Received: 16 July 2021 | Accepted: 18 February 2022

DOI: 10.1111/2041-210X.13905

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

RESEARCH ARTICLE

A continuous-score occupancy model that incorporates uncertain machine learning output from autonomous biodiversity surveys

Tessa A. Rhinehart[1] | Daniel Turek[2] | Justin Kitzes[1]

[1]Department of Biological Sciences, University of Pittsburgh, Pittsburgh, PA, USA

[2]Department of Mathematics & Statistics, Williams College, Williamstown, MA, USA

Correspondence
Tessa A. Rhinehart
Email: tessa.rhinehart@pitt.edu

Funding information
Department of Biological Sciences, University of Pittsburgh; National Fish and Wildlife Foundation

Handling Editor: Torbjørn Ergon

Abstract

1. Ecologists often study biodiversity by evaluating species occupancy and the relationship between occupancy and other covariates. Occupancy models are now widely used to account for false absences in field surveys and to reduce bias in estimates of covariate relationships. Existing occupancy models take as inputs binary detection/non-detection observations of species at each visit to each site. However, autonomous sensing devices and machine learning models are increasingly used to survey biodiversity, generating a new type of observation record (i.e. continuous-score data) that reflects the model's confidence a species is present in each autonomously sensed file, instead of binary detection/non-detection data. These data are not directly compatible with traditional binary occupancy modelling methods.
2. Here, we develop a new occupancy model that models continuous scores on a visit level as a Gaussian mixture, combining a distribution of scores for files that do contain the species of interest and a distribution of scores for files that do not. The model takes as input continuous scores for each autonomously sensed and classified file, along with an optional small number of binary, manually verified detection and non-detection annotations.
3. We present a simulation study that shows that over a range of empirically realistic parameters, our model outperforms traditional occupancy models that are based on binary annotation alone. We also apply this new model to an empirical case study using data generated from five machine learning classifiers applied to autonomous acoustic recordings gathered in the eastern United States.
4. Because our occupancy model generalizes allowable input data beyond binary observations, it is particularly well-suited to the increasing volume of machine learning classified data in ecology and conservation.

KEYWORDS
acoustic recorders, machine learning, occupancy modelling, remote sensing

1 | INTRODUCTION

Ecologists and conservation biologists are increasingly in need of tools to efficiently survey biodiversity at large scales. Scientists often assess occupancy, the proportion of sites at which a species occurs, to identify species at risk or evaluate conservation efforts. Occupancy itself can suggest the degree to which a population persists or is threatened across its range (Field et al., 2005; MacKenzie, 2005) and is used in studies of metapopulation dynamics (e.g. Chandler et al., 2015; Hanski, 1994; MacKenzie et al., 2003). Occupancy is also used to predict how species respond to covariates such as habitat characteristics, climate change, disturbances or restoration efforts (e.g. Cavada et al., 2019; Chambert, Kendall, et al., 2015; Johnson, 2007). Accurately estimating species occupancy is thus of critical importance to advancing ecological understanding and conserving species.

To estimate species occupancy, surveys are often conducted to attempt to detect a species of interest at particular sites. However, detection/non-detection data do not themselves provide knowledge of site occupancy, which cannot generally be measured directly in the field because observers detect species imperfectly (MacKenzie et al., 2002). Most significantly, field surveys usually include false absences, where a species is present at a site but undetected (Martin et al., 2005). Field surveys may also produce false presences, in which a species does not occupy a site but is mistakenly identified as present, for example, due to misidentifications (Iknayan et al., 2014). Imperfect detection can cause bias and reduce power in studies measuring species occurrence or the effects of ecological covariates on species occurrence, especially where covariates also influence detection probability (Field et al., 2005; Gu & Swihart, 2004; Mackenzie, 2006; Martin et al., 2005; Tyre et al., 2003).

Over the past two decades, occupancy modelling has gained popularity among ecologists as a method of accounting for measurement errors in surveys and reducing bias in estimation of covariates' effects on occupancy (MacKenzie et al., 2018). Single-species occupancy models traditionally leverage multiple binary detection/non-detection surveys at each sampled site to partition the probability that a species is present from the probability that a species is detected given its presence. More recently, models have been developed to account for the rate of false positives as well (Chambert, Miller, & Nichols, 2015; MacKenzie et al., 2018; Miller et al., 2011, 2013; Royle & Link, 2006).

While traditional occupancy models were developed for use with binary observational data generated from human detection/non-detection surveys, researchers are increasingly collecting similar survey data with autonomous sensors. These sensors, including automated acoustic recorders and camera traps, have several advantages for biodiversity surveys. Autonomous sensors may enable scientists to collect data over larger spatial scales (Darras et al., 2019; Shonfield & Bayne, 2017), increase detection of nocturnal, rare or secretive species (e.g. Bobay et al., 2018; Wrege et al., 2017), more easily survey remote environments (Buxton & Jones, 2012; Gray et al., 2019), provide a verifiable and long-lasting record of species (Darras et al., 2019; Newson et al., 2017), and generate many repeat 'surveys' for potential use in occupancy modelling (Abrahams & Geary, 2020; Chambert et al., 2018).

Given the large spatial and temporal scales at which these sensors can collect images and audio, much effort has been put into creating automated algorithms capable of analysing large volumes of sensor data, such as machine learning classifiers and signal processing algorithms (Priyadarshani et al., 2018; Tabak et al., 2019). Deep learning algorithms such as convolutional neural networks are becoming increasingly popular for processing these data, and in recent years have been shown to outcompete other algorithms for the analysis of complex data (e.g. avian soundscape data, Kahl et al., 2019). These and other automated recognizer algorithms (e.g. signal processing algorithms, Lapp et al., 2021) produce continuous scores reflecting the model's confidence that each file contains the species they are designed to identify (though these scores should not be confused for calibrated probabilities of presence within a file, especially in deep learning; see Guo et al., 2017).

Here, we develop an occupancy model that takes as inputs continuous scores of the type returned by many machine learning classifiers. Instead of 'visits', which in traditional applications of occupancy models are a series of physical visits to each field site, our individual surveys are 'files', the numerous autonomously sensed files generated by autonomous sensors and scored by a machine learning algorithm. We examine the model's performance on both simulated and real classifier scores and compare this to the performance of a traditional occupancy model that uses only human annotated data, without the additional information provided by classifier scores. We show that the continuous-score model recovers unbiased estimates of true occupancy across a wide range of plausible parameter values. In an empirical case study, the continuous-score model is found to outperform the traditional model in 52% of cases, provide comparable performance in an additional 28% of cases and produce a slightly less accurate estimate of occupancy in 20% of cases.

2 | MATERIALS AND METHODS

2.1 | Standard occupancy model

Static single-species occupancy models are designed to estimate true occupancy for a species based on multiple visits to a site. For i in $1, \ldots, T$ sites, let z_i be a Bernoulli random variable that equals 1 if a species is present at site i and is otherwise 0. Assuming a species' occupancy probability, ψ, does not vary across sampled locations (see Supporting Information for a generalization), standard occupancy models estimate latent presence of a species as

$$z_i \sim \text{Bernoulli}(\psi).$$

Let y_{ij} be an observation of site i on visit j in $1, \ldots, J_i$ visits, where $y_{ij} = 1$ if the species is detected and $y_{ij} = 0$ if the species is not detected. Traditional occupancy models assume that if a site is occupied, the species is detected during a visit with probability p, while visits to an unoccupied site always result in non-detections (i.e., no false positives). The probability of detecting a species at a site i conditional on its true presence z_i is then

$$y_{ij} \mid z_i \sim \text{Bernoulli}(z_i p).$$

This model can be applied to data collected using autonomous sensors if a human annotator labels a subset of files from each site i with detection or non-detection of the species. If annotators do not falsely detect a species in a file when it is absent in the file, these data can be used directly in the model framework above.

2.2 | Continuous-score model design

Consider now the case in which each sample unit or 'site' i in $1, \ldots, T$ sites is surveyed by one autonomous sensor such as an acoustic recorder or a camera trap. Define a file f_{i,k_i} as a short audio clip or single camera trap image where the second subscript denotes k_i in $1, \ldots, K_i$ files generated by the sensor at site i. Each file is scored by a machine learning model, generating a score s_{i,k_i} that is a continuous real number representing the algorithm's confidence that a species of interest is present in the file. Scores are not required to sum to 1 across files or across multiple species in a multispecies classifier. The model also optionally allows for a second observation type, in which a subset of files receive a binary detection/non-detection annotation from a human observer.

Although many distributions may be used to describe s_{i,k_i}, empirical data (see below) suggest that a Gaussian mixture is appropriate for such classifier output. For machine learning algorithms that output scores in the range [0, 1], real number valued scores can be produced by applying the logit transform to the raw scores.

As above, we presume that the species occurs at site i with occupancy probability ψ, such that

$$z_i \sim \text{Bernoulli}(\psi).$$

We assume that files collected at an occupied site i contain the species with probability θ_i, representing a 'call rate' for acoustic surveys or an 'appearance rate' for camera trap surveys. Assuming that θ_i is equal for all occupied sites, the presence of a species in a file f_{i,k_i} is given by

$$f_{i,\,k_i} \sim \text{Bernoulli}(z_i\theta),$$

such that if the site is unoccupied ($z_i = 0$) the species will not be present in any file. Mechanistically, θ can be interpreted to include all the factors leading to the presence of a species in a file, including not only a species' rate of calling or appearing, but also its availability for being detected by the autonomous sensor using a given survey (see Brack et al., 2018; Dénes et al., 2015; Martin et al., 2005).

If a file f_{i,k_i} does not contain the species, the score for that file is assumed to be assigned by a classifier as

$$s_{i,\,k_i} \mid f_{i,\,k_i} = 0 \sim N(\mu_0, \sigma_0),$$

where μ_0 and σ_0 are the mean and standard deviation of scores returned by the classifier for files that do not contain the species. Conversely, for files that do contain the species,

$$s_{i,\,k_i} \mid f_{i,\,k_i} = 1 \sim N(\mu_1, \sigma_1),$$

where μ_1 and σ_1 are the mean and standard deviation of scores returned by the classifier for files that do contain the species.

These parameters provide measures of classifier performance. Classifiers with larger $\mu_1 - \mu_0$ are better able to discriminate between files that do and do not contain the species (Figure 1a). Throughout, we assume that $\mu_1 > \mu_0$, such that the classifier is, at a minimum, capable of assigning higher scores on average to files in which a species is present. Likewise, classifiers with smaller σ_1 and σ_0 can discriminate with higher precision between files that do and do not contain the species (Figure 1a).

The overall distribution of scores within or across sites can thus be described as a two-component Gaussian mixture (Table 1). One component of the mixture consists of files in which the species is absent and the second component consists of files in which the species is present (Figure 1b). Within an occupied site i, both components will be present in the mixture with a weight of θ assigned to the positive component. When a site is not occupied, the positive component will not appear and the score distribution will be simply the Gaussian distribution $N(\mu_0, \sigma_0)$. If scores across all occupied and unoccupied sites are aggregated, the weight of the positive component will be the product of appearance rate and occupancy, $\theta * \psi$.

As shown below, this model can be used as specified above without any additional data. However, to improve performance, the model can be extended to also incorporate binary detection/non-detection human annotations for a small random subset of files at each site. When annotated data are available, every file f_{i,k_i} can be in one of three states: $h_{i,k_i} = 0$ if the file is confirmed to not contain the species by an annotator, $h_{i,k_i} = 1$ if the file is confirmed to contain the species by an annotator, or $h_{i,k_i} = 2$ if the file is not reviewed. Scores s_{i,k_i} are still available for all files. Human annotators are presumed to correctly annotate whether a file does or does not contain the species (see Section 2.3 below).

For this expanded model, the probability π_{i,k_i} of observing a score s_{ij} given the site's true occupancy state z_i and the human observation category h_{i,k_i} for that file is

$$\pi_{i,\,k_i} = P(s_{i,\,k_i} \mid z_i, h_{i,\,k_i}),$$

the parameterization of which is given in Table 1.

Given the vectors of scores S and of human annotations H the likelihood of the model is expressed as

$$\mathcal{L}\left(\psi, \mu_0, \sigma_0, \mu_1, \sigma_1, \theta \mid S, H\right) = \prod_{i=1}^{T}\left[(1-\psi)\prod_{k_i=1}^{K_i}\pi(s_{i,k_i} \mid z_i = 0, h_{i,k_i}) + \psi\prod_{k_i=1}^{K_i}\pi(s_{i,\,k_i} \mid z_i = 1h_{i,\,k_i})\right].$$

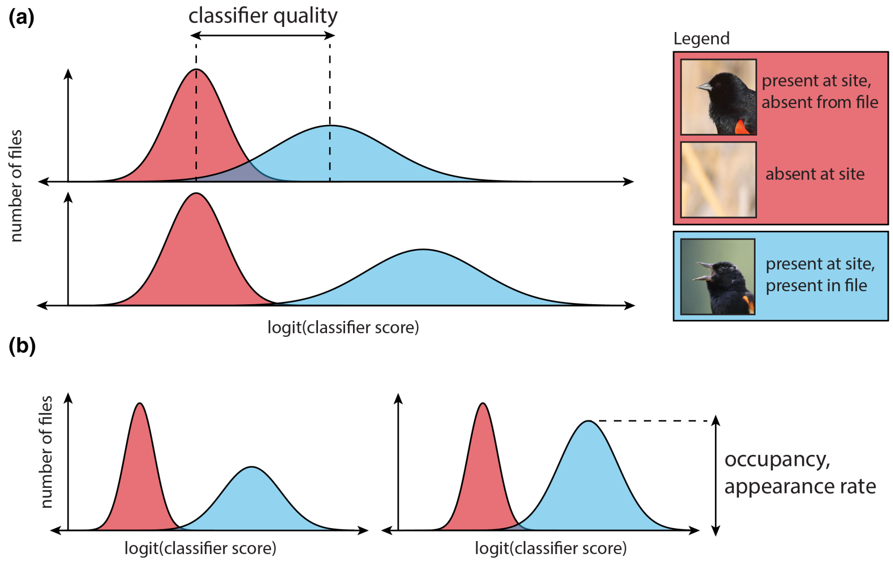

FIGURE 1 The outputs of a machine learning classifier applied to remote sensing files form a mixture of two Normal distributions. The logit function is applied to transform machine learning scores from [0, 1] to the real numbers. The mixture's upper and lower components are formed by scores from files that, respectively, did or did not contain the species of interest. (a) The amount of overlap between the two components is a measure of classifier quality. Higher quality classifiers may increase the mean of the positive distribution, reducing overlap between the two components representing an improved ability to differentiate positive and negative files. (b) Higher site occupancy and appearance rates increase the relative size of the upper component of the mixture. Note that when scores are examined on a site-by-site basis, the upper component will not be present for sites where the species does not occur

TABLE 1 The probability $\pi_{i,k_i} = P\left(s_{i,k_i} \mid z_i, h_{i,k_i}\right)$ of recording a machine learning score s_{i,k_i} given the site's true occupancy state z_i and the human observation for the file h_{i,k_i}. In this table, $f\left(s_{i,k_i} \mid \mu, \sigma\right)$ is the probability density function of the Normal distribution with mean μ and variance σ^2. The means and standard deviations of score distributions in files that do not contain the species and do contain the species are, respectively, μ_0, σ_0 and μ_1, σ_1. The probability that a species appears in a file at an occupied site is θ

	File annotated, species not detected, $h_{i,k_i} = 0$	**File annotated, species detected, $h_{i,k_i} = 1$**	**File not annotated, $h_{i,k_i} = 2$**
Site unoccupied, $z_i = 0$	$f\left(s_{i,k_i} \mid \mu_0, \sigma_0\right)$	0	$f\left(s_{i,k_i} \mid \mu_0, \sigma_0\right)$
Site occupied, $z_i = 1$	$(1-\theta) f\left(s_{i,k_i} \mid \mu_0, \sigma_0\right)$	$\theta f\left(s_{i,k_i} \mid \mu_1, \sigma_1\right)$	$(1-\theta) f\left(s_{i,k_i} \mid \mu_0, \sigma_0\right) + \theta f\left(s_{i,k_i} \mid \mu_1, \sigma_1\right)$

The full hierarchical specification of the model and additional details about the likelihood function are in Supporting Information.

We used the NIMBLE package in R to conduct a simulation study and an empirical study using this model and a traditional occupancy model (based off an implementation by Kéry & Royle, 2016) in a Bayesian framework (simulations: NIMBLE v0.9.0, R v4.0.0; empirical study: NIMBLE v0.11.0, R v3.6.3; de Valpine et al., 2017; NIMBLE Development Team, 2021). We used uninformative priors of Uniform(0, 1) for ψ and θ, $N\left(0, 10^4\right)$ for the mean of the distribution of scores for files where the species was absent, $N\left(1, 10^4\right)$ for the mean of the distribution of scores for files where the species was present, and Uniform(0, 10^4) for the standard deviations of the both of the score distributions. In fitting the traditional model, we used uninformative priors of Uniform(0, 1) for ψ and p. We estimated all posterior distributions using Markov-chain Monte Carlo (MCMC) simulation using 1 chain of 10,000 samples (10,000 burn-in, no thinning or adaptation, NIMBLE binary and RW samplers). We chose MCMC settings by visually inspecting plots of sample chains for several examples. We conducted graphical validation of sample chains to assess the convergence of a subset of models using v.0.19-4 of the CODA package for R (Plummer et al., 2006; Robert & Casella, 2010).

The model above has similarities to existing occupancy models that account for false positives in binary detection history (Chambert, Miller, & Nichols, 2015; Mackenzie et al., 2006; Miller et al., 2011; Royle & Link, 2006). Of these, our model is most similar to the 'multiple detection methods' site-confirmation design, which makes use of ambiguous detections that are easier to collect and unambiguous detections that are harder to collect (Miller et al., 2011). It also incorporates features of the 'observation confirmation' method, which allows definitive confirmation of species identification post-survey (Chambert, Miller, & Nichols, 2015).

Simultaneous to the development of our continuous-score model, Kéry and Royle (2021, p. 431) described an occupancy model that also treats the continuous output of a classification model as a Gaussian mixture. The model proposed by Kéry and Royle (2021) is structured more similarly to the model proposed by Chambert et al. (2018), while our continuous-score model above is most similar in structure to the 'multiple detection methods' site-confirmation design of Miller et al. (2011). Aside from presenting a complementary approach to continuous-score modelling, our work contributes the first tests of such a continuous-score model using realistic ranges of empirical parameter values and empirical machine learning classified data, while comparing the model performance to traditional occupancy models. Our analysis also examines the effects of varying levels of file annotation, a question that Kéry and Royle (2021) noted as particularly important to address.

2.3 | Methodological considerations

Before addressing the performance of our continuous-score model, we note several methodological considerations that are relevant to the use of this model in practice.

First, like static single season occupancy models (MacKenzie et al., 2002, 2003; Mackenzie & Royle, 2005), we expect that the continuous-score model described above may have difficulty estimating parameters that are very near the boundaries of their range, such as rare or infrequently detected species for which parameters ψ and θ may be expected to be near zero. In the empirical example below, which contains continuously sampled songbird vocalizations in a dawn chorus (Chronister et al., 2021), there are many species for which both parameters are far from zero. However, θ may be expected to be lower in some experimental designs, such as camera trapping studies in which a large proportion of photos are 'empty'. A detection model may be used prior to classification to remove empty files (Beery et al., 2019), although it is unclear how this would affect the underlying score distribution, or surveys may be restricted to times of day when a species is known to be more detectable.

Second, both static single season occupancy models and our continuous-score model above assume that detections in visits or files are independent of each other (Chambert et al., 2018; MacKenzie et al., 2002), and thus temporal autocorrelation should be avoided to the extent possible. This could be accomplished by empirically determining a minimum interval between files (e.g. Brook et al., 2012; Wang et al., 2015) or creating a survey-specific covariate to indicate whether a species has been detected at a site in prior surveys (e.g. MacKenzie et al., 2002). We note that the empirical study below does not attempt to correct for temporal autocorrelation, as the purpose of this study is to evaluate model performance under real-world parameter values and classifier performance, not to infer population parameters.

Third, the model above assumes that human annotators are able to correctly annotate the presence or absence of a species within a file. This assumption only requires such accuracy at the level of a file, and is not equivalent to stating that human annotation will always detect that a site is occupied when $z_i = 1$. Multiple human annotators could re-review a file as needed to ensure such accuracy. Future extensions of this model could be designed to account for scenarios where human annotators produce false negatives in particular types of files, such as recordings in which a vocalization is quiet or photos in which an animal is distant. If such false negatives tend to receive low scores from the classifier (i.e. both humans and the classifier are more likely to 'miss' the same files), the mixture model above could be extended to a three-component mixture. Alternative approaches could add an additional hierarchical layer that models the annotation process separately from θ to distinguish between detection and availability, which in traditional occupancy models are not important to distinguish.

Fourth, we note that the model above assumes that a species will be present in at least one file if it is present at a site, which once again suggests that our model will be most successful when θ is reasonably far from zero. As an illustration, a species present at a site with $\theta = 0.1$, a value much lower than that found in our empirical case study, will have less than a 0.02% chance of being absent in every file in a 60 file sample. In our analysis below, we approximate the probability that a species appears in zero files at an occupied site to be zero. In cases where θ is expected to be low, the approaches described above or an increase in the number of files per site can increase the appropriateness of this assumption.

2.4 | Simulation study

We first conducted a simulation study to test the ability of our model to estimate occupancy and compare its performance to a traditional occupancy model. Each simulation included 100 sites with 720 autonomously sensed 'files' collected at each site, equivalent to the number of 5 s audio clips that would be generated from 1 hr of autonomous acoustic recording.

We generated 100 simulations of each of 375 scenarios representing combinations of parameter values that were chosen to span a range of empirically realistic values. Each site was simulated to be occupied by a species of interest with probability ψ ($\psi = 0.05, 0.20, 0.5, 0.8, 0.95$). Files at occupied sites were simulated to contain the species with probability θ ($\theta = 0.05, 0.10, 0.25, 0.50, 0.75$). Differences in classifier performance were simulated by variation in μ_1 ($\mu_1 = 0.5, 3, 6$) while holding the other parameters describing classifier distributions constant ($\mu_0 = 0, \sigma_0 = 1$, and $\sigma_1 = 2$). These values of means and standard deviations are comparable to those of classifiers in our empirical examples. Finally, we simulated varying levels of human annotation using the true file-level presence or absence of a species in N files per site as the 'human annotation' for each file, corresponding to a range of zero to 10% file annotation. In addition to these analyses, we performed an additional demonstration of the model's ability to incorporate covariates on both θ and ψ (see Supporting Information).

While examining initial results from model fitting, we found there were cases in which the fitted model appeared unable to distinguish the two components of the mixture model, with estimates of μ_0 and μ_1 nearly equal. Approximately 5% of the total runs (1,742 of the 37,500 simulations) produced this type of failure. We consider this issue, diagnosed by fits in which the estimates of the two classifier means differed by less than 0.1, to be a type of model failure. For additional details, see the Supporting Information.

In this simulation study, we assessed models' bias through three methods: using the median error of the estimates, checking whether the interquartile range of estimates for the scenario contained the true population value of occupancy and calculating root-mean-square-error across the estimates for the scenario. We assessed precision by comparing the sizes of the interquartile ranges for estimated model parameters.

2.5 | Empirical study

We also tested our model empirically using a dataset of annotated autonomous recording unit (ARU) recordings collected at Powdermill Nature Reserve in Rector, PA. The goal of this study was to evaluate the models' performance under real-world parameter values and classifier performance. The source recordings totalled 385 min of recordings from 4 ARUs deployed during spring migration and the breeding season in an avian community of temperate eastern North America. Expert annotators determined the exact time, frequency and species identity of almost every bird vocalization in the recordings (Chronister et al., 2021).

We trained five machine learning classifiers to provide file-level scores for five species common in this dataset: Eastern Towhee *Pipilo erythrophthalmus*, Wood Thrush *Hylocichla mustelina*, Black-throated Green Warbler *Setophaga virens*, Black-capped Chickadee *Poecile atricapillus* and Northern Cardinal *Cardinalis cardinalis*. We split each recording into 5 s clips and determined whether each clip contained each species based on expert annotations. This process created 4,620 labelled 5 s clips. We used recordings from one ARU containing 840 5 s clips to train a ResNet18 convolutional neural network classifier for each species on spectrograms of the 5 s clips. We excluded recordings from this ARU from the remainder of the analysis. Classifiers were trained using OpenSoundscape v0.4.7 (Kitzes et al., 2020, Supporting Information). We then used these classifiers to produce file-level scores for species presence in the remaining 3,780 5 s files in the dataset. We split these remaining clips into 63 ‘temporal sites’, each corresponding to a continuous 5-min recording. We considered each 5-min recording, each containing 60 5 s files, to be a site.

We defined true ψ at temporal sites as the sample-at-hand occupancy (see also Chambert et al., 2018). Similar to Chambert et al. (2018), we estimated a site's sample-at-hand occupancy by whether the species was detected in at least one of the 60 files annotated for the site. We defined true θ as the percentage of files in which a species was detected divided by the total number of files taken at sites that were determined to be occupied. The proportion of sites occupied ψ and call rate θ at occupied sites for each species was ($\psi = 0.98$, $\theta = 0.79$) for Eastern Towhee, (0.38, 0.61) for Wood Thrush, (0.44, 0.57) for Black-throated Green Warbler, (0.59, 0.39) for Northern Cardinal and (0.46, 0.41) for Black-capped Chickadee. We considered annotation levels ranging from zero annotations to annotation of every file ($N = 0, 1, 2, 4, 6, 8, 32, 60$). Clips were randomly selected in a nested fashion such that annotations for each higher annotation level included clips annotated in lower annotation levels. We fit our continuous-score occupancy model to all annotation levels and the traditional occupancy model to all annotation levels except $N = 0$ and $N = 1$, as traditional occupancy models require multiple surveys per site.

3 | RESULTS

3.1 | Simulation study

After removing the approximately 5% of simulations which failed (see Section 2), our results demonstrate that the continuous-score model successfully produces unbiased estimates of occupancy. Across all succeeding simulations, the mean and median root mean squared error (RMSE) of the continuous-score model's estimates were both 0.04. The true value of occupancy probability ψ was within 0.05 of the median estimate of the simulations for 373 out of 375 scenarios (Figure 2; Figures S1 and S1). When the 5% of simulations which failed were removed, only 1 scenario showed notable bias in the interquartile range of the estimated occupancy, where the interquartile range did not contain the true value of ψ ($\mu_1 = 6$, 0 annotations, $\psi = 0.8$, $\theta = 0.05$). In this scenario, most of the simulations had been removed due to model failures (see Supporting Information; Figure S1). When model failures are not removed, an additional 8 scenarios of the 375 demonstrated interquartile ranges that did not contain the true value of ψ (Figures S3–S5).

Although nearly all scenarios produced unbiased estimates of ψ on average, scenarios varied in the number of outlier estimates of ψ. When excluding model failures, outliers were all associated with the use of poor classifiers to identify a rare or infrequently calling species when no files were annotated. When including model failures, outliers were more evenly spread across all parameter combinations (Figures S3–S5). When excluding model failures, the precision of occupancy estimates increased slightly with increasing ψ for our models but was otherwise consistent over annotation levels, call rates and classifier qualities (Figure 2; Figures S1 and S1). With model failures included, estimates were notably less precise in scenarios in which $\mu_1 = 6$. We stress again that model failures are easily diagnosable by $|\mu_1 - \mu_0| < 0.1$. Fitting the continuous-score model to 37,500 simulated datasets in a cluster environment took approximately 17,000 core-hours.

As compared to our continuous-score model when failures are removed, the traditional occupancy model produced notably biased and less precise estimates of ψ. The traditional occupancy model gave results comparable to the continuous-score model across all parameter combinations only in the 72-annotation scenario, in which 10% of all files had been assumed to be annotated (Figure 2). Below

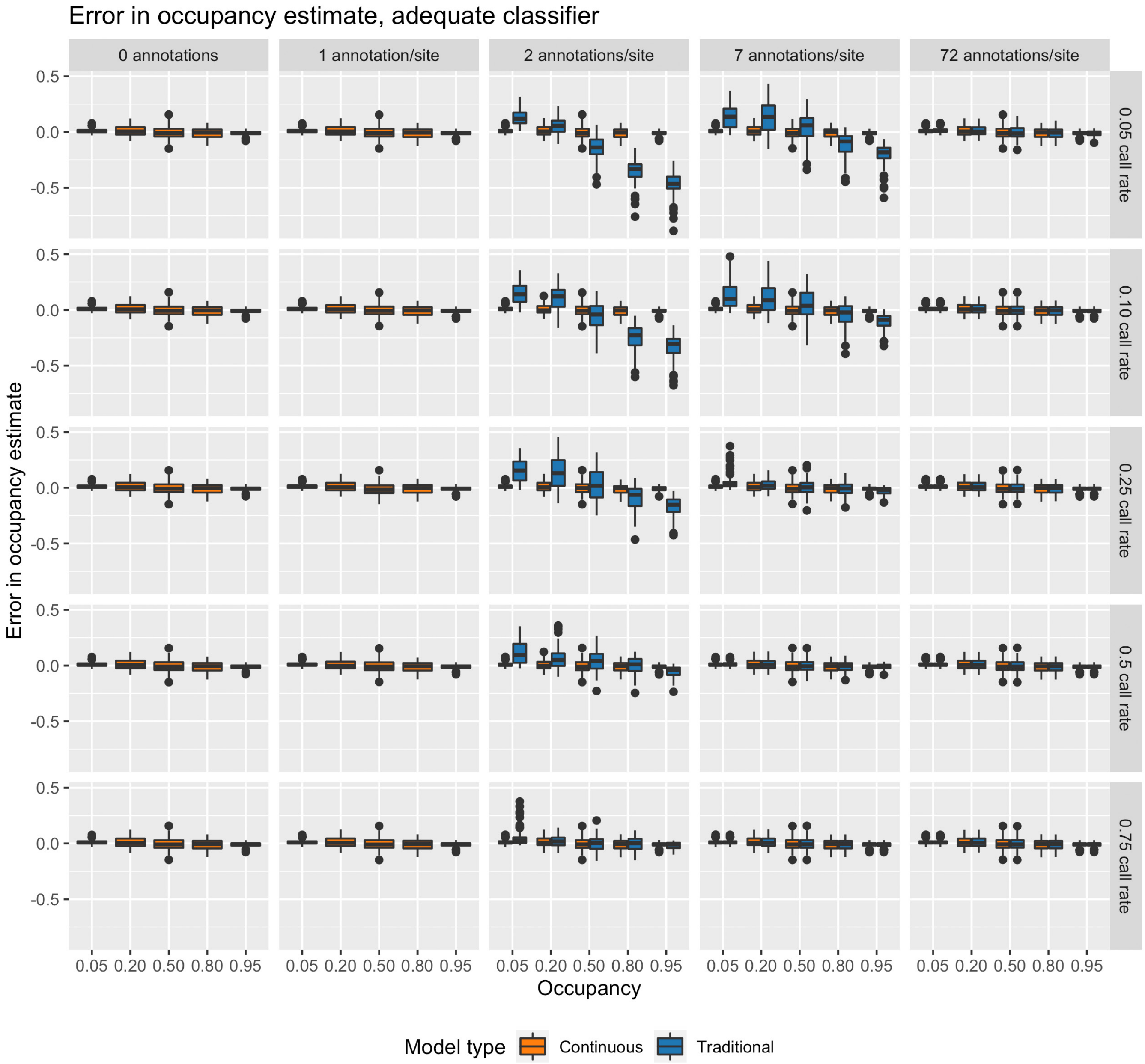

FIGURE 2 A simulation study finds that the continuous-score model's occupancy estimates are unbiased and are more precise than the estimates of a traditional occupancy model. The continuous-score model can also be applied to data with fewer than 2 annotations per site, unlike a traditional occupancy model. Simulations were performed over 5 annotation levels, ranging from no annotations to 72 annotations per site (10% of all data annotated), 5 call rate levels, 5 occupancy levels and 3 classifier quality levels. For both the continuous-score model and the traditional model, this figure shows the difference between each model's occupancy estimate and the true occupancy in the 'adequate classifier' simulations. Simulations in which the continuous-score model failed to distinguish the components of the Gaussian mixture are removed

this annotation level, the traditional occupancy model was unable to match the performance of the continuous-score model in any scenario in terms of bias and precision. Across all scenarios, the mean RMSE of the traditional model's estimates was 0.1 and the median RMSE was 0.05. In particular, the traditional occupancy model was biased at lower appearance rate θ and lower annotation levels, a phenomenon noted in other applications of occupancy modelling (MacKenzie et al., 2002, 2003; Mackenzie & Royle, 2005). The direction and amount of the bias and the precision of the estimate was influenced by the number of annotations per site, the occupancy level and the call rate (Figure 2).

Our second simulation study demonstrated that our model provided unbiased estimates of covariates on both θ and ψ in 11 of 12 scenarios; in the 12th scenario, a combination of using a good classifier with no annotations, most of the simulations did not converge (Figures S6–S9).

3.2 | Empirical study

Figures 3 and 4 show the results of the empirical study for the five species. Detailed figures showing the two models' posterior estimates of ψ for each species are provided in Figures S10–S14. Summaries of the other parameters are provided in Figures S15–S19. Fitting these models with empirical data generated no failures of the type observed in the simulation tests. Both the continuous-score model's estimates and the traditional model's estimates generally improved as the number of annotations increased (Figure 3), with both models' estimates found within 0.02 of the sample-at-hand ('true') occupancy at the maximum number of annotated files. The continuous-score model's estimate of occupancy is closer to sample-at-hand occupancy in 52% of cases (13/25), is comparable to (within 0.001 of) the traditional model's estimate in 28% of cases (7/25) and was outperformed by the traditional model in 20% of cases (5/25). In four of the five remaining cases, the errors of the two models' occupancy estimates were very similar, differing by less than 0.03. In one case (Northern Cardinal, 8 annotations per site), the traditional occupancy model's estimate was closer to sample-at-hand true occupancy by 0.05. Unlike the traditional occupancy model, the continuous-score model was able to produce occupancy estimates with no annotations or only one annotation per site, and the continuous-score model required fewer annotations than the traditional model to reach a particular accuracy level.

The classifier-generated score distributions generally did not strikingly deviate from our assumption that they were two-component Gaussian mixtures except for the Black-capped Chickadee Classifier (Figure 4). The score distribution for files containing Black-capped Chickadees was bimodal due to the classifier assigning low scores to files containing the 'twitter' vocalization of Black-capped Chickadees, which was not present in the training dataset used to create the classifier. Importantly, this bimodal score distribution did not strongly impact the accuracy of the occupancy estimation.

At low annotation levels, the fitted two-component mixture differed from the actual score distribution for two species, Eastern Towhee and Wood Thrush, but in both cases improved to closely match the true distributions with more annotations (Figure 4). For the Eastern Towhee, this poor fit is due to a poor classifier that only weakly distinguished between Eastern Towhee positive and negative files, owing to the training data containing almost no negative files from this highly vocal species, leading to underestimations of call rate and occupancy (Figure 3; Figure S15). In contrast, the Wood Thrush model demonstrated the opposite problem, where at lower annotation levels, the model overestimated the call rate and occupancy for most annotation levels (Figure 3; Figure S15).

4 | DISCUSSION

The continuous-score occupancy model presented above provides a framework that takes as input continuous machine learning classifier scores and, optionally, a small number of human-annotated files. When used to predict occupancy probabilities, a simulation study found that even with no annotated data, the model can recover

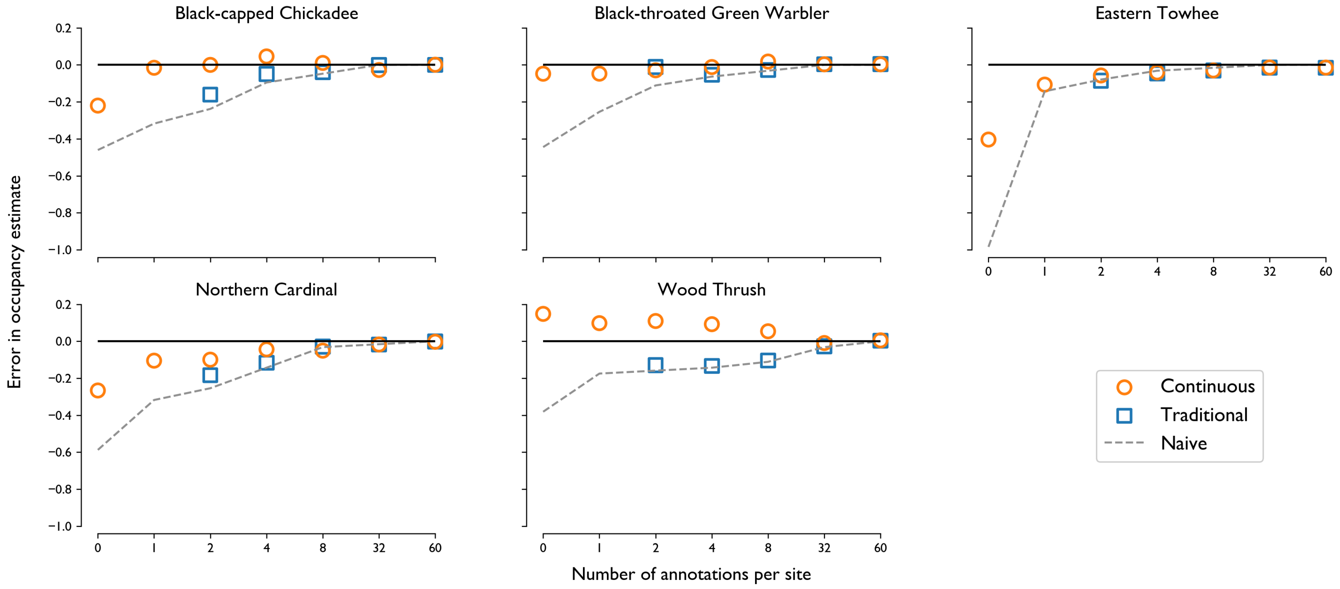

FIGURE 3 Error in estimate of occupancy, ψ, for the five species in the empirical study. Each point represents the difference between an occupancy estimate and the sample-at-hand occupancy for one annotation level for one of three occupancy estimates: The median of the posterior distribution of ψ estimated by the continuous-score model, the median of the posterior distribution of ψ estimated by the traditional model, and the 'naive' estimate, the number of sites with presence confirmed through annotation alone. The black line in the figure at $y = 0$ represents the position at which the estimate is equal to the sample-at-hand occupancy. Note that the x-axis represents the number of annotations per site and is not a linear scale

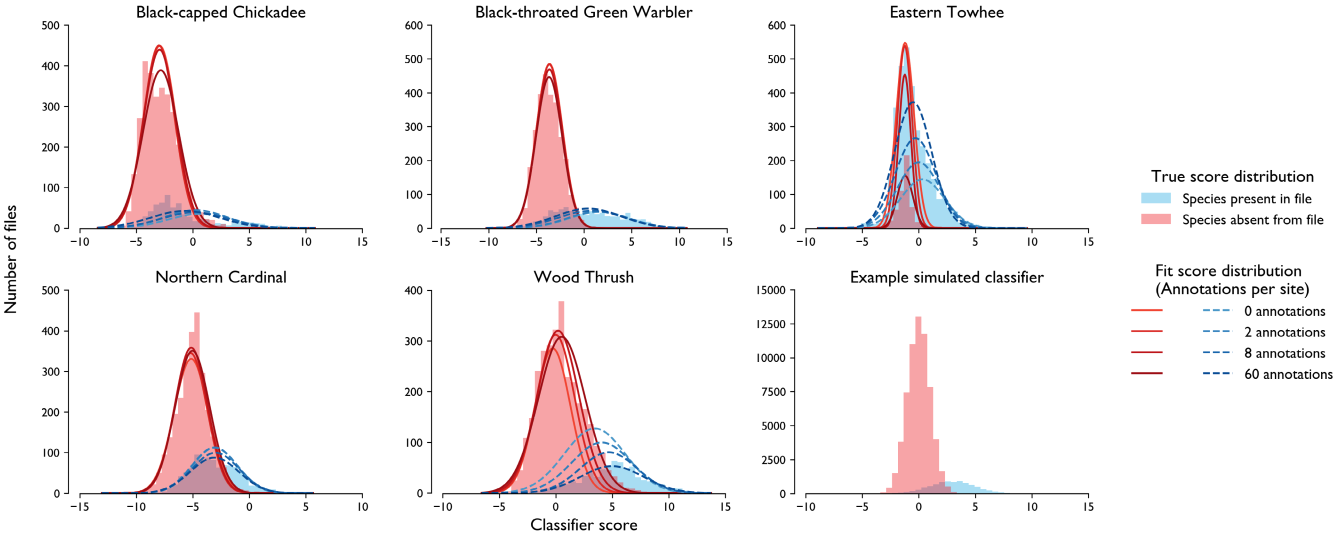

FIGURE 4 Comparison of machine learning scores of classifiers on autonomous acoustic recorder data with scores from an example simulated classifier. Overlaid on the empirical score distributions are the distributions fit by the continuous-score model at four levels of annotation, representing either 0, 2, 8 or 60 files annotated per site. The simulated scores were generated with an 'adequate' classifier where the species occupied 50% of sites and calls in 25% of the files at occupied sites

unbiased estimates of occupancy under a range of empirically realistic parameters, whereas traditional occupancy models applied to annotated data produced highly biased estimates of occupancy at the same annotation levels. An empirical study found that the model reached a level of accuracy at one annotation per site that the traditional occupancy model required four annotations per site to reach.

In our simulation study, we found that in about 5% of simulations, our model failed to produce valid results, as indicated by a failure to distinguish the parameters μ_0 and μ_1. This result was somewhat dependent on initial values. For example, using initial values of $\mu_0 = 0.5$ and $\mu_1 = 0.5$ resulted in approximately 10% of simulations failing to produce valid results, but when using initial values of $\mu_0 = 0$, $\mu_1 = 1$, only 5% of simulations exhibited this failure. No such issues were observed in the empirical study. In practice, these errors are easy to diagnose by comparing the estimates of μ_0 and μ_1, and might be addressed where they occur by estimating the means of the components through tests of classifier performance and using these means as initial values.

Our empirical study tested the continuous-score model's performance using real-world data and classifier output. Four of the five species appeared to match the model's assumptions, but we observed bimodality of the positive component of the Gaussian mixture in the scores produced by the Black-capped Chickadee classifier (Figure 4). Differences in classifier score distributions could be addressed by replacing the Gaussian mixture of the model with a more flexible distribution. Even when the score distributions do not violate the assumptions, however, fitting can still be poor if the machine learning model does not strongly differentiate positive and negative files, as observed in the case of the Eastern Towhee.

Our simulation experiment and empirical case study have a notable difference in scale from field ARU studies in practice. First, modern ARUs may generate hundreds of thousands of files over hundreds of sites, whereas our studies were limited to 720 files per site and 60 files per site for the simulation and empirical study respectively. Scaling up the studies above to larger datasets would require more computational resources to fit models but is also likely to lead to more precise estimates of model parameters. We also note that in larger field studies, the ability of the continuous-score model to decrease the required number of annotated files per site, as compared to a traditional occupancy model, will become more valuable.

Aside from the continuous-score model developed above, there are two broad alternative approaches that could be used to estimate occupancy using autonomously collected data. First, human annotators can review a sample of available files, ignoring the remainder of unreviewed files, and use traditional occupancy models to analyse this sample. This approach is straightforward to implement, is analogous to analysis methods used for human survey data and is still commonly used in the literature. As demonstrated above, our continuous-score model substantially outperforms such an approach.

Second, continuous-score data from a machine learning classifier can be converted to binary 0/1 data through a choice of a threshold. As the resulting binary data are nearly certain to contain false positives and false negatives, these binary data should be used with an occupancy model that accounts for these false positives and false negatives (e.g. Chambert, Miller, & Nichols, 2015; MacKenzie et al., 2018; Miller et al., 2011, 2013). Many of these 'false-positive occupancy models' require a second type of observation that is not subject to false positives, unlike our continuous-score model above which benefits from such additional verification but does not require it.

While we did not explicitly compare our continuous-score model to this approach of thresholding data for use with a false-positive occupancy model, such a comparison is an important future direction of research. Several structural designs for false-positive occupancy models (e.g. Chambert et al., 2018; Miller et al., 2011, Balantic & Donovan, 2019) could be used for this evaluation. We note that when making such comparisons, false-positive and false-negative rates will be determined by the choice of threshold, and thus the accuracy and precision of parameter estimates in such models is likely to vary with the choice of threshold as well. As there are likely to be systematic differences between the dataset used to train/test the machine learning classifier and the dataset used for model fitting (Knight & Bayne, 2018), we stress that false-positive and false-negative rates will need to be estimated within the model itself, rather than assumed based on performance on training or validation data.

We anticipate several particularly useful applications of our model. First, the model may be applicable to studies over large spatiotemporal scales where gathering annotated data is time-consuming. Examples include estimation of occupancy over many sites or assessment of changes in occupancy over the course of a season. This may also apply to studies of multiple species in which annotating all species is time-consuming, requires additional expertise or requires annotation during multiple survey periods (e.g. diurnal vs. nocturnal species). Second, the model may potentially be applicable to studies where call rate or occupancy are thought to be low. Traditional occupancy models exhibit a strong bias in these scenarios that is only mediated with high levels of annotation. Third, the model may be useful in studies where call rate is a parameter of interest. An additional useful extension to occupancy models that utilize machine learning classifier-produced scores would be using classifier scores to direct listening effort toward higher-scoring files that are more likely to contain the species of interest.

AUTHORS' CONTRIBUTIONS

T.A.R. and J.K. conceived the ideas and designed the methodology; T.A.R. and D.T. implemented the simulation code and continuous-score model; T.A.R. created the machine learning model and ran the simulation experiment and empirical study; T.A.R. and J.K. led the final writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

ACKNOWLEDGEMENTS

We thank Samuel Lapp for contributions to identifying the underlying distributions of machine learning classifier outputs. We are grateful for the collaboration of the Powdermill Avian Research Center in producing the data used for the empirical study. We are appreciative of Torbjørn Ergon and two anonymous reviewers for their comments which greatly improved the manuscript. This work was financially supported by the University of Pittsburgh Department of Biological Sciences and the National Fish and Wildlife Foundation Central Appalachia Habitat Stewardship Program. This research was supported in part by the University of Pittsburgh Center for Research Computing through the resources provided.

CONFLICT OF INTEREST

The authors declare that there are no competing interests.

DATA AVAILABILITY STATEMENT

Implementations of continuous-score and traditional models, scripts for data simulation, fitting models to simulated data, and classifier training, and the trained machine learning models are available at Dryad Digital Repository https://doi.org/10.5061/dryad.ns1rn8ptd (Rhinehart et al., 2022). Data used to train machine learning models are available from Dryad Digital Repository https://doi.org/10.5061/dryad.d2547d81z (Chronister et al., 2021).

ORCID

Tessa A. Rhinehart https://orcid.org/0000-0002-4352-3464
Daniel Turek https://orcid.org/0000-0002-1453-1908

REFERENCES

Abrahams, C., & Geary, M. (2020). Combining bioacoustics and occupancy modelling for improved monitoring of rare breeding bird populations. *Ecological Indicators*, *112*, 106131.

Balantic, C., & Donovan, T. (2019). Dynamic wildlife occupancy models using automated acoustic monitoring data. *Ecological Applications*, *29*, e01854.

Beery, S., Morris, D. & Yang, S. (2019). Efficient pipeline for camera trap image review. ArXiv:1907.06772 [Cs]. Retrieved from http://arxiv.org/abs/1907.06772

Bobay, L. R., Taillie, P. J., & Moorman, C. E. (2018). Use of autonomous recording units increased detection of a secretive marsh bird. *Journal of Field Ornithology*, *89*, 384–392.

Brack, I. V., Kindel, A., & Oliveira, L. F. B. (2018). Detection errors in wildlife abundance estimates from Unmanned Aerial Systems (UAS) surveys: Synthesis, solutions, and challenges. *Methods in Ecology and Evolution*, *9*(8), 1864–1873. https://doi.org/10.1111/2041-210X.13026

Brook, L. A., Johnson, C. N., & Ritchie, E. G. (2012). Effects of predator control on behavior of an apex predator and indirect consequences for mesopredator suppression. *Journal of Applied Ecology*, *49*(6), 1278–1286.

Buxton, R. T., & Jones, I. L. (2012). Measuring nocturnal seabird activity and status using acoustic recording devices: Applications for Island restoration: Acoustic monitoring of nocturnal seabirds. *Journal of Field Ornithology*, *83*, 47–60.

Cavada, N., Worsøe Havmøller, R., Scharff, N., & Rovero, F. (2019). A landscape-scale assessment of tropical mammals reveals the effects of habitat and anthropogenic disturbance on community occupancy. *PLoS ONE*, *14*, e0215682.

Chambert, T., Kendall, W. L., Hines, J. E., Nichols, J. D., Pedrini, P., Waddle, J. H., Tavecchia, G., Walls, S. C., & Tenan, S. (2015). Testing hypotheses on distribution shifts and changes in phenology of imperfectly detectable species. *Methods in Ecology and Evolution*, *6*, 638–647.

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.

Chambert, T., Waddle, J. H., Miller, D. A. W., Walls, S. C., & Nichols, J. D. (2018). A new framework for analysing automated acoustic species detection data: Occupancy estimation and optimization of recordings post-processing. *Methods in Ecology and Evolution*, *9*, 560–570.

Chandler, R. B., Muths, E., Sigafus, B. H., Schwalbe, C. R., Jarchow, C. J., & Hossack, B. R. (2015). Spatial occupancy models for predicting metapopulation dynamics and viability following reintroduction. *Journal of Applied Ecology*, *52*, 1325–1333.

Chronister, L. M., Rhinehart, T. A., Place, A., & Kitzes, J. (2021). An annotated set of audio recordings of eastern north American birds containing frequency, time, and species information. *Ecology*, *102*, e03329.

Darras, K., Batáry, P., Furnas, B. J., Grass, I., Mulyani, Y. A., & Tscharntke, T. (2019). Autonomous sound recording outperforms human observation for sampling birds: A systematic map and user guide. *Ecological Applications*, *29*, e01954.

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., & Bodik, R. (2017). Programming with models: Writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, *26*, 403–413.

Dénes, F. V., Silveira, L. F., & Beissinger, S. R. (2015). Estimating abundance of unmarked animal populations: Accounting for imperfect detection and other sources of zero inflation. *Methods in Ecology and Evolution*, *6*, 543–556.

Field, S. A., Tyre, A. J., Thorn, K. H., O'Connor, P. J., & Possingham, H. P. (2005). Improving the efficiency of wildlife monitoring by estimating detectability: A case study of foxes (*Vulpes vulpes*) on the Eyre peninsula, South Australia. *Wildlife Research*, *32*, 253.

Gray, P. C., Fleishman, A. B., Klein, D. J., McKown, M. W., Bézy, V. S., Lohmann, K. J., & Johnston, D. W. (2019). A convolutional neural network for detecting sea turtles in drone imagery. *Methods in Ecology and Evolution*, *10*, 345–355.

Gu, W., & Swihart, R. K. (2004). Absent or undetected? Effects of non-detection of species occurrence on wildlife–habitat models. *Biological Conservation*, *116*, 195–203.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *arXiv:1706.04599* [cs].

Hanski, I. (1994). Patch-occupancy dynamics in fragmented landscapes. *Trends in Ecology & Evolution*, *9*, 131–135.

Iknayan, K. J., Tingley, M. W., Furnas, B. J., & Beissinger, S. R. (2014). Detecting diversity: Emerging methods to estimate species diversity. *Trends in Ecology & Evolution*, *29*, 97–106.

Johnson, M. D. (2007). Measuring habitat quality: A review. *The Condor*, *109*, 489–504.

Kahl, S., Stoter, F.-R., Goeau, H., Glotin, H., Vellinga, W.-P., & Joly, A. (2019). Overview of BirdCLEF 2019: Large-scale bird recognition in soundscapes. 9.

Kéry, M., & Royle, J. A. (2016). *Applied hierarchical modeling in ecology: Analysis of distribution, abundance and species richness in R and BUGS*. Elsevier.

Kéry, M., & Royle, J. A. (2021). *Applied hierarchical modeling in ecology: Analysis of distribution, abundance and species richness in R and BUGS*. Elsevier.

Kitzes, J., Rhinehart, T., Moore, B., & Lapp, S. (2020). OpenSoundscape.org. Retrieved from https://opensoundscape.org

Knight, E. C., & Bayne, E. M. (2018). Classification threshold and training data affect the quality and utility of focal species data processed with automated audio-recognition software. *Bioacoustics*, *28*, 539–554.

Lapp, S., Wu, T., Richards-Zawacki, C., Voyles, J., Rodriguez, K. M., Shamon, H., & Kitzes, J. (2021). Automated detection of frog calls and choruses by pulse repetition rate. *Conservation Biology*, *0*, 10.

MacKenzie, D. I. (2005). What are the issues with presence-absence data for wildlife managers? *Journal of Wildlife Management*, *69*, 849–860.

MacKenzie, D. I. (2006). *Occupancy estimation and modeling: Inferring patterns and dynamics of species*. Elsevier.

MacKenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L. L., & Hines, J. E. (2006). Extensions to basic approaches. In *Occupancy estimation and modeling: Inferring patterns and dynamics of species* (pp. 255–274). Elsevier.

MacKenzie, D. I., Nichols, J. D., Hines, J. E., Knutson, M. G., & Franklin, A. B. (2003). Estimating site occupancy, colonization, and local extinction rates when a species is detected imperfectly. *Ecology*, *84*, 2200–2207.

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Andrew Royle, J., & Langtimm, C. A. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, *83*, 2248–2255.

MacKenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L. L., & Hines, J. E. (2018). *Occupancy estimation and modeling*. Elsevier.

Mackenzie, D. I., & Royle, J. A. (2005). Designing occupancy studies: General advice and allocating survey effort. *Journal of Applied Ecology*, *42*, 1105–1114.

Martin, T. G., Wintle, B. A., Rhodes, J. R., Kuhnert, P. M., Field, S. A., Low-Choy, S. J., Tyre, A. J., & Possingham, H. P. (2005). Zero tolerance ecology: Improving ecological inference by modelling the source of zero observations: Modelling excess zeros in ecology. *Ecology Letters*, *8*, 1235–1246.

Miller, D. A., Nichols, J. D., McClintock, B. T., Grant, E. H. C., Bailey, L. L., & Weir, L. A. (2011). Improving occupancy estimation when two types of observational error occur: Non-detection and species misidentification. *Ecology*, *92*, 1422–1428.

Miller, D. A. W., Nichols, J. D., Gude, J. A., Rich, L. N., Podruzny, K. M., Hines, J. E., & Mitchell, M. S. (2013). Determining occurrence dynamics when false positives occur: Estimating the range dynamics of wolves from public survey data. *PLoS ONE*, *8*, e65808.

Newson, S. E., Bas, Y., Murray, A., & Gillings, S. (2017). Potential for coupling the monitoring of bush-crickets with established large-scale acoustic monitoring of bats. *Methods in Ecology and Evolution*, *8*, 1051–1062.

NIMBLE Development Team. (2021). NIMBLE: MCMC, particle filtering, and programmable hierarchical modeling. *Zenodo*.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News*, *6*(1), 7–11.

Priyadarshani, N., Marsland, S., & Castro, I. (2018). Automated birdsong recognition in complex acoustic environments: A review. *Journal of Avian Biology*, *49*, jav-01447.

Rhinehart, T., Turek, D., & Kitzes, J. (2022). Supplementary information for: A continuous-score occupancy modeling framework for incorporating uncertain machine learning output in autonomous biodiversity surveys. *Dryad Digital Repository*, https://doi.org/10.5061/dryad.ns1rn8ptd

Robert, C., & Casella, G. (2010). *Introducing Monte Carlo methods with R*. Springer New York.

Royle, J. A., & Link, W. A. (2006). Generalized site occupancy models allowing for false positive and false negative errors. *Ecology*, *87*, 835–841.

Shonfield, J., & Bayne, E. M. (2017). Autonomous recording units in avian ecological research: Current use and future applications. *Avian Conservation and Ecology*, *12*(1), 14.

Tabak, M. A., Norouzzadeh, M. S., Wolfson, D. W., Sweeney, S. J., Vercauteren, K. C., Snow, N. P., Halseth, J. M., Salvo, P. A. D., Lewis, J. S., White, M. D., Teton, B., Beasley, J. C., Schlichting, P. E., Boughton, R. K., Wight, B., Newkirk, E. S., Ivan, J. S., Odell, E. A., Brook, R. K., … Miller, R. S. (2019). Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods in Ecology and Evolution*, *10*, 585–590.

Tyre, A. J., Tenhumberg, B., Field, S. A., Niejalke, D., Parris, K., & Possingham, H. P. (2003). Improving precision and reducing bias in biological surveys: Estimating false-negative error rates. *Ecological Applications*, *13*, 1790–1801.

Wang, Y., Allen, M. L., & Wilmers, C. C. (2015). Mesopredator spatial and temporal responses to large predators and human development in the Santa Cruz Mountains of California. *Biological Conservation*, *190*, 23–33.

Wrege, P. H., Rowland, E. D., Keen, S., & Shiu, Y. (2017). Acoustic monitoring for conservation in tropical forests: Examples from forest elephants. *Methods in Ecology and Evolution*, *8*, 1292–1301.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Rhinehart, T. A., Turek, D., & Kitzes, J. (2022). A continuous-score occupancy model that incorporates uncertain machine learning output from autonomous biodiversity surveys. *Methods in Ecology and Evolution*, *13*, 1778–1789. https://doi.org/10.1111/2041-210X.13905