Targeted Review for AI-Assisted Biodiversity Surveys: Active Continuous-Score Occupancy Modeling

Timm Haucke
MIT
haucke@mit.edu

Lauren Harrell
Google Research

Justin Kay
MIT

Mary Clapp
The Institute for Bird Populations

Sara Beery
MIT

Abstract

We increasingly use machine learning to label scientific datasets. The models we develop and deploy are improving all the time, but they are not and will likely never be perfect. Mistakes matter, as errors can propagate into our scientific understanding, particularly when systematically biased. Very reasonably, scientists thus review *substantial* proportions of ML-generated labels to verify or correct mistakes in pursuit of ensuring their scientific findings are not biased by ML. In this work, we focus on helping scientists allocate review effort efficiently and in alignment with their scientific goals. We focus on a specific class of scientists (ecologists) and a specific, widespread, and impactful modeling target (occupancy modeling, which estimates where species are likely to occur, conditioned on environmental factors). We introduce **A**ctive **Co**ntinuous-Sco**r**e Occupancy Model**i**n**g** (ACORN), a method that incorporates ML predictions into occupancy models and strategically selects samples for expert review by expected information gain for downstream ecological analysis. Across camera-trap and bioacoustic datasets, our method recovers ecological conclusions close to those obtained from fully human-labeled data, while requiring substantially fewer expert reviews than non-targeted review policies. Our results suggest that ML-assisted scientific workflows should optimize expert effort for downstream inference, rather than for classifier accuracy alone, especially when human review budget is limited. Our code is available at https://github.com/timmh/acorn

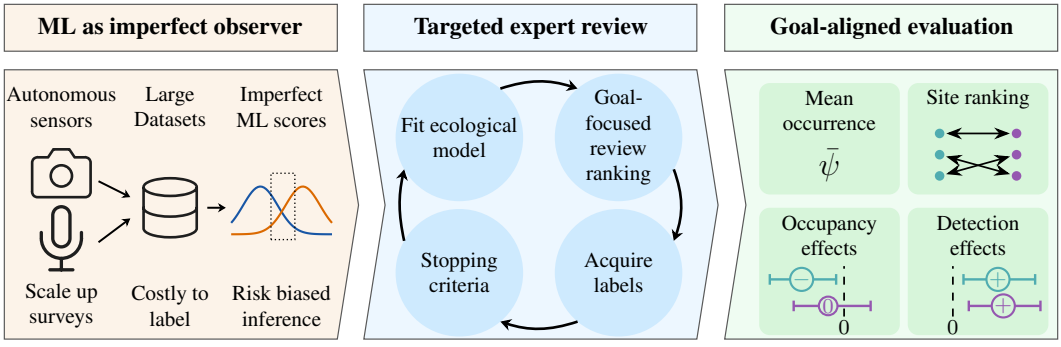


Figure 1: ACORN turns imperfect ML classifications into targeted ecological inference. Large-scale camera-trap and bioacoustic surveys produce datasets that are costly to label exhaustively, but naively relying on ML predictions can bias occupancy estimates and covariate conclusions. ACORN combines continuous classifier scores with a hierarchical occupancy model, actively prioritizes expert review using goal-directed Bayesian experimental design, and evaluates progress using ecological quantities that matter for downstream decision-making: occurrence probabilities, site rankings, and occupancy and detection effects.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

Machine learning (ML) is rapidly changing how ecological data is processed [62]. Data from camera traps and passive acoustic sensors, which passively collect images and audio *in situ*, are increasingly collected at such massive scales that they are often infeasible to review manually. ML models can classify species per example (e.g. image or audio window) far quicker than human experts [47, 60, 32]. However, in many ecological studies, the final objective is not per-example classification (e.g. identifying every species in every image of the dataset). Rather, the goal is often to use per-example classifications for population-level inference, such as estimating whether a species is present at a site (*occupancy*), how its detectability varies with survey conditions, and relating environmental factors (covariates) to the probability of a species occurring in a given place [69, 45].

Population-level statistical ecological models typically assume the per-example classifications they are built on are (by construction) correct. However, when we move to ML-generated classifications, this assumption breaks. Small error rates in per-example classification, i.e. even a single missed detection or false positive, can have substantial effects on downstream inference depending on the place, time, and type of mistake, biasing estimates and distorting estimated covariate relationships [39, 55, 57]. As a result, many ecologists are reasonably reluctant to rely on ML-generated observations in ecological workflows [64]. In practice, classifiers are frequently used to filter obvious negatives or otherwise de-prioritize uninformative samples, while experts continue to manually review large fractions of data [57]. This review burden is expensive and slow, limiting the extent to which data collected from large-scale sensor networks translates into timely ecological insights or conservation actions [46, 24].

Instead, we seek to use ML outputs directly in the downstream ecological model while explicitly accounting for classifier error. As an illustrative example we focus on occupancy models, hierarchical latent variable models that disentangle the probability of species presence from the probability of observing it. Because occupancy models facilitate unbiased inference from noisy observations, their use is widespread in ecology and conservation. They are explicitly recognized in IUCN Red List guidance as an appropriate tool for quantitative extinction-risk analysis [31], and have been applied to invasive species monitoring [2], identifying high-priority conservation areas [13], and used to inform conservation management [21]. However, our approach is adaptable to, and could be similarly impactful for, other population-level inference tasks built from ML-generated classifications.

Recent work has shown that it is possible to combine human reviews and ML predictions to estimate occupancy more efficiently than with human reviews alone [53]. But the question remains: *which datapoints are most useful to review?* Building heuristics for prioritizing which reviews should be collected is extremely difficult, since the information each reviewed example provides depends on where and when it was captured, the ecological context, and how redundant it is with respect to previously acquired reviews. Bayesian experimental design [38, 11] provides a natural information-theoretic framework to prioritize the collection of new observations (in our case, human reviews) that maximize information gained about a quantity of interest.

We propose **A**ctive **Co**ntinuous-Sco**r**e Occupancy Model**i**n**g** (ACORN), a method that combines noisy machine learning classifications with targeted review to maximize information gain towards a user-specified ecological inference target, based on Bayesian experimental design. Rather than targeting review to optimize classifier performance, as you would in active learning, our review acquisition policy is constructed to align with the posterior quantities that matter for ecological analysis, such as occupancy parameters, detection parameters, and derived occurrence estimates. We additionally propose a stopping criterion to determine when additional expert effort is no longer worthwhile, motivated by practical deployment needs. We systematically benchmark our method on two representative ecological datasets capturing observations of species across deployed networks of static sensors, one capturing species observations through images and one through bioacoustic recordings. We compare methods that only use verified reviews for inference and methods which additionally make use of ML scores on unverified data, and across both inference methods compare five different review policies. Evaluated across four metrics that measure recovery of ecological conclusions, we find that our method, ACORN which uses ML scores across the data to inform a targeted information gain review policy, generally converges more quickly towards estimates consistent with fully labeled datasets. Overall, this work emphasizes the importance of developing systems that enable scientists to work *with* imperfect models to improve *targeted* scientific understanding.

2 Related work

ML-based species identification and occupancy modeling. ML models can label camera-trap and acoustic biodiversity data at scale, but deployment remains limited by frequent errors related to geographic or visual shifts, rare species, and poorly-calibrated predictions [47, 60, 67, 4, 5, 32, 23]. ML scores are often used to speed up human labeling via thresholding, filtering, or active learning [47, 66, 46, 8, 36, 33, 40, 25]. However, nearly all such pipelines are still classifier-centric: they choose thresholds or queried examples to improve classification performance, domain adaptation, or annotation efficiency, rather than to directly reduce uncertainty in a downstream ecological posterior.

Occupancy modeling with imperfect detection and uncertain labels. Occupancy models are a natural target when moving beyond per-example classification toward population-level inference. They separate latent site occupancy from imperfect detection, allowing a species to be present at a site even if it is not detected during one or more repeated surveys, or "replicates", and they allow occupancy and detection probabilities to depend on ecological covariates [39]. Classical occupancy models account for false negatives through imperfect detection, but ML-derived observations also introduce false positives. Ignoring these additional error processes can materially bias inference [55, 41, 18, 12, 68], though occupancy inference has been shown to be robust to classifier errors in some settings [6, 61, 34]. Recently, a joint continuous-score model was proposed which directly models ML score distributions conditional on latent detection states, optionally anchored by human labels [53]. This retains more information than thresholding and provides a probabilistic interface between ML outputs and ecological inference. However, that framework does not address the practical allocation problem created by limited expert review: which samples should be prioritized by experts, and when can review stop? Prior design work for occupancy studies has optimized field visits and survey schedules [39, 27]. In contrast, we treat expert verification itself as the design problem: given an existing set of ML-scored observations, which sample should be reviewed next?

Prediction-assisted inference under scarce labels. Methods that use ML predictions to inform downstream statistical inference with few verified labels include post-prediction inference [65], semi-supervised and surrogate-outcome methods [26], and prediction-powered inference (PPI), which combines a small labeled sample with many ML predictions to obtain tighter confidence intervals for low-dimensional population parameters [1]. Standard PPI formulations are typically fixed-design: labeled and unlabeled samples are representative draws from the target population, and labels serve as a correction or calibration sample. Follow-up work extends this paradigm to actively collected labels [70]. Our approach is complementary but distinct: labels are acquired intentionally non-representatively to reduce uncertainty in an entire ecological posterior.

Bayesian experimental design, Bayesian optimal design, and Bayesian active learning. Bayesian experimental design underpins our decision-theoretic approach for active review, prioritizing experiments, measurements, or labels by maximizing expected utility under the current posterior and predictive distribution [38, 11, 48, 56, 52]. In the canonical information-theoretic formulation, the utility is expected information gain, or expected posterior entropy reduction. Classical works distinguish parameter-, prediction-, and decision-focused utilities, which matters when the target is not classifier accuracy but an ecological summary [11]. Bayesian active learning methods such as BALD select labels informative about the model posterior [30]. Bayesian experimental design underlies related work in active testing and model selection [37, 35]. We adapt this view to biodiversity surveys: the experiment is asking an expert to verify one ML-scored replicate, the outcome is a binary detection / non-detection label, and the utility targets ecological conclusions rather than classifier performance.

3 Methods

Active **Co**ntinuous-Sco**r**e Occupancy Model**i**n**g** treats expert review prioritization for ML-supported species occupancy inference as an iterative Bayesian experimental design problem. Two main design components of ACORN are the ecological model (Section 3.1) and the review policy (Section 3.2).

3.1 The ecological model

The classic Bernoulli occupancy model [39] models latent occupancy of site $i \in S$ as

$$z_i \sim \text{Bernoulli}(\psi_i) \tag{1}$$

where ψ_i is the probability of occupancy and is modeled using logistic regression:

$$\log\left(\frac{\psi_i}{1-\psi_i}\right) = \beta_0 + X_i^\top \beta \tag{2}$$

To account for the fact that species can be present even if unobserved, we further model a detection probability p_{ij} for replicate $j \in \{1, ..., J_i\}$:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_0 + W_{ij}^\top \alpha \tag{3}$$

Occupancy covariates X_i vary per site and include environmental variables like average temperature or elevation. Detection covariates W_{ij} vary per site and per replicate and include factors that influence how easy a species is to detect, such as time of day. The full set of covariates used across acoustic and camera trap models can be found in Appendix A.

Expert-verified binary labels f_{ij} indicate whether the focal species was detected at site i and replicate j. These detections are then modeled conditional on the occupancy status:

$$f_{ij}|z_i \sim \text{Bernoulli}(z_i p_{ij}) \tag{4}$$

Our decoupled continuous-score occupancy model.

We now introduce our decoupled continuous-score occupancy model, which builds upon the classic Bernoulli occupancy model to incorporate expert-reviewed observations and upon the joint continuous-score occupancy model [53] to utilize noisy ML classifier scores. To do so, we specify a score model based on two class-conditional Normal densities,

$$s_{ij} \mid f_{ij} = 0 \sim \mathcal{N}(\mu_0, \sigma_0^2), \qquad s_{ij} \mid f_{ij} = 1 \sim \mathcal{N}(\mu_1, \sigma_1^2), \qquad \mu_1 > \mu_0, \tag{5}$$

where the lower component corresponds to observations without the focal species and the upper component to replicates containing the focal species.

The original joint formulation [53] estimates the score parameters $(\mu_0, \sigma_0, \mu_1, \sigma_1)$ and the ecological parameters in one posterior. This is statistically appealing when the Normal score model is well specified, because occupancy structure and score separation can inform one another. In our data, however, the score distribution is often much better identified by the marginal classifier-score distribution and the reviewed labels than by the ecological process itself. If the score mixture is misspecified, joint estimation can therefore create a problematic feedback loop: the occupancy and detection model can pull the score components toward a configuration that explains the spatial pattern of scores, while the distorted score components then feed back into occupancy inference.

We therefore use a "decoupled" version of the Normal score model. At each review round, we first fit the two-component score model to the finite classifier scores and the currently reviewed labels. Unreviewed scores contribute through the marginal mixture likelihood, while reviewed scores are assigned to the component indicated by the expert label. This produces a posterior $q^{(t)}(\eta)$ over the score-calibration parameters $\eta = (\pi, \mu_0, \sigma_0, \mu_1, \sigma_1)$, where π is the marginal prevalence of true detections in the score distribution. Crucially, this posterior is conditioned only on scores and reviewed labels, not on the occupancy likelihood. The ecological model subsequently receives the resulting score evidence, but it cannot update the calibration parameters. In this sense, information flows from the score calibration into the occupancy model, while there is no feedback from occupancy into calibration.

For each score, the calibrated evidence is summarized as a log Bayes factor

$$\ell_{ij} = \log \mathbb{E}_{\eta\sim q^{(t)}}[p(s_{ij} \mid f_{ij} = 1, \eta)] - \log \mathbb{E}_{\eta\sim q^{(t)}}[p(s_{ij} \mid f_{ij} = 0, \eta)]\,. \tag{6}$$

A value $\ell_{ij} = 0$ is neutral evidence, $\ell_{ij} > 0$ means the score is more typical of a true detection, and $\ell_{ij} < 0$ means the score is more typical of a non-detection. For acquisition functions, we also retain a small ensemble of calibration draws $\ell_{ij}^{(r)}$ so that uncertainty in score calibration contributes to uncertainty about candidate labels.

Conditional on an occupied site, an unreviewed score adds directly to the log odds that replicate j contains the focal species:

$$\log\left(\frac{\Pr(f_{ij} = 1 \mid z_i = 1, s_{ij})}{1 - \Pr(f_{ij} = 1 \mid z_i = 1, s_{ij})}\right) = \log\left(\frac{p_{ij}}{1-p_{ij}}\right) + \ell_{ij}. \tag{7}$$

Equivalently, an unreviewed score contributes the replicate-level occupancy evidence $(1 - p_{ij}) + p_{ij} \exp(\ell_{ij})$ after the common negative-score density has been absorbed into the normalizing constant. In other words, a high score is not treated as ultimate proof that the site is occupied, rather, it is treated as evidence that this particular replicate may contain the species, which in turn makes site occupancy more plausible. When a replicate is reviewed, we use the expert label directly in the same Bernoulli detection model as above. Intuitively, this replaces the probabilistic evidence of the score with definite evidence of the reviewed human label. The reviewed score remains useful for calibrating future scores, but once the label is available, the score is not also counted as separate evidence about whether the site is occupied.

3.2 The review policy

We interpret choosing the next replicate to review as a Bayesian experimental design problem. Let a denote a candidate review action and let θ denote the set of model parameters and latent variables. A natural utility is the expected information gain

$$\mathrm{EIG}(a) = H\Big[p(\theta \mid \mathcal{D}^{(t)})\Big] - \mathbb{E}_{f_a \sim p(f_a \mid \mathcal{D}^{(t)})}\Big[H\Big[p(\theta \mid \mathcal{D}^{(t)}, f_a)\Big]\Big], \tag{8}$$

where $\mathcal{D}^{(t)}$ denotes all information available after t review rounds. This criterion asks which review is expected to reduce posterior uncertainty across the entire posterior the most.

However, ecological analyses often care about a more specific set of quantities rather than the full posterior. **We therefore propose a targeted expected information gain criterion (which we call Target EIG)** that focuses on the specific set of quantities we care about. We define a target vector $\phi = g(\theta)$ containing the site-level occupancy probabilities and the occupancy and detection regression coefficients, and score a candidate by

$$\mathrm{TargetEIG}(a) = H\Big[p(\phi \mid \mathcal{D}^{(t)})\Big] - \mathbb{E}_{f_a \sim p(f_a \mid \mathcal{D}^{(t)})}\Big[H\Big[p(\phi \mid \mathcal{D}^{(t)}, f_a)\Big]\Big]. \tag{9}$$

Thus, Target EIG prioritizes reviews expected to reduce uncertainty in the ecological quantities evaluated in Section 4.3. We approximate it with posterior draws by estimating the entropy of ϕ with a Gaussian approximation. For each candidate review, the hypothetical positive and negative outcomes reweight the existing posterior draws according to the candidate-specific posterior probabilities $\Pr(f_a = 1 \mid \theta, \mathcal{D}^{(t)})$ and $\Pr(f_a = 0 \mid \theta, \mathcal{D}^{(t)})$, avoiding a full model refit for every possible label. Because ϕ includes many site-level occupancy probabilities, we compute this entropy after standardizing the target summaries and projecting them to a low-dimensional principal-component representation that preserves most posterior variation. To anchor the positive score distribution, Target EIG and BALD initially select max-score batches, switching to their respective acquisition criteria once at least one positive label has been verified. The resulting policy is greedy, batched and uses approximate utilities. As a consequence, it has no guarantee of global optimality over a finite review budget.

4 Experiments

4.1 Data

We evaluate on one dataset of passive-acoustic recordings and one camera-trap dataset, using the same downstream review loop in both settings. **Acoustic:** We use the acoustic forest soundscape dataset [49] together with Perch v2 [63] outputs. **Camera trap:** We use the iWildCam 2022 camera-trap dataset [3] together with SpeciesNet [22] outputs. Each camera trap is a site, and each day of collected data represents a replicate. Both datasets are represented as single-species, single-season occupancy problems. We fit individual models across the 50 most prevalent bird species in the acoustic data and across the 25 most prevalent species in the camera trap data. This eliminates a long tail of extremely rare species that are difficult to model, even with complete labels, and thus difficult to evaluate. The acoustic and camera-trap experiments otherwise differ only in how data is filtered, how replicate-level scores and labels are constructed, and which covariates are used. More details can be found in Appendix A. Considerations and practical guidance for applying ACORN to other datasets is provided in Appendix I.

4.2 Baselines

Ecological models. We compare our proposed decoupled continuous-score occupancy model to two baseline models: the original Bernoulli model fit only using reviewed labels and no ML classifier scores (Section 3.1), and the original joint continuous-score model [53].

Review policies. We compare our proposed Target EIG policy to another information-theoretic review policy: BALD [30], a computationally simpler equivalent to the expected information gain (Section 3.2) that does not require estimating posterior distributions for each candidate review:

$$\mathrm{BALD}(a) = I(f_a; \theta \mid \mathcal{D}^{(t)}) = H\Big[p(f_a \mid \mathcal{D}^{(t)})\Big] - \mathbb{E}_{\theta \sim p(\theta \mid \mathcal{D}^{(t)})}\Big[H\Big[p(f_a \mid \theta, \mathcal{D}^{(t)})\Big]\Big], \tag{10}$$

where H denotes Shannon entropy. It favors observations whose reviewed label is both uncertain under the full current posterior and diagnostically useful for reducing that uncertainty.

We furthermore compare with three simpler baseline policies. *Random* review samples uniformly across reviewable replicates without replacement. *Max-score* selects the reviewable replicates with maximum classifier scores and is a heuristic commonly deployed in ecological practice (since positive observations are frequently rare and thus particularly valuable). *Posterior predictive uncertainty* uses

$$\mathrm{PPU}(a) = H\Big[p(f_a \mid \mathcal{D}^{(t)})\Big] \tag{11}$$

and selects the m reviewable replicates with largest utility. This captures ambiguity in the next reviewed label without explicitly asking whether that ambiguity is informative about the posterior.

4.3 Evaluation

Our comparison target for models fit on partially reviewed data is a Bernoulli occupancy model (Section 3.1) fit on all ground-truth dataset labels and without ML scores (the *oracle*). In our evaluation, we focus on metrics that directly evaluate the agreement in ecological conclusions with the oracle. We denote posteriors after t reviews using superscripts (t) and the corresponding oracle posteriors using superscripts of O. For site i, let $\bar{\psi}_i^{(t)}$ and $\bar{\psi}_i^{O}$ be the posterior mean occupancy probabilities, let S be the set of sites, and write $\bar{\psi}_S^{(t)}$ and $\bar{\psi}_S^{O}$ for the corresponding vectors over S. For a coefficient ω, define the 90% credible-interval conclusion using the quantile function q:

$$c(\omega) = \begin{cases} +, & q_{0.05}(\omega) > 0, \\ -, & q_{0.95}(\omega) < 0, \\ 0, & \text{otherwise}, \end{cases} \tag{12}$$

where $+$, $-$, and 0 denote positive, negative, and uncertain effects, respectively.

Site-level recovery is measured both on the probability scale and by the induced ranking of sites:

$$A_\psi = 1 - \frac{\|\bar{\psi}^{(t)} - \bar{\psi}^{O}\|_1}{|S|}, \qquad \rho_\psi = \rho\left(R\left[\bar{\psi}^{(t)}\right], R\left[\bar{\psi}^{O}\right]\right) \tag{13}$$

The occupancy agreement A_ψ compares posterior mean occupancy probabilities directly to the oracle and thus assesses overall prevalence, while the site-rank Spearman correlation ρ_ψ asks whether the current posterior ranks high- and low-occupancy sites in the same order as the oracle and thus has implications for prioritizing sites for conservation and restoration. For covariate conclusions, we compute the fraction of coefficients with the same positive, negative, or uncertain conclusion as the oracle, separately for occupancy and detection covariates:

$$A_\beta = \mathrm{mean}\left(\mathbf{1}\Big\{c(\beta^{(t)}) = c(\beta^{O})\Big\}\right), \qquad A_\alpha = \mathrm{mean}\left(\mathbf{1}\Big\{c(\alpha^{(t)}) = c(\alpha^{O})\Big\}\right) \tag{14}$$

We furthermore denote the average of $A_\psi,\ \rho_\psi,\ A_\beta,\ A_\alpha$ as $\overline{A}$.

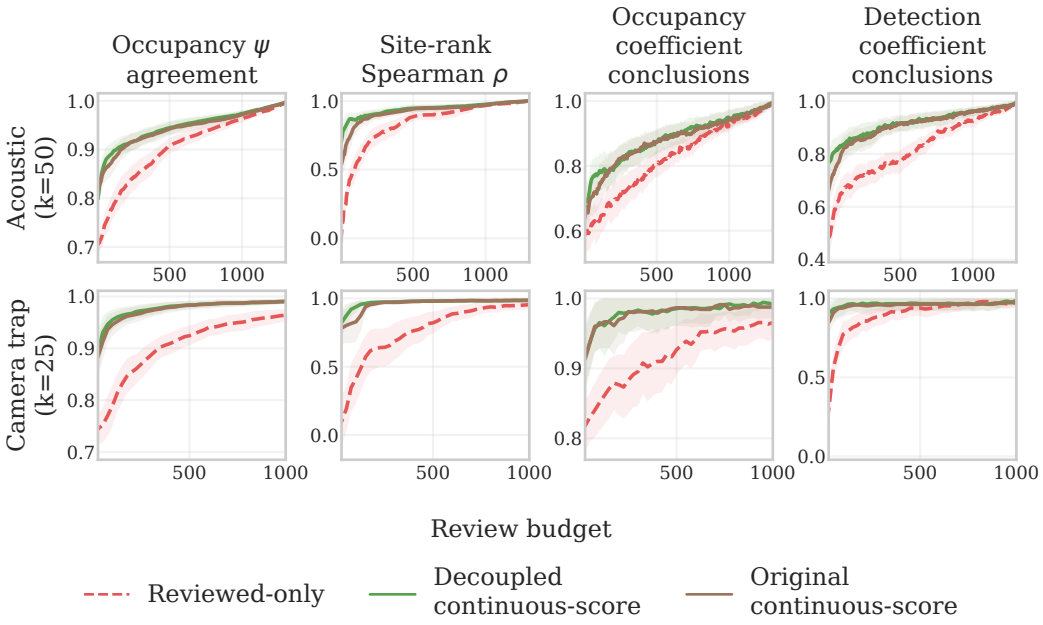


Figure 2: **Using ML scores as noisy evidence enables continuous-score models to start off closer to the ecological conclusions made by the oracle model, and converge more quickly.** The x-axis shows the total number of reviews, while the y-axis shows agreement with the oracle averaged over review policies and species (higher is better). Shaded bands show pointwise 95% normal-approximation confidence intervals (CIs) across species, after averaging seeds and then shared review policies (BALD and Target EIG) within each species. Acoustic reviews go up to 100% of available data, while camera trap reviews are capped at 1,000 reviews.

4.4 Stopping criteria

We perform a post-hoc analysis of three stopping criteria on the Target EIG review trajectories. For a candidate stopping point t, we compute regret separately for the four oracle-agreement metrics (Section 4.3). The plotted regret is the mean difference between the best value achieved by that trajectory and the value at t, averaged over these four metrics.

We evaluate each criterion at different operating points by sweeping over a set of criterion-specific thresholds. Let $\bar{R}^{(t)}$ be the MCMC $\hat{R}$ diagnostic at step t, $G^{(t)}$ be the maximum acquisition score among reviewable samples, and $\Delta^{(t)}$ be the mean posterior distance from the previous review step. The *MCMC diagnostic* stops at the first step with $\bar{R}^{(t)} \leq \tau_R$. The *expected-gain* criterion stops after two consecutive steps with $G^{(t)}/G_0 \leq \tau_G$. The *posterior-shift* criterion stops after two consecutive steps with $\Delta^{(t)}/\Delta_0 \leq \tau_\Delta$.

4.5 Review loop

At each review step t, we select m replicates for review according to the review policy. We use $m = 5$ for the acoustic and $m = 25$ for the camera trap data, balancing computational cost and update frequency (Appendix D). We reveal the true labels corresponding to the m replicates to simulate review and refit the occupancy model with this added information, yielding a new posterior at $t + m$. We run each review trajectory with three random seeds per species and configuration.

5 Results and discussion

5.1 The positive impact of incorporating ML scores

The decoupled continuous-score model (which benefits from information contained in classifier scores) begins much closer to the oracle posterior on average across species than the reviewed-only

Table 1: **ACORN reaches oracle agreement thresholds with few reviews.** Median number of reviews required per species to reach dataset-specific agreement thresholds. Results of the best-performing method(s) are bolded. Thresholds are shown beside each metric name and are higher for camera-trap datasets, where lower thresholds are frequently reached at zero reviews by continuous-score models. The speedup column reports the decoupled continuous-score max-score median review count divided by the ACORN median review count.

	Reviewed-only Bernoulli				Decoupled continuous-score				
Metric	Random	Max-score	BALD	Target EIG	Random	Max-score	BALD	ACORN (ours)	Speedup vs. max-score
Acoustic									
$A_\psi \geq 0.90$	340	125	295	205	305	130	65	**35**	3.7×
$\rho_\psi \geq 0.90$	365	120	210	175	100	25	20	**15**	1.7×
$A_\beta = 1$	565	180	325	265	400	175	105	**60**	2.9×
$A_\alpha = 1$	525	695	370	615	165	35	**25**	45	0.8×
Camera trap									
$A_\psi \geq 0.95$	825	750	425	325	375	125	75	**50**	2.5×
$\rho_\psi \geq 0.95$	600	400	250	300	200	75	**50**	**50**	1.5×
$A_\beta = 1$	275	25	225	175	**0**	**0**	**0**	**0**	–
$A_\alpha = 1$	100	850	50	100	**0**	**0**	**0**	**0**	–

Bernoulli model (Figure 2). Even before reviews, mean agreement across the four ecological-conclusion metrics (Section 4.3) increases from 0.441 to 0.735 for acoustic and from 0.422 to 0.886 for camera traps when comparing the reviewed-only model to the decoupled continuous-score model. This gain enables our targeted review to start from a stronger posterior. Interestingly, early oracle agreement is only moderately associated with classifier quality, suggesting that even imperfect predictions can be useful early on. After review, oracle agreement is consistently high irrespective of classifier quality (Appendix H).

How scores enter the ecological model also matters. Within the continuous-score family, the original joint formulation [53] benefits from classifier scores, but is less robust when the score model is misspecified (Section 3.1). The underlying cause for this misspecification is score-distribution heterogeneity: nominal non-detections at occupied sites had higher scores than non-detections at likely unoccupied sites (sites without detections) for the majority of species. Factors that contribute this heterogeneity include label errors (Figure 8) and potential classifier bias. Under joint estimation, this misspecification can create a feedback loop where the occupancy and detection model pull the shared score components toward a configuration that explains spatial score patterns, and the distorted score components then feed back into occupancy inference. Decoupling cuts that feedback by fitting score calibration outside the occupancy posterior and particularly improves oracle agreement at low review counts (Figure 2). This improvement is particularly strong among acoustic species in the top quartile of heterogeneity: agreement improved by 0.115 at zero reviews and 0.059 at 100 reviews, compared with 0.059 and 0.025 among the remaining species. We discuss this phenomenon in detail in Appendix F.

5.2 Targeting information gain in review increases efficiency

When combined with the decoupled continuous-score occupancy model, Target EIG results in the quickest overall convergence to the fully reviewed oracle, particularly in low- and medium-budget regimes (Figure 3). Its advantage over BALD comes from target alignment: BALD favors hard-to-predict labels, but difficult replicates are not necessarily useful for ecological inference. They may be already-resolved, redundant, or unlikely to change the posterior. Target EIG instead scores expected entropy reduction in ecological target quantities, so it can choose lower-uncertainty labels if they better resolve site states, detection processes, or covariate conclusions. Acquisition diagnostics support this interpretation: Target EIG selections are neither simply the highest-BALD nor highest-score reviews, often choosing lower-entropy points than BALD while avoiding many nearly certain high-score detections. Interestingly, focusing EIG on a single task does not consistently improve over the combined objective (Appendix C).

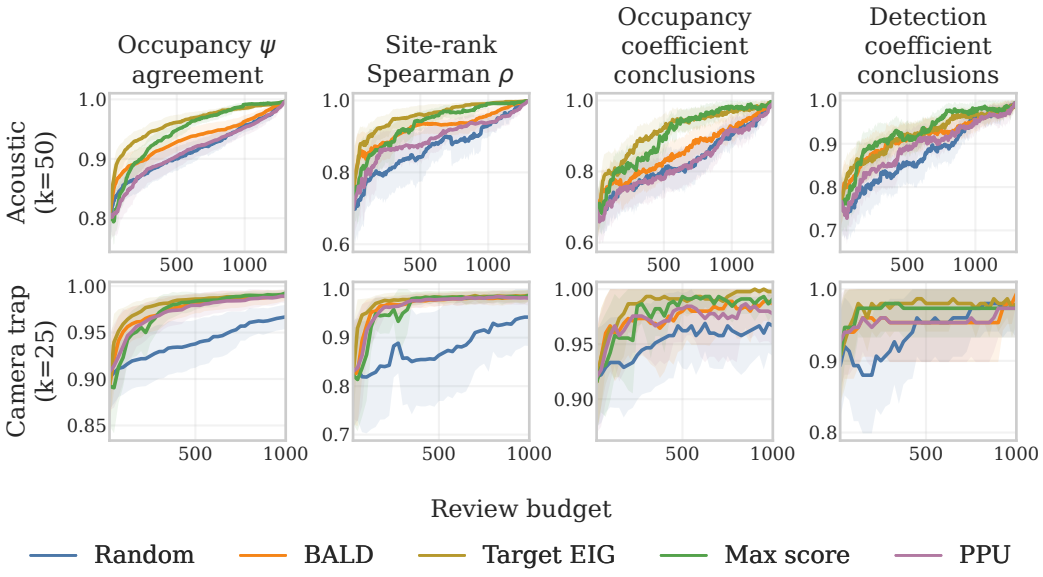


Figure 3: **Target EIG accelerates oracle agreement, especially at low review budgets.** Agreement in ecological conclusions over the number of reviews across review policies. We measure agreement in ecological conclusions by comparing the overall occupancy probability (Eq. (13)), the agreement in ranking of individual sites by occupancy probability (Eq. (13)), the significance and sign of occupancy coefficients (Eq. (14)), and significance and sign of detection coefficients (Eq. (14)). Results are averaged over all species and shaded bands show pointwise 95% normal-approximation CIs across species after averaging seeds within species. Colors indicate the review policy (PPU = posterior predictive uncertainty).

Despite its simplicity, max-score is a surprisingly strong baseline. Confirmed detections are rare in our data, and positive reviews can anchor site occupancy, inform detection probability, and supply signal for covariate effects. Because classifier scores already rank likely detections well, max-score prioritizes many labels that occupancy models need. This effectiveness of positive-focused review is consistent with prior findings in active learning of classification models for conservation [25]. The weakness of max-score review is redundancy: the highest-score labels are often predictable from the continuous-score model and can repeat information across visits or sites. Thus Target EIG leads at small budgets, while max-score catches up, and sometimes overtakes Target EIG on particular thresholds, once positive-label accumulation outweighs redundancy.

Overall, our results imply substantial expert time savings. For each metric in Table 1, ACORN's median review requirement is at most 60 acoustic or 50 camera-trap reviews. The corresponding maxima for max-score review are 175 and 125 reviews, respectively. For acoustic data, where each replicate is a ten-minute recording, ACORN reduces review time from 217.0 hours for all 1,302 recordings to 10.0 hours, saving 207.0 expert-hours per species. Relative to max-score, it saves 115 reviews, or 19.2 hours. This assumes a conservative 1 minute of annotation effort for 1 minute of audio [58]. For camera traps, the iWildCam location-day replicates used in our sweep contain 35.4 images on average. Assuming one expert-hour per 500 images [17], ACORN reduces expected review time from 125.4 hours to 3.5 hours, saving 121.9 expert-hours per species. Relative to max-score, it saves 75 reviews, or 5.3 hours per species. Reviewing only the highest-scoring example can reduce effort further, with near-equal results on camera traps but substantially different results on acoustics (Appendix E).

5.3 Our stopping criterion balances cost and value

To be useful in practice, ACORN needs to know when to stop reviewing. Crucially, we cannot rely on the oracle model, since that requires datasets that are labeled upfront, thereby defeating the purpose of active review. We therefore evaluate three intuitive stopping criteria based on how well they recover most of the benefit of a full review process (Figure 4). The tradeoff curves show that low $\hat{R}$ is best

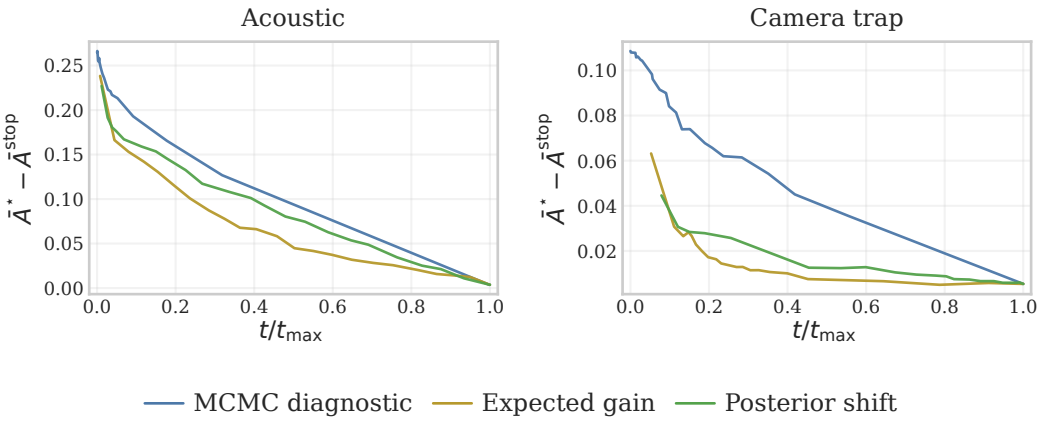


Figure 4: **Expected gain and posterior shift are better stopping criteria than MCMC diagnostics.** Post-hoc stopping-rule tradeoff for the continuous-score Target EIG trajectories. Each line shows, for one stopping criterion and one dataset, the mean fraction of the full review budget used versus agreement regret relative to the best observed step on that same trajectory, averaged across the four agreement metrics used in Figure 3. Criteria farther down and to the left are preferable.

interpreted as a screen for whether a fitted model is usable, not as evidence that additional labels have stopped changing the ecological conclusions. The expected-gain and posterior-shift rules more directly track whether review remains scientifically useful: the former asks whether any unreviewed sample still has substantial acquisition value, while the latter asks whether successive review steps continue to move the ecological posterior.

These results suggest that stopping should be treated as a decision problem rather than as a convergence diagnostic. Conservative thresholds spend more of the review budget but reduce agreement regret, while aggressive thresholds save review effort at the cost of stopping before coefficient and detection conclusions have stabilized. Without an oracle, practitioners can set tolerances for changes in occupancy probabilities, site rankings, and coefficient conclusions across review batches, then choose expected-gain or posterior-shift thresholds accordingly, allowing species-specific stopping counts.

6 Conclusion

We propose **A**ctive **Co**ntinuous-Sco**r**e Occupancy Model**i**n**g** (ACORN), a method that combines a novel occupancy model formulation that robustly incorporates ML classifier scores with a novel information-theoretic active review policy that prioritizes reviews by estimated information gain about ecological quantities of interest. Across a systematic comparison to alternative combinations of ecological models and review policies on datasets covering two ecological data modalities (camera trap images and acoustic recordings), ACORN both provides a much stronger initial starting point and better prioritizes expert reviews, resulting in models that recover oracle-like ecological posteriors with substantially less manual review than unguided strategies.

Limitations and future work. Future work could extend targeted review beyond occupancy to abundance estimation with N-mixture models [54] or distance sampling [9], and population-size or survival estimation with capture–recapture models [44]. Adapting the ecological model and the information gathered by expert review to these goals could support a wider range of conservation questions. Multi-species and spatial models could further enable more efficient review across species and locations. Examining the sensitivity to temporal and spatial aggregation would clarify how survey design affects both ecological inference and review efficiency. Another promising direction is to determine when expert effort is better used for fine-tuning classifiers or directly informing downstream models, and how these uses can be combined adaptively. Finally, convergence guarantees could complement our empirical evidence that ACORN converges to the oracle solution.

Acknowledgements and disclosures

This work was supported in part by the Schwarzman College of Computing fellowship, NSF fellowship award No. 2313998, the AI and Biodiversity Change (ABC) Global Center, which is funded by the US National Science Foundation under Award No. 2330423 and Natural Sciences and Engineering Research Council of Canada under Award No. 585136, NSF CAREER Award No. 2441060, the MIT Social and Ethical Responsibilities of Computing Initiative, and the MIT-Google program for Computing Innovation and Planetary Health. The authors declare no competing interests.

References

[1] Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023.

[2] Alex W Bajcz, Wesley J Glisson, Jeffrey W Doser, Daniel J Larkin, and John R Fieberg. A within-lake occupancy model for starry stonewort, nitellopsis obtusa, to support early detection and monitoring. *Scientific reports*, 14(1):2644, 2024.

[3] Sara Beery, Arushi Agarwal, Elijah Cole, and Vighnesh Birodkar. The iwildcam 2021 competition dataset. *arXiv preprint arXiv:2105.03494*, 2021.

[4] Sara Beery, Grant Van Horn, Oisin Mac Aodha, and Pietro Perona. The iwildcam 2018 challenge dataset. *arXiv preprint arXiv:1904.05986*, 2019.

[5] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018.

[6] Peggy A Bevan, Omiros Pantazis, Holly AI Pringle, Guilherme Braga Ferreira, Daniel J Ingram, Emily K Madsen, Liam Thomas, Dol Raj Thanet, Thakur Silwal, Santosh Rayamajhi, et al. Deep learning-based ecological analysis of camera trap images is impacted by training data quality and quantity. *Remote Sensing in Ecology and Conservation*, 2025.

[7] Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul A. Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep universal probabilistic programming. *J. Mach. Learn. Res.*, 20:28:1–28:6, 2019.

[8] Ludwig Bothmann, Lisa Wimmer, Omid Charrakh, Tobias Weber, Hendrik Edelhoff, Wibke Peters, Hien Nguyen, Caryl Benjamin, and Annette Menzel. Automated wildlife image classification: An active learning tool for ecological applications. *Ecological Informatics*, 77:102231, 2023.

[9] Stephen T Buckland, David R Anderson, Kenneth Paul Burnham, Jeffrey Lee Laake, David Louis Borchers, and Leonard Thomas. *Introduction to distance sampling: estimating abundance of biological populations*. Oxford university press, 2001.

[10] Center For International Earth Science Information Network-CIESIN-Columbia University. Gridded population of the world, version 4 (gpwv4): Population density, revision 11, 2017.

[11] Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical science*, pages 273–304, 1995.

[12] Thierry Chambert, Evan H Campbell Grant, David AW Miller, James D Nichols, Kevin P Mulder, and Adrianne B Brand. Two-species occupancy modelling accounting for species misidentification and non-detection. *Methods in Ecology and Evolution*, 9(6):1468–1477, 2018.

[13] Amielle A De Wan, Patrick J Sullivan, Arthur J Lembo, Charles R Smith, John C Maerz, James P Lassoie, and Milo E Richmond. Using occupancy models of forest breeding birds to prioritize conservation planning. *Biological Conservation*, 142(5):982–991, 2009.

[14] Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

[15] Ralph Dubayah, Michelle Hofton, James Blair, John Armston, Hao Tang, and Scott Luthcke. Gedi l2a elevation and height metrics data global footprint level v002, 2021.

[16] Tom G Farr, Paul A Rosen, Edward Caro, Robert Crippen, Riley Duren, Scott Hensley, Michael Kobrick, Mimi Paller, Ernesto Rodriguez, Ladislav Roth, et al. The shuttle radar topography mission. *Reviews of geophysics*, 45(2), 2007.

[17] Mitchell Fennell, Christopher Beirne, and A Cole Burton. Use of object detection in camera trap image identification: Assessing a method to rapidly and accurately classify human and animal detections for research and application in recreation ecology. *Global Ecology and Conservation*, 35:e02104, 2022.

[18] Paige FB Ferguson, Michael J Conroy, and Jeffrey Hepinstall-Cymerman. Occupancy models for data with false positive and false negative errors and heterogeneity across sites and surveys. *Methods in Ecology and Evolution*, 6(12):1395–1406, 2015.

[19] Stephen E Fick and Robert J Hijmans. Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas. *International journal of climatology*, 37(12):4302–4315, 2017.

[20] Mark Friedl and Damien Sulla-Menashe. Modis/terra+aqua land cover type yearly l3 global 500m sin grid v061, 2022.

[21] Angela K Fuller, Daniel W Linden, and J Andrew Royle. Management decision making for fisher populations informed by occupancy modeling. *The Journal of Wildlife Management*, 80(5):794–802, 2016.

[22] Tomer Gadot, Ștefan Istrate, Hyungwon Kim, Dan Morris, Sara Beery, Tanya Birch, and Jorge Ahumada. To crop or not to crop: Comparing whole-image and cropped classification on a large dataset of camera trap images. *IET Computer Vision*, 18(8):1193–1208, 2024.

[23] Burooj Ghani, Tom Denton, Stefan Kahl, and Holger Klinck. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. *Scientific Reports*, 13(1):22876, 2023.

[24] Rory Gibb, Ella Browning, Paul Glover-Kapfer, and Kate E. Jones. Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. *Methods in Ecology and Evolution*, 10(2):169–185, 2019.

[25] Lucia Gordon, Nikhil Behari, Samuel Collier, Elizabeth Bondi-Kelly, Jackson A. Killian, Catherine Ressijac, Peter Boucher, Andrew Davies, and Milind Tambe. Find rhinos without finding rhinos: active learning with multimodal imagery of south african rhino habitats. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23, 2023.

[26] Jessica L Gronsbell and Tianxi Cai. Semi-supervised approaches to efficient evaluation of model prediction performance. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(3):579–594, 2018.

[27] Gurutzeta Guillera-Arroita, Martin S Ridout, and Byron JT Morgan. Two-stage bayesian study design for species occupancy estimation. *Journal of Agricultural, Biological, and Environmental Statistics*, 19(2):278–291, 2014.

[28] Timm Haucke, Lauren Harrell, and Sara Beery. Biolith: Bayesian ecological modeling in python, March 2026.

[29] Matthew D Hoffman, Andrew Gelman, et al. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

[30] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

[31] IUCN Standards and Petitions Committee. Guidelines for using the IUCN red list categories and criteria. version 16. Technical report, International Union for Conservation of Nature, Gland, Switzerland and Cambridge, UK, March 2024. Prepared by the Standards and Petitions Committee of the IUCN Species Survival Commission.

[32] Stefan Kahl, Connor M Wood, Maximilian Eibl, and Holger Klinck. Birdnet: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61:101236, 2021.

[33] Hannes Kath, Patricia P Serafini, Ivan B Campos, Thiago S Gouvêa, and Daniel Sonntag. Leveraging transfer learning and active learning for data annotation in passive acoustic monitoring of wildlife. *Ecological Informatics*, 82:102710, 2024.

[34] Lydia KD Katsis, Tessa A Rhinehart, Elizabeth Dorgay, Emma E Sanchez, Jake L Snaddon, C Patrick Doncaster, and Justin Kitzes. A comparison of statistical methods for deriving occupancy estimates from machine learning outputs. *Scientific Reports*, 15(1):14700, 2025.

[35] Justin Kay, Grant Van Horn, Subhransu Maji, Daniel Sheldon, and Sara Beery. Consensus-driven active model selection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4594–4604, 2025.

[36] Mahnoosh Kholghi, Yvonne Phillips, Michael Towsey, Laurianne Sitbon, and Paul Roe. Active learning for classifying long-duration audio recordings of the environment. *Methods in Ecology and Evolution*, 9(9):1948–1958, 2018.

[37] Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Thomas Rainforth. Active surrogate estimators: An active learning approach to label-efficient model evaluation. *Advances in Neural Information Processing Systems*, 35:24557–24570, 2022.

[38] Dennis V Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

[39] Darryl I MacKenzie, James D Nichols, Gideon B Lachman, Sam Droege, J Andrew Royle, and Catherine A Langtimm. Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83(8):2248–2255, 2002.

[40] Zhongqi Miao, Ziwei Liu, Kaitlyn M Gaynor, Meredith S Palmer, Stella X Yu, and Wayne M Getz. Iterative human and automated identification of wildlife images. *Nature Machine Intelligence*, 3(10):885–895, 2021.

[41] David A Miller, James D Nichols, Brett T McClintock, Evan H Campbell Grant, Larissa L Bailey, and Linda A Weir. Improving occupancy estimation when two types of observational error occur: Non-detection and species misidentification. *Ecology*, 92(7):1422–1428, 2011.

[42] I. Misztal. Reliable computing in estimation of variance components. *Journal of Animal Breeding and Genetics*, 125(6):363–370, 2008.

[43] Radford M Neal. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, pages 47–95, 2011.

[44] James D. Nichols. Capture-recapture models. *BioScience*, 42(2):94–102, 1992.

[45] James D Nichols and Byron K Williams. Monitoring for conservation. *Trends in ecology & evolution*, 21(12):668–673, 2006.

[46] Mohammad Sadegh Norouzzadeh, Dan Morris, Sara Beery, Neel Joshi, Nebojsa Jojic, and Jeff Clune. A deep active learning system for species identification and counting in camera trap images. *Methods in ecology and evolution*, 12(1):150–161, 2021.

[47] Mohammad Sadegh Norouzzadeh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S Palmer, Craig Packer, and Jeff Clune. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, 115(25):E5716–E5725, 2018.

[48] Mike Oaksford and Nick Chater. Optimal data selection: Revision, review, and reevaluation. *Psychonomic Bulletin & Review*, 10(2):289–318, 2003.

[49] Pooja Panwar, Wyatt J. Cummings, Sharon Martinson, David A. Lutz, Hannah ter Hofstede, Aaron Weed, Matthew P. Ayres, and Laurel B. Symes. Large-scale, stratified, fully annotated acoustic forest soundscape dataset of avian vocalizations from eastern north america, 2025.

[50] Brent S. Pease, Krishna Pacifici, Roland Kays, and Brian Reich. What drives spatially varying ecological relationships in a wide-ranging species? *Diversity and Distributions*, 28(9):1752–1768, 2022.

[51] Du Phan, Neeraj Pradhan, and Martin Jankowiak. Composable effects for flexible and accelerated probabilistic programming in numpyro. *arXiv preprint arXiv:1912.11554*, 2019.

[52] Tom Rainforth, Adam Foster, Desi R Ivanova, and Freddie Bickford Smith. Modern bayesian experimental design. *Statistical Science*, 39(1):100–114, 2024.

[53] Tessa A Rhinehart, Daniel Turek, and Justin Kitzes. A continuous-score occupancy model that incorporates uncertain machine learning output from autonomous biodiversity surveys. *Methods in Ecology and Evolution*, 13(8):1778–1789, 2022.

[54] J Andrew Royle. N-mixture models for estimating population size from spatially replicated counts. *Biometrics*, 60(1):108–115, 2004.

[55] J Andrew Royle and William A Link. Generalized site occupancy models allowing for false positive and false negative errors. *Ecology*, 87(4):835–841, 2006.

[56] Elizabeth G Ryan, Christopher C Drovandi, James M McGree, and Anthony N Pettitt. A review of modern computational algorithms for bayesian optimal design. *International Statistical Review*, 84(1):128–154, 2016.

[57] Simone Santoro, Santiago Gutiérrez-Zapata, Javier Calzada, Nuria Selva, Diego Marín-Santos, Sara Beery, Kate Brandis, Iñaki Fernández de Viana, Paul Meek, Alessio Mortelliti, et al. Essential tools but overlooked bias: Artificial intelligence and citizen science classification affect camera trap data. *Methods in Ecology and Evolution*, 16(11):2638–2652, 2025.

[58] Taylor Shaw, Sina-Rebekka Schönamsgruber, João M Cordeiro Pereira, and Grzegorz Mikusiński. Refining manual annotation effort of acoustic data to estimate bird species richness and composition: The role of duration, intensity, and time. *Ecology and Evolution*, 12(11):e9491, 2022.

[59] US Geological Survey. 3d elevation program 1-meter resolution digital elevation model, 2019.

[60] Michael A Tabak, Mohammad S Norouzzadeh, David W Wolfson, Steven J Sweeney, Kurt C VerCauteren, Nathan P Snow, Joseph M Halseth, Paul A Di Salvo, Jesse S Lewis, Michael D White, et al. Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods in Ecology and Evolution*, 10(4):585–590, 2019.

[61] D. Thornton, D. Morris, T. King, L. Perera-Romero, A. Anderson, R. Garcia-Anleu, S. Fitkin, and C. Vynne. Identification of camera trap images by artificial intelligence and human experts produces similar multi-species occupancy models. *Journal of Applied Ecology*, 2026.

[62] Devis Tuia, Benjamin Kellenberger, Sara Beery, Blair R Costelloe, Silvia Zuffi, Benjamin Risse, Alexander Mathis, Mackenzie W Mathis, Frank Van Langevelde, Tilo Burghardt, et al. Perspectives in machine learning for wildlife conservation. *Nature communications*, 13(1):792, 2022.

[63] Bart van Merriënboer, Vincent Dumoulin, Jenny Hamer, Lauren Harrell, Andrea Burns, and Tom Denton. Perch 2.0: The bittern lesson for bioacoustics, 2026.

[64] Juliana Vélez, William McShea, Hila Shamon, Paula J Castiblanco-Camacho, Michael A Tabak, Carl Chalmers, Paul Fergus, and John Fieberg. An evaluation of platforms for processing camera-trap data using artificial intelligence. *Methods in Ecology and Evolution*, 14(2):459–477, 2023.

[65] Siruo Wang, Tyler H. McCormick, and Jeffrey T. Leek. Methods for correcting inference based on outcomes predicted by machine learning. *Proceedings of the National Academy of Sciences*, 117(48):30266–30275, 2020.

[66] Robin C Whytock, Jędrzej Świeżewski, Joeri A Zwerts, Tadeusz Bara-Słupski, Aurélie Flore Koumba Pambo, Marek Rogala, Laila Bahaa-el din, Kelly Boekee, Stephanie Brittain, Anabelle W Cardoso, et al. Robust ecological analysis of camera trap data labelled by a machine learning model. *Methods in Ecology and Evolution*, 12(6):1080–1092, 2021.

[67] Marco Willi, Ross T. Pitman, Anabelle W. Cardoso, Christina Locke, Alexandra Swanson, Amy Boyer, Marten Veldthuis, and Lucy Fortson. Identifying animal species in camera trap images using deep learning and citizen science. *Methods in Ecology and Evolution*, 10(1):80–91, 2019.

[68] Wilson J. Wright, Kathryn M. Irvine, Emily S. Almberg, and Andrea R. Litt. Modelling misclassification in multi-species acoustic data when estimating occupancy and relative activity. *Methods in Ecology and Evolution*, 11(1):71–81, 2020.

[69] Nigel G Yoccoz, James D Nichols, and Thierry Boulinier. Monitoring of biological diversity in space and time. *Trends in ecology & evolution*, 16(8):446–453, 2001.

[70] Tijana Zrnic and Emmanuel Candes. Active statistical inference. In *Forty-first International Conference on Machine Learning*, 2024.

A Additional details on datasets and covariates

We use different preprocessing pipelines for the acoustic forest soundscape and iWildCam camera-trap data before representing each focal species as a single-season occupancy problem.

Acoustic. We use the acoustic forest soundscape dataset [49] with Perch v2 classifier outputs [63]. This dataset is openly available [49] and licensed under the Creative Commons Attribution 4.0 International license. We treat each recording (typically 10 minutes long) as one occupancy replicate. We keep recordings from the ACAD, MABI, and SIMR subcollections and assign a positive label when the ground truth labels for a file contain the focal species. The classifier score is the maximum target-species Perch v2 logit over an entire recording. Each of the 50 focal bird species is modeled on the same 104 recorder sites and 1,302 reviewable recordings, with at most 28 recordings per site. Across these species and recording labels, there are 8,962 positive detections. The number of per-species detections ranges from 20 to 1,012, with median 116.

The acoustic occupancy model uses three site-level environmental covariates summarized around each recorder: mean tree-canopy cover percentage from MODIS [20], median canopy height from GEDI [15], and mean elevation from USGS 3DEP [59], sampled using Google Earth Engine. The detection covariates are hours since local sunrise, day of year, and binary indicators for the MABI and SIMR subcollections, with ACAD as the reference level. Covariates are median-imputed and z-score normalized.

Camera trap. We use iWildCam 2022 [3] with SpeciesNet outputs [22] for our camera-trap analysis. iWildCam 2022 is available at `https://lila.science/datasets/iwildcam-2022/` and licensed under the Community Data License Agreement (CDLA). We use only the training split, since the test split does not contain ground truth labels. We discard sites that lack geographic locations, and discard images whose capture datetime cannot be parsed. The occupancy site is an iWildCam location and the replicate is a location-day. For each focal species, a location-day is positive when any image from that location-day has the target category, and its classifier score is the maximum SpeciesNet target logit over images from that location-day. We summarize daily effort as the number of unique image sequences on the location-day and use $\log(\max(\text{sequence count}, 1))$ as a detection covariate.

Because sites in iWildCam are distributed across the globe and our goal is to model regional occupancy processes at fine rather than continental scales, we apply species-specific geographic filtering before fitting the occupancy model. We form deterministic connected-component clusters of locations using a 250 km Haversine-distance radius and, for each species, keep only clusters that contain at least one positive location-day. This removes distant negative-only clusters that are unlikely to be informative about the same regional occupancy process as the target detections. The 25 focal camera-trap species have 29 to 98 sites per species (median 42), 692 to 2,365 reviewable location-days per species (median 2,133), and at most 24 to 93 location-days per retained site. Across these species-specific location-day labels there are 44,231 samples and 4,622 positives. The per-species positive count ranges from 41 to 479, with median 151.

The camera-trap occupancy model uses five site-level covariates loosely based on [50]: elevation from SRTM [16], forest cover from MODIS [20], annual mean temperature and annual mean precipitation from WorldClim [19], and mean human population density from GPW 2020 [10], sampled using Google Earth Engine. The camera-trap detection model uses the log daily sequence-count covariate described above as a proxy for sampling effort and trigger sensitivity. Just as for the acoustic data, covariates are median-imputed and z-score normalized.

B Experimental details

B.1 Model fitting

The Bernoulli and continuous-score occupancy models are implemented using the Biolith library [28]. To sample from the posterior, we use the No-U-Turn Sampler (NUTS, [29]), a variant of Hamiltonian Monte Carlo [43, 14], implemented using NumPyro [51, 7]. We use three chains, 500 warmup steps and 500 retained samples per chain. Separate score calibration uses one chain with 100 warmup and 100 retained samples, eight calibration draws for acquisition, and a prevalence prior with mean 0.01 and concentration 100.

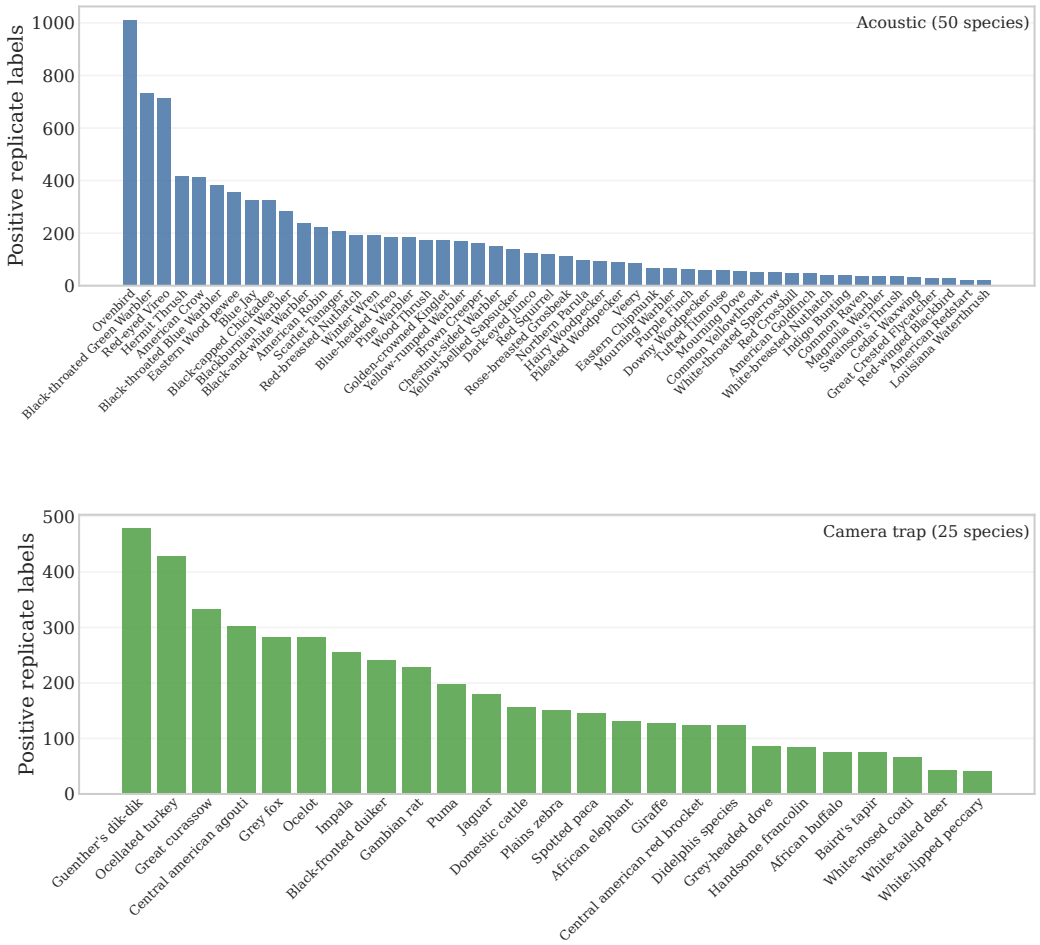


Figure 5: **Class distribution of focal species used in the real-data experiments.** Bars show the number of positive replicate-level labels for the 50 acoustic species and 25 iWildCam camera-trap species used in the main sweep, sorted from most to least common within each dataset. Acoustic replicates are recordings. Camera-trap replicates are location-days after train-split, geocoding, and positive-geographic-cluster filtering.

B.2 Computational complexity

Each active-review iteration has two computationally relevant stages: refitting the current ecological model and ranking candidate reviews. As noted in prior work in reliable statistical computation [42], the choice of estimation algorithm can involve substantial tradeoffs between computational efficiency and reliability. In our setting, the refitting stage dominates due to computationally intensive MCMC sampling. For the decoupled continuous-score model, the reported fitting time combines the occupancy-model fit with the separate score-mixture calibration fit. For the original continuous-score model, the score-mixture parameters are fit jointly inside the occupancy-model sampler. Candidate ranking is much cheaper in comparison. Random and max-score selection require only simple ordering or sampling over reviewable replicates, PPU and BALD additionally average over posterior draws, and Target EIG is the most expensive ranking rule because it recomputes a target-posterior entropy approximation for each candidate label outcome. Even for Target EIG, however, acquisition is a few seconds per iteration and remains smaller than posterior fitting (Table 2).

All experiments were executed without GPUs on a Slurm cluster with one job per species. Each job ran on 16 cores of an x86_64 AMD EPYC 9554 64-Core CPU with 32 GB of memory.

Table 2: Mean wall-clock time per active-review iteration, with three seeds pooled within each species, averaged over species-experiment trajectories after excluding terminal iterations with no selected reviews. Model fitting is the occupancy-model MCMC time plus score-mixture calibration for the decoupled continuous-score model. Ranking is the time used to score and order candidate reviews. Total includes recorded per-iteration overhead in addition to these two components. Entries show mean $\pm$ 95% CI half-width, using the normal approximation across species-experiment trajectory means.

Dataset	Model	Review method	Species	Model fitting (s, mean $\pm$ 95% CI)	Ranking (s, mean $\pm$ 95% CI)	Total (s, mean $\pm$ 95% CI)
Acoustic	Reviewed-only Bernoulli	Random	50	8.38 ± 0.57	< 0.01	8.64 ± 0.58
		Max-score	50	8.74 ± 0.75	< 0.01	9.00 ± 0.75
		PPU	50	10.39 ± 0.92	< 0.01	10.66 ± 0.93
		BALD	50	8.78 ± 0.66	< 0.01	9.05 ± 0.67
		Target EIG	50	8.99 ± 0.47	1.84 ± 0.08	11.10 ± 0.53
	Decoupled continuous-score	Random	50	10.14 ± 0.62	< 0.01	10.84 ± 0.62
		Max-score	50	10.16 ± 0.57	< 0.01	10.85 ± 0.58
		PPU	50	10.11 ± 0.63	< 0.01	10.80 ± 0.64
		BALD	50	10.05 ± 0.60	< 0.01	10.74 ± 0.60
		Target EIG	50	9.26 ± 0.56	2.29 ± 0.08	12.26 ± 0.59
	Original continuous-score	BALD	50	20.72 ± 1.45	< 0.01	21.03 ± 1.46
		Target EIG	50	19.93 ± 1.43	1.98 ± 0.07	22.22 ± 1.46
Camera trap	Reviewed-only Bernoulli	Random	25	6.15 ± 0.63	0.01 ± 0.00	6.40 ± 0.66
		Max-score	25	6.31 ± 0.65	< 0.01	6.50 ± 0.67
		PPU	25	6.56 ± 0.71	< 0.01	6.75 ± 0.73
		BALD	25	6.58 ± 0.80	< 0.01	6.76 ± 0.82
		Target EIG	25	6.46 ± 0.73	1.65 ± 0.09	8.30 ± 0.83
	Decoupled continuous-score	Random	25	8.59 ± 0.85	< 0.01	9.27 ± 0.96
		Max-score	25	8.78 ± 0.85	< 0.01	9.46 ± 0.97
		PPU	25	8.69 ± 0.88	< 0.01	9.35 ± 1.00
		BALD	25	8.71 ± 0.85	< 0.01	9.38 ± 0.96
		Target EIG	25	8.06 ± 0.83	2.28 ± 0.25	10.99 ± 1.14
	Original continuous-score	BALD	25	23.91 ± 3.41	< 0.01	24.15 ± 3.44
		Target EIG	25	24.59 ± 4.77	1.69 ± 0.23	26.52 ± 4.98

Across the three seeds, runtime for the all configurations averaged 10.7 hours per acoustic species-seed and 1.5 hours per camera-trap species-seed, totaling approximately 1,718 job-hours. With P concurrent jobs, the estimated wall-clock time to reproduce the current 50 acoustic and 25 camera-trap experiments is

$$\frac{1{,}718 \text{ job-hours}}{P}.$$

While this is the approximate time required to reproduce our experiments given the specified hardware, the full research project required more resources in total.

B.3 Uncertainty quantification

Confidence intervals (CIs) are pointwise 95% intervals, without multiple-comparison correction. Normal-approximation CIs use the mean ± 1.96 times its standard error, with the averaging unit specified in each caption. Unless noted otherwise, seeds are averaged within species before estimating uncertainty. Species-bootstrap CIs use 2,000 resamples of species: percentile CIs use the 2.5th and 97.5th percentiles of bootstrap means, whereas bootstrap-normal CIs use the bootstrap standard error. Posterior credible bounds describe within-fit uncertainty, not uncertainty across species.

C Ecological tasks and characteristics of selected data

We investigate task-specific differences across methods and compare ACORN with random, max-score, BALD, and PPU review under the same decoupled continuous-score model, and with EIG focused separately on site occupancy, occupancy coefficients, or detection coefficients. Table 3 shows differences in average oracle agreement over each full review trajectory. Site-occupancy agreement averages A_ψ and ρ_ψ, coefficient tasks use A_β or A_α, and the combined score is $\overline{A}$ (Section 4.3). Seeds are averaged within species before resampling species for confidence intervals.

Over full trajectories, ACORN improves combined agreement over max-score by 1.08 percentage points on acoustics and 0.90 on camera traps. Averaging agreement over the first 100 reviews instead gives larger gains of 4.79 points (95% species-bootstrap percentile CI: $[2.32, 7.36]$) and 2.44 points ($[0.74, 4.03]$), respectively, computed as the area under the agreement curve divided by the review budget. The largest gains over random review across full trajectories are for acoustic occupancy coefficients (8.48 points) and camera-trap site occupancy (6.88 points). Gains are not

Table 3: **Average agreement over full review trajectories by ecological task.** Entries are paired differences in trajectory-average oracle agreement, in percentage points, shown as mean ± 95% CI half-width, using the species-bootstrap normal approximation after averaging paired differences across seeds within species. Positive entries (green) mean ACORN performs better while negative entries (red) mean ACORN performs worse. Task-specific ACORN focuses EIG on the evaluated task.

	ACORN vs.				
Evaluation task	Random	Max-score	BALD	PPU	Task-specific ACORN
Acoustic					
Combined	$+6.50 \pm 1.85$	$+1.08 \pm 1.02$	$+3.11 \pm 0.95$	$+5.65 \pm 1.52$	–
Site occupancy	$+6.57 \pm 2.10$	$+1.85 \pm 1.11$	$+2.97 \pm 0.91$	$+5.44 \pm 1.52$	$+0.02 \pm 0.05$
Occupancy coefficients	$+8.48 \pm 3.81$	$+1.27 \pm 2.02$	$+5.82 \pm 1.88$	$+8.93 \pm 2.76$	-0.05 ± 0.26
Detection coefficients	$+4.37 \pm 2.90$	-0.64 ± 1.86	$+0.69 \pm 1.12$	$+2.81 \pm 1.75$	-0.01 ± 0.95
Camera trap					
Combined	$+5.18 \pm 1.83$	$+0.90 \pm 0.44$	$+1.21 \pm 1.36$	$+1.36 \pm 1.60$	–
Site occupancy	$+6.88 \pm 3.61$	$+1.18 \pm 0.69$	$+0.81 \pm 0.54$	$+1.00 \pm 0.80$	-0.10 ± 0.13
Occupancy coefficients	$+3.38 \pm 2.31$	$+0.84 \pm 0.54$	$+1.16 \pm 1.41$	$+1.42 \pm 1.70$	$+0.13 \pm 0.23$
Detection coefficients	$+3.58 \pm 3.02$	$+0.40 \pm 0.83$	$+2.06 \pm 3.32$	$+2.04 \pm 3.36$	$+1.68 \pm 2.68$

Table 4: **Characteristics of selected replicates over the first 100 reviews.** Entries average all selected replicates over the first 100 reviews. Bold marks the highest displayed value in each column within each dataset. *New site* is the percentage of sites with no prior reviewed label. *Score rank* and *covariate extremeness* are mean percentiles within the current unreviewed candidate pool. Extremeness is squared Mahalanobis distance in the occupancy or detection covariates. *Label confidence* is $100 \max(p, 1-p)$ for the ecological model's posterior predictive label probability p.

Review policy	New site	Score rank	Label confidence	Occupancy covariate extremeness	Detection covariate extremeness
Acoustic					
Random	56.0	52.6	**96.3**	52.8	48.2
Max-score	41.1	**99.8**	89.5	47.2	47.8
BALD	48.8	87.1	82.0	50.2	54.0
PPU	38.2	84.0	70.7	49.3	47.2
ACORN (combined)	57.7	87.7	88.9	**53.7**	58.5
EIG: site occupancy	57.9	87.5	89.1	**53.7**	58.4
EIG: occupancy coefficients	**58.0**	87.6	89.0	**53.7**	58.5
EIG: detection coefficients	45.9	87.4	82.6	49.9	**59.8**
Camera trap					
Random	51.4	48.9	**98.9**	**50.5**	51.5
Max-score	48.3	**99.2**	96.9	47.7	52.2
BALD	57.1	90.7	89.9	48.9	56.0
PPU	47.7	86.4	86.0	49.2	53.1
ACORN (combined)	64.1	91.1	91.9	49.6	55.4
EIG: site occupancy	**66.5**	90.3	92.2	49.8	55.4
EIG: occupancy coefficients	65.9	90.8	92.1	49.9	55.7
EIG: detection coefficients	56.4	91.1	90.2	48.9	**58.5**

uniform: acoustic detection-coefficient agreement is 0.64 points lower than max-score on average. Interestingly, focusing EIG on the evaluated task changes mean agreement by less than 0.14 points, except for camera-trap detection-only targeting, which falls 1.68 points behind combined targeting. Thus, task specialization does not consistently improve performance. We hypothesize this is because our occupancy modeling tasks are inherently coupled: errors in the estimation of the occupancy process force the detection process to compensate, and vice-versa.

We furthermore investigate characteristics of the data that is prioritized by each method for the first 100 reviewed replicates in Table 4. Combined-target ACORN does not simply follow classifier-score order: mean current-pool score percentiles are 87.7 for acoustics and 91.1 for camera traps, compared with 99.8 and 99.2 for max-score. Detection-focused EIG has the highest mean visit-covariate

extremeness percentile in both datasets (59.8 vs. 58.5 for combined targeting in acoustics, 58.5 vs. 55.4 in camera traps). Site-occupancy targeting selects more reviews from previously unreviewed sites than detection targeting (57.9% vs. 45.9% for acoustics, 66.5% vs. 56.4% for camera traps).

Different targets can nevertheless select distinct review sets: after 100 acoustic reviews, the mean Jaccard overlap (intersection divided by union) is 0.43 between site-occupancy and detection-coefficient targeting, versus 0.95 between site-occupancy and occupancy-coefficient targeting. Similar ecological agreement can therefore arise from different sets of reviewed observations.

Dataset characteristics also matter. For example, acoustic gains in averaged oracle agreement over max-score correlate with positive prevalence, site occupancy, and empirical detection probability (Spearman $\rho = 0.45$, 0.46, and 0.39), and decrease when positives are concentrated at fewer sites ($\rho = -0.39$). Mean classifier average precision is higher for camera traps (0.95 versus 0.74 for acoustics). These post-hoc associations do not establish why gains differ between datasets.

Table 5: **Batch-size sensitivity and computational cost for Target EIG.** Comparisons use the finest cadence instead of the oracle ($m = 1$ acoustic, $m = 5$ camera trap). Occupancy differences are mean absolute site-probability differences in percentage points (pp). Early comparisons use 20 acoustic or 50 camera-trap reviews. Coefficient agreement at 100 reviews uses the positive, negative, or uncertain conclusions defined in Eq. 14. Bold batch sizes are the default.

Dataset	m	Fits	Time (%)	Occupancy difference (pp) Early	At 100	Coeff. agreement (%) Occupancy	Detection
Acoustic	1	101	100.0	0.00	0.00	100.0	100.0
	2	51	51.0	1.58	0.92	98.2	96.4
	5	21	21.5	3.51	1.30	98.0	96.5
	10	11	10.8	5.91	2.05	96.3	94.8
	20	6	6.1	10.21	3.13	94.0	92.1
Camera trap	5	21	100.0	0.00	0.00	100.0	100.0
	10	11	55.9	0.60	0.68	99.3	100.0
	25	5	27.2	2.27	1.26	98.7	98.7
	50	3	17.2	9.67	3.20	97.6	98.7
	100	2	10.5	–	6.45	95.1	98.0

D Sensitivity to review batch size

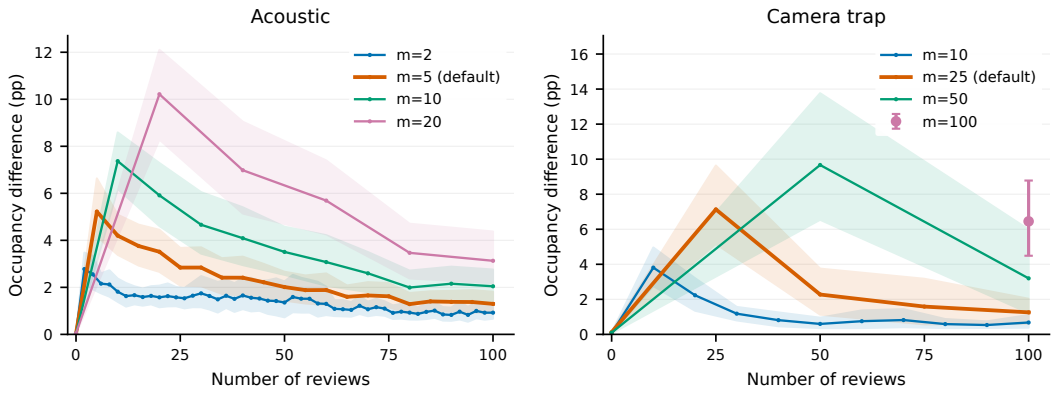
The batch size m controls how many expert reviews are collected before refitting the model and updating the review priorities. We use $m = 5$ for acoustic data and $m = 25$ for camera traps. To characterize the tradeoff between computational cost and responsiveness to new labels, we fit the decoupled continuous-score model with four policies (Target EIG, BALD, random, and max-score), and a budget of 100 reviews. We evaluate $m \in \{1, 2, 5, 10, 20\}$ for acoustic data and $m \in \{5, 10, 25, 50, 100\}$ for camera traps.

To assess sensitivity without relying on a fully reviewed oracle, we compare each trajectory with the finest cadence for the same species, policy, and seed ($m_{\min} = 1$ acoustic, $m_{\min} = 5$ camera trap). At a shared review count R, we measure

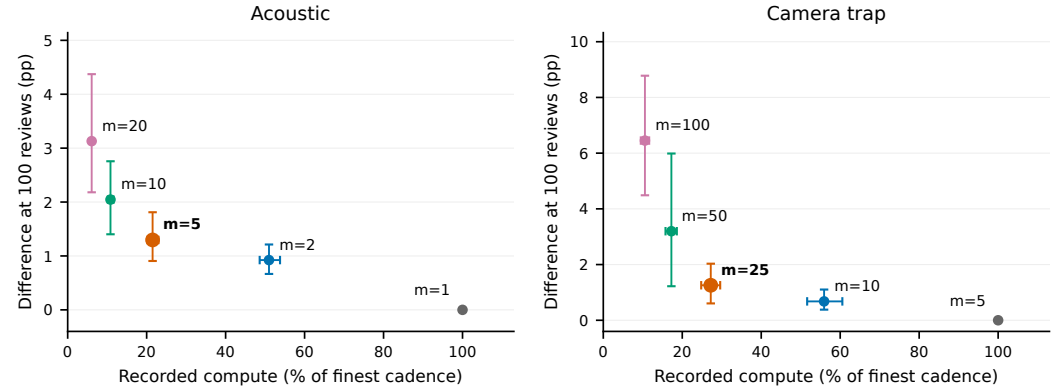
$$\Delta_{\psi}^{(m)}(R) = \frac{1}{|S|} \sum_{i \in S} \left| \bar{\psi}_i^{(m)}(R) - \bar{\psi}_i^{(m_{\min})}(R) \right|. \tag{15}$$

We also compare coefficient conclusions using the definitions in Eq. 14, with the finest cadence replacing the oracle.

Figure 6 and Table 5 show the results for Target EIG. Larger batches have an early impact: after 20 acoustic reviews, mean occupancy differences from the finest cadence are 3.51 percentage points for $m = 5$ and 10.21 for $m = 20$. After 50 camera-trap reviews, they are 2.27 points for $m = 25$ and 9.67 for $m = 50$. By 100 reviews, the differences at the defaults decrease to 1.30 and 1.26 points, respectively. The defaults require 21.5% and 27.2% of the recorded computation of the finest cadence. Thus, they offer substantial computational savings while retaining more frequent feedback than the coarser alternatives. Interestingly, larger batches do not uniformly increase disagreement between review policies. Relative to $m = 5$, using $m = 25$ increases camera-trap Target EIG–BALD occupancy disagreement at 50 reviews by 1.1 percentage points (95% species-bootstrap percentile CI: $[0.2, 2.1]$), but using $m = 50$ reduces it by 1.9 points ($[-2.7, -1.2]$).



(a) Occupancy differences over the review trajectory, relative to the finest cadence ($m = 1$ acoustic, $m = 5$ camera trap).



(b) Recorded compute vs. occupancy differences at 100 reviews. Bold labels mark our defaults ($m = 5$ acoustic, $m = 25$ camera trap).

Figure 6: **Larger review batches reduce computation but change early posterior trajectories.** All panels show Target EIG, comparing each m with the finest cadence instead of the oracle. Occupancy differences are mean absolute differences in site occupancy probabilities, in percentage points (pp). Bands and error bars show pointwise 95% species-bootstrap percentile CIs after averaging seeds within species.

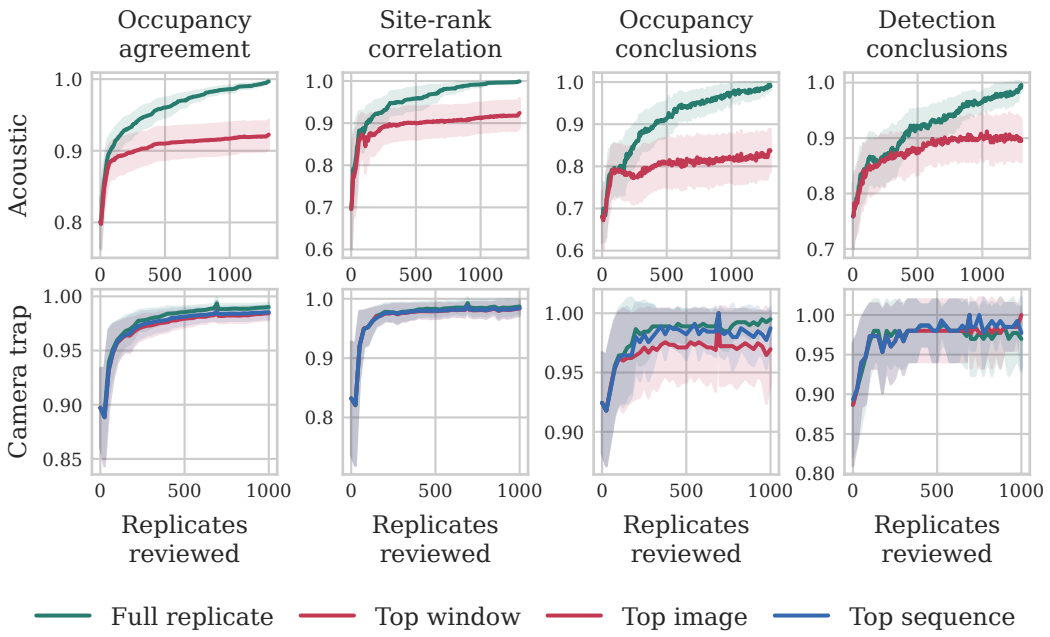


Figure 7: **Highest-scoring review saves time, but can impact oracle agreement.** We compare different review granularities with the default full-replicate oracle. The x-axis shows the total number of reviews, while the y-axis shows agreement with the oracle (higher is better). Shaded bands show pointwise 95% normal-approximation CIs across species after averaging seeds within species.

E Sensitivity to review granularity

In the main experiments, we simulate expert review across all samples in a selected replicate (an acoustic recording or a camera trap location-day). However, when expert time is scarce and the classifier scores true observations highly, this might be wasteful.

We therefore compare default ACORN using full-replicate labels with labels from only the highest-scoring five-second acoustic window, camera image, or camera sequence (the sequence containing the highest-scoring image within that location-day). Crucially, if the highest-scoring sample is a negative, the entire replicate is treated as negative, potentially missing true detections that are lower-scoring. We compare the impact of this highest-scoring review strategy using the same full-replicate oracle described in Section 3.

Figure 7 shows the resulting review trajectories. Early camera-trap agreement is nearly unchanged, although final occupancy-coefficient agreement is 2.2 percentage points lower for top-image review. For acoustics, mean agreement after all 1,302 reviews is 0.895 with top windows versus 0.996 with full recordings. Therefore, in our experiments, the highest-scoring review strategy is more promising for camera traps than for acoustics, likely because classifier performance is higher on camera traps (Appendix H).

At 100 reviews, highest-scoring review reduces the length of reviewed audio by approximately 120-fold and camera-trap image counts by 64-fold (top image) or nine-fold (top sequence).



(a) Top-scoring nominal negative sample for Impala (SpeciesNet logit for Impala: 14.09, Ground-truth label: Plains zebra, Image ID: 9214d000-21bc-11ea-a13a-137349068a90)



(b) Top-scoring nominal negative sample for Günther's dik-dik (SpeciesNet logit for Günther's dik-dik: 17.08, Ground-truth label: Grévy's zebra, Image ID: 8fbd0584-21bc-11ea-a13a-137349068a90)

Figure 8: Example camera trap images corresponding to obvious ground-truth label errors surfaced by highly scoring nominal negatives across two species.

Table 6: **Nominal negative score distributions can differ between occupied and likely unoccupied sites.** Values are mean raw logits. Occupied sites have at least one positive observation ($\exists j : f_{ij} = 1$), whereas likely unoccupied sites are defined as sites without positive observations ($f_{ij} = 0 \forall j$). The table shows species with large positive gaps between negative replicates at occupied and likely unoccupied sites within each dataset.

Dataset	Species	Positive Mean	Negative Mean at Occupied Sites	Negative Mean at Likely Unoccupied Sites
Acoustic	Scarlet Tanager	10.96	8.36	7.26
	Ovenbird	12.48	8.25	7.23
	Rose-breasted Grosbeak	10.00	7.31	6.39
Camera trap	Impala	14.00	6.74	5.00
	Puma	18.92	4.46	2.80
	Jaguar	17.25	4.53	2.94

F Score distribution heterogeneity

The continuous-score model assumes that classifier scores are generated from a mixture of normal score distributions that are shared between all sites. In practice, this approximation is not equally good for all species. In post-hoc experiments on both acoustic and iWildCam camera-trap data, several species were well described by a simple two-component score model, but others showed clear score-distribution heterogeneity: scores from non-detection replicates at occupied sites were not always distributed like scores from truly unoccupied sites. Some high-confidence "no-detection" days in iWildCam then appeared on manual inspection to contain real detections missed by the annotation process (see Figure 8), creating a high-score tail among nominal negatives (see Table 6 for illustration). Another contributing factor may be classifier bias caused by spurious correlations: visual or acoustic backgrounds that are more likely to contain the focal species might be scored higher, even if the focal species is not present.

This heterogeneity makes the original joint continuous-score formulation more fragile than the calibrated variant used in our main experiments. In the original continuous-score occupancy model formulation [53], the score-mixture parameters and ecological parameters are estimated together, so misspecification of the score likelihood can feed back into occupancy and detection estimates. The sampler can also encounter weakly identified mixture geometries, because shifts in the score

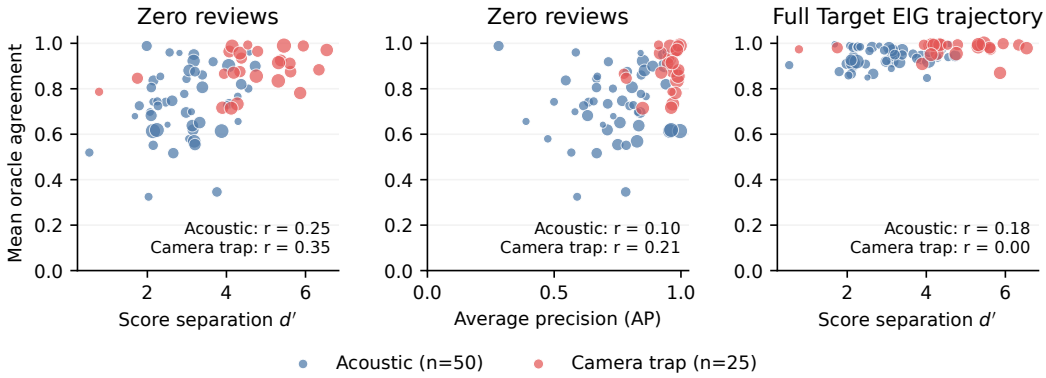


Figure 9: **Species-level classifier metrics only partially explain downstream ecological agreement.** Each point is one species. Agreement is the mean of the four oracle-agreement metrics (Section 4.3). Left: zero-review agreement versus score separation d'. Middle: zero-review agreement versus average precision (AP). Right: Target EIG reviewed-trajectory agreement versus d'. Point size is proportional to empirical site occupancy prevalence.

distributions, detection probability, and occupancy probability can explain similar patterns in the observed scores. Decoupling blocks this feedback by fitting score calibration outside the occupancy sampler and passing in fixed log Bayes factors. Moreover, once a replicate is reviewed, the reviewed binary label replaces the score contribution in the occupancy likelihood, preventing a small number of reviewed high-score outliers from directly distorting the ecological posterior through the score model.

We experimented with several alternative joint models to address this misspecification directly, including a separate false-score distribution for occupied sites, noisy-annotation models, and annotation-primary variants with a missed-label score component. These variants sometimes improved agreement with the oracle in particular review regimes, especially at higher review counts, but no single alternative was robust across species and review budgets. The more flexible models often introduced weakly identified mixture components and convergence problems, while fixed-tail annotation-primary models required species- and review-count-specific tuning. We therefore use the calibrated continuous-score model for the main analysis and treat models that allow richer score heterogeneity, particularly for difficult-to-label species, as an important direction for future work.

G Covariate shift

New surveys may differ in habitat composition, potentially affecting both occupancy and classifier scores. We ask whether decoupling makes ecological inference more robust to these changes. We compare joint and decoupled refits under covariate shift alone and when classifier scores also change with habitat.

We compared both models without reviewed labels in 30 paired simulations (100 sites, ten visits per site). We shifted $X \sim \mathcal{N}(0, 1)$ to $\mathcal{N}(1.5, 1)$ while keeping occupancy and detection relationships fixed. A second condition added $0.75X$ to both class-conditional score means, introducing habitat-dependent score bias.

Covariate shift alone shows no clear difference in site-occupancy RMSE (decoupled minus joint: -0.007, 95% CI: $[-0.023, 0.009]$). When scores also shift, decoupling improves occupancy-slope recovery (absolute-error difference -0.204, CI: $[-0.312, -0.096]$) but worsens site predictions (RMSE difference 0.035, CI: $[0.011, 0.058]$). In these simulations, decoupling can therefore help improve coefficient recovery without improving predictions. We note that these simulations are necessarily artificial, and that we leave investigation into more realistic shifts for future work.

H Impact of classifier quality

We summarize classifier quality in two ways. First, we use score separation d' as the distance between the positive- and negative-score calibration means in units of standard-deviation. Second, we use average precision (AP), which is the mean precision achieved as true positive replicates are swept from high to low classifier score. Species with larger d' tended to start closer to the oracle at zero reviews, but the association was only moderate (Figure 9). AP was an even weaker predictor of zero-review agreement. After review using the Target EIG policy (Section 3.2), agreement was high regardless of classifier quality. This indicates that improvements in classification accuracy translate into improvements in low-review regimes, but review remains valuable across varying classifier quality.

I Application to custom datasets

ACORN requires repeated samples across sites, suitable site- and replicate-level covariates, expert review capacity, and a classifier covering the focal species. ACORN models one species at a time, with occupancy assumed constant within the survey season. Multi-species, multi-season and spatial random-effects models are promising future work.

As discussed in Appendix H, imperfect classifiers can still supply useful evidence, but regional differences in backgrounds or sensors may shift the class-conditional score distributions. While decoupling reduces feedback from score-model misspecification, calibration diagnostics and sensitivity analyses remain critical when applying ACORN to new datasets. Finally, stopping thresholds should be selected according to the cost of expert review and the desired stability of the ecological conclusions, rather than being interpreted as universal convergence criteria.

J Parameter trajectories

To illustrate qualitatively how parameter estimates converge to the oracle estimates, we visualize species-specific parameter trajectories in Figures 10 and 11.

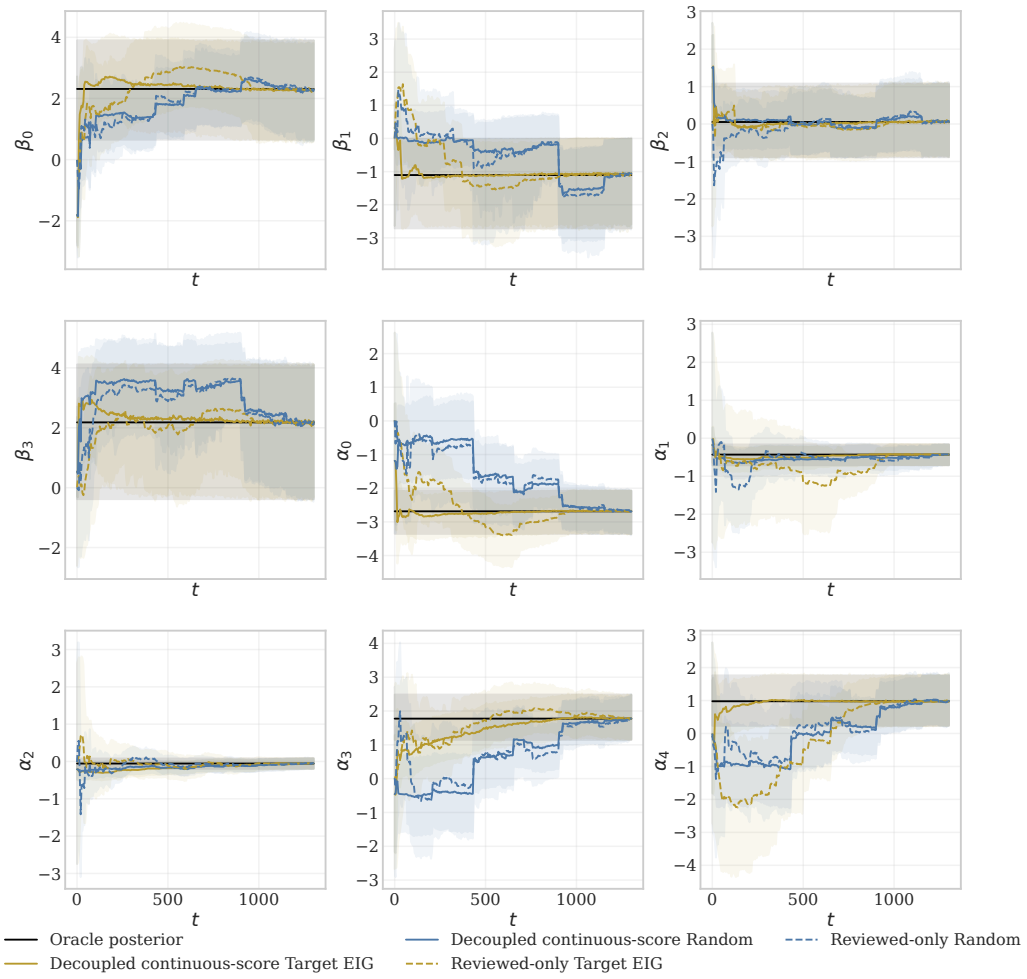


Figure 10: Parameter estimates for the Scarlet Tanager. Shaded bands show within-seed 90% posterior credible bounds (5th and 95th percentiles), averaged across three seeds.

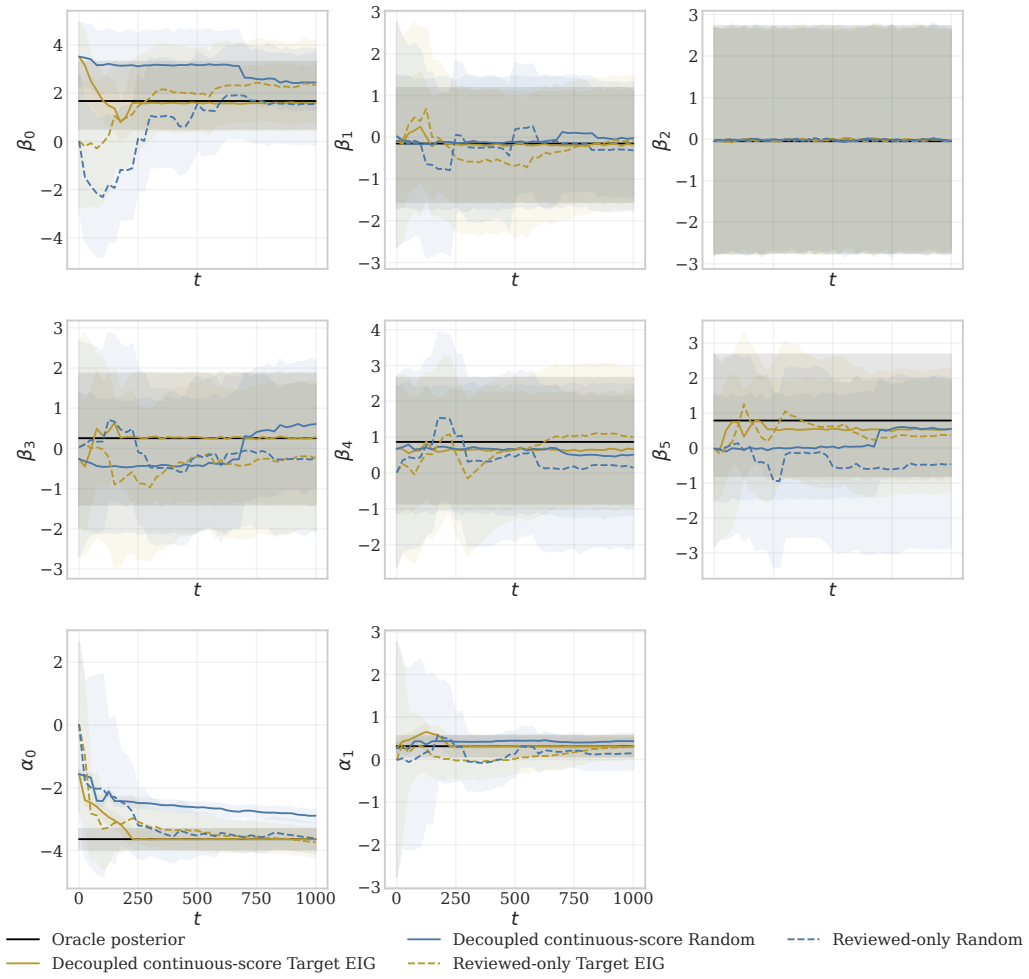


Figure 11: Parameter estimates for the White-tailed deer. Shaded bands show within-seed 90% posterior credible bounds (5th and 95th percentiles), averaged across three seeds.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the abstract and introduction are supported by evidence presented in Section 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are explicitly highlighted and discussed in Section 6.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: We do not derive theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper does detail the experimental setup. We furthermore provide code and use publicly available datasets to enable reproducibility.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The two datasets used for our experiments are openly accessible [49, 3]. We furthermore provide source code together with instructions on how to run that code to reproduce our experiments and figures.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Details about our experimental setup can be found in Section 4. Details about data can be found in Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Where applicable, we show 95% confidence intervals in figures.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Details on compute resources needed to reproduce our experiments are provided in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We do not perform research on human subjects and perform research on existing publicly available datasets. We furthermore do not foresee exceptional negative societal impact or other harmful consequences arising from this work.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper discusses the importance of occupancy modeling to global ecology and conservation, emphasizing the positive societal impact of the work. We do not foresee any substantial negative societal impacts. Potential harms arising from the method giving potentially incorrect results are discussed as limitations in Section 6.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not introduce data or models that have a high risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators of datasets used in our experiments are properly credited and their licenses are explicitly mentioned and respected (see Section 4.1 and Appendix A).

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide a code archive to reproduce our experiments as part of the supplementary material. Documentation is provided as part of the code archive.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: We do not perform crowdsourcing experiments or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Our paper does not involve LLMs.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.