Leveraging an observed-data likelihood improves the use of machine learning labels in a Bayesian hierarchical model for bioacoustic data

Jacob K. Oram; Katharine M. Banner; Christian Stratton; Andrew Hoegh; Kathryn M. Irvine

The Annals of Applied Statistics
2025, Vol. 19, No. 4, 2957–2980
https://doi.org/10.1214/25-AOAS2096

LEVERAGING AN OBSERVED-DATA LIKELIHOOD IMPROVES THE USE OF MACHINE LEARNING LABELS IN A BAYESIAN HIERARCHICAL MODEL FOR BIOACOUSTIC DATA

BY JACOB K. ORAM[1,a], KATHARINE M. BANNER[1,b], CHRISTIAN STRATTON[2,d], ANDREW HOEGH[1,c] AND KATHRYN M. IRVINE[3,e]

[1]*Department of Mathematical Sciences, Montana State University,* [a]*jacoboram@montana.edu,* [b]*katharine.banner@montana.edu,* [c]*andrew.hoegh@montana.edu*

[2]*Department of Mathematics and Statistics, Middlebury College,* [d]*cstratton@middlebury.edu*

[3]*U.S. Geological Survey Northern Rocky Science Center,* [e]*kirvine@usgs.gov*

Classification of massive datasets by machine learning (ML) algorithms is promising for many scientific domains, especially wildlife monitoring programs that rely on passive acoustic surveys for detecting species. However, treating ML-predicted class labels (e.g., species identity) as truth biases inferences of focal parameters within common modeling frameworks. One solution is to model the misclassification process explicitly using human-validated true-class labels for a subset of observations. Validation by experts can present a substantial bottleneck in otherwise efficient workflows that use ML predictions. Bioacoustics practitioners seek guidance on both the quantity and process for selecting ML-labeled data to validate by an expert. We derive an alternative model formulation that jointly models human-validated and ML-predicted class labels with an observed-data likelihood (ODL) and use empirically informed simulations motivated by a real-data application to explore different probability designs for selecting class labels for validation. Simulation results suggest that with smaller validation sets the ODL formulation increases computational speed and reduces estimation error compared to a default MCMC data augmentation routine. Our methodology is transferable to applications that treat predictions from classification algorithms as the response variable of interest.

1. Introduction. Machine learning (ML) algorithms, and their ability to quickly process large quantities of data, hold immense promise for scientific domains that rely on automated data streams to efficiently gather information (Carleo et al. (2019)). Often, the use of machine learning for classification is not only advantageous but entirely necessary for data processing (Deiana et al. (2022)). Widespread adoption of these powerful computing tools has occurred in fields as diverse as water quality evaluation (Zhu et al. (2022)) and proteomics (Bludau et al. (2022)). We focus on wildlife monitoring programs, such as eBird (Sullivan et al. (2009)) and the North American Bat Monitoring Program (NABat; Loeb et al. (2015)), the latter of which motivates our work. The ultimate goal of these programs is to inform conservation management decisions by assessing population status and trends for multiple species simultaneously at varying spatial extents (continental, country-level, regional, state/provincial, local).

For programs that monitor taxa that are rare or cryptic, such as bats, certain species of birds, anurans and insects (Loeb et al. (2015), Brauer et al. (2016), Jeliazkov et al. (2016), Zhong et al. (2021)), "passive" acoustic methods that use autonomous recording units (ARUs) enable researchers to collect data from multispecies assemblages (i.e., a group of species that exist within the same geographic region; Stroud et al. (2015)) with lower deployment costs

Received July 2024; revised September 2025.
Key words and phrases. Missing data, machine learning, Bayesian hierarchical model, population monitoring.

than "active" survey methods requiring human observers (Walters et al. (2012), Loeb et al. (2015), Gibb et al. (2019)). When paired with a probabilistic sampling design and appropriate statistical model, ARU data from multiple species provide researchers with a rich source of information to estimate ecologically relevant parameters with uncertainty for all species (e.g., Chambert et al. (2018), Stiffler, Anderson and Katzner (2018), Wright et al. (2020)).

The vast quantity of observations collected during spatially extensive annual surveys necessitates an automated process for assigning species labels to observations (i.e., proprietary classification software or open source ML algorithms) before fitting statistical models. While classification algorithms for ecological data have improved with increased engagement from the ML community (Gibb et al. (2019), Khalighifar et al. (2022)), the single-species labels assigned to ARU observations (hereafter, autoIDs) are still prone to error with cross-species misclassification resulting in false-positive and associated false-negative detections (Wright et al. (2020), Koslovsky et al. (2024)). For large-scale monitoring programs, the concern is that failure to account for misclassification errors in autoIDs may lead to biased statistical estimates, and consequently, set off adverse downstream effects (e.g., incorrect management decisions, "take" of a vulnerable species, or failure to adequately address serious conservation issues; Russo and Voigt (2016), Rydell et al. (2017), Irvine et al. (2022)). As such, a desirable statistical model must account for both false-negative and false-positive detections due to the misclassification process when estimating parameters of interest.

Wright et al. (2020) proposed the *count-detection* model, a statistical framework that accounts for imperfect detection and explicitly models the misclassification process inherent to ARU datasets. The count-detection model uses a confusion matrix to describe the probability of one species being misidentified as another, making the explicit assumption that a false-positive detection for one species is due to the misclassification of another (Wright et al. (2020)), as opposed to the presence of ambient noise in the environment (Chambert et al. (2018)). This assumption makes it the most appropriate framework for modeling species assemblages with bat-acoustic data to date (Irvine et al. (2022), Stratton et al. (2022)). The focal parameters of the count-detection model are occurrence probabilities and relative activity rates for all species in a species assemblage. Relative activity is defined as the expected number of detections per sampling occasion for a species, given its presence and the other species in the assemblage; it is often of interest to monitoring programs because it can serve as an early-stage indicator of species declines (Stratton et al. (2025)). Recent extensions of this modeling framework include a multiyear observation-level formulation with morphospecies designations (Spiers et al. (2022)) and a scalable zero-inflated Dirichlet-multinomial model that incorporates a misclassification indicator and covariates at the observation level (Koslovsky et al. (2024)).

A fundamental challenge of modeling false-negative and false-positive detections resulting from the misclassification process is that parameters are not identifiable from the sample data alone; additional information is required to inform classification parameters (Royle and Link (2006), Chambert et al. (2018), Wright et al. (2020)). One potential solution is to estimate misclassification parameters using a separate auxiliary dataset, which Ruiz-Gutierrez, Hooten and Campbell Grant (2016) refer to as a "calibration model" in their taxonomy of false-positive occupancy models. However, calibration models require the assumption that the auxiliary dataset is representative of the sample in hand, which can be difficult to assess in practice (Stratton et al. (2022)). Another solution is to construct informative priors for the misclassification rates when such information is available (e.g., Spiers et al. (2022), Stratton et al. (2022)). Alternatively, one may use *coupled classification*, wherein a subset of the machine-predicted autoIDs in the observed data are validated by human experts, yielding a pair of true and ML-predicted species labels. Coupled classification can be viewed as a special case of supervised learning (Spiers et al. (2022)). When paired with the count-detection

model framework, coupled classification (or "observation model confirmation" in the taxonomy of Ruiz-Gutierrez, Hooten and Campbell Grant (2016)) can yield narrower posterior intervals than auxiliary data (Stratton et al. (2022)), but also requires human experts to manually verify a probabilistic sample of observed recordings. The requirement that recordings be manually reviewed for each annual survey presents a hidden, and potentially uncontrollable cost in an otherwise efficient data processing stream. As such, a key question facing wildlife monitoring programs is how to tailor the autoID selection process, or *validation design*, to reduce the total number of autoIDs requiring manual validation while simultaneously ensuring statistical inference for focal parameters is accurate and precise enough to inform management decisions.

We use the NABat acoustic processing workflow and empirical data from Oregon and Washington, USA to motivate and contextualize our work. We consider three validation designs either used or suggested for use as cost-efficient ways to implement coupled classification within the count-detection model. We highlight practical challenges associated with each design to motivate the development of an alternative model formulation that may improve estimation with reduced validation effort. In the first validation design, the Northwestern Bat Hub, a regionally organized partnership within the NABat program, tried to reduce validation effort by randomly selecting 20% of autoIDs from the first sampling occasion at each location (for bats a "site-night"). As a second example, Stratton et al. (2022) used simulation to explore the implications of a design that assumed all recordings (100%) from a subset of site-nights were validated. Although this validation design has the advantage of yielding a computationally efficient count-detection model formulation, Stratton et al. (2022) noted that the number of recordings per species available for estimating the classification parameters was stochastic. As a result, only prohibitively high levels of validation effort could guarantee stable estimation of model parameters, limiting the effectiveness of exhaustive validation within sampling occasions. Based on these findings, Stratton et al. (2022) proposed using the third validation design we consider: a stratified random sample from the entire collection of single-species ML-labeled recordings with strata defined by autoID label. We refer to this design as a stratified-by-species design. This design could potentially allow monitoring programs to control the costs of validation while also ensuring Markov Chain Monte Carlo (MCMC) convergence by prioritizing validation of particular autoID labels.

We recast the count-detection model as a missing-data problem and show that the stratified-by-species design can lead to true labels that are missing at random, making the validation design ignorable (Gelman et al. (2013), p. 202). The observation- or recording-level specification of the count-detection model proposed by Spiers et al. (2022) used a complete-data likelihood (CDL) approach by imputing the missing true-species labels within the MCMC algorithm (data augmentation; Link and Barker (2010), p. 165). We obtain a partially-integrated CDL model specification by integrating over the missing true-class labels; for brevity and because we model only the observed true-species labels and autoIDs, we refer to our specification as the observed-data likelihood (ODL). We use empirically informed simulation studies to show that the ODL converges more quickly than the CDL and results in unbiased estimation with nominal coverage for all the parameters in the count-detection model across a range of data-generating values and reduced-effort validation designs. The ODL model provides a promising solution to the costly bottleneck in many bioacoustic data streams—including NABat's—that has prevented the widespread adoption of the count-detection model for status and trend monitoring. We introduce the modeling framework and derive the ODL in the following section.

2. Statistical methods. We adopt the observation-level (i.e., recording or call file level) count-detection model specification originally proposed by Spiers et al. (2022).

2.1. *Model specification.* Let $i = 1, 2, \ldots, n$ index sites and $k = 1, 2, \ldots K$ index species. The model assumes a latent occurrence state of species k at site i, denoted Z_{ik}, that takes on value 1 when species k is present at site i and zero otherwise. Assuming that occurrence is independent across sites i and species k, we model each Z_{ik} as a Bernoulli random variable with occurrence probability ψ_{ik}:

$$Z_{ik} \sim \text{Bernoulli}(\psi_{ik}). \tag{1}$$

In this work, we assume that ψ_k is constant for all sites i. However, one could extend the model via a logit or probit regression to accommodate covariates at the site level. For ease of notation, let $\mathbf{Z}_i = (Z_{i1}, Z_{i2}, \ldots, Z_{iK})'$ denote the vector of species occurrence indicators at location i.

Let $j = 1, 2, \ldots, J$ index a detector night (or visit). At locations where species k is present (i.e., $Z_{ik} = 1$), we model the number of recordings from that species on visit j to site i as a Poisson random variable with rate λ_{ijk}, representing the relative activity of species k on visit j to site i:

$$Y_{ijk} | Z_{ik} = 1 \sim \text{Poisson}(\lambda_{ijk}).$$

We assume independence of Y_{ijk} across species. Furthermore, we assume that $\lambda_{ijk} = \lambda_k$ for all sites i and visits j so that each species has a single relative activity rate that does not vary across sites or visits. Although this assumption simplifies our derivations, extending this model to accommodate covariates in a generalized linear model framework would be straightforward with an appropriate link function.

The assumption of independence across species—especially at the level of individual call files—is difficult to assess and is common in the multispecies occupancy literature. However, some authors (e.g., Rota et al. (2016)) have developed occupancy models that allow estimation of dependency among species' presence, but this has not been extended to models for both occupancy and relative activity, as in our framework. In the context of our model, the consequence of assuming independence is that we are potentially not capturing associations in counts due to the attractive or repulsive influence of one species on another.

In the presence of an imperfect classification algorithm, we do not directly observe Y_{ijk}. However, we do observe the total number of calls from the site-night, which is denoted as $Y_{ij\cdot} = \sum_{k=1}^{K} Y_{ijk}$. We assume independence across visits j and model $Y_{ij\cdot}$ as a mixture distribution with components defined by the species occurrence indicators $\mathbf{Z}_i$:

$$Y_{ij\cdot} | \mathbf{Z}_i \sim \begin{cases} 0 \text{ with probability } \prod_k (1 - \psi_{ik}) \text{ when } \mathbf{Z}_i = \mathbf{0}, \\ \text{Poisson}(\mathbf{Z}_i' \boldsymbol{\lambda}) \text{ with probability } \left(1 - \prod_k (1 - \psi_{ik})\right) \text{ otherwise.} \end{cases} \tag{2}$$

In equation (2), $\boldsymbol{\lambda}$ denotes a $K \times 1$ vector where the kth entry contains the relative activity parameter for species k.

At locations where observations were made (i.e., at least one species was present and recorded so that $Y_{ij\cdot} > 0$), let $l_{ij} = 1, 2, \ldots, Y_{ij\cdot}$ index an individual observation within a site-night (i, j). Further, let $\mathbf{T}_{l_{ij}}$ and $\mathbf{A}_{l_{ij}}$ denote the $K \times 1$ vectors indicating the true and autoID species labels, respectively. We sometimes write $T_{l_{ij}} = k$ to denote $\mathbf{T}_{l_{ij}} = (T_{l_{ij}1}, T_{l_{ij}2}, \ldots, T_{l_{ij}k}, \ldots, T_{l_{ij}K})' = (0, 0, \ldots, 0, 1, 0, \ldots, 0)'$, where the only nonzero entry is the kth element. After manually validating a recording, the true species label $\mathbf{T}_{l_{ij}}$ is known. We assume that when selected, recordings can be validated and that the resulting true-species labels are always correct. We refer to the subset of validated recordings with known true labels as the validation set. In practice, the assumption that human-annotated labels are always correct may be violated, which will bias inference for target parameters due to the lingering

presence of false-positive and associated false-negative detections. For this reason, validation is carried out by expert biologists who have undergone extensive training following standardized protocols that are intended to reduce errors by human observers (Loeb et al. (2015), Reichert et al. (2018)). A similar standard would be required to make this a reasonable assumption for other types of bioacoustic data.

Due to properties of independent Poisson random variables, and conditional on the observed total count of recordings $Y_{ij.} = y_{ij.} > 0$, the number of recordings due to true species k is a Multinomial random variable with probability vector $\mathbf{p}_i$ (defined below). Assuming independence among recordings, we model the true-species label of an individual recording, $\mathbf{T}_{l_{ij}}$ as a Multinomial random variable with the same probability vector $\mathbf{p}_i$:

$$\begin{aligned} \mathbf{T}_{l_{ij}}|\mathbf{Z}_i, \boldsymbol{\lambda} &\sim \text{Multinomial}(1, \mathbf{p}_i), l_{ij} = 1, 2, \ldots, y_{ij.} \quad \text{where} \\ \mathbf{p}_i &= (p_{i1}, p_{i2}, \ldots p_{iK})', \quad \text{and} \\ p_{ik} &= \frac{Z_{ik}\lambda_k}{\sum_{m=1}^{K} Z_{im}\lambda_m}. \end{aligned} \tag{3}$$

The kth entry in $\mathbf{p}$, p_{ik}, corresponds to the probability that a recording observed at location i is attributable to species k. The straightforward interpretation of each p_{ik} as the proportion of expected activity due to the presence of species k at site i results from our assumption that the total counts of recordings on a detector-night follow a Poisson distribution. To model overdispersed data, one could use the modeling framework of Koslovsky et al. (2024), which shares many similar features to the framework here and has shown benefits when overdispersion is present.

Conditional on the true species label, we model the classification process through a classification matrix Θ, where entry $\{k, k^\star\} = \theta_{kk^\star}$ denotes the probability that a recording from true species k is assigned autoID label $k^\star$. Note that diagonal entries in this classification matrix correspond to the probability of correct classification and the rows of Θ sum to 1: $\sum_m \theta_{km} = 1$. Given the true species label for recording l_{ij}, the autoID $\mathbf{A}_{l_{ij}}$ is modeled as a multinomial random variable with the probability vector determined by row of Θ corresponding to the true species that generated the recording:

$$\mathbf{A}_{l_{ij}}|\mathbf{T}_{l_{ij}}, \Theta \sim \text{Multinomial}\big(1, \mathbf{T}'_{l_{ij}}\Theta\big). \tag{4}$$

As with $\mathbf{T}_{l_{ij}}$, we sometimes use $A_{l_{ij}} = m$ as a notational shorthand for $\mathbf{A}_{l_{ij}} = (A_{l_{ij}1}, \ldots, A_{l_{ij}m-1}, A_{l_{ij}m}, A_{l_{ij}m+1}, \ldots, A_{l_{ij}K}) = (0, \ldots, 0, 1, 0, \ldots, 0)'$, where the 1 is in the mth entry.

To summarize, the unnormalized posterior distribution of our hierarchical construction is given by

$$\begin{aligned} p(\boldsymbol{\psi}, \boldsymbol{\lambda}, \Theta, \mathbf{Z}|\mathbf{T}, \mathbf{A}) \propto \prod_{i=1}^{n} \Bigg[\prod_{k=1}^{K} p(Z_{ik}|\psi_k) \prod_{j=1}^{J} \Bigg[p(Y_{ij.}|\mathbf{Z}_i, \boldsymbol{\lambda}) \\ \times \prod_{l_{ij}=1}^{Y_{ij.}} p(\mathbf{A}_{l_{ij}}|\mathbf{T}_{l_{ij}}, \Theta) p(\mathbf{T}_{l_{ij}}|\mathbf{Z}_i, \boldsymbol{\lambda}) \Bigg] \Bigg] \times \big[p(\boldsymbol{\psi}) p(\boldsymbol{\lambda}) p(\Theta)\big]. \end{aligned} \tag{5}$$

In equation (5), $p(\boldsymbol{\psi})$, $p(\boldsymbol{\lambda})$ and $p(\Theta)$ are appropriate prior distributions for the parameters of interest. In our simulations and application, we follow Stratton et al. (2022) and use independent Beta(1, 1) priors for ψ_k, independent half-normal priors $N_+(0, \sigma = 100)$ for each λ_k, and Dirichlet($\boldsymbol{\alpha}$) priors on each row of Θ. We set the concentration vector $\boldsymbol{\alpha} = 1/K \times \mathbf{1}$, the reference distance prior recommended by Stratton et al. (2022). To avoid drawing missing true-species labels at each MCMC iteration, we use the observed-data likelihood, which we introduce next.

2.2. *Missing data notation, types of missingness and the observed-data likelihood.* Modeling data gathered under probability designs with even moderate complexity requires consideration of the data collection mechanism (Gelman et al. (2013), Irvine et al. (2018), Little and Rubin (2020), Zachmann et al. (2022)). In Bayesian models for survey data, the data collection mechanism is often formulated in terms of a "missing data problem." Missing data problems are typically categorized according to a three-type scheme, which we describe informally here. In the mildest case, observations are missing completely at random (MCAR) and missingness is independent of all observed and unobserved variables. In the second case, observations are missing at random (MAR; Rubin (1976)), meaning the missingness depends only on the observed components of the data that are available to the analyst. Finally, if observations are missing not at random (MNAR), the missingness of an observation is influenced by the values of unknown data. For a full treatment of missing data assumptions, see Little (2021). We consider only validation designs where observations are selected based on a characteristic that is known for all observations (e.g., the autoID labels).

For a stratified-by-species validation design, recordings are selected probabilistically using fully observed autoIDs, and observations are missing at random, with the probability of missingness determined by the autoID and the priorities of the monitoring program. To illustrate how one would construct the ODL model, we follow Gelman et al. (2013) (Chapter 8, pp. 199–202) and introduce missing-data notation.

Let $\mathbf{X} = (\mathbf{X}_{11}, \mathbf{X}_{12}, \mathbf{X}_{21}, \ldots, \mathbf{X}_{1J}, \ldots, \mathbf{X}_{n(J-1)}, \mathbf{X}_{nJ})'$ denote a matrix of potential data that is partitioned by site-night. For example, if 2 visits are made to each of 3 sites, $\mathbf{X}$ has the following form:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_{11} \\ \mathbf{X}_{12} \\ \mathbf{X}_{21} \\ \mathbf{X}_{22} \\ \mathbf{X}_{31} \\ \mathbf{X}_{32} \end{bmatrix}.$$

Each $\mathbf{X}_{ij}$, $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, J$ is itself a matrix of dimension $Y_{ij\cdot} \times 2$. Each row in $\mathbf{X}_{ij}$ is a 2-vector containing the true and autoID species labels for observation l_{ij}, respectively. For example, we might have a site-night data matrix that appears as

$$\mathbf{X}_{ij} = \begin{array}{c} \\ l_{ij} = 1 \\ l_{ij} = 2 \\ \vdots \\ l_{ij} = Y_{ij\cdot} \end{array} \begin{array}{c} \begin{array}{cc} T_{l_{ij}} & A_{l_{ij}} \end{array} \\ \begin{bmatrix} 1 & 1 \\ \text{NA} & 2 \\ \vdots & \vdots \\ 2 & 2 \end{bmatrix} \end{array}.$$

Now let $\mathbf{I} = (\mathbf{I}_{11}, \ldots, \mathbf{I}_{nJ})'$ denote a matrix of the dimensions as $\mathbf{X}_{ij}$, with each entry taking on value 1 if the corresponding entry in $\mathbf{X}$ is observed and zero otherwise. For instance, for the example matrix $\mathbf{X}_{ij}$ above, we would have the corresponding indicator matrix:

$$\mathbf{I}_{ij} = \begin{array}{c} \\ l_{ij} = 1 \\ l_{ij} = 2 \\ \vdots \\ l_{ij} = Y_{ij\cdot} \end{array} \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix}.$$

Note that the ML-predicted autoIDs are observed for all recordings, so that the second column of $\mathbf{I}_{ij}$ is a vector of 1s. Based on this feature of the data, partitioning the potential data

into observed and missing components (i.e., into $\mathbf{X}_{ij}^{(\text{obs})}$ and $\mathbf{X}_{ij}^{(\text{mis})}$) is determined solely by the first column of $\mathbf{I}_{ij}$. The entry in the first column of $\mathbf{I}_{ij}$ corresponding to recording l_{ij} we denote as the scalar $I_{l_{ij}}$. As an additional consequence of always observing the ML-predicted labels, the missing data $\mathbf{X}_{ij}^{(\text{mis})} = \{(T_{l_{ij}}, A_{l_{ij}}) : I_{l_{ij}} = 0\}$ is the set of observations with unknown true-species labels. With these definitions for the missing and observed components at the site-night level, we construct the missing and observed components of the data across all sites and nights as

$$\mathbf{X}^{(\text{obs})} = \begin{bmatrix} \mathbf{X}_{11}^{(\text{obs})} \\ \mathbf{X}_{12}^{(\text{obs})} \\ \vdots \\ \mathbf{X}_{nJ}^{(\text{obs})} \end{bmatrix} \quad \text{and} \quad \mathbf{X}^{(\text{mis})} = \begin{bmatrix} \mathbf{X}_{11}^{(\text{mis})} \\ \mathbf{X}_{12}^{(\text{mis})} \\ \vdots \\ \mathbf{X}_{nJ}^{(\text{mis})} \end{bmatrix}.$$

Modeling data from complex sampling designs may involve jointly modeling the responses $\mathbf{X}$ and the inclusion indicators $\mathbf{I}$. In the context of our problem, the joint distribution is factored as

$$p(\mathbf{X}, \mathbf{I}|\boldsymbol{\lambda}, \boldsymbol{\psi}, \boldsymbol{\Theta}, \boldsymbol{\phi}) = p(\mathbf{X}|\boldsymbol{\lambda}, \boldsymbol{\psi}, \boldsymbol{\Theta})\, p(\mathbf{I}|\mathbf{X}, \boldsymbol{\phi}), \tag{6}$$

where $\boldsymbol{\phi}$ is a vector of parameters indexing the missingness mechanism that will vary depending on the validation design. For instance, if a monitoring program used a stratified-by-species approach to validation, $\boldsymbol{\phi}$ would be a $K \times 1$ vector containing pre-specified probabilities of selecting a recording for validation, given its autoID label. Note that equation (6) assumes that the data $\mathbf{X}$ are conditionally independent of the missingness parameters $\boldsymbol{\phi}$ given the parameters associated with the complete data, $(\boldsymbol{\lambda}, \boldsymbol{\psi}, \boldsymbol{\Theta})$, and that $\mathbf{I}$ is conditionally independent of $(\boldsymbol{\lambda}, \boldsymbol{\psi}, \boldsymbol{\Theta})$ given $\boldsymbol{\phi}$. In other words, the parameters indexing the missingness mechanism provide no additional information about the data labels once the data-generating parameters are known, and the data-generating parameters provide no additional knowledge about the probability of missingness once we know the elements of $\boldsymbol{\phi}$. We assume that the inferential goal is to model $(\boldsymbol{\lambda}, \boldsymbol{\psi}, \boldsymbol{\Theta})$. Equation (6) is sometimes called the *complete-data likelihood* (CDL; Link and Barker (2010), p. 188).

Gelman et al. (2013) note that the CDL can be useful for initially setting up a model, but advocate for use of the *observed-data likelihood*, which marginalizes out the missing values (p. 201). In our case, the missing values are the true-species labels of observations not selected for validation:

$$p(\mathbf{X}^{(\text{obs})}, \mathbf{I}|\boldsymbol{\lambda}, \boldsymbol{\psi}, \boldsymbol{\Theta}, \boldsymbol{\phi}) = \int p(\mathbf{X}, \mathbf{I}|\boldsymbol{\lambda}, \boldsymbol{\psi}, \boldsymbol{\Theta}, \boldsymbol{\phi})\, d\mathbf{X}^{(\text{mis})}. \tag{7}$$

In the context of our study, the elements of $\boldsymbol{\phi}$ are known by design and are distinct from the data parameters $(\boldsymbol{\lambda}, \boldsymbol{\psi}, \boldsymbol{\Theta})$. As such, we can obtain posterior inference with the observed-data likelihood by specifying prior distributions on $(\boldsymbol{\lambda}, \boldsymbol{\psi}, \boldsymbol{\Theta})$. Next, we fully translate the observed data likelihood (7) into the context of the individual-level count-detection model outlined in Section 2.1.

2.3. *Deriving the ODL.* In the context of the individual-level count-detection model, a response $\mathbf{X}_{l_{ij}} = (T_{l_{ij}}, A_{l_{ij}})$, where $T_{l_{ij}}$ and $A_{l_{ij}}$ are the scalar indices corresponding to the true and autoID species labels, respectively. Applying equations (6) and (7), we factor the joint distribution of the indicators, true labels and autoIDs assuming a stratified-by-species design. Similar steps could be used for alternative validation designs. Translating equation (6) into our model notation and dropping the (i, j) subscripts, the joint probability of the indicators, true and autoID labels can be factored as $p(I, \mathbf{X}|\boldsymbol{\lambda}, \boldsymbol{\psi}, \Theta, \boldsymbol{\phi}) = p(I_l, \mathbf{A}_l, \mathbf{T}_l|\boldsymbol{\lambda}, \boldsymbol{\psi}, \Theta, \boldsymbol{\phi}) =$

$p(I_l|\mathbf{A}_l, \boldsymbol{\phi})p(\mathbf{A}_l|\mathbf{T}_l, \boldsymbol{\lambda}, \boldsymbol{\psi}, \Theta)p(\mathbf{T}_l|\boldsymbol{\lambda}, \boldsymbol{\psi}, \Theta)$. The first term in this factorization holds because inclusion in the validation set is conditionally independent of the true label and all model parameters, given the autoID. Notice that under the stratified-by-species design, $I_{l_{ij}}$ depends only on the fully-observed autoIDs, making this design MAR with respect to our model. Together with the fact that $\boldsymbol{\phi}$ and $(\boldsymbol{\lambda}, \boldsymbol{\psi}, \boldsymbol{\Theta})$ are independent *a priori*, the stratified-by-species validation design is ignorable in our Bayesian model (Rubin (1976), Little (2021)).

With this factorization, we partition the product $\prod_{l_{ij}=1}^{Y_{ij\cdot}} p(\mathbf{A}_{l_{ij}}|\mathbf{T}_{l_{ij}}, \Theta)p(\mathbf{T}_{l_{ij}}|\mathbf{Z}_i, \boldsymbol{\lambda})$ in equation (5) into two components corresponding to the likelihood contribution of the recordings with observed and unobserved true-species labels but do not formally model the missing data indicators because the validation design is ignorable:

$$\begin{aligned}(8)\qquad &\prod_{l_{ij}:I_{l_{ij}}=1} p(\mathbf{A}_{l_{ij}}|\mathbf{T}_{l_{ij}}, \Theta)p(\mathbf{T}_{l_{ij}}|\mathbf{Z}_i, \boldsymbol{\lambda}) \\ &\quad\times \prod_{l_{ij}:I_{l_{ij}}=0} p(\mathbf{A}_{l_{ij}}|\mathbf{Z}_i, \boldsymbol{\lambda}, \Theta).\end{aligned}$$

The first product in equation (8) is the same as in equation (5). The second product is the likelihood contributions of the observations with unknown true-species labels marginalized out to obtain the observed-data likelihood as follows.

Recall that in our notation $T_{l_{ij}} = k \iff \mathbf{T}_{l_{ij}} = (0, 0, \ldots, 1, \ldots, 0)'$, where the 1 is in the kth entry, with a similar convention for $A_{l_{ij}} = k^\star$. Note that for a single observation with autoID $A_{l_{ij}} = k^\star$, we have

$$\begin{aligned}(9)\qquad \Pr(A_{l_{ij}} = k^\star|\mathbf{Z}_i, \boldsymbol{\lambda}, \Theta) &= \sum_{k=1}^{K} \Pr(A_{l_{ij}} = k^\star|T_{l_{ij}} = k, \Theta)\Pr(T_{l_{ij}} = k|\mathbf{Z}_i, \boldsymbol{\lambda}) \\ &= \sum_{k=1}^{K}[\text{Multinomial}(1, \mathbf{T}'_{l_{ij}}\Theta) \times \text{Multinomial}(1, \mathbf{p})] \\ &= \sum_{k=1}^{K}[\theta_{k1}^{A_{l_{ij}1}}\theta_{k2}^{A_{l_{ij}2}}\ldots\theta_{kK}^{A_{l_{ij}K}} \times p_{k1}^{T_{l_{ij}1}}p_{k2}^{T_{l_{ij}2}}\cdots p_{kK}^{T_{l_{ij}K}}] \\ &= \sum_{k=1}^{K}\theta_{kk^\star} \times p_{ik} \\ &= \sum_{k=1}^{K}\left[\theta_{kk^\star} \times \frac{Z_{ik}\lambda_k}{\sum_m Z_{im}\lambda_m}\right].\end{aligned}$$

The last line of equation (9) is a mixture of discrete distributions with component weights $p_{ik} = Z_{ik}\lambda_k/\sum_m Z_{im}\lambda_m$. This result matches the fact that the set of ambiguous autoIDs with label $k^\star$ is a mixture of calls from all species present at location i. In the third line, the equality holds because $1! = 0! = 1$ in the multinomial likelihood. The fourth line follows because only the $k^\star$th entry of $\mathbf{A}_{l_{ij}}$ is nonzero (and has value one). Similarly, only the kth entry of $\mathbf{T}_{l_{ij}}$ is nonzero, with value one.

Let $\delta_{k^\star} = \sum_{k=1}^{K}[\theta_{kk^\star} \times Z_{ik}\lambda_k/\sum_m Z_{im}\lambda_m]$, and let $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots \delta_K)'$ denote the $K \times 1$ vector of marginal probabilities for each autoID label if the recording is not validated. Note that $\sum_{m=1}^{K}\delta_m = 1$ and we can write $p(\mathbf{A}_{l_{ij}} = k^\star|\mathbf{Z}_i, \boldsymbol{\lambda}, \Theta) = \delta_1^{A_{l_{ij}1}}\delta_2^{A_{l_{ij}2}}\ldots\delta_{k^\star}^{A_{l_{ij}k^\star}}\ldots\delta_K^{A_{l_{ij}K}}$, which is the PMF of the Multinomial$(1, \boldsymbol{\delta})$ distribution.

With the unobserved true-species labels marginalized out, the unnormalized posterior distribution based on the observed-data likelihood is

$$
\begin{aligned}
p(\boldsymbol{\psi}, \boldsymbol{\lambda}, \Theta, \mathbf{Z}|\mathbf{T}, \mathbf{A}) \propto \prod_{i=1}^{n} \Bigg[\prod_{k=1}^{K} p(Z_{ik}|\psi_k) \prod_{j=1}^{J} \Bigg[p(Y_{ij\cdot}|\mathbf{Z}_i, \boldsymbol{\lambda}) \\
\times \prod_{l_{ij}:I_{l_{ij}}=1} p(\mathbf{A}_{l_{ij}}|\mathbf{T}_{l_{ij}}, \Theta) p(\mathbf{T}_{l_{ij}}|\mathbf{Z}_i, \boldsymbol{\lambda}) \\
\times \prod_{l_{ij}:I_{l_{ij}}=0} p(\mathbf{A}_{l_{ij}}|\mathbf{Z}_i, \boldsymbol{\lambda}, \Theta) \Bigg] \Bigg] \\
\times \left[p(\boldsymbol{\psi}) p(\boldsymbol{\lambda}) p(\Theta) \right],
\end{aligned}
\tag{10}
$$

with $p(\mathbf{A}_{l_{ij}}|\mathbf{Z}_i, \boldsymbol{\lambda}, \Theta)$ as in equation (9) above.

2.4. *Simulation design.* To explore the performance of our method, we conduct two simulation studies. The goals of these studies are (1) to demonstrate the advantages of using the observed-data likelihood by comparing the posterior inference resulting from this likelihood with a model that uses data augmentation to impute unknown true-species labels; and (2) to show that our model recovers the data generating parameter values with varying validation designs regardless of the species assemblage characteristics.

For simulation study 1, we compare the performance of the ODL model with a complete-data model similar to the one proposed by Spiers et al. (2022) that imputes the unobserved true-species labels using data augmentation (the "CDL model"). In some settings, categorical random variables can be imputed using a latent multivariate normal random variable (Schafer (1997), Sections 6.3–6.4). Song et al. (2009) showed these procedures can bias inference and provided a correction factor for low-dimensional settings where the probability distribution of the observed categorical variable is known. More often, data augmentation schemes sample missing class labels from their conditional predictive distribution during each MCMC iteration (Schafer (1997), p. 72). In our case, missing labels are drawn using the default `categorical_sampler` in the probabilistic programming language NIMBLE (de Valpine et al. (2017)). We defined performance in terms of average estimation error, average 95% posterior interval widths and the coverage rate of 95% posterior intervals. We only compare results from fitted models that exhibited evidence of convergence. We defined evidence of convergence as the fitted model having $\hat{R}$ (Gelman and Rubin (1992), Vehtari et al. (2021)) values less than 1.1 for the ODL model and 1.2 for the CDL model. We increased the convergence threshold because CDL model fits frequently achieved $\hat{R} \leq 1.1$ for all model parameters except λ_3, which often failed to converge. Because the Gelman–Rubin statistic alone is not sufficient to assess convergence (Vats and Knudson (2021)), we chose a threshold of $\hat{R} \leq 1.2$ for the CDL model after inspecting traceplots and effective sample sizes for λ_3 in a subset of simulations. More discussion of convergence results are provided in Section 3.1.

We assumed that data were obtained from an assemblage of five species across 100 sites with 4 visits to each. We used empirically informed data-generating values based on posterior estimates for bats surveyed in British Columbia, Canada (Stratton et al. (2022)). The assumed species assemblage included one highly active and prevalent species (species 1, which we refer to as the "common" species; $\lambda_1 = 28$ and $\psi_1 = 0.9$) and one rare and relatively inactive species (species 4; $\lambda_4 = 1$, $\psi_4 = 0.1$). We compare the performance of the ODL model with the CDL model under four stratified-by-species validation scenarios (Table 1) and under a simple random sample of all recordings.

TABLE 1

Data generating values and validation scenarios for the five-species assemblage assumed in simulation study 1. We explore four stratified validation scenarios, with the level of effort for the "common" species fixed at 10% and the level of effort for species 4 (the "rare" species) fixed at 100%. These choices reflect a design where the greatest conservation concern is placed on the rare species and researchers desire an efficient way to reduce the programmatic costs of validation for species that are frequently encountered. We also consider a simple random sample (SRS) of 50% of all observations

Species	ψ	λ	Scenario 1 (100%)	Scenario 2 (75%)	Scenario 3 (50%)	Scenario 4 (25%)	Scenario 5 (SRS)
1	0.90	28.0	0.1	0.10	0.10	0.10	SRS of
2	0.70	2.4	1.0	0.75	0.50	0.25	50% of all
3	0.24	11.9	1.0	0.75	0.50	0.25	autoIDs
4	0.10	1.0	1.0	1.00	1.00	1.00	
5	0.61	4.0	1.0	0.75	0.50	0.25	

In each of the four stratified-by-species validation scenarios, we assumed that the species of greatest conservation concern was the rare species (species 4), which received 100% validation. To reduce the overall number of validated recordings, we fixed the level of validation effort for the common species (i.e., species 1) at 10% for all scenarios. To explore how further reductions in validation effort affected posterior inference under the two models, we varied the level of validation effort for the three remaining species from 100% to 25% over the four stratified-by-species scenarios. We assigned the same level of effort for species 2, 3 and 5, because their expected number of calls (the product $\psi_k \times \lambda_k$) was neither very high nor very low. In practice, these assignments require some prior knowledge of the species assemblage under consideration. The final validation scenario we considered for comparison between models was a simple random sample of 50% of all observations.

For each validation scenario, we compared the posterior inference from each model under two classification scenarios. The first classification scenario for simulation study 1, designated $\mathbf{\Theta}_1$, assumes that the classifier is *well behaved* with high classification probabilities along the diagonal:

$$\mathbf{\Theta}_1 = \begin{array}{c} \\ \text{True} = 1 \\ \text{True} = 2 \\ \text{True} = 3 \\ \text{True} = 4 \\ \text{True} = 5 \end{array} \begin{array}{c} \begin{array}{ccccc} \text{AutoID} = 1 & \text{AutoID} = 2 & \text{AutoID} = 3 & \text{AutoID} = 4 & \text{AutoID} = 5 \end{array} \\ \begin{bmatrix} 0.95 & 0.02 & 0.01 & 0.01 & 0.01 \\ 0.05 & 0.9 & 0.02 & 0.02 & 0.01 \\ 0.01 & 0.02 & 0.95 & 0.01 & 0.01 \\ 0.02 & 0.03 & 0.05 & 0.85 & 0.05 \\ 0.04 & 0.01 & 0.01 & 0.01 & 0.93 \end{bmatrix} \end{array}.$$

The second classification scenario (designated $\mathbf{\Theta}_2$) assumes that the classifier features low probability of correct identification for the common species (species 1, $\theta_{11} = 0.55$) and comparatively high probabilities of classifying calls from species 1 to other species (e.g., $\theta_{14} = 0.15$). Under our second scenario, the total count of calls attributed to species 2–5 will be contaminated with a large number of false-positive detections that truly came from species 1. Note that a nontrivial proportion of detections from species 1 are incorrectly labeled as

TABLE 2

The three-species assemblages assumed in simulation study 2, which explores a subset of all possible combinations of low, medium and high values for occurrence probabilities ψ and relative activity parameters λ

Assemblage	ψ	λ
1	0.15	2
	0.50	12
	0.90	40
2	0.15	2
	0.90	12
	0.50	40
3	0.50	2
	0.15	12
	0.90	40
4	0.50	2
	0.90	12
	0.15	40
5	0.90	2
	0.15	12
	0.50	40
6	0.90	2
	0.50	12
	0.15	40

species 4, resulting in a high rate of false-positives for the rare species.

$$\boldsymbol{\Theta}_2 = \begin{array}{c} \\ \text{True} = 1 \\ \text{True} = 2 \\ \text{True} = 3 \\ \text{True} = 4 \\ \text{True} = 5 \end{array} \begin{array}{c} \begin{array}{ccccc} \text{AutoID} = 1 & \text{AutoID} = 2 & \text{AutoID} = 3 & \text{AutoID} = 4 & \text{AutoID} = 5 \end{array} \\ \begin{bmatrix} \mathbf{0.55} & 0.1 & 0.1 & \mathbf{0.15} & 0.1 \\ 0.05 & 0.9 & 0.02 & 0.02 & 0.01 \\ 0.03 & 0.02 & 0.87 & 0.04 & 0.04 \\ 0.02 & 0.03 & 0.05 & 0.85 & 0.05 \\ 0.04 & 0.01 & 0.01 & 0.01 & 0.93 \end{bmatrix} \end{array}.$$

Consequently, we refer to $\boldsymbol{\Theta}_2$ as the *common-to-rare* classifier scenario because of the direction of false-positives.

For simulation study 2, we assumed assemblages of three species to reduce computational demand. Assemblages of three species are small compared with those encountered in practice, but are still at least as large as those assumed in simulations by Wright et al. (2020), Spiers et al. (2022) and others. For each assemblage shown in Table 2, we simulated 50 datasets, assuming 4 visits to each of 100 sites. Each of the 50 datasets was subjected to three validation designs. The first of these was a simple random sample (SRS) of 50% of all recordings. The second scenario followed the "fixed effort" (FE) design used for validation of the Oregon/Washington bat acoustic data, but with 50% of recordings from the first visit to a site validated. We also explore a stratified-by-species validation design (Stratified), where the most common species received 30% validation, the second most common species received 50% validation, and rare species autoIDs were always validated.

For simulation study 2, we held the classifier constant and used values based on testing by the Montana Natural Heritage Program of classifiers using hand-released bats (written communication from Dan Bachen, Montana Natural Heritage Program, 2024). The empirically

motivated classifier for simulation study 2 is given by

$$\boldsymbol{\Theta} = \begin{array}{c} \\ \text{True} = 1 \\ \text{True} = 2 \\ \text{True} = 3 \end{array} \begin{array}{c} \begin{array}{ccc} \text{AutoID} = 1 & \text{AutoID} = 2 & \text{AutoID} = 3 \end{array} \\ \begin{bmatrix} 0.91 & 0.01 & 0.08 \\ 0.15 & 0.54 & 0.31 \\ 0.11 & 0.03 & 0.86 \end{bmatrix} \end{array}.$$

In both studies, we simulated a complete dataset then randomly sampled according to the validation design. All observations that were not selected into the validation set had their true species label masked, simulating an ambiguous detection. For each simulated dataset, we tracked the runtime required to fit the model, the posterior mean, 95% posterior interval width, evidence of convergence for each parameter, and whether the 95% posterior interval captured the true data-generating value. All MCMC algorithms were implemented in NIMBLE (de Valpine et al. (2017, 2022)) in R (version 4.3.2, R Core Team (2021)). Although not necessary for fitting our model, we chose to expedite our simulation studies by performing computations on the Tempest High Performance Computing System, operated and supported by University Information Technology Research Cyberinfrastructure at Montana State University. We assigned each validation scenario to a node, with three cores used per node to fit chains in parallel. All nodes had 256 CPUs and 1024 GB of memory available, which is far beyond the requirements of fitting this model. We provide code to fit the ODL model on a desktop computer in Supplementary Material B (Oram et al. (2025)).

3. Simulation results.

3.1. *Comparison of complete-data and the observed-data likelihood.* Our first simulation study compared coverage and average estimation error obtained from the observed-data likelihood (ODL) model with that from a complete-data model that uses data augmentation to draw missing true-species labels of recordings not selected into the validation set (i.e., the "CDL model"). In general, the ODL model achieved near-nominal coverage and minimal estimation error in all validation scenarios, including under a simple random sample of 50% of all recordings (Figure 1). Additionally, the ODL model matched or outperformed the CDL model, even when observations were selected for validation according to an SRS. We noted an increase in model performance (i.e., coverage that was closer to nominal levels and decreased estimation error) using the observed-data likelihood especially when the classifier was poor (i.e., under Θ_2; refer to Supplementary Material A, Oram et al. (2025), Figures 4–6).

We observed particularly large estimation error and an associated decrease in coverage for large relative activity parameters (e.g., λ_1 and λ_3) under the CDL model as the level of effort decreased across validation scenarios 1–4 (Figure 1). In contrast, the ODL model exhibited minimal estimation error, with similar posterior interval widths and near-nominal coverage. For relative activity parameters with smaller true values, such as λ_2 and λ_5, inference was nearly identical under the two models, with little variation among validation scenarios (Supplementary Material A, Oram et al. (2025), Figures 1 and 3). These findings are consistent with the two-species and one morphospecies simulation studies conducted by Spiers et al. (2022), which set small values for relative activity parameters, and with the simulation results reported in the Supplementary Material of Koslovsky et al. (2024).

Underestimation of relative activity parameters by the CDL model corresponded with a proportionally large overestimate of that species' probability of occurrence. For instance, in validation scenario 4 (25%, Figure 1), which had the lowest overall effort, we observed severe underestimation of λ_3 and substantial overestimation of ψ_3. We noted that validation scenarios with the most severe estimation error for (λ_1, ψ_1) and (λ_3, ψ_3) were often accompanied by

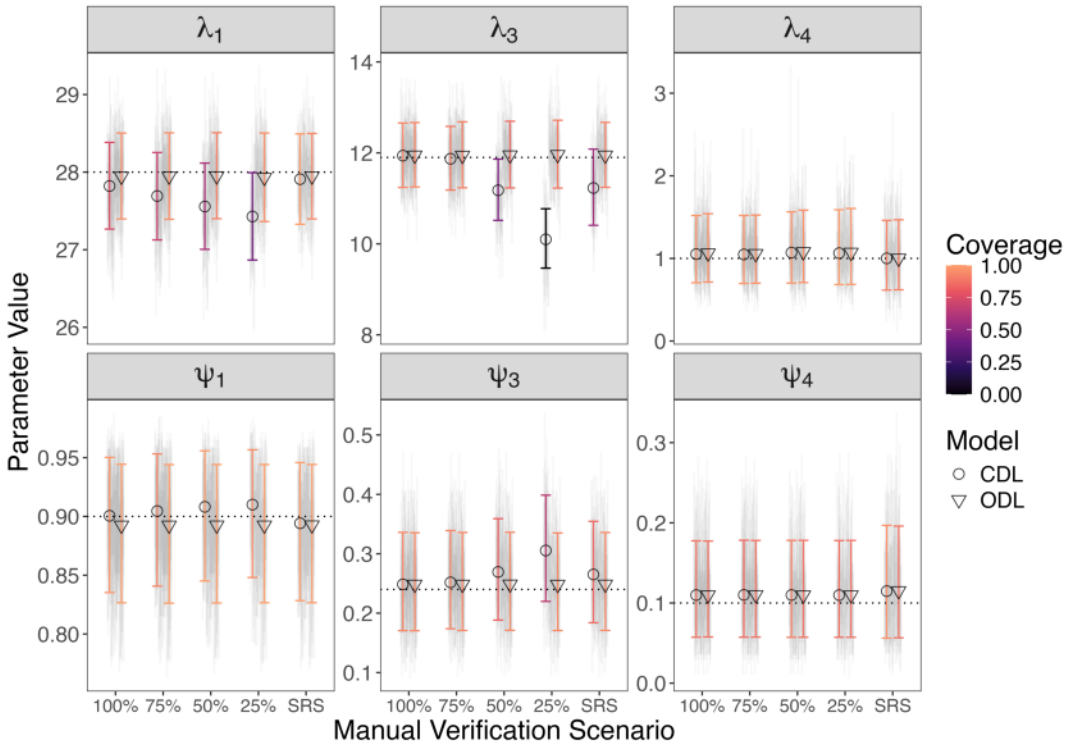

FIG. 1. *Average posterior means and 95% posterior intervals for relative activity parameters under the well-behaved classifier* (Θ_1). *Circles and triangles indicate the average posterior means for the complete-data (CDL) and observed-data likelihood (ODL) models, respectively. The x-axis corresponds to the validation design. Percentage labels on the x-axis correspond to the % of calls validated for species* 2, 3 *and* 5 *under the stratified design, where the level of effort for species* 1 (*"common" species*) *is fixed at* 10% *and the level of effort for species* 4 (*"rare" species*) *is fixed at* 100%. *Background error bars show posterior intervals from an individual fit of a simulated dataset that showed evidence of convergence* (*i.e., Gelman–Rubin statistic* $\hat{R} < 1.1$ *for the ODL model and* $\hat{R} < 1.2$ *for the CDL model*). *Large error bars indicate average posterior intervals across all simulated datasets that converged, with color indicating coverage. The dotted lines indicate the true parameter values for each parameter.*

estimation error in θ_{13}, the probability that a recording generated by species 1 was classified as species 3 by the ML algorithm (Supplementary Material A, Oram et al. (2025), Figure 3). All other misclassification parameters were estimated with minimal error and near-nominal coverage by both models (Supplementary Material, Oram et al. (2025), Figures 4–6).

The severity of estimation error under the CDL model varied with the characteristics of the classifier. In general, we observed much larger estimation error under the poor classifier (Θ_2) than under the well-behaved classifier, especially for problematic parameter pairs such as (ψ_1, λ_1) and (ψ_3, λ_3). Although we consistently observed far greater error in estimates obtained using the CDL model under the poor classifier, there was little change in interval widths, resulting in coverage far below nominal levels for some relative activity and occurrence parameters (e.g., (ψ_1, λ_1) and (ψ_3, λ_3); Figure 1). Estimates obtained from the ODL model, in contrast, were not as strongly affected by the accuracy of the classifier; performance gains (in terms of coverage and estimation error) were marginal with the well-trained classifier (comparing Figure 1 with Supplementary Material A, Oram et al. (2025), Figures 4–5). For example, in Scenario 4 (25%), which had the lowest total validation effort, we documented a decrease in the CDL model's coverage of λ_5 from 0.96 under the well-trained classifier to 0.63 under the poor classifier (Supplementary Material A, Oram et al. (2025)). For this same parameter, the ODL model's coverage changed from 0.96 to 0.9, indicating

the ODL model is less affected by large off-diagonal elements in Θ than the CDL model for fixed validation effort.

In our simulations, we found that the ODL model converged quickly (within 2000 iterations), while the CDL model often required substantially more MCMC iterations to reach convergence ($> 20{,}000$). This result aligns with existing knowledge that integrating out latent variables from complex models can improve the efficiency and stability of MCMC algorithms (Liu, Wong and Kong (1994), Yackulic et al. (2020)). In our case, even with a 10-fold increase in MCMC iterations, many more of the CDL model fits failed to meet our convergence criterion and were excluded (Table 1, Supplementary Material A, Oram et al. (2025)). We excluded all model fits that did not reach MCMC convergence. The low convergence rates in the CDL model were primarily driven by λ_3, which often exhibited substantial autocorrelation in traceplots and correspondingly small effective sample sizes. Often, all parameters in a CDL model fit would have $\hat{R} \approx 1$ except for λ_3, even after a vast number of MCMC iterations. For the sake of comparing results across the two models, we softened the threshold required for a CDL model fit to be considered "converged" from 1.1 to 1.2. Although we did not observe a substantial difference in the number of MCMC iterations per minute, the reduced number of iterations required of the ODL model to reach $\hat{R} < 1.1$ for all parameters substantially reduced computation time (e.g., ≈ 5 minutes versus ≈ 56 minutes for a dataset of $\approx 13{,}000$ observations).

For practitioners, the results of our first simulation study have two main implications. First, the ODL model provides minimal estimation error and near-nominal coverage with a stratified-by-species validation design at much lower levels of effort than the CDL model. For example, the performance of the CDL model was substantially worse, especially for λ_1 and λ_3, unless at least 75% species 2, 3 and 5 autoIDs were validated (Figure 1). Second, the difference in performance is expected to increase if the classifier frequently misclassifies common species as rare ones, as in Θ_2.

3.2. *Comparing across assemblages with the ODL model.* Our second simulation study focused on comparing results across different species assemblages within each of three validation designs: a stratified sample, with 100%, 50% and 30% validation for rare, medium and common species, respectively; a fixed effort design, wherein the first sampling occasion (visit) of each sampling unit (site) had a fixed percentage (50%) validated; and a simple random sample of 50% of all observations validated.

Our results were remarkably consistent across assemblages with high convergence rates, near-nominal coverage for 95% posterior intervals, and minimal estimation error for ecologically relevant parameters such as relative activity rates and occurrence probabilities (Figure 2 and Figure 7, Supporting Information 1, Oram et al. (2025)). Uncertainty interval widths for relative activity parameters appear to change across assemblages in Figure 2; however, across assemblages these results are consistent in the sense that the relative activity parameter that exhibited the widest uncertainty intervals always belonged to the species with the lowest probability of occurrence.

Often, greater uncertainty (relative to the same parameter value in other assemblages) in model estimates of λ_k was matched by greater uncertainty in the misclassification parameters $\boldsymbol{\theta}_k$ (Figures 8–10, Supplementary Material A, Oram et al. (2025)). This is to be expected because rareness in a species' occurrence means there are fewer opportunities to observe calls, providing less information to estimate the rate at which that species generates recordings and the rate at which those recordings are subsequently misclassified. The patterns observed across assemblages persisted regardless of whether observations were validated according to a simple random sample, stratified random sample or a fixed effort design. We illustrate the use of our method on real data using a fixed-effort validation design in the following section.

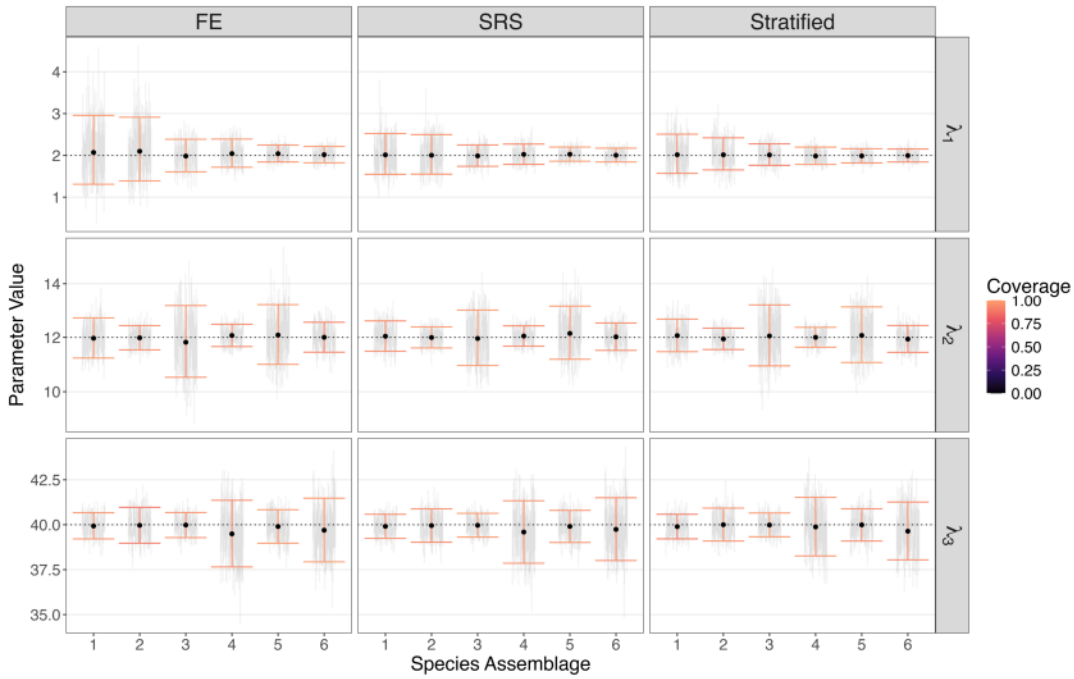

FIG. 2. *Simulation results across validation designs (facet columns) for relative activity parameters (facet rows) in six different species assemblages. "FE," "SRS" and "Stratified" denote fixed effort, simple random sample and stratified validation designs, respectively. The dotted lines indicate the true parameter values, gray intervals show individual 95% posterior intervals from fitted models, and colored error bars indicate average 95% posterior intervals with color indicating coverage. Black points indicate average posterior means.*

4. Application: Bat acoustic data in Oregon and Washington. We applied our method to the motivating acoustic dataset obtained from Oregon and Washington during summertime surveys in 2019. These data were gathered following the NABat spatially balanced master sample (Reichert et al. (2021)). In brief, the study region was divided into nonoverlapping 10×10 kilometer grid cells, which were then assigned a priority ordering according to the Generalized Tessellation Randomized Stratified sampling (GRTS) algorithm (Stevens and Olsen (2004)). An advantage of the GRTS algorithm is that any subset of grid cells, when visited in priority order, will yield a spatially balanced sample—a desirable quality when monitoring a species over large spatial extents. In our case, the empirical data were obtained from the first 10 grid cells in Washington, and the first 10 grid cells in Oregon. One grid cell in Oregon had no recordings with single-species autoIDs and was discarded. Within each grid cell, between 2 and 4 acoustic detectors were deployed according to NABat protocols (Rodriguez et al. (2019)); these detectors typically remained in place for one night. When detectors were deployed at the same location for more than one night, we modeled only the first and last night to reduce temporal correlation in nightly detections. With this scheme, a site-night corresponds to a detector night at a specific location within a GRTS cell. When acoustic surveys were complete, raw audio files were filtered and classified to species using the proprietary classification software SonoBat (Sonobat (2019)). A fixed-effort validation design was attempted, with 20% of the call files from the first recording night at each grid cell randomly selected for validation. Additionally, researchers ensured that at least 50 recordings were validated from the first detector night at each grid cell; if fewer than 50 recordings were obtained on the first detector night, all recordings received validation. The true species identity of many selected calls could not be confirmed, leading to between 2 and 70% of recordings with known true-species labels at sites with $Y_{ij\cdot} \geq 10$ recordings. We treated these

TABLE 3
Species codes for the eight focal species. Any observed bat species not listed here was given a label of OTHER

Species code	Latin name	Common name
EPFU	*Eptesicus fuscus*	Big brown bat
LACI	*Lasiurus cinereus*	Hoary bat
LANO	*Lasionycteris noctivagans*	Silver-haired bat
MYCA	*Myotis californicus*	California Myotis
MYCI	*Myotis ciliolabrum*	Western small-footed bat
MYEV	*Myotis evotis*	Long-eared Myotis
MYLU	*Myotis lucifugus*	Little brown bat
MYYU	*Myotis yumanensis*	Yuma Myotis

recordings as ambiguous data, effectively assuming confirmability is independent of the true-species label. Throughout, we refer to species using their four-letter abbreviations, which are shown in Table 3. Although our derivations and simulations in Sections 2 and 3 have advocated for a stratified-by-species validation design, the Oregon/Washington data comprise the only published bat acoustic dataset that contains partial verification of recordings within a detector night because model formulations prior to Spiers et al. (2022) limited the flexibility of validation designs available to researchers.

Using the ODL model formulation, we obtained posterior estimates for occurrence probabilities and relative activity with sufficiently narrow posterior interval widths to be practically useful for researchers (Figures 3 and 4). The results obtained from our comparatively simple model, which assumes that occurrence probabilities and relative activity rates are constant across sites and visits, are generally consistent with other published results and domain knowledge. For example, using data gathered in British Columbia, Canada, Stratton et al. (2022) found similar estimates for occurrence probability of *Lasiurus cinereus*, *Myotis evotis* and *Myotis lucifugus* (abbreviated as LACI, MYEV and MYLU, respectively; Table 3) using a unique dataset with all true and autoID labels known. We estimated higher probability of

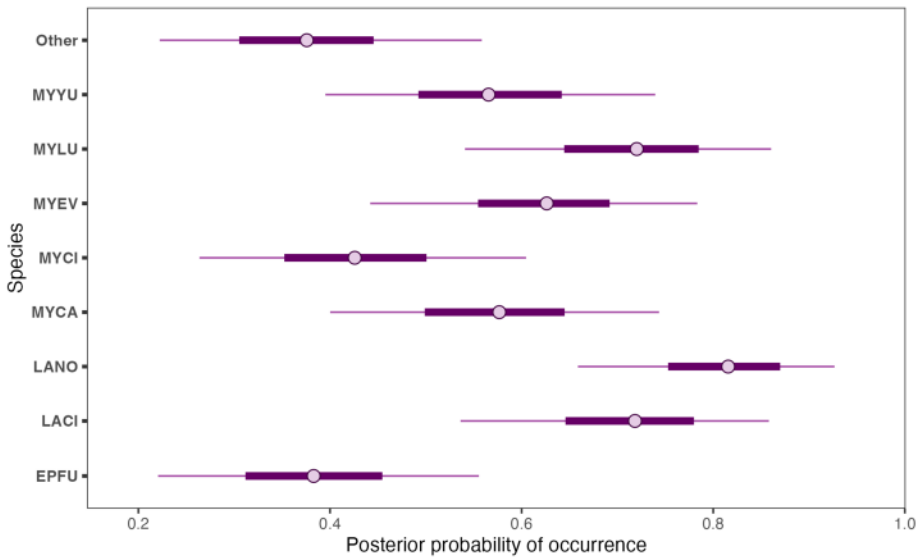

FIG. 3. *Posterior intervals for occurrence probabilities ψ_k obtained from the* 2019 *Oregon/Washington, USA, bat acoustic data. Thick lines are* 50% *posterior intervals, thin lines are* 95% *posterior intervals and points are posterior mean estimates.*

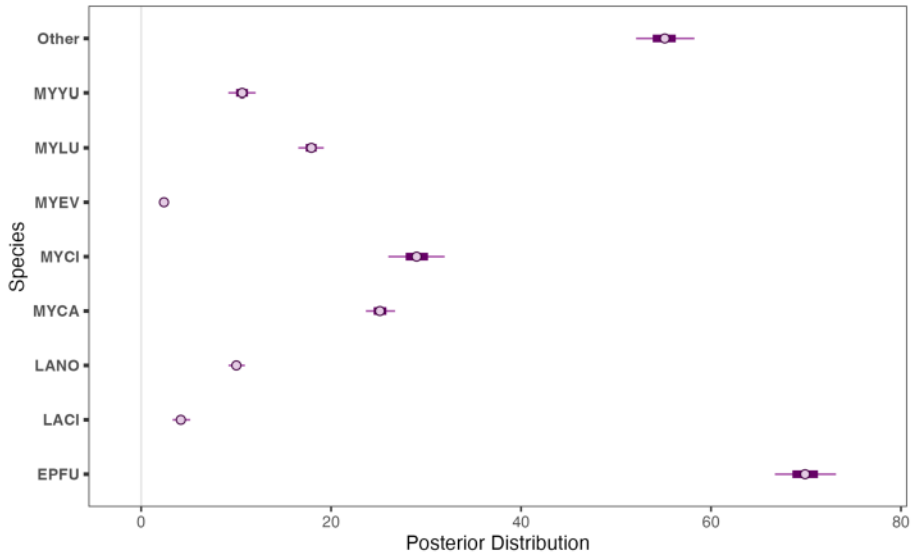

FIG. 4. *Posterior intervals for relative activity rates λ_k for each focal species considered in the* 2019 *Oregon/Washington, USA, bat acoustic data. Nonfocal species were modeled using the "other" category. Thick lines are* 50% *posterior intervals, thin lines are* 95% *posterior intervals and points are posterior mean estimates.*

occurrence for their fourth focal species (*Myotis ciliolabrum* or MYCI; 0.24 vs. 0.42 in our analysis), but variation in some species' parameter estimates is expected because our data are from different regions within each species' range.

Our estimates of relative activity are also in line with existing results. For instance, we found that MYEV and LACI had the lowest relative activity levels, consistent with Stratton et al. (2022). We also found that the species with the highest relative activity level was *Eptesicus fuscus* (EPFU; Figure 4), which others have documented as a highly active species (Christophersen and Kuntz (2003), Miller (2010)). Of note, naïve counts of EPFU autoIDs are not particularly large, with the exception of one GRTS cell. However, we found that there was a nontrivial probability (0.53) that recordings from EPFU were mistakenly classified as *Lasionycteris noctivagans* (LANO; Figure 5), which did have a large number of autoIDs. The high estimates for off-diagonal classification parameters underscore the importance of modeling the ML classification process; if we did not account for misclassification and only used naïve counts, we would severely underestimate the relative activity rate for EPFU. The high value for θ_{13} is expected because LANO and EPFU often share call characteristics, potentially making them hard to distinguish for automated algorithms (Bachen et al. (2018)).

5. Discussion. To assist the population monitoring efforts of bat researchers, we have introduced an observed-data model that displays computational and estimation advantages compared to relying on a data augmentation approach to handle the unobserved true-species labels. Although either approach should accommodate ignorable validation designs, the ODL framework facilitates monitoring programs to iterate more quickly on model-building and sensitivity checks as computational time was decreased in our simulations.

Our ODL model specification provides an opportunity for researchers to use optimal design criteria to obtain validation designs that meet the goals and objectives of their study. As a first step in this process, the researcher would define the design space and the study goal in terms of an optimal design criteria, such as minimizing the volume of the 95% credible region ellipsoid for model parameters (i.e., Bayesian D-optimality; Huan, Jagalur and Marzouk (2024)) or expected information gain (Bernardo (1979), Rainforth et al. (2024)). The design space in the case of acoustic bat monitoring would be a set of possible validation designs

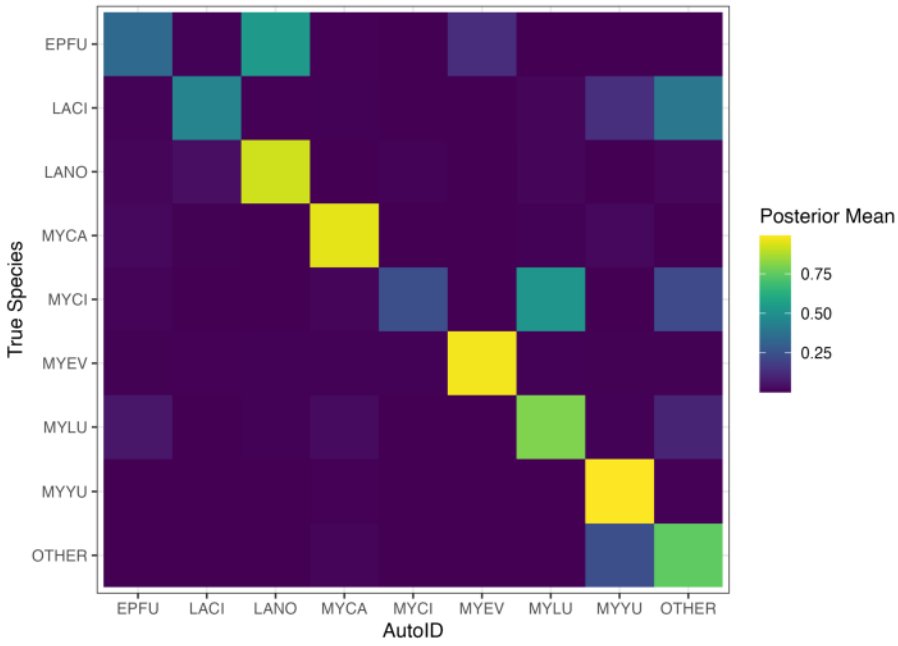

FIG. 5. *Posterior means for classification parameters. Lighter colors correspond with higher probabilities. Refer to Table 3 for species codes.*

that induce missingness at random and meet budget constraints, together with the number of sites and balanced visits to be sampled. Some care would be needed in defining the design space, however, because the optimal design may change depending on the characteristics of the species assemblage and the classifier. By providing researchers with unbiased estimates and near-nominal coverage of target parameters from flexible and cost-effective validation designs inducing missingness at random, we anticipate that the ODL model will substantially lower the barriers that have prevented widespread adoption of the count-detection modeling framework.

We showed in Section 3 that for some validation designs (such as the stratified-by-species approach in Section 3.1), using the observed-data likelihood improves the efficiency of manual verification. The increased efficiency is visible in Figure 1 because reliable estimates of model parameters can be obtained for substantially lower effort when using the ODL model compared to the CDL model. Across all of our simulation scenarios, the ODL model outperformed the CDL model: the average widths of posterior intervals were comparable for the two models, but with substantially less estimation error when using the observed-data likelihood. This combination of characteristics led to near-nominal coverage of posterior intervals with the ODL model, even when the number of validated records was low. In contrast, the CDL model exhibited very low coverage (even zero in some scenarios) for larger relative activity parameters (e.g., λ_1 and λ_3). We noted persistent estimation error for the CDL model even when a simple random sample was taken (i.e., scenario SRS in simulation 1). We did not expect these results because Stratton et al. (2022) did not document estimation error when validating all calls from a simple random sample of site-nights, but note that Koslovsky et al. ((2024), Supplementary Material A, Section 3) used a simple random sample to select recordings and found the greatest estimation error occurred for large relative abundance parameters in their model framework.

In comparing data augmentation with the ODL model, it is evident that with a stratified-by-species validation design, the ODL model requires far fewer recordings to accurately estimate occurrence probabilities and relative activity levels. For instance, assuming a well-behaved

classifier like Θ_1, if reliable inference is desired for a species with characteristics similar to species 3 in our assemblage, we would need to use a design similar to simulation scenario 1, where all recordings with autoIDs other than the extremely common and active species are validated (Table 1). For our assumed assemblage, this corresponded to around 4200 recordings, a high enough level of validation effort to ensure that recordings from all sites are validated, which helps to ensure occurrence probabilities and relative activity levels are identifiable. In contrast, with the ODL approach, we could obtain comparable inference under simulation scenario 4, which requires only 25% of recordings with autoID labels 2, 3 and 5 be validated. This level of effort decreases the required number of validated recordings to approximately 1900 recordings. If one assumes that it is not more costly or time-consuming to validate rare species autoIDs than very common autoIDs, then using the observed-data likelihood with stratified-by-species validation can roughly halve the costs incurred by coupled classification. Clearly, this is an attractive feature for monitoring programs that would like to use coupled classification because it allows for validation to be assigned according to program objectives while simultaneously reducing the effect of the validation step and increasing the overall efficiency of the program.

Our second set of simulations found that our results do not change substantially with the characteristics of the species assemblage (Section 3.2). Although we did not alter the hierarchical form of the model, we varied the characteristics of the parameter values because preliminary simulations indicated that some combinations of parameters (e.g., low occurrence probability ψ_k and high relative activity λ_k) were particularly challenging when imputing missing true-species labels. Although we did not investigate alternative model forms in this paper, altering our ODL approach to fit other model specifications would be straightforward in principle as long as the validation design is ignorable.

We applied our method to the motivating bat acoustic data obtained from Oregon and Washington in 2019. Ours is the first published analysis of these data using the count-detection framework, largely because the Poisson formulation of the likelihood introduced by Wright et al. (2020) and further investigated by Stratton et al. (2022) could not accommodate the partial verification of autoID labels within a given site-night. By adopting the individual-level model proposed by Spiers et al. (2022) and marginalizing over the missing true-species labels, we successfully estimated occurrence probabilities and relative activity rates for eight species of interest. The model we fit in our application is comparatively simple because it assumes that the occurrence probabilities for each species are constant across sites. We also assume that the relative activity rates for each species do not vary across sites or visits; modeling these data with a generalized linear model framework would allow for mapping of each species distribution across space. In addition, one could consider exploring an efficient validation design that selects recordings according to the confidence score output by the ML algorithm. However, many bat acoustic classifiers do not make this information available to users. If confidence scores are available, one could model species occupancy and relative activity using a model similar to Rhinehart, Turek and Kitzes (2022) while selecting recordings according to confidence scores. In this case, it would likely be necessary to include the confidence scores in the model, even if they are not of scientific interest to researchers (Collins, Schafer and Kam (2001), Nakagawa (2015), Rabe-Hesketh and Skrondal (2023)). Similar caution is warranted if only high-quality recordings are retained for manual review, as often happens with other bat acoustic datasets where the goal is to identify "voucher" recordings indicative of species presence (Reichert et al. (2018)).

Our work has connections to other developments in the ML and statistics literature. For instance, coupled classification can be viewed as a special case of Bayesian Data Reconciliation (Hanks, Hooten and Baker (2011)); we assume the ecological process is observed through a rigorous gold-standard data source, which is then used to update inference made

from less accurate data. In the case of coupled classification, validated recordings provide the gold-standard data used to improve inference from the remaining, ambiguous subset of observations. Additionally, an ODL formulation of the Koslovsky et al. (2024) framework can be obtained by partitioning the product in the likelihood for each individual recording into observed and unobserved components as we do in equation (8). The true class labels can then be marginalized out of the unobserved component to obtain the ODL (Oram et al. (2025)). While not required in the framework of Koslovsky et al. (2024), they note that validated observations can improve estimation performance. By marginalizing out latent true class labels, we anticipate that an ODL formulation could further increase the speed and stability of their model (Liu, Wong and Kong (1994), Yackulic et al. (2020), Park and Lee (2022)).

In the ML literature, Angelopoulos et al. (2023) recently proposed *prediction-powered inference*, a framework that uses a small auxiliary (or "gold-standard") dataset to inform the classification error from a pre-trained ML algorithm. The input for prediction-powered inference is a set of predicted and true-class labels for observations in the gold-standard dataset, which are then used to bias-correct (or "rectify") the predictions in a large unlabeled dataset for which only the predicted labels are available. Finally, all of the observations in both datasets are used to generate confidence intervals for quantities of interest that are centered on the bias-corrected point estimate. Angelopoulos et al. (2023) acknowledge that to obtain good results using prediction-powered inference, the features of the gold-standard dataset must follow approximately the same distribution as those in the unlabeled dataset. In other words, the auxiliary dataset must be representative of the sample in hand; if they are not, it is possible that the unlabeled data suffer from "covariate shift." Furthermore, if the proportion of each label in the gold-standard dataset is different from the proportion of each label in the unlabeled dataset, then there is said to be "label shift."

In essence, the steps of prediction-powered inference are the same general approach as used by Chambert et al. (2018), Wright et al. (2020) and Spiers et al. (2022). All of these models rely on a large number of observations with potentially erroneous species labels generated by an imperfect classifier; these correspond to the unlabeled dataset in prediction-powered inference. Where prediction-powered inference uses auxiliary gold-standard data to rectify the point estimate and correct for bias induced by the imperfect classifier, the models of Chambert et al. (2018), Wright et al. (2020) and Spiers et al. (2022) instead introduce a parametric model for the misclassification process that is informed by the auxiliary data. These authors' models, as well as our ODL model formulation, can be used with coupled classification to inform the validation set so that the gold-standard dataset is a subset of the observed sample data. By using the sample in hand to model misclassification, coupled classification with a probabilistic validation design can help ensure that the gold-standard dataset is representative of the unlabeled data, alleviating concerns of covariate shift. However, if the validation design preferentially samples some labels, the validation design may induce label shift, and consideration of the process giving rise to the validation set is necessary. When the selection process is known and data are missing at random with respect to the model, ignoring the validation design and using an ODL formulation such as the one we derive can yield accurate inference.

Machine learning classification algorithms have immense promise for improving the efficiency of scientific research. The methods presented here improve the efficiency of a specific workflow designed for bat acoustic data in North America, so that validation effort can be carefully tailored to the programmatic constraints and conservation priorities of a particular monitoring program. However, the principles behind our methods—namely, using ignorable validation designs and modeling the observed data—are broadly applicable to many scientific domains. By thoughtfully incorporating machine-generated labels into statistical models, researchers can leverage the high-thoroughput nature of ML workflows while simultaneously modeling the misclassification process to obtain reliable inference regarding ecological parameters of interest such as species occurrence and relative activity.

Acknowledgments. We are grateful to Cori Lausen and Jason Rae of Wildlife Conservation Society Canada for providing their perspectives on early results of our simulation study, and to Will Janousek of the U.S. Geological Survey Northern Rocky Mountain Science Center for insightful comments on a draft of this manuscript. We appreciate the pilot effort conducted by the Northwest Bat Hub led by Roger Rodriguez, Ben Neece and Tom Rodhouse in 2019 that resulted in the subset of the Oregon/Washington bat dataset that we used for demonstration. We would also like to thank two anonymous reviewers for their constructive comments that substantially improved the quality of this paper.

Any use of trade, firm or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Funding. This material is based upon work over the last 3 years that has been supported by the U.S. Geological Survey under Grant/Cooperative Agreement Nos. G20AC00406-02 (JO supported), G23AC00118-00 (CS supported, KMB partially supported), G21AC10748-00 (JO and AH supported, KMB partially supported).

Access to the computing cluster and partial support for KMB's participation was supported by the AI Institute for Dynamical Systems at the University of Washington, NSF: Award Number 2112085.

KMI's participation was secured through sustained support from the U.S. Geological Survey's Ecosystems Mission Area and funding for White-nose Syndrome and the North American Bat Monitoring Program.

SUPPLEMENTARY MATERIAL

Full simulation results (DOI: 10.1214/25-AOAS2096SUPPA; .pdf). A collection of tables and figures showing simulation results for all scenarios, species and assemblages.

Code (DOI: 10.1214/25-AOAS2096SUPPB; .pdf). An overview of the data structure used in simulations, and code to fit the ODL model using NIMBLE.

Partial integration of the CDL in the framework of Koslovsky et al. (2024) (DOI: 10.1214/25-AOAS2096SUPPC; .pdf). This supplement derives a partially integrated form of the Koslovsky et al. (2024) model that is similar to the ODL model we consider.

REFERENCES

ANGELOPOULOS, A. N., BATES, S., FANNJIANG, C., JORDAN, M. I. and ZRNIC, T. (2023). Prediction-powered inference. *Science* **382** 669–674. MR4672910

BACHEN, D., MCEWAN, A., BURKHOLDER, B., HILTY, S., BLUM, S. and MAXELL, B. (2018). Bats of Montana: Identification and natural history. Technical Report, Montana Natural Heritage Program.

BERNARDO, J.-M. (1979). Expected information as expected utility. *Ann. Statist.* **7** 686–690. MR0527503

BLUDAU, I., WILLEMS, S., ZENG, W.-F., STRAUSS, M. T., HANSEN, F. M., TANZER, M. C., KARAYEL, O., SCHULMAN, B. A. and MANN, M. (2022). The structural context of posttranslational modifications at a proteome-wide scale. *PLoS Biol.* **20** 1–23. https://doi.org/10.1371/journal.pbio.3001636

BRAUER, C. L., DONOVAN, T. M., MICKEY, R. M., KATZ, J. and MITCHELL, B. R. (2016). A comparison of acoustic monitoring methods for common anurans of the northeastern United States. *Wildl. Soc. Bull.* **40** 140–149. https://doi.org/10.1002/wsb.619

CARLEO, G., CIRAC, I., CRANMER, K., DAUDET, L., SCHULD, M., TISHBY, N., VOGT-MARANTO, L. and ZDEBOROVÁ, L. (2019). Machine learning and the physical sciences. *Rev. Modern Phys.* **91** 045002. https://doi.org/10.1103/RevModPhys.91.045002

CHAMBERT, T., WADDLE, J. H., MILLER, D. A. W., WALLS, S. C. and NICHOLS, J. D. (2018). A new framework for analysing automated acoustic species detection data: Occupancy estimation and optimization of recordings post-processing. *Methods Ecol. Evol.* **9** 560–570. https://doi.org/10.1111/2041-210X.12910

CHRISTOPHERSEN, R. G. and KUNTZ, R. C. II (2003). A survey of bat species composition, distribution and relative abundance. Technical Report, North Cascades National Park.

COLLINS, L. M., SCHAFER, J. L. and KAM, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol. Methods* **6** 330–351.

DE VALPINE, P., PACIOREK, C., TUREK, D., MICHAUD, N., ANDERSON-BERGMAN, C., OBERMEYER, F., WEHRHAHN CORTES, C., RODRÍGUEZ, A., TEMPLE LANG, D. et al. (2022). NIMBLE User Manual R package manual version 0.12.2. https://doi.org/10.5281/zenodo.1211190

DE VALPINE, P., TUREK, D., PACIOREK, C. J., ANDERSON-BERGMAN, C., TEMPLE LANG, D. and BODIK, R. (2017). Programming with models: Writing statistical algorithms for general model structures with NIMBLE. *J. Comput. Graph. Statist.* **26** 403–413. MR3640196 https://doi.org/10.1080/10618600.2016.1172487

DEIANA, A. M., TRAN, N., AGAR, J., BLOTT, M., DI GUGLIELMO, G., DUARTE, J., HARRIS, P., HAUCK, S., LIU, M. et al. (2022). Applications and techniques for fast machine learning in science. *Front. Big Data* **5**. https://doi.org/10.3389/fdata.2022.787421

GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2013). *Bayesian Data Analysis*, 3rd ed. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. MR3235677

GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472. https://doi.org/10.1214/ss/1177011136

GIBB, R., BROWNING, E., GLOVER-KAPFER, P. and JONES, K. E. (2019). Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. *Methods Ecol. Evol.* **10** 169–185. https://doi.org/10.1111/2041-210X.13101

HANKS, E. M., HOOTEN, M. B. and BAKER, F. A. (2011). Reconciling multiple data sources to improve accuracy of large-scale prediction of forest disease incidence. *Ecol. Appl.* **21** 1173–1188. https://doi.org/10.1890/09-1549.1

HUAN, X., JAGALUR, J. and MARZOUK, Y. (2024). Optimal experimental design: Formulations and computations. *Acta Numer.* **33** 715–840. MR4793684 https://doi.org/10.1017/S0962492924000023

IRVINE, K. M., BANNER, K. M., STRATTON, C., FORD, W. M. and REICHERT, B. E. (2022). Statistical assessment on determining local presence of rare bat species. *Ecosphere* **13** e4142. https://doi.org/10.1002/ecs2.4142

IRVINE, K. M., RODHOUSE, T. J., WRIGHT, W. J. and OLSEN, A. R. (2018). Occupancy modeling species–environment relationships with non-ignorable survey designs. *Ecol. Appl.* **28** 1616–1625. https://doi.org/10.1002/eap.1754

JELIAZKOV, A., BAS, Y., KERBIRIOU, C., JULIEN, J.-F., PENONE, C. and LE VIOL, I. (2016). Large-scale semi-automated acoustic monitoring allows to detect temporal decline of bush-crickets. *Glob. Ecol. Conserv.* **6** 208–218. https://doi.org/10.1016/j.gecco.2016.02.008

KHALIGHIFAR, A., GOTTHOLD, B. S., ADAMS, E., BARNETT, J., BEARD, L. O., BRITZKE, E. R., BURGER, P. A., CHASE, K., CORDES, Z. et al. (2022). NABat ML: Utilizing deep learning to enable crowdsourced development of automated, scalable solutions for documenting North American bat populations. *J. Appl. Ecol.* **59** 2849–2862. https://doi.org/10.1111/1365-2664.14280

KOSLOVSKY, M. D., KAPLAN, A., TERRANOVA, V. A. and HOOTEN, M. B. (2024). A unified Bayesian framework for modeling measurement error in multinomial data. *Bayesian Anal.* 1–31. https://doi.org/10.1214/24-BA1477

LINK, W. A. and BARKER, R. J. (2010). *Bayesian Inference: With Ecological Applications*. Academic Press, Boston, MA.

LITTLE, R. J. (2021). Missing data assumptions. *Annu. Rev. Stat. Appl.* **8** 89–107. MR4243540 https://doi.org/10.1146/annurev-statistics-040720-031104

LITTLE, R. J. A. and RUBIN, D. B. (2020). *Statistical Analysis with Missing Data*, 3rd ed. *Wiley Series in Probability and Statistics*. Wiley-Interscience, Hoboken, NJ.

LIU, J. S., WONG, W. H. and KONG, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81** 27–40. MR1279653 https://doi.org/10.1093/biomet/81.1.27

LOEB, S. C., RODHOUSE, T. J., ELLISON, L. E., LAUSEN, C. L., REICHARD, J. D., IRVINE, K. M., INGERSOLL, T. E., COLEMAN, J. T. H., THOGMARTIN, W. E. et al. (2015). A plan for the North American Bat Monitoring Program (NABat). Technical Report No. SRS-208, USDA Forest Service.

MILLER, B. W. (2010). Revised relative abundance estimates and temporal activity of bats at three Great Lakes national parks based on acoustic data. Technical Report, U.S. National Park Service.

NAKAGAWA, S. (2015). Missing data: Mechanisms, methods and messages. In *Ecological Statistics: Contemporary Theory and Application* (G. A. Fox, S. Negrete-Yankelevich and V. J. Sosa, eds.) **10** 81–105. Oxford Academic.

ORAM, J. K., BANNER, K. M., STRATTON, C., HOEGH, A. and IRVINE, K. M. (2025). Supplement to "Leveraging an observed-data likelihood improves the use of machine learning labels in a Bayesian hierarchical model for bioacoustic data." https://doi.org/10.1214/25-AOAS2096SUPPA, https://doi.org/10.1214/25-AOAS2096SUPPB, https://doi.org/10.1214/25-AOAS2096SUPPC

PARK, T. and LEE, S. (2022). Improving the Gibbs sampler. *Wiley Interdiscip. Rev.: Comput. Stat.* **14** Paper No. e1546, 11. MR4396896 https://doi.org/10.1002/wics.1546

R CORE TEAM (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

RABE-HESKETH, S. and SKRONDAL, A. (2023). Ignoring non-ignorable missingness. *Psychometrika* **88** 31–50. MR4554871 https://doi.org/10.1007/s11336-022-09895-1

RAINFORTH, T., FOSTER, A., IVANOVA, D. R. and BICKFORD SMITH, F. (2024). Modern Bayesian experimental design. *Statist. Sci.* **39** 100–114. MR4718529 https://doi.org/10.1214/23-sts915

REICHERT, B., LAUSEN, C., LOEB, S., WELLER, T., ALLEN, R., BRITZKE, E., HOHOFF, T., SIEMERS, J., BURKHOLDER, B. et al. (2018). A guide to processing bat acoustic data for the North American Bat Monitoring Program (NABat) Technical Report No. 2018-1068, United States Geologic Survey. https://doi.org/10.3133/ofr20181068

REICHERT, B. E., BAYLESS, M., CHENG, T. L., COLEMAN, J. T. H., FRANCIS, C. M., FRICK, W. F., GOTTHOLD, B. S., IRVINE, K. M., LAUSEN, C. et al. (2021). NABat: A top-down, bottom-up solution to collaborative continental-scale monitoring. *Ambio* **50** 901–913.

RHINEHART, T. A., TUREK, D. and KITZES, J. (2022). A continuous-score occupancy model that incorporates uncertain machine learning output from autonomous biodiversity surveys. *Methods Ecol. Evol.* **13** 1778–1789. https://doi.org/10.1111/2041-210X.13905

RODRIGUEZ, R. M., RODHOUSE, T. J., BARNETT, J., IRVINE, K., BANNER, K. M., LONNEKER, J. and ORMSBEE, P. C. (2019). North American Bat Monitoring Program regional protocol for surveying with stationary deployments of echolocation recording devices: Narrative version 1.0, Pacific Northwestern U.S. Technical Report, Human & Ecosystem Resiliency & Sustainability Lab's Northwestern Bat Hub, Oregon State Univ.—Cascades.

ROTA, C. T., FERREIRA, M. A. R., KAYS, R. W., FORRESTER, T. D., KALIES, E. L., MCSHEA, W. J., PARSONS, A. W. and MILLSPAUGH, J. (2016). A multispecies occupancy model for two or more interacting species. *Methods Ecol. Evol.* **7** 1164–1173. https://doi.org/10.1111/2041-210X.12587

ROYLE, J. A. and LINK, W. A. (2006). Generalized site occupancy models allowing for false positive and false negative errors. *Ecology* **87**. 835–841. https://doi.org/10.1890/0012-9658(2006)87[835:GSOMAF]2.0.CO;2

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. With comments by R. J. A. Little and a reply by the author. MR0455196 https://doi.org/10.1093/biomet/63.3.581

RUIZ-GUTIERREZ, V., HOOTEN, M. B. and CAMPBELL GRANT, E. H. (2016). Uncertainty in biological monitoring: A framework for data collection and analysis to account for multiple sources of sampling bias. *Methods Ecol. Evol.* **7** 900–909. https://doi.org/10.1111/2041-210X.12542

RUSSO, D. and VOIGT, C. C. (2016). The use of automated identification of bat echolocation calls in acoustic monitoring: A cautionary note for a sound analysis. *Ecol. Indicators* **66** 598–602. https://doi.org/10.1016/j.ecolind.2016.02.036

RYDELL, J., NYMAN, S., EKLÖF, J., JONES, G. and RUSSO, D. (2017). Testing the performances of automated identification of bat echolocation calls: A request for prudence. *Ecol. Indicators* **78** 416–420. https://doi.org/10.1016/j.ecolind.2017.03.023

SCHAFER, J. L. (1997). *Analysis of Incomplete Multivariate Data. Monographs on Statistics and Applied Probability* **72**. CRC Press, London. MR1692799 https://doi.org/10.1201/9781439821862

SONG, R., MCDAVID HARRISON, K., HANSON, D. L. and HALL, H. I. (2009). Correction of bias in imputing missing values of categorical variables. *Comm. Statist. Theory Methods* **39** 350–362. MR2654883 https://doi.org/10.1080/03610920902750061

SONOBAT (2019). Sonobat software. Ver. 3. Arcata, CA.

SPIERS, A. I., ROYLE, J. A., TORRENS, C. L. and JOSEPH, M. B. (2022). Estimating species misclassification with occupancy dynamics and encounter rates: A semi-supervised, individual-level approach. *Methods Ecol. Evol.* **13** 1528–1539. https://doi.org/10.1111/2041-210X.13858

STEVENS, D. L. JR. and OLSEN, A. R. (2004). Spatially balanced sampling of natural resources. *J. Amer. Statist. Assoc.* **99** 262–278. MR2061889 https://doi.org/10.1198/016214504000000250

STIFFLER, L. L., ANDERSON, J. T. and KATZNER, T. E. (2018). Occupancy modeling of autonomously recorded vocalizations to predict distribution of rallids in tidal wetlands. *Wetlands* **38** 605–612. https://doi.org/10.1007/s13157-018-1003-z

STRATTON, C., IRVINE, K. M., BANNER, K. M., ALMBERG, E. S., BACHEN, D. and SMUCKER, K. (2025). Joint spatial modeling bridges the gap between disparate disease surveillance and population monitoring efforts informing conservation of at-risk bat species. *J. Agric. Biol. Environ. Stat.* **30** 120–145. MR4875745 https://doi.org/10.1007/s13253-023-00593-8

STRATTON, C., IRVINE, K. M., BANNER, K. M., WRIGHT, W. J., LAUSEN, C. and RAE, J. (2022). Coupling validation effort with in situ bioacoustic data improves estimating relative activity and occupancy for multi-

ple species with cross-species misclassifications. *Methods Ecol. Evol.* **13** 1288–1303. https://doi.org/10.1111/2041-210X.13831

STROUD, J. T., BUSH, M. R., LADD, M. C., NOWICKI, R. J., SHANTZ, A. A. and SWEATMAN, J. (2015). Is a community still a community? Reviewing definitions of key terms in community ecology. *Ecol. Evol.* **5** 4757–4765.

SULLIVAN, B. L., WOOD, C. L., ILIFF, M. J., BONNEY, R. E., FINK, D. and KELLING, S. (2009). eBird: A citizen-based bird observation network in the biological sciences. *Biol. Conserv.* **142** 2282–2292. https://doi.org/10.1016/j.biocon.2009.05.006

VATS, D. and KNUDSON, C. (2021). Revisiting the Gelman-Rubin diagnostic. *Statist. Sci.* **36** 518–529. MR4323050 https://doi.org/10.1214/20-sts812

VEHTARI, A., GELMAN, A., SIMPSON, D., CARPENTER, B. and BÜRKNER, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC (with discussion). *Bayesian Anal.* **16** 667–718. Includes comments and discussions by seven discussants and a rejoinder by the authors. MR4298989 https://doi.org/10.1214/20-ba1221

WALTERS, C. L., FREEMAN, R., COLLEN, A., DIETZ, C., BROCK FENTON, M., JONES, G., OBRIST, M. K., PUECHMAILLE, S. J., SATTLER, T. et al. (2012). A continental-scale tool for acoustic identification of European bats. *J. Appl. Ecol.* **49** 1064–1074. https://doi.org/10.1111/j.1365-2664.2012.02182.x

WRIGHT, W. J., IRVINE, K. M., ALMBERG, E. S., LITT, A. R. and YOCCOZ, N. (2020). Modelling misclassification in multi-species acoustic data when estimating occupancy and relative activity. *Methods Ecol. Evol.* **11** 71–81.

YACKULIC, C. B., DODRILL, M., DZUL, M., SANDERLIN, J. S. and REID, J. A. (2020). A need for speed in Bayesian population models: A practical guide to marginalizing and recovering discrete latent states. *Ecol. Appl.* **30** e02112. https://doi.org/10.1002/eap.2112

ZACHMANN, L. J., BORGMAN, E. M., WITWICKI, D. L., SWAN, M. C., MCINTYRE, C. and HOBBS, N. T. (2022). Bayesian models for analysis of inventory and monitoring data with non-ignorable missingness. *J. Agric. Biol. Environ. Stat.* **27** 125–148. MR4391026 https://doi.org/10.1007/s13253-021-00473-z

ZHONG, M., TAYLOR, R., BATES, N., CHRISTEY, D., BASNET, H., FLIPPIN, J., PALKOVITZ, S., DODHIA, R. and LAVISTA FERRES, J. (2021). Acoustic detection of regionally rare bird species through deep convolutional neural networks. *Ecol. Inform.* **64** 101333. https://doi.org/10.1016/j.ecoinf.2021.101333

ZHU, M., WANG, J., YANG, X., ZHANG, Y., ZHANG, L., REN, H., WU, B. and YE, L. (2022). A review of the application of machine learning in water quality evaluation. *Eco-Environ. Health* **1** 107–116. https://doi.org/10.1016/j.eehl.2022.06.001