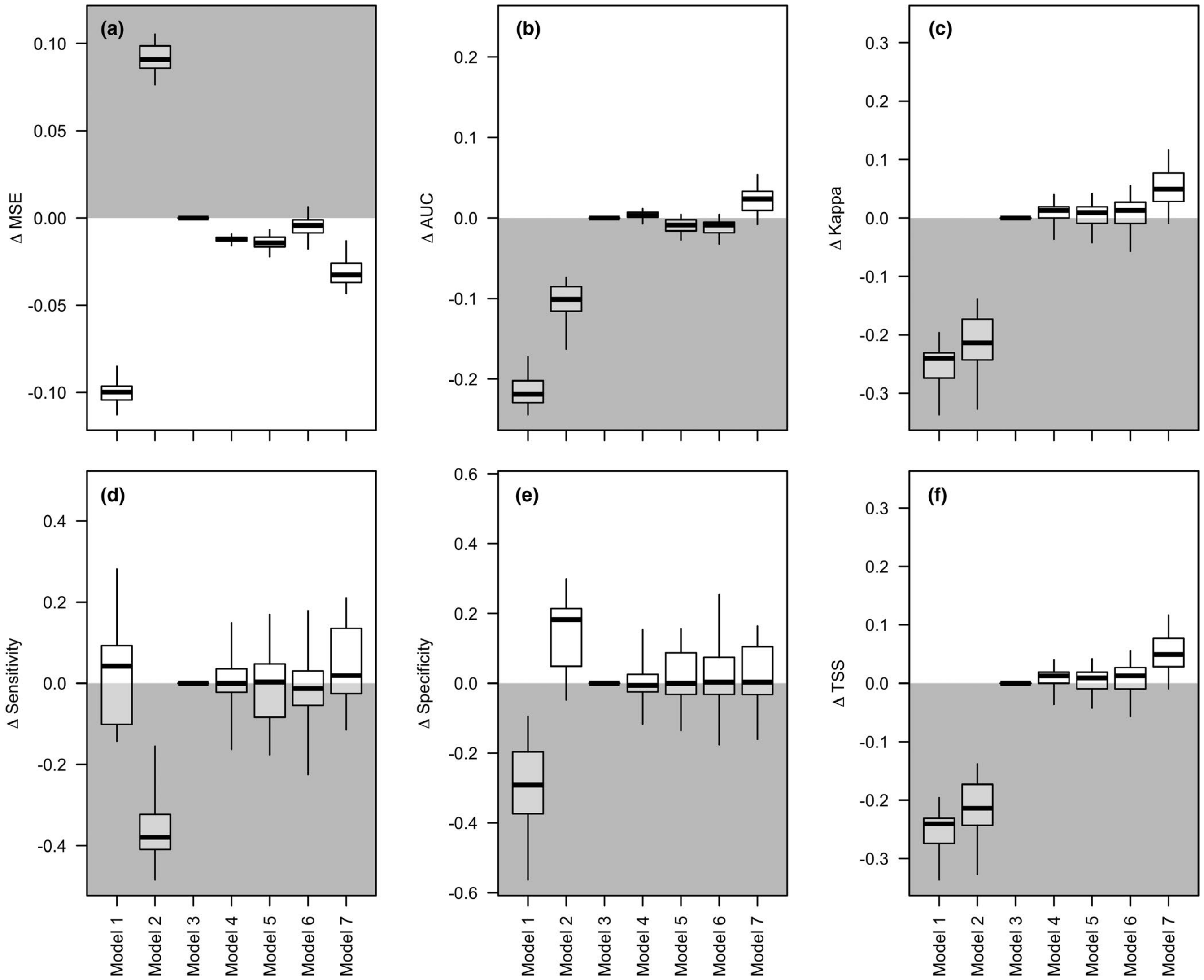


FIGURE 3 Differences in predictive performance metrics for the wood thrush encounter rate models 1–7 against balanced and subsampled eBird data from 2017. Metrics are compared to the performance from model 3, and the y-axis values show differences relative to model 3. The white halves of the plots indicate where model performance is *better* than model 3. The grey halves of the plots indicate where model performance is *worse* than model 3. Model 3 uses all the data in a random forest encounter rate model. Model 7 is the random forest encounter rate model using complete checklists, spatial subsampling, effort variable filters and effort variables as covariates. The validation metrics are calculated for 25 different model runs. For details of models 1–7, see Table 1 and the text. Boxes show the median and the interquartile range, and whisker ends denote the extremes of the distributions

between estimates (Figures S5 and S16). The notable outlier for occupancy models was model 7, when effort covariates were included. This led to correlated, but larger absolute estimates of occupancy (Figures S7 and S17), and slightly improved model performance by some metrics (Figures S8 and S19). With these training and validation datasets, therefore, we could not strongly identify improvements resulting from most of the data processing steps, but including effort covariates describing heterogeneity in detectability was an important improvement (Figures S6 and S18).

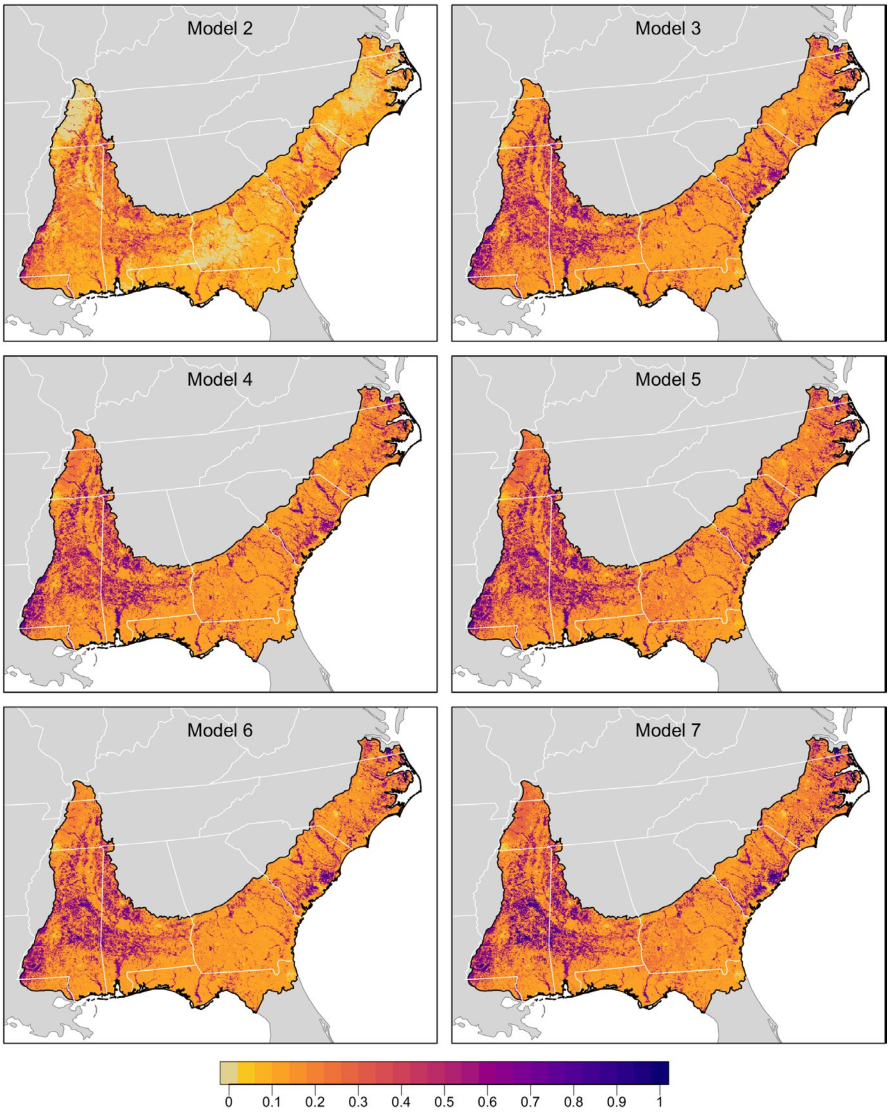
3.3 | Varying sample size

Model 7 (with all data refinement steps) was better than model 3 (no data refinement) (Figure 3). However, the benefits of using model 7 were reduced at smaller sample sizes (Figure 5 and Figure S9). This may be because reducing the dataset size by filtering (Figure 1) also has a cost when there are fewer data. However, we find that even with the smallest datasets, there is no disadvantage to using model 7—it performs equivalent to or better than model 3 across all sample sizes that we tested (Figure 5 and Figure S9).

4 | DISCUSSION

Community science datasets are becoming increasingly valuable research tools for ecology and conservation due to their increasing prevalence (Pocock et al., 2017) and broad spatio-temporal scope (Chandler et al., 2017). For example, eBird data have been used to study phenology, species distributions, population trends,

FIGURE 4 Estimated occupancy of wood thrush across the BCR27 region for occupancy models 2–7 calculated with data processing steps (ii) to (v). The occupancy is the expected probability that cells are occupied by wood thrush



evolution and behaviour and to inform conservation (Lang et al., 2019; MacPherson et al., 2018; Mattsson et al., 2018; Mayor et al., 2017; Seeholzer et al., 2017). However, CS data generally have more errors, assumptions and biases associated with them, often a result of relatively unconstrained survey design and a highly heterogeneous observation process. Here we demonstrate how thoughtful combinations of data filtering and analysis can remove relatively uninformative data and control for much of the statistical noise in CS data.

In our example, spatial subsampling did not result in large changes to the model performance. Spatial subsampling is designed to reduce the impact of spatial bias on the environmental relationships and subsequent species distribution. In line with our results, previous studies have found that spatial bias can have surprisingly little impact on estimated species distributions (Beck et al., 2014; Higa et al., 2014; Johnston et al., 2020). This may be particularly true where there are high data volumes, good coverage of environmental space, sampling that covers the species' environmental niche and stationarity of the species distribution across the region (Johnston et al., 2020), all of which are true in our example datasets. Accordingly, we did not see any impact of the spatial subsampling on the results. In general, we expect the impact of spatial subsampling would vary in different situations and with different subsampling parameters. For example, there may be a greater impact of spatial subsampling when estimating population trends or other processes that show spatial non-stationarity (Kamp et al., 2016; Zbinden et al., 2014).

Our results suggest that where effort data are not available, in some situations occupancy models may be a more robust modelling approach. Including information on the observation process has generally been shown to produce more accurate and robust results (Johnston et al., 2018; Isaac et al., 2014). In our analyses, the advantages of effort variables were important for chuck-will's-widow occupancy models, but were less apparent for wood thrush occupancy models. We also recognize that our occupancy model validation scheme was less robust, and further study is needed.

We found that model performance was poorer when we degraded the data to reflect two common types of CS data: to detections only (presence-only data) and to incomplete checklists only. There are clear limitations to the ecological insights that can be gained from presence-only data (Aranda & Lobo, 2011;

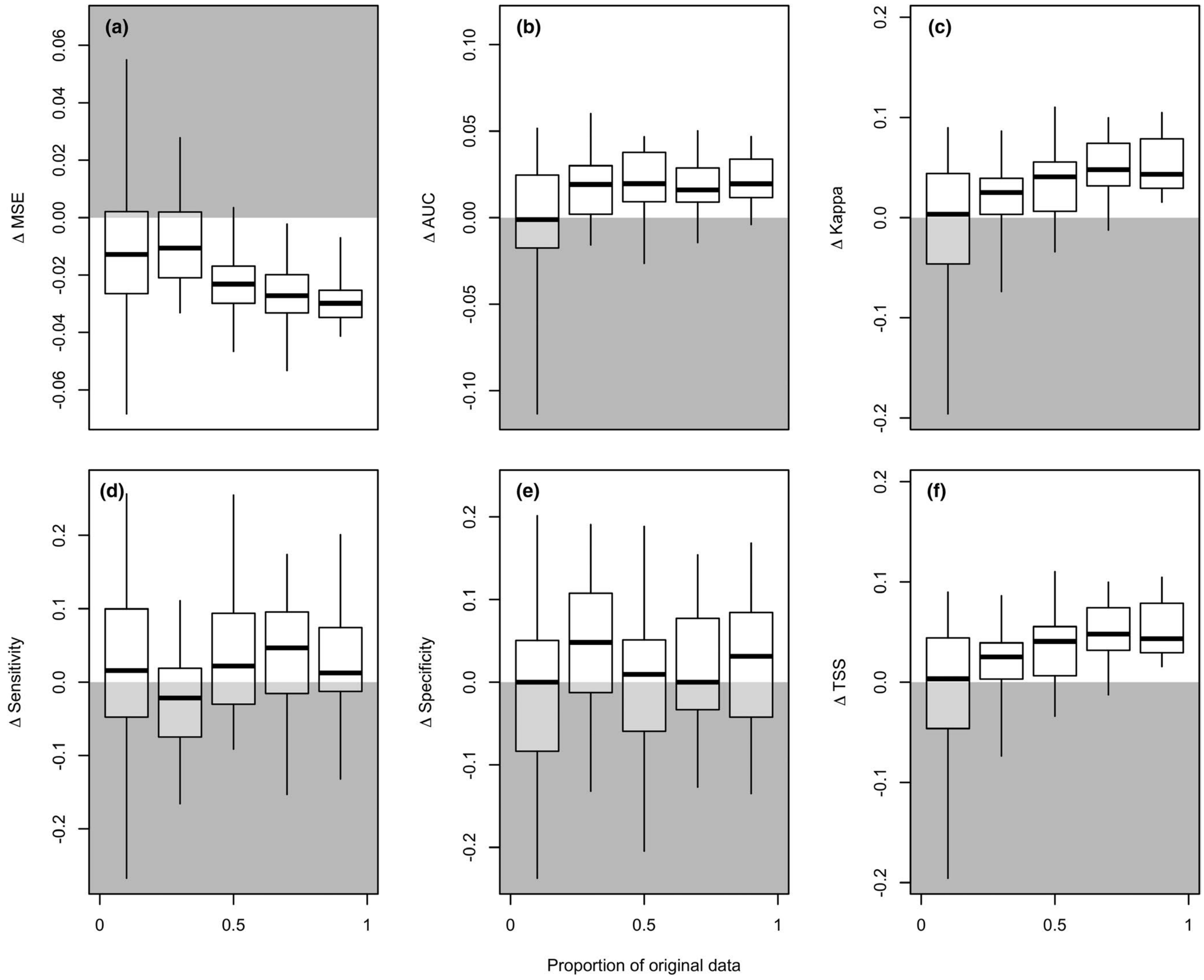


FIGURE 5 Effect of sample size on differences in predictive performance metrics for the wood thrush encounter rate models 3 and 7. Differences were computed between the metrics as (model 7 - model 3); the y-axis values show differences relative to model 3. The white halves of the plots indicate where model 7 performance is *better* than model 3. The grey halves of the plots indicate where model 7 performance is *worse* than model 3. The test dataset was balanced and subsampled eBird data from 2017. The datasets were random subsampled to 0.75 of the original checklists. Then, they were further reduced to a proportion of this dataset: 0.1, 0.3, 0.5, 0.7 and 0.9. This process was repeated 25 times to produce 25 paired comparisons of model performance for each dataset size. Each paired comparison between model 3 and model 7 used the same randomly subsampled test and train datasets. See Table 1 for further details of model 3 and model 7. Panels show the following performance metrics: A mean squared error (*MSE*); B area under the curve (AUC); C kappa; D sensitivity; E specificity; and F true skill statistic (TSS). Boxes show the median, the interquartile range and the extremes of the distributions

Václavík & Meentemeyer, 2009). As a result, multiple approaches have been suggested for inferring non-detection events when data are stored in a presence-only format (Hill, 2012; van Strien et al., 2013). Our case study strongly supports the importance of complete checklists and the value of retaining this information in analyses.

Our general recommendation is that both filtering and modelling variation in effort are important analytical tools, although their benefits will vary across datasets and modelling objectives. In our examples, we find that analysing complete checklists and using effort variables as covariates made the largest difference to the model quality. However, the raw data, the volume of data, the model type and the modelling objective will all affect the relative benefit of the data processing steps that we describe. In the two metrics and two species we investigated, the refinements we made to the data and models either had no negative effect or notably improved model performance. As such, we suggest these refinements should be implemented as a general practice; however, the impact of these filtering and modelling practices should be further evaluated for different datasets and ecological questions. Here, we investigated and recommend current best practices for using semi-structured CS data to estimate species occurrence. However, for other ecological questions the trade-offs related to data quantity and refinement may lead to different optimal data processing steps. Most importantly, we encourage other researchers to carefully consider and test appropriate data processing for their own questions.

While we have focused on aspects of the observation process that create false-negative errors, data can also contain false-positive errors. These false positives occur when an observer falsely records a bird as present, which is usually a result of misidentification of another species. An increasing number of studies demonstrate the importance of accounting for false positives (Pillay et al. 2014; Chambert et al. 2015) and in some cases even a low rate of false positives can create biased estimates of species distributions (Miller et al. 2011). We have not discussed the treatment of false positives in this paper, and eBird data do not contain required information to estimate false-positive error rates. Additionally, due to the eBird review process, unlikely species records require additional evidence to enter the publicly accessible data, so false positives in the eBird data will only be species that could be plausibly detected in those places. Therefore, false positives in eBird should not affect estimates of species ranges, but could bias estimates of occurrence or relative abundance within a species' range.

There are numerous CS programmes in the world today, but only a limited number of them collect the information needed to infer non-detections (Pocock et al., 2017). eBird provides evidence that information on observer effort and completeness of species lists can be collected while maintaining high participation. While we focused on modelling species distributions, many other types of ecological inference and conservation planning will also benefit from these data processing steps. In combination, the approaches outlined here for collecting, processing and modelling CS data can inform ways to improve existing and future programmes, while increasing our current capacity to conduct robust analyses using growing volumes of community science data.

ACKNOWLEDGEMENTS

We thank the thousands of participants in the eBird project. We are grateful to Frank La Sorte, Marshall Iliff, journal editors and reviewers for discussions and comments that improved this manuscript. Funding for this work came from the Wolf Creek Foundation, The Leon Levy Foundation and the National Science Foundation (grants ITR-0427914, DBI-0542868, IIS-0612031, ABI-1356308, CCF-1522054 and ICER-1927646). ICR-1927646 funding to Cornell University was through the 2017-2018 Belmont Forum and BiodivERsA joint call for research proposals, under the BiodivScen ERA-Net COFUND programme, with financial support from the Academy of Finland (AKA, Univ. Turku: 326327, Univ. Helsinki: 326338), the Swedish Research Council (Formas, SLU: 2018-02440, Lund Univ.: 2018-02441), the Research Council of Norway (Forskningsrådet, NINA: 295767) and the U.S. National Science Foundation.

AUTHOR CONTRIBUTIONS

All authors conceived the ideas and designed methodology. AJ analysed the data, based on preliminary analyses conducted with WMH, VRG and OR. AJ, WMH, VRG, ETM and OR wrote the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/ddi.13271.

DATA AVAILABILITY STATEMENT

We provide the data and code for the analyses in the github repository: https://github.com/ali-johnston/ebird_sdms_DD_paper/tree/v1.0 with https://doi.org/10.5281/zenodo.4642528. eBird data for other species or regions can be downloaded from here: https://ebird.org/data/download.

ORCID

Alison Johnston https://orcid.org/0000-0001-8221-013X
Orin J. Robinson https://orcid.org/0000-0001-8935-1242

REFERENCES

Altwegg, R., & Nichols, J. D. (2019). Occupancy models for citizen-science data. *Methods in Ecology and Evolution*, *10*(1), 8–21. https://doi.org/10.1111/2041-210X.13090

Aranda, S. C., & Lobo, J. M. (2011). How well does presence-only-based species distribution modelling predict assemblage diversity? A case study of the Tenerife flora. *Ecography*, *34*(1), 31–38. https://doi.org/10.1111/j.1600-0587.2010.06134.x

Beck, J., Böller, M., Erhardt, A., & Schwanghart, W. (2014). Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics*, *19*, 10–15. https://doi.org/10.1016/j.ecoinf.2013.11.002

Bird, T. J., Bates, A. E., Lefcheck, J. S., Hill, N. A., Thomson, R. J., Edgar, G. J., Stuart-Smith, R. D., Wotherspoon, S., Krkosek, M., Stuart-Smith, J. F., Pecl, G. T., Barrett, N., & 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

Botts, E. A., Erasmus, B. F. N., & Alexander, G. J. (2011). Geographic sampling bias in the South African Frog Atlas Project: Implications for conservation planning. *Biodiversity and Conservation*, *20*(1), 119–139. https://doi.org/10.1007/s10531-010-9950-6

Casanovas, P., Lynch, H. J., & Fagan, W. F. (2014). Using citizen science to estimate lichen diversity. *Biological Conservation*, *171*, 1–8. https://doi.org/10.1016/j.biocon.2013.12.020

Chambert, T., Miller, D. A. W., & Nichols, J. D. (2015). Modeling false positive detections in species occurrence data under different study designs. *Ecology*, *96*, 332–339. https://doi.org/10.1890/14-1507.1

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

Dennis, R. L. H., & Thomas, C. D. (2000). Bias in butterfly distribution maps: The influence of hot spots and recorder's home range. *Journal of Insect Conservation*, *4*(2), 73–77.

Diefenbach, D. R., Marshall, M. R., Mattice, J. A., & Brauning, D. W. (2007). Incorporating availability for detection in estimates of bird abundance. *The Auk*, *124*(1), 96–106. https://doi.org/10.1093/auk/124.1.96

Ellis, M. V., & Taylor, J. E. (2018). Effects of weather, time of day, and survey effort on estimates of species richness in temperate woodlands. *Emu*, *118*(2), 183–192. https://doi.org/10.1080/01584197.2017.1396188

Fiske, I., & Chandler, R. (2011). unmarked: An R package for fitting hierarchical models of wildlife occurrence and abundance. *Journal of Statistical Software*, *43*(10), 1–23.

Friedl, M., & Sulla-Menashe, D. (2015). MCD12Q1 MODIS/Terra+ Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V006. NASA EOSDIS Land Processes DAAC.

Fuller, R. J., & Langslow, D. R. (1984). Estimating numbers of birds by point counts: How long should counts last? *Bird Study*, *31*(3), 195–202. https://doi.org/10.1080/00063658409476841

Higa, M., Yamaura, Y., Koizumi, I., Yabuhara, Y., Senzaki, M., & Ono, S. (2014). 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

Hijmans, R. J., Garrett, K. A., Huaman, Z., Zhang, D. P., Schreuder, M., & Bonierbale, M. (2000). Assessing the geographic representativeness of genebank collections: The case of Bolivian wild potatoes. *Conservation Biology*, *14*(6), 1755–1765. https://doi.org/10.1046/j.1523-1739.2000.98543.x

Hill, M. O. (2012). Local frequency as a key to interpreting species occurrence data when recording effort is not known. *Methods in Ecology and Evolution*, *3*(1), 195–205. https://doi.org/10.1111/j.2041-210X.2011.00146.x

Hochachka, W. M., Alonso, H., Gutiérrez-Expósito, C., Miller, E., & Johnston, A. (2021). Regional variation in the impacts of the COVID-19 pandemic on the quantity and quality of data collected by the project eBird. *Biological Conservation*, *254*, 108974. https://doi.org/10.1016/j.biocon.2021.108974

Hochachka, W. M., Winter, M., & Charif, R. A. (2009). Sources of variation in singing probability of Florida Grasshopper Sparrows, and implications for design and analysis of auditory surveys. *Condor*, *111*(2), 349–360. https://doi.org/10.1525/cond.2009.080086

Howard, E., Aschen, H., & Davis, A. K. (2010). Citizen science observations of monarch butterfly overwintering in the southern United States. *Psyche*, *2010*, 689301.

Isaac, N. J. B., Van Strien, A. J., August, T. A., de 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

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*, 88–97. https://doi.org/10.1111/2041-210X.12838

Johnston, A., Moran, N., Musgrove, A., Fink, D., & Baillie, S. R. (2020). Estimating species distributions from spatially biased citizen science data. *Ecological Modelling*, *422*, 108927. https://doi.org/10.1016/j.ecolmodel.2019.108927

Kadmon, R., Farber, O., & Danin, A. (2004). Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications*, *14*(2), 401–413. https://doi.org/10.1890/02-5364

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., Johnston, A., Bonn, A., Fink, D., Ruiz-Gutierrez, V., Bonney, R., Fernandez, M., Hochachka, W. M., Julliard, R., Kraemer, R., & Guralnick, R. (2019). Using semi-structured surveys to improve citizen science data for monitoring biodiversity. *BioScience*, *69*(3), 170–179. https://doi.org/10.1093/biosci/biz010

Kelling, S., Johnston, A., Hochachka, W. M., Iliff, M., Fink, D., Gerbracht, J., Lagoze, C., La Sorte, F. A., Moore, T., Wiggins, A., Wong, W. K., Wood, C., & 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

Lang, S. D. J., Mann, R. P., & Farine, D. R. (2019). Temporal activity patterns of predators and prey across broad geographic scales. *Behavioral Ecology*, *30*(1), 172–180. https://doi.org/10.1093/beheco/ary133

MacPherson, M. P., Jahn, A. E., Murphy, M. T., Kim, D. H., Cueto, V. R., Tuero, D. T., & Hill, E. D. (2018). Follow the rain? Environmental drivers of *Tyrannus* migration across the New World. *The Auk*, 881–894.

Mair, L., & Ruete, A. (2016). Explaining spatial variation in the recording effort of citizen science data across multiple taxa. *PLoS One*, *11*(1), e0147796. https://doi.org/10.1371/journal.pone.0147796

Mattsson, B. J., Dubovsky, J. A., Thogmartin, W. E., Bagstad, K. J., Goldstein, J. H., Loomis, J. B., Diffendorfer, J. E., Semmens, D. J., Wiederholt, R., & López-Hoffman, L. (2018). Recreation economics to inform migratory species conservation: Case study of the northern pintail. *Journal of Environmental Management*, *206*, 971–979. https://doi.org/10.1016/j.jenvman.2017.11.048

Mayor, S. J., Guralnick, R. P., Tingley, M. W., Otegui, J., Withey, J. C., Elmendorf, S. C., Andrew, M. E., Leyk, S., Pearse, I. S., & Schneider, D. C. (2017). Increasing phenological asynchrony between spring green-up and arrival of migratory birds. *Scientific Reports*, *7*(1), 1902. https://doi.org/10.1038/s41598-017-02045-z

Miller, D. A., Nichols, J. D., McClintock, B. T., Campbell Grant, E. H., Bailey, L. L., & Weir, L. A. (2011). Improving occupancy estimation when two types of observational error occur: non-detection and species misidentification. *Ecology*, *92*(7), 1422–1428. https://doi.org/10.1890/10-1396.1

Miller, D. A. W., Pacifici, K., Sanderlin, J. S., & Reich, B. J. (2019). The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution*, *10*(1), 22–37. https://doi.org/10.1111/2041-210X.13110

Newson, S. E., Evans, H. E., & Gillings, S. (2015). A novel citizen science approach for large-scale standardised monitoring of bat activity and distribution, evaluated in eastern England. *Biological Conservation*, *191*, 38–49. https://doi.org/10.1016/j.biocon.2015.06.009

NABCI: North American Bird Conservation Initiative (2000). *Bird conservation region descriptions: a supplement to the North American Bird Conservation Initiative bird conservation regions map*. US NABCI Committee.

Oliveira, C. V., Olmos, F., dos Santos-Filho, M., & Bernardo, C. S. S. (2018). Observation of diurnal soaring raptors in northeastern Brazil depends on weather conditions and time of day. *The Journal of Raptor Research*, *52*(1), 56–65. https://doi.org/10.3356/JRR-16-102.1

Pacifici, K., Reich, B. J., Miller, D. A. W., Gardner, B., Stauffer, G., Singh, S., McKerrow, A., & Collazo, J. A. (2017). Integrating multiple data sources in species distribution modeling: A framework for data fusion. *Ecology*, *98*(3), 840–850. https://doi.org/10.1002/ecy.1710

Phillips, S. (2016). Maxnet: Fitting "maxent" species distribution models with "glmnet".

Pillay, R., Miller, D. A. W., Hines, J. E., Joshi, A. A., Madhusudan, M. D. (2014). Accounting for false positives improves estimates of occupancy from key informant interviews. *Biodiversity Research*, *20*(2), 223–225. https://doi.org/10.1111/ddi.12151

Pocock, M. J. O., Tweddle, J. C., Savage, J., Robinson, L. D., & Roy, H. E. (2017). The diversity and evolution of ecological and environmental citizen science. *PLoS One*, *12*(4), e0172579. https://doi.org/10.1371/journal.pone.0172579

Pya, N. (2013). scam: Shape constrained additive models.

R Core Team (2018). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Sauer, J. R., Pardieck, K. L., Ziolkowski, D. J., Smith, A. C., Hudson, M.-A.- R., Rodriguez, V., Berlanga, H., Niven, D. K., & Link, W. A. (2017).

The first 50 years of the North American Breeding Bird Survey. *The Condor*, *119*(3), 576–593. https://doi.org/10.1650/CONDOR-17-83.1

Seeholzer, G. F., Claramunt, S., & Brumfield, R. T. (2017). Niche evolution and diversification in a Neotropical radiation of birds (Aves: Furnariidae). *Evolution*, *71*(3), 702–715. https://doi.org/10.1111/evo.13177

Strimas-Mackey, M., Miller, E., & Hochachka, W. (2017). auk: eBird Data Extraction and Processing with AWK. *R Package Version*.

Sullivan, B. L., Aycrigg, J. L., Barry, J. H., Bonney, R. E., Bruns, N., Cooper, C. B., Damoulas, T., Dhondt, A. A., Dietterich, T., Farnsworth, A., Fink, D., Fitzpatrick, J. W., Fredericks, T., Gerbracht, J., Gomes, C., Hochachka, W. M., Iliff, M. J., Lagoze, C., La Sorte, F. A., … 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

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

Tulloch, A. I. T., Possingham, H. P., Joseph, L. N., Szabo, J., & Martin, T. G. (2013). Realising the full potential of citizen science monitoring programs. *Biological Conservation*, *165*, 128–138. https://doi.org/10.1016/j.biocon.2013.05.025

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

Václavík, T., & Meentemeyer, R. K. (2009). Invasive species distribution modeling (iSDM): Are absence data and dispersal constraints needed to predict actual distributions? *Ecological Modelling*, *220*(23), 3248–3258. https://doi.org/10.1016/j.ecolmodel.2009.08.013

Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2018). block CV: An r package for generating spatially or environmentally separated folds for k -fold cross-validation of species distribution models. *Methods in Ecology and Evolution*, *67*, 617.

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

Vianna, G. M. S., Meekan, M. G., Bornovski, T. H., & Meeuwig, J. J. (2014). Acoustic telemetry validates a citizen science approach for monitoring sharks on coral reefs. *PLoS One*, *9*(4), e95565. https://doi.org/10.1371/journal.pone.0095565

Wiggins, A., & Crowston, K. (2011). From Conservation to Crowdsourcing: A Typology of Citizen Science. 2011 44th Hawaii International Conference on System Sciences, 1–10.

Wright, M., & Ziegler, A. (2017). ranger: A fast implementation of Random Forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*(1), 1–17.

Zbinden, N., Kéry, M., Häfliger, G., Schmid, H., & Keller, V. (2014). A resampling-based method for effort correction in abundance trend analyses from opportunistic biological records. *Bird Study*, *61*(4), 506–517. https://doi.org/10.1080/00063657.2014.969679

BIOSKETCH

The authors work together at the Lab of Ornithology, working on statistical analysis and conservation applications of eBird data. They create analytical approaches for eBird data and other ecological data, designed to enable robust ecological inference.

SUPPORTING INFORMATION

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

How to cite this article: Johnston A, Hochachka WM, Strimas-Mackey ME, et al. Analytical guidelines to increase the value of community science data: An example using eBird data to estimate species distributions. *Divers Distrib*. 2021;27:1265–1277. https://doi.org/10.1111/ddi.13271