Methodological choices influence ecological inference in passive acoustic monitoring of a Neotropical nightjar

Liliana Piatti[1,*], Daiene L. H. de Sousa[2], Beatriz dos S. Oliveira[1], Alyson V. de Melo[3], Larissa S. M. Sugai[4,↑], Diogo B. Provete[1,2,3,5↑]

[1] Instituto de Biociências, Universidade Federal de Mato Grosso do Sul, Campo Grande, Mato Grosso do Sul, 79002-970, Brazil.

[2] Programa de Pós-Graduação em Ecologia e Conservação, Instituto de Biociências, Universidade Federal de Mato Grosso do Sul, Campo Grande, Mato Grosso do Sul, 79002-970, Brazil.

[3] Programa de Pós-Graduação em Biologia Animal, Instituto de Biociências, Universidade Federal de Mato Grosso do Sul, Campo Grande, Mato Grosso do Sul, 79002-970, Brazil.

[4] K. Lisa Yang Center for Conservation Bioacoustics, Cornell University, Ithaca, New York, USA.

[5] Gothenburg Global Biodiversity Centre, Göteborg, Box 100, S 405 30, Sweden.

* Corresponding author, email: liliana.piatti@ufms.br

[↑]Joint senior authors

Methodological choices influence ecological inference in passive acoustic monitoring of a Neotropical nightjar

ABSTRACT

Passive acoustic monitoring (PAM) is increasingly used to investigate species activity and habitat use through occupancy analyses. Yet, the complex analytical workflow, from automated detector choice to confidence thresholds and statistical modeling framework, is amongst the factors that influence ecological inference, and the extent these decisions affect modelling outputs is poorly debated. Here, we use acoustic data from a Neotropical nightjar (*Nyctidromus albicollis*) in the Pantanal wetlands, Brazil, to evaluate how these decisions lead to different conclusions about the environmental drivers of nocturnal vocal activity. We processed 97,906 minutes of audio files from seven sites using two classification algorithms (a locally trained custom model and a pre-built global model), each evaluated at two confidence thresholds based on precision and F1 metrics, yielding four detection datasets. We modeled vocal activity in relation to temperature, relative humidity, and lunar illumination using both conventional regression and Bayesian occupancy models that explicitly separate the contributions of ecological and observation processes. All models agreed on the direction of environmental effects: vocal activity declined with increasing temperature and humidity and was weakly associated with lunar illumination. However, the magnitude of the effects differed substantially across detector–threshold combinations. Detections based on the prebuilt model overestimated the effects of all variables, with predicted occupancy at low temperatures ranging from over 90% (prebuilt) to below 25% (custom). High-precision thresholds were far better calibrated, highlighting the importance of prioritizing precision for more reliable inference under hierarchical modeling. The naïve regression approach produced attenuated effect sizes and narrower uncertainty intervals compared to the occupancy framework, with discrepancies depending on the detector used. Our results demonstrate how methodological decisions across the analytical workflow affect quantitative ecological inference, emphasizing the need to explicitly address the impact of different decisions during the analytical process and incorporate detection uncertainty through hierarchical modeling.

INTRODUCTION

Biodiversity monitoring is essential for understanding ecological processes and informing conservation in a rapidly changing world. However, traditional survey methods are often constrained by limited temporal coverage, logistical costs, and observer bias, restricting their ability to capture ecological dynamics across broad spatial and temporal scales (Moussy et al. 2021; Valdez et al. 2023). These limitations have motivated the development of scalable approaches capable of providing consistent data on species occurrence and activity (Connor et al. 2024; Nguyen Chi et al. 2025; Riede and Balakrishnan 2025).

Passive acoustic monitoring (PAM) has emerged as a powerful tool to address these challenges (Ross et al. 2023). By continuously recording soundscapes, PAM enables non-invasive monitoring of vocal species and provides detailed information on temporal patterns of activity (Sugai et al. 2019). This is particularly valuable for taxa that are difficult to observe directly, such as nocturnal birds. As a result, PAM has become increasingly common in ecological studies, especially for investigating species activity and its environmental drivers (Chirino et al. 2025, Salustio-Gomes et al. 2026, Sugai et al. 2019).

Despite its advantages, PAM generates large volumes of audio data, creating challenges for data processing and interpretation (Teixeira et al. 2024). Manual annotation is often impractical, leading to the growing use of automated detection methods based on machine learning (Kahl et al. 2021). These approaches have improved the efficiency of species identification, but their performance remains context-dependent, varying with environmental conditions, species-specific vocal behavior, and characteristics of the acoustic environment (Pérez-Granados et al. 2026). In particular, models trained on generalized datasets may perform

59 inconsistently when applied to new soundscapes, often requiring local validation
60 (Funosas et al. 2026).
61 Automated detection tools, such as BirdNET have become widely used due to
62 their accessibility and broad taxonomic coverage (Wood & Kahl 2024). However,
63 their performance is not uniform across regions and species, with recent global
64 evaluations showing substantial variation in detection accuracy depending on
65 ecological context and species coverage (Funosas et al. 2025). These results
66 highlight the need for context-specific evaluation and careful calibration of detection
67 parameters when applying such tools to ecological studies (Schiavo et al. 2025).
68 Importantly, PAM workflows involve multiple analytical steps, including
69 species detection and classification, data filtering, and statistical modeling. Each of
70 these processes requires methodological decisions that influence how acoustic data
71 are translated into information ecologically relevant (Gelman & Loken 2018, Breznau
72 et al. 2022). Detector choice, confidence thresholds, and validation procedures
73 determine the structure of detection data used in subsequent analyses. However,
74 these decisions are often evaluated in isolation and seldom fully reported, despite
75 their impact on downstream inference.
76 A further challenge arises from imperfect detection. The probability of
77 detecting a species depends not only on its activity, but also on environmental
78 conditions, recording system characteristics, and behavioral variability (Wood &
79 Peery 2022). In passive acoustic monitoring, this heterogeneity can be particularly
80 pronounced, with detection probability varying substantially across sites and
81 sampling occasions (Rhinehart et al. 2026). Ignoring such variation can lead to
82 biased estimates of occupancy and misleading relationships with environmental
83 drivers (Rhinehart et al. 2026). Hierarchical modeling approaches, such as

84 occupancy models, explicitly account for this issue by separating the ecological and
85 observation processes (MacKenzie et al. 2018). Moreover, the application of
86 occupancy models to acoustic data remains inconsistent, with many studies failing to
87 adequately incorporate detection uncertainty, potentially leading to biased ecological
88 inference (Martins et al. 2025).

89 Reliable ecological inference also depends on consistent sampling across the
90 environmental gradients of interest. In many ecological studies, it is assumed that
91 variation in observed responses reflects underlying environmental drivers rather than
92 differences in sampling effort or observation processes. However, in passive
93 acoustic monitoring, this assumption can be challenged. Variation in recording
94 conditions, microphone sensitivity, temporal coverage, and data processing
95 decisions can introduce additional heterogeneity into detection data. As a result,
96 apparent patterns along environmental gradients may partly reflect inconsistencies in
97 the observation process rather than true ecological responses.

98 Together, these considerations highlight a key methodological question: how
99 do decisions in the PAM analytical workflow influence ecological inference?
100 Evidence from psychology, ecology and evolutionary biology shows that different
101 analytical choices applied to the same dataset can lead to substantial variation in
102 estimated effect sizes and conclusions (Gould et al. 2025). In PAM, this issue may
103 be further amplified, as multiple sequential decisions jointly shape the data used for
104 inference. Despite advances in both automated detection and hierarchical modeling,
105 their integration remains limited, and there is still little consensus on how to
106 incorporate detection uncertainty derived from acoustic classifiers into ecological
107 models (Martins et al. 2025).

Nocturnal birds provide a useful system in which to explore these issues. Their vocal activity is often influenced by environmental factors such as temperature, humidity, and lunar illumination, while their detectability can vary with both behavior and recording conditions (Knight et al. 2017, Schaaf et al. 2023, Salustio-Gomes et al. 2026). A representative example is the Common Pauraque, *Nyctidromus albicollis*, a widely distributed nocturnal bird inhabiting a variety of environments from southern Texas to northeastern Argentina (Latta & Howell, 2020). The literature indicates that its vocal activity is restricted to the nocturnal period, peaks just after dusk and just before dawn, and is detected year-round, with vocal output being higher during full moon nights than during new moon nights, and with the diel pattern of vocal activity also varying with moon phase (Pérez-Granados & Schuchmann 2020b; Pérez-Granados et al. 2022). Here, we use passive acoustic monitoring data from a Neotropical nightjar in the Pantanal wetlands, Brazil, to evaluate how different analytical decisions influence ecological inference. We focus on decisions related to acoustic classifiers (pre-trained versus regional custom classifier), confidence score threshold levels (informed either by F1 or precision), and modelling framework (hierarchical occupancy models and conventional regression-based approaches). We compare inferences based on the relationship between vocal activity and temperature, relative humidity, and lunar illumination.

In terms of ecological inference, we a priori expected negative effects of temperature and relative humidity on detection probability based on the thermal sensitivity of caprimulgid aerial insectivory and the suppressive effect of precipitation on vocal activity (Schaaf et al. 2023; Salustio-Gomes et al. 2026). For lunar illumination, we expected a positive effect on detection probability, as moonlight stimulates vocal activity in Neotropical nightjars by enhancing visual foraging

conditions (Pérez-Granados & Schuchmann, 2020a; Pérez-Granados & Schuchmann, 2020b; Pérez-Granados et al. 2022). By evaluating these components within a single analytical framework, our study illustrates how methodological choices in passive acoustic monitoring can shape inference about species activity.

MATERIAL AND METHODS

Study area and Sampling design

The study was conducted in the Pantanal wetlands of Mato Grosso do Sul, central Brazil. Monitoring sites were distributed along an approximately 180-km environmental gradient between the Pantanal floodplain and the plateaus of the Serra da Bodoquena and Serra de Maracajú, encompassing a heterogeneous mosaic of flooded *campos*, gallery forests, and upland Cerrado vegetation (Rolim & Theodorovick 2012, Junk et al. 2014). The climate in the region is strongly seasonal, with a rainy season from October to March and a dry season from April to September, driving marked annual fluctuations in water availability, habitat structure, and wildlife activity (Heckman 1999).

Seven monitoring sites were established across this gradient (Fig. 1). This sampling design captures variation in hydrological regimes and habitat structure, providing a suitable environmental gradient for examining patterns of species activity (Table S1; Fig. 1).

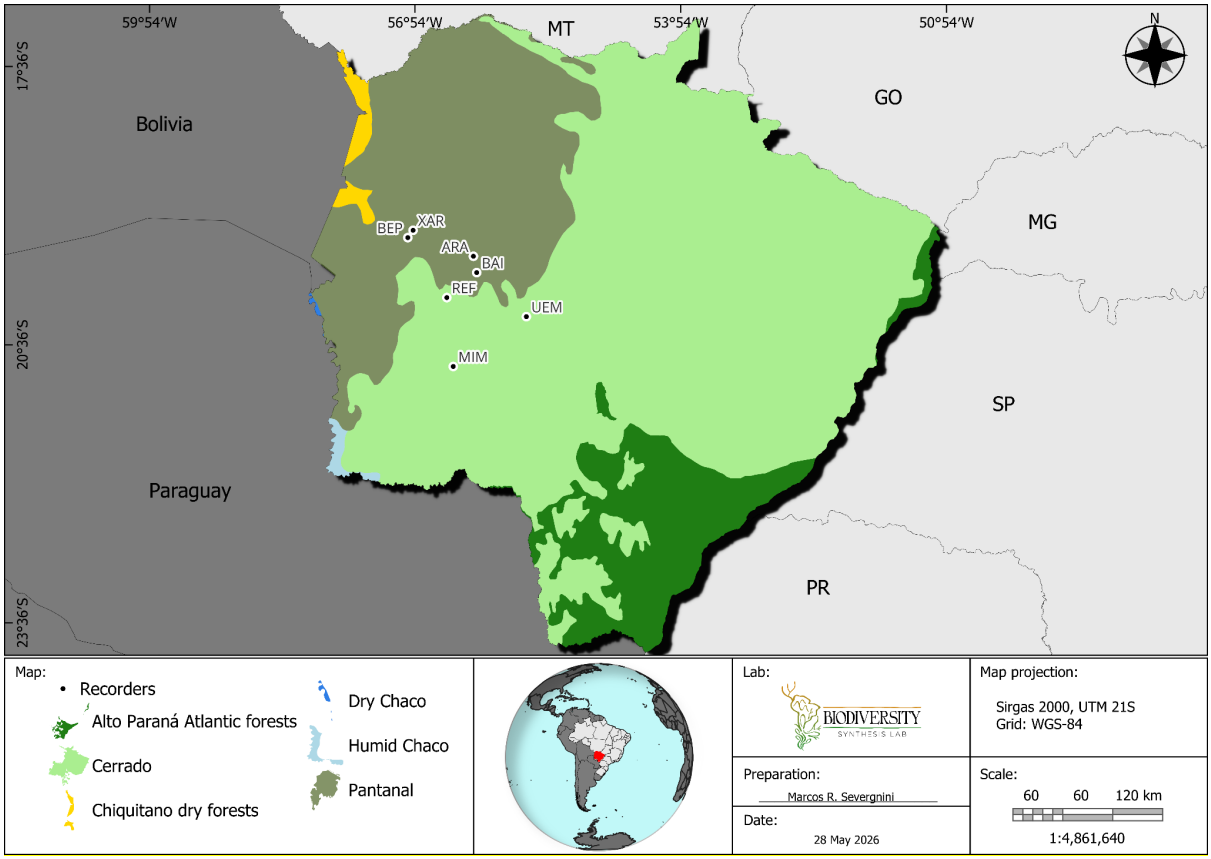


Figure 1. Location of the study area and sampling sites in the South Pantanal wetland and Cerrado, central Brazil.

At each site, one autonomous recording unit (SongMeter SM4, Wildlife acoustics, Maynard, Massachusetts, USA) was installed adjacent to a lentic water body. Recorders were programmed to record for one minute every 15 minutes throughout the entire night, at a sampling rate of 44.1 kHz with 16-bit depth, stereo channel configuration, 16 dB pre-amplification, in uncompressed WAV format. Monitoring began on 21 September 2021 and ended on Feb 2025 across the entire sampling. All audio files were transferred to a Network-Attached Storage (NAS) server with redundant backup.

Climatic variables

166 At each site, a HOBO U23 Pro V2 (Onset Computer Corporation) temperature (ºC)
167 and relative humidity (%) datalogger was installed adjacent to the recording unit.
168 Dataloggers were programmed to log at 15-minute intervals. Because dataloggers
169 were deployed after the acoustic sensors were deployed and to ensure acoustic and
170 environmental information matched, subsequent analyses were restricted to
171 recordings for which paired microclimatic data were available. Nightly values of
172 temperature and relative humidity were derived by averaging datalogger readings
173 across the nocturnal recording window for each site-night combination. Lunar
174 illumination was calculated as the fraction of the lunar disc illuminated at local
175 midnight for each site-night, obtained with the R package `lunar` (Śmielak 2023).
176
177 *Acoustic data processing and detection workflow*
178 After filtering for recordings with available climatic data, a total of 97,906 audio files
179 were retained. These recordings span between 2023-11-26 and 2025-02-27,
180 covering the full nocturnal recording window at each site (from 4 pm to 6 am next
181 day). Temporal coverage and sampling effort were not uniform among sites,
182 reflecting differences in recorder deployment duration and data availability (Fig. S1).
183 To detect the occurrence of the focal species (*Nyctidromus albicollis*) in these
184 recordings, we applied two automated classification approaches (Fig. S2): (1) a pre-
185 trained global model available in BirdNET-Analyzer GUI (Kahl et al. 2021) and (2) a
186 custom classifier using region-specific acoustic data. For the first approach, all audio
187 files were processed using BirdNET-Analyzer v2.4.0 (model:
188 BirdNET_GLOBAL_6K_V2.4_Model_FP32.tflite). For the second approach, we
189 trained a custom classification model using the training interface available in
190 BirdNET-Analyzer (see Supplementary Information).

To evaluate the performance of the two classifiers, we built an independent ground-truth test dataset comprising approximately 1% of the full dataset (1,199 audio files, Fig 2). The selection followed a stratified approach in which the number of files per site was proportional to the total number of recordings available at each site within the filtered dataset, to ensure that the test dataset reflected the spatial distribution and sampling effort of the data used in subsequent analyses.

Figure 2. Summary of automated detection outputs in the test dataset, across study sites. The upper panel shows the total number of audio files and the number of detections obtained using the BirdNET and custom classifiers at each site. The lower panel presents the distribution of confidence scores for detections generated by each algorithm across sites

Outputs generated by the pre-trained BirdNET model and the custom-trained classifier were independently compared against manual annotations from the test dataset using the evaluation interface available in BirdNET-Analyzer (https://birdnet-team.github.io/BirdNET-Analyzer/best-practices/evaluation-tool.html) and results were exported using the “Download data table” option. These outputs were subsequently used to construct confusion matrices for each classifier, allowing calculation of precision [TP / (TP + FP)], recall [TP / (TP + FN)], and F1-score [2 × (precision × recall) / (precision + recall)]. These metrics were then used to define confidence thresholds for validating detections, taking into account the requirements of the subsequent modeling approaches (see Authors 2026).

With the threshold metrics selected, the complete acoustic dataset was processed using both automated approaches (global BirdNET model and custom-trained classifier). The models were configured to detect the target species using a sensitivity setting of 1.0, a bandpass frequency range of 1–4000 Hz, and a minimum confidence threshold of 0.1. Detection tables generated by each classifier were then filtered according to confidence thresholds derived from performance analyses. Specifically, thresholds corresponding to maximum F1-score (“Max F1”) and maximum precision (“Max Precision”) were identified and applied following the analytical combinations described in the workflow diagram (Fig. S2).

The Max F1 threshold was used for naïve regression-based analyses, because it provides a balance between omission and commission errors by simultaneously maximizing precision and recall. In contrast, occupancy models were primarily fitted using detections filtered with the Max Precision threshold, as this criterion minimizes false-positive detections and therefore better satisfies its assumptions. Additionally, occupancy models were also fitted using detections

filtered under the Max F1 threshold. This complementary analysis was conducted to facilitate comparisons among modeling approaches, while maintaining a consistent detection threshold across analytical workflows.

Data Analysis

Naïve models (regression-based approach)

Our naïve modelling approach fitted nightly detection as a binary response variable (presence = 1 if the species was acoustically detected during the recording session; absence = 0 otherwise) using a Bernoulli Generalized Linear Mixed-effects Model (GLMM) with a logit link function. This approach was applied identically and independently to two detection datasets derived from the same acoustic recordings: one produced by a locally trained custom classifier (hereafter Custom) and one by the BirdNET pre-built classifier (hereafter PreBuilt; Kahl et al. 2021). Each observational unit was a site-night combination, yielding 1,753 initial observations across seven monitoring sites for each detector. In both cases, we used detections based on MaxF1.

Fixed-effect environmental predictors (Fig. S3) comprised air temperature, relative humidity, and moon illumination (proportion, 0–1). All predictors were standardized to zero mean and unit SD prior to analysis. Sampling effort (n_minutes, the number of unique one-minute recordings per night) was nearly constant across sessions (median = 56 min; 99% of nights ≥ 52 min) and was therefore not included as a predictor or offset term.

Because recordings were collected over consecutive nights within each site, we assessed first-order temporal dependence before fitting models to either dataset to address temporal pseudoreplication. For each detector, we computed the empirical transition matrix of nightly detection states, that is, the conditional probability of detecting the species on night t given its detection status on night $t-1$.

For the Custom dataset, the transition matrix revealed a fourfold difference in detection probability depending on the previous night's state: P(detected | detected yesterday) = 0.58 *vs.* P(detected | absent yesterday) = 0.14 (Table S2). For the PreBuilt dataset, this contrast was stronger: P(detected | detected yesterday) = 0.84 *vs.* P(detected | absent yesterday) = 0.29, suggesting greater temporal clustering of detections with the pre-built model. In both cases, this marked day-to-day persistence indicated substantial first-order temporal autocorrelation.

We therefore included “lag_presence” as an additional fixed effect in both models, treating detection as a first-order transition process conditional on environmental covariates (Zeger & Qaqish 1988; Diggle et al. 2002; see also Zuur et al. 2009). Observations immediately following monitoring gaps (non-consecutive nights) were excluded prior to modelling, yielding a final analytical dataset of 1,742 site-nights for each detector.

Non-independence of repeated observations within sites was addressed with a site-level random effect. To guide the choice of random effects structure, we computed autocorrelation functions (ACF) of randomized quantile residuals from a random-intercept-only baseline model (Hartig 2024) for each detector independently. In the Custom dataset, persistent positive autocorrelation at MIM and transient patterns at BEP and BAI indicated heterogeneous temporal dependence across sites. In the PreBuilt dataset, the same sites showed residual autocorrelation, but

patterns were generally more pronounced given the higher overall detection rate.

In both cases, we modelled the coefficient of lag_presence with a random slope in addition to a random intercept, allowing the strength of day-to-day temporal persistence to vary among sites. The full model structure was identical for both detectors:

$$presence_{it} \sim Bernoulli(p_{it})$$

$$logit(p_{it}) = \beta_0 + \beta_1 \times Temp_{it} + \beta_2 \times RH_{it} + \beta_3 \times Moon_{it} + \beta_4 \times lag\_presence_{it} + u_{0i} + u_{1i} \times lag\_presence_{it}$$

$$(u_{0i}, u_{1i})^T \sim N(0, \Sigma)$$

where *i* indexes site and *t* indexes night.

For each detector, we compared the random-slope model against a random-intercept-only alternative using a likelihood ratio test (LRT) and information criteria (AIC and BIC). In both cases the random-slope model was preferred on the basis of ΔAIC and LRT significance (see Results). Although ΔBIC was marginally positive in both comparisons, this reflects the total number of observations (n = 1,742) rather than the number of sites (n = 7) at which the random slope is estimated. We therefore consider BIC overly conservative for this component and retained the random-slope model.

Model fit was assessed using randomized quantile residuals (`DHARMa`; Hartig 2024). We evaluated overall uniformity (Kolmogorov–Smirnov test), dispersion, and leverages. Temporal autocorrelation in residuals was examined with site-specific Durbin–Watson tests applied to each site's `DHARMa` residual series ordered by date.

We quantified the relative contribution of each fixed-effect predictor to model-explained variance using hierarchical partitioning of the marginal R^2 (Nakagawa & Schielzeth 2013; `glmm.hp` package, Lai et al. 2022). Because lag_presence

represents temporal system state rather than an environmental covariate, we report its contribution separately from the hierarchical partitioning of temperature, relative humidity, and moon illumination. Semi-partial R^2 analogues were also computed as the difference in marginal R^2 between the full model and models with each predictor removed.

To aid biological interpretation, marginal predicted detection probabilities were estimated across the observed range of each predictor, with remaining predictors held at their mean and random effects marginalized (re.form = NA; `ggeffects` package, Lüdecke 2018). Confidence intervals were computed via the delta method. All analyses were performed in R version 4.5.2 (R Core Team 2025) using the packages `lme4` v2.0.1 (Bates et al. 2015), `DHARMa` v. 0.4.7 (Hartig 2024), `glmm.hp` v1.0.0 (Lai et al. 2022), `ggeffects` v2.3.2 (Lüdecke 2018), and `easystats` v.0.7.5 (Lüdecke et al. 2022).

Occupancy modeling

To account for imperfect detection, we fitted single-season occupancy models using a Bayesian framework implemented in the R package `sp0ccupancy` v0.9.0 (Doser et al. 2022). Specifically, we used a Pólya-Gamma data augmentation scheme for efficient posterior sampling from binary occupancy models (Polson et al. 2013). In this framework, each observational unit (*site*) was defined as a unique site-night combination, i.e. the same unit used in the regression-based approach. Within each site-night, the individual one-minute recordings constituted repeated surveys (detection replicates), yielding a binary detection matrix y of dimension $J \times K$, where J is the number of site-nights and K is the maximum number of one-minute

recordings within a night. Site-nights with fewer recordings than the maximum were padded with NA values to produce a rectangular matrix.

The occupancy sub-model included the same three environmental predictors used in the regression-based approach: air temperature, relative humidity, and lunar illumination. Lunar illumination was log-transformed prior to standardization (Fig. S3). All three covariates were z-standardized (zero mean, unit SD). These covariates were treated as site-level (night-level) variables, constant across the repeated surveys within a given site-night.

The detection sub-model included two recording-level acoustic summaries as covariates: the Bioacoustic Index (BIO; Boelman et al. 2007) and the Acoustic Complexity Index (ACI; Pieretti et al. 2011), both z-standardized. These indices capture variation in the acoustic environment across individual one-minute recordings and were included to account for heterogeneity in detection probability associated with background soundscape conditions. The BIO measures the accumulated acoustic energy within a defined frequency band and is commonly used as a proxy for biological acoustic activity. We calculated BIO using the same frequency range targeted by the automated classifiers and encompassing the vocalization bandwidth of the focal species (0–4,000 Hz). The ACI quantifies temporal variation in sound intensity across frequency bins and is often interpreted as an indicator of acoustic activity and soundscape complexity. Unlike BIO, ACI was calculated across the full frequency range captured by the recordings. Acoustic indices were calculated in R using the `soundecology` package (Villanueva-Rivera & Pijanowski 2018). In addition, we included a random intercept for the physical recorder site (i.e. the seven monitoring locations) in the detection sub-model, to

absorb site-level variation in detection probability attributable to recorder placement, local acoustic conditions, and equipment characteristics.

The model structure can be summarized as:

$$Occupancy: logit(\psi_j) = \beta_0 + \beta_1 \times Temp_j + \beta_2 \times RH_j + \beta_3 \times \log(Moon)_j$$

$$Detection: logit(p_{jk}) = \alpha_0 + \alpha_1 \times BIO_{jk} + \alpha_2 \times AIC_{jk} + n_{site(j)}$$

where ψ_j denotes the occupancy probability of site-night *j*, p_jk denotes the detection probability for the *k*-th survey within site-night *j*, and η_site(j) is a normally distributed random intercept for the recorder site associated with site-night *j*.

We note that in this framework, 'occupancy' (ψ) is interpreted as the probability that the species is *vocally active* during a given *site-night*, rather than the *probability of spatial occupancy* at a monitoring location. This interpretation follows the 'use' framework of Nichols et al. (2008), in which the parameter of interest is the probability that a site is used during a primary sampling occasion. The closure assumption — that the species' vocal availability does not change within a night — is reasonable for a nocturnally active resident species whose vocal bouts typically span the entire nocturnal period.

We applied this model independently to four detection datasets defined by the factorial combination of two classifiers (Custom, PreBuilt) and two confidence thresholds (MaxF1, MaxPrecision), yielding four parallel occupancy analyses (Fig. S2). We specified weakly informative normal priors (mean = 0, variance = 2.72) for all occupancy (β) and detection (α) regression coefficients. For the variance of the detection random effect, we placed an inverse-gamma prior with shape = 0.1 and scale = 0.1, following the default parameterization in `spOccupancy` (Doser et al. 2022). These priors place minimal constraint on posterior estimates, as confirmed by the ESS and convergence diagnostics.

For each of the four models, we ran four parallel MCMC chains of 20,000 iterations each, discarding the first 4,000 as burn-in and thinning by a factor of 10, yielding 6,400 posterior samples per model (1,600 per chain). Convergence was assessed using the Gelman–Rubin diagnostic ($\hat{R}$; Gelman & Rubin 1992) and effective sample size (ESS) for all occupancy and detection parameters. We considered $\hat{R} < 1.05$ and ESS > 400 as evidence of adequate convergence. Traceplots were visually inspected for all parameters. Model predictive performance was compared using the Widely Applicable Information Criterion (WAIC; Watanabe 2010) for each of the four models. Marginal occupancy probabilities were derived as posterior predictive distributions across the observed range of each environmental covariate, holding remaining covariates at their mean (zero on the standardized scale). Analyses were performed in R package `coda` v0.19-4 (Plummer et al. 2006). Posterior predictive checks (PPC; Kéry and Royle 2016) were computed using 250 posterior draws to reduce computational burden associated with the large detection matrix ($J \times K \approx 98{,}000$ cells per model).

Comparison of approaches

To assess how modeling framework and detector choice jointly influence ecological inference, we compared the regression-based (GLMM) and occupancy modeling approaches along three axes. First, we compared the direction, magnitude, and statistical support of the environmental effects (temperature, relative humidity, and lunar illumination) estimated under each approach. However, the two approaches estimate fundamentally different quantities: the GLMM estimates the log-odds of *acoustic detection* on a given night, conflating occupancy and detectability into a

single process, whereas the occupancy model estimates the log-odds of *true occupancy* (the probability that the species is vocally active), with detection probability modelled as a separate process (Tredennick et al. 2021). Consequently, we do not treat the coefficient comparison as a test of which model is 'correct', but rather as a diagnostic of how separating the observation process from the ecological process alters inference about environmental effects. Specifically, we compare the direction, relative importance, and uncertainty structure of estimated effects between frameworks, rather than treating the numerical estimates as directly interchangeable.

Second, we examined whether the two classifiers (Custom and PreBuilt) led to qualitatively different ecological conclusions under each modeling framework. Specifically, we compared whether the sign, relative importance, and credible/confidence interval overlap of environmental effects were consistent between classifiers within the same model class, and whether any discrepancies were amplified or attenuated by the choice of modeling framework.

Third, we compared model-level summaries of explanatory power. For the GLMM, we used marginal and conditional R^2 (Nakagawa & Schielzeth 2013) and hierarchical partitioning of the marginal R^2 (Lai et al. 2022). For the occupancy models, we used WAIC and k-fold cross-validation deviance as measures of predictive accuracy. Although these metrics are not directly comparable across model classes, they provide complementary perspectives on how well each approach captures the variation in the data. Together, this comparison framework allows us to disentangle the influence of three methodological decisions on inference: (1) the choice of automated detector, (2) the confidence threshold applied to detector output, and (3) the statistical modeling framework used to relate detections to environmental drivers.

RESULTS

Overall, both classifiers showed good performance in distinguishing target vocalizations from non-target recordings (Fig. 3). However, the Custom classifier had consistently higher performance for across confidence thresholds, with a substantially greater area under the curve (AUC = 0.971) compared to the BirdNET pre-built model (AUC = 0.869), reflecting improved classification accuracy under local acoustic conditions. Across confidence thresholds, classifiers performance varied substantially between the BirdNET pre-built model and the custom-trained classifier (Fig. 3, Tab. S3). The custom classifier consistently achieved higher recall and F1-scores over most threshold values, while also maintaining high precision at intermediate and high confidence thresholds. In contrast, the global BirdNET model showed relatively high precision even at low thresholds, but recall declined rapidly as confidence thresholds increased, resulting in lower overall F1-scores.

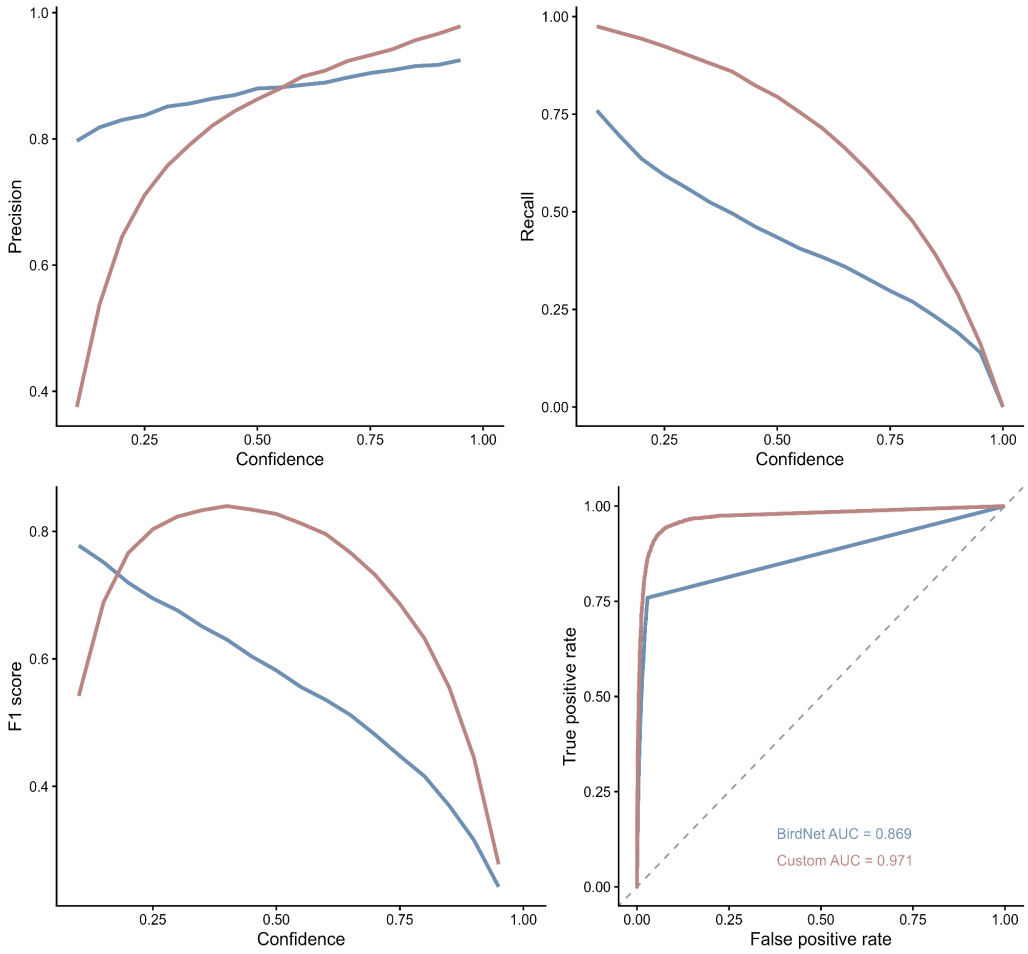


Figure 3. Performance comparison between BirdNET and the custom classifier across detection thresholds. Panels show changes in precision, recall, F1 score, a ROC curve summarizes the overall discrimination performance of each classifier. Blue lines represent BirdNET predictions, whereas red lines represent the custom classifier.

Nightly detection rates differed markedly between detectors. The Custom classifier produced a nightly presence rate of 25.3% (444 presence nights out of 1,753), whereas the BirdNET pre-built model yielded 64.5% (1,130 out of 1,753; Table 1, Fig. S4). Detection rates varied among sites under both approaches but were consistently higher under the PreBuilt detector at every site (Table 1, Fig S4). Performance metrics also differed substantially between classifiers and thresholding strategies (Table S3). The custom-trained classifier achieved the highest F1-score

when using a confidence threshold of 0.40 (F1 = 0.84), compared to the pre-built BirdNET model at its corresponding Max F1 threshold of 0.10 (F1 = 0.78). Under the Max Precision strategy, both classifiers used a threshold of 0.95, with the custom classifier achieving slightly higher precision (0.98 vs. 0.92), although recall remained low in both cases.

Table 1. Number of nights with detections per site and detector before threshold application.

Site	Number of nigths	BirdNET pre-built	Custom classifier
ARA	338	223	100
BAI	227	139	56
BEP	141	60	14
MIM	352	173	118
REF	226	202	68
UEM	223	131	29
XAR	246	202	59

Naïve approach (GLMM-based regression)

Residual diagnostics revealed site-specific temporal autocorrelation in baseline models, which were handled with a site-specific random slope structure (see Supplementary Material).

Variance on the random intercept for 'Site' was higher for the PreBuilt classifier (SD = 0.83) compared to the Custom classifier (SD=0.34), indicating stronger baseline differences in detection probability among sites with the PreBuilt classifier. The random slope variance for lag_presence was 0.61 for PreBuilt and 0.2 for the Custom classifier, underscoring that the PreBuilt dataset had highly variable

temporal persistence among sites. The correlation between the random intercept and slope was low for the PreBuilt (-0.05) and strongly negative (-0.91) for the Custom classifier, indicating that higher site-level baseline detection probability is associated with weaker temporal persistence effects only for the Custom classifier dataset. Overall, compared to the Custom classifier, the PreBuilt classifier showed greater spatial heterogeneity in baseline detection rates but lower variation among sites in the effect of temporal dependence.

Fixed-effect estimates differed notably between detectors in both magnitude and direction for some predictors (Fig. 4, Table S4.). Humidity had the strongest environmental effect for both classifiers, with differences in magnitude order. Under the PreBuilt classifier, with detection probability declined markedly from approximately 98% at the lowest observed humidity values to approximately 42% at the highest (Fig. 4; OR = 0.55, 95% CI = 0.470.64), whereas the Custom classifier's detection probabilities declined sharply with increasing humidity, dropping from approximately 55% at the lowest observed values (~30% RH) to approximately 10% at the highest (~100% RH), with other predictors held at their means (Fig. 4; OR = 0.6695% CI = 0.58–0.75).

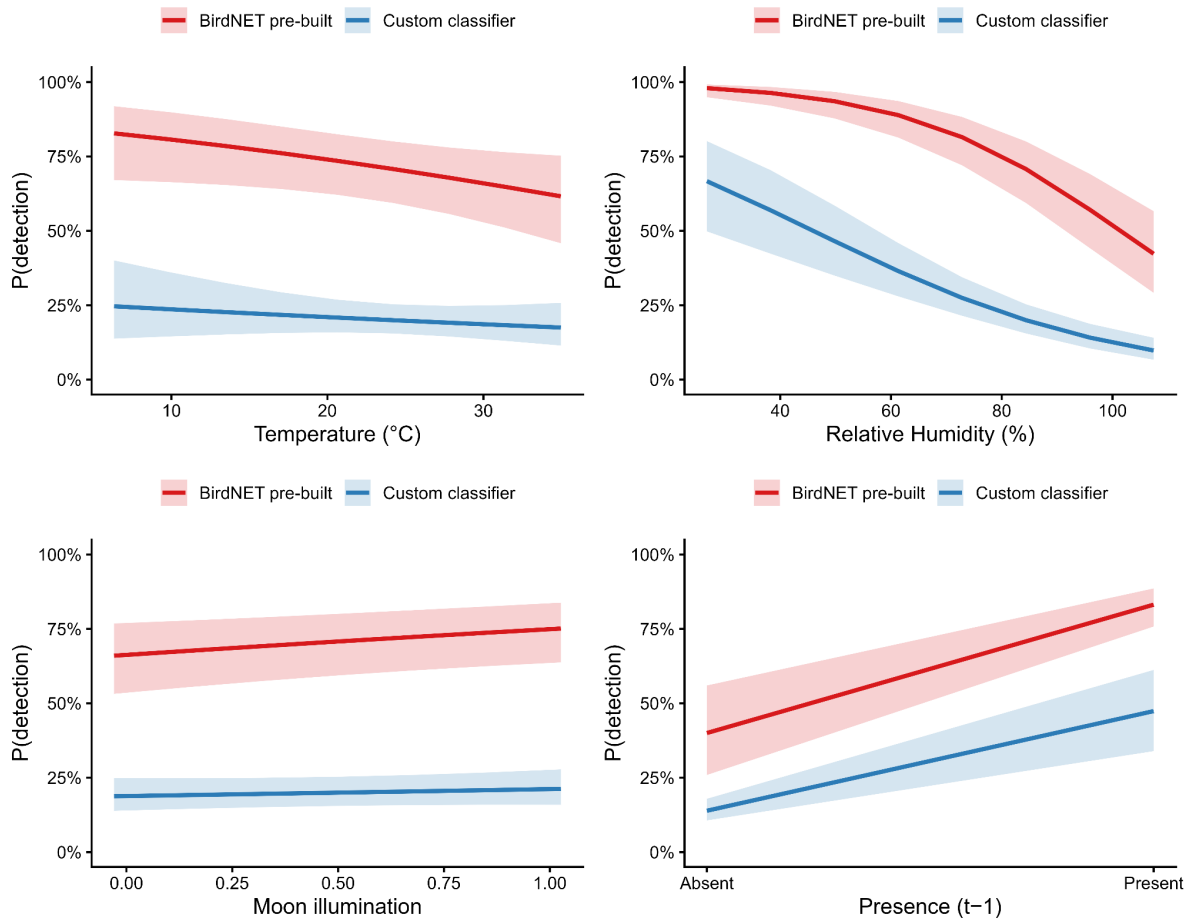


Figure 4. Marginal predicted detection probabilities for the Custom classifier and the BirdNET pre-built model in relation to temperature, relative humidity, moon illumination, and detection status on the previous survey night (presence at $t - 1$). Solid lines represent model predictions and shaded ribbons indicate 95% confidence intervals.

Temperature had a moderate negative effect on PreBuilt's detection probability, decreasing from approximately 83% on the coolest nights to approximately 62% on the warmest (OR = 0.87, 95% CI = 0.76–1.00; Fig. 4). However, the effect was non-significant and weaker on the Custom classifier's detection probabilities (OR = 0.95, 95% CI = 0.83–1.08; Table S4). Moon illumination showed a weak positive association for the PreBuilt's dataset, with detection probability increasing from approximately 66% under the darkest

conditions to approximately 75% under the brightest (OR = 1.16, 95% CI = 1.02–1.31; Fig. 4). In contrast, detection probabilities from the Custom classifier were weakly positively associated with moon illumination (OR = 1.16, 95% CI = 1.02–1.31, Table S4).

The lagged detection covariate was the strongest overall predictor for both datasets: PreBuilt's detection probabilities were substantially higher following nights with previous detections (OR = 7.40, 95% CI = 5.25–10.43; Fig. 4), whereas Custom classifier's detection probabilities increased from approximately 15% following a night of absence to approximately 45% following a night of presence (OR = 5.58, 95% CI = 3.24–9.62; Fig. 4).

The direction of environmental effects generally agreed between detectors, but effect sizes were larger under the PreBuilt model, particularly for humidity and lagged presence, consistent with the generally higher detection rates produced by the PreBuilt classifier (Table S4).

Hierarchical partitioning of the marginal R^2 revealed consistent qualitative patterns across detectors, but with notable quantitative differences (Fig. S5). In both models, lag_presence dominated the explained variance, accounting for 81.58% and 79.70%of the marginal R^2 under the PreBuilt and Custom classifiers, respectively. Among the three environmental predictors, relative humidity contributed the largest share under both detectors (16.26% vs 19.09%), followed by temperature (0.98% vs 0.87%) and moon illumination (1.18% vs 0.34%). The full models yielded marginal R^2 values of 0.30 and 0.21, and conditional R^2 values of 0.39 and 0.25, for the PreBuilt and Custom classifiers, respectively. The larger marginal R^2 under the PreBuilt classifier suggests that the higher overall detection rate provided greater statistical power to detect environmental effects.

Occupancy models

All occupancy models achieved adequate convergence, with well-mixed MCMC chains and no evidence of instability (see Supplementary Material). However, model fit differed markedly among detector vs threshold combinations. Models based on MaxPrecision thresholds showed substantially better calibration and lower WAIC values than those based on MaxF1 thresholds, indicating greater compatibility between detection data and the occupancy-model assumptions.

The direction of environmental effects on occupancy was consistent across all four models, but their magnitude varied substantially with detector and threshold choice (Fig. 5; Table S7).

Relative humidity had a negative effect on occupancy probability across all models. The effect was strongest under PreBuilt — MaxPrecision (posterior mean -1.65) and weakest under Custom — MaxF1 (-0.51). Temperature showed a similarly consistent negative effect, ranging from -0.12 posterior mean under Custom — MaxF1 to -1.810 under PreBuilt — MaxPrecision. For both covariates, the 95% credible intervals excluded zero in all four models (Table S7), providing strong evidence of negative associations with vocal activity.

Lunar illumination showed generally weak and inconsistent associations with occupancy across models (Fig. 5; Table S7). Posterior means ranged from −0.050 to 0.309 on the log-odds scale. Credible intervals overlapped zero in two of the four model configurations (Custom — MaxF1 and PreBuilt — MaxPrecision), indicating limited support for a consistent moonlight effect across detectors and thresholding strategies. In contrast, both the Custom — MaxPrecision model (mean = 0.309, 95%

CI = 0.040–0.599) and the BirdNET — MaxF1 model (mean = 0.192, 95% CI = 0.063–0.319) showed positive associations between lunar illumination and occupancy, with credible intervals excluding zero.

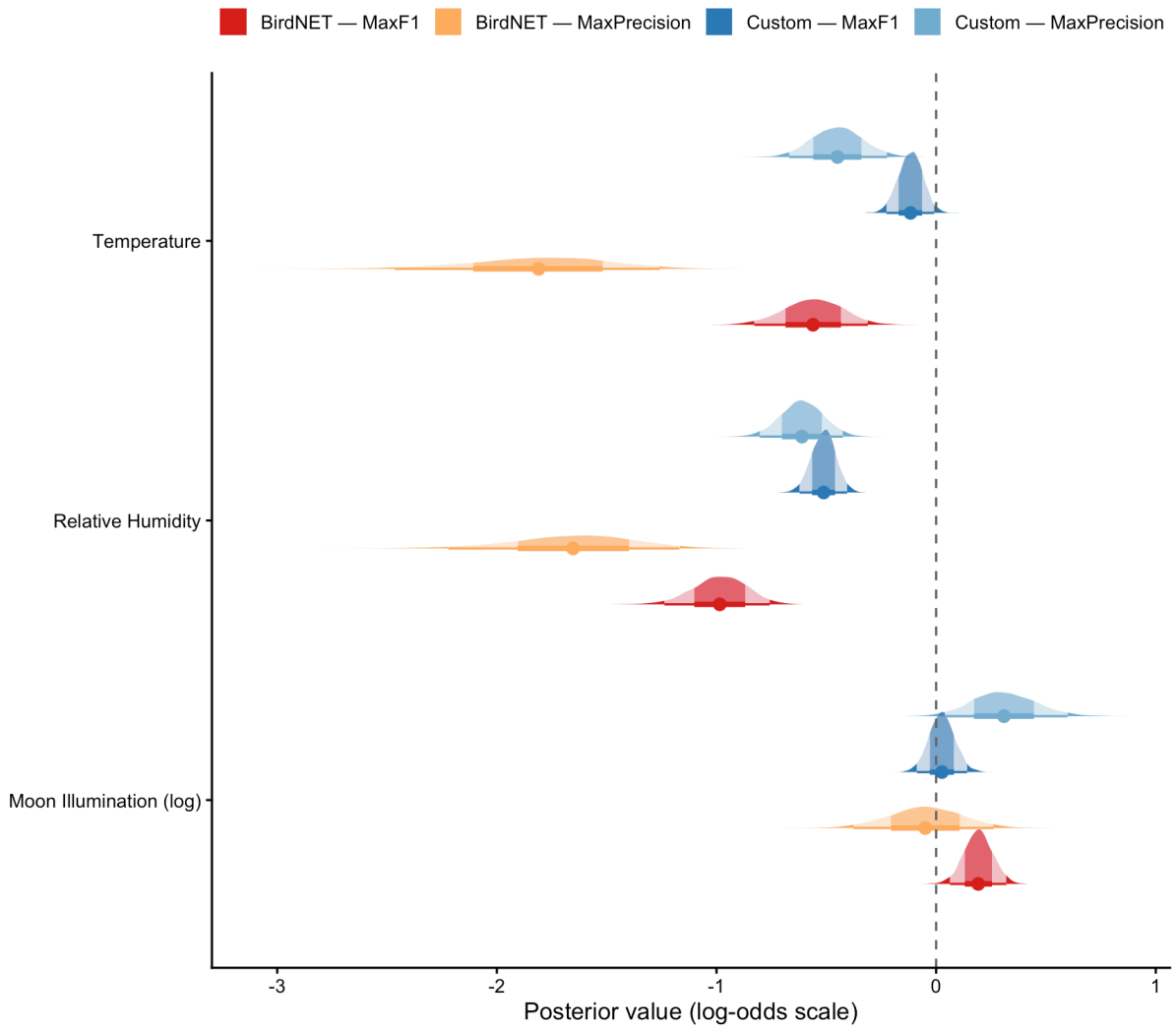


Figure 5. Posterior means ± 95% credible intervals for occupancy coefficients. Dashed line at 0 (null effect). Coefficients are z-standardized; positive values indicate higher occupancy probability.

A notable pattern emerged across threshold choices: MaxPrecision models consistently estimated larger absolute effect sizes than MaxF1 models for both

temperature and relative humidity, while also exhibiting wider posterior distributions (Fig. 5). This likely reflects the lower detection rates under MaxPrecision thresholds, which reduce sample size but also reduce false-positive contamination, allowing the occupancy model to estimate environmental effects with less bias but greater uncertainty.

Detection coefficients

The Bioacoustic Index (BIO) had a positive effect on detection probability in all four models (Fig. 6), indicating that recordings with higher bioacoustic activity were more likely to contain detectable vocalizations of *N. albicollis*. This effect was strongest under MaxPrecision for both models and weakest under PreBuilt — MaxF1.

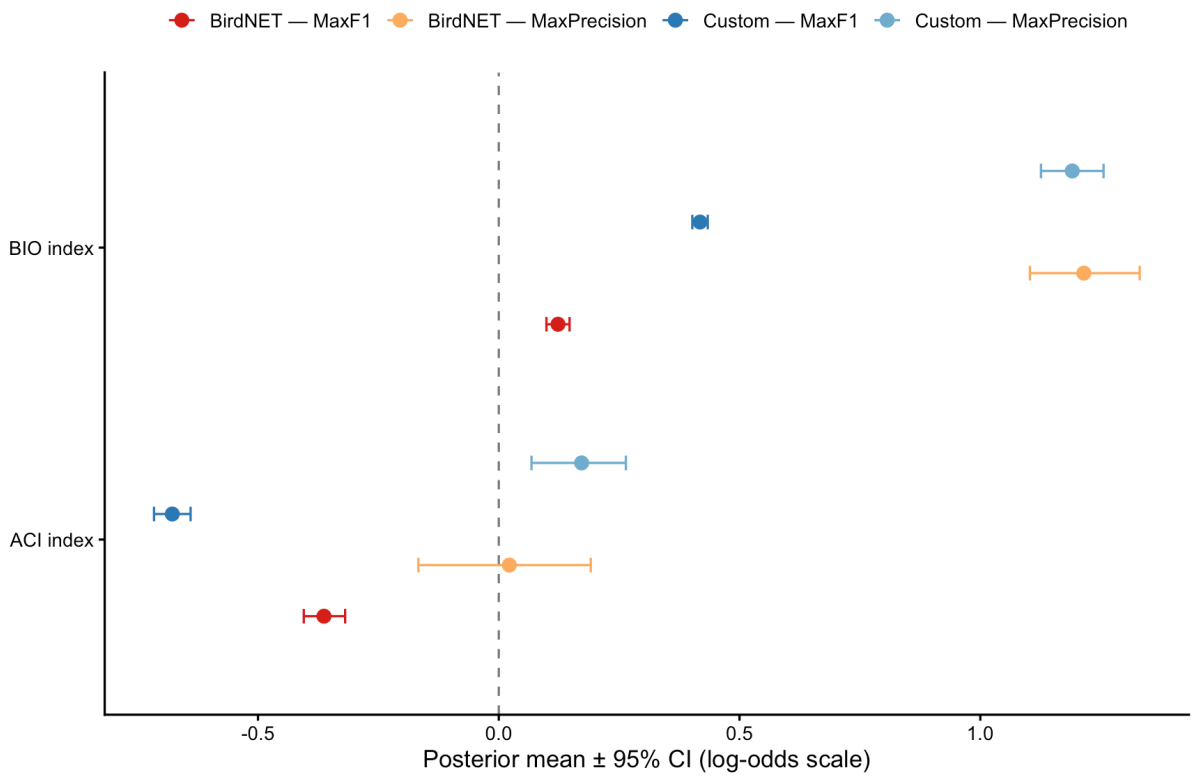
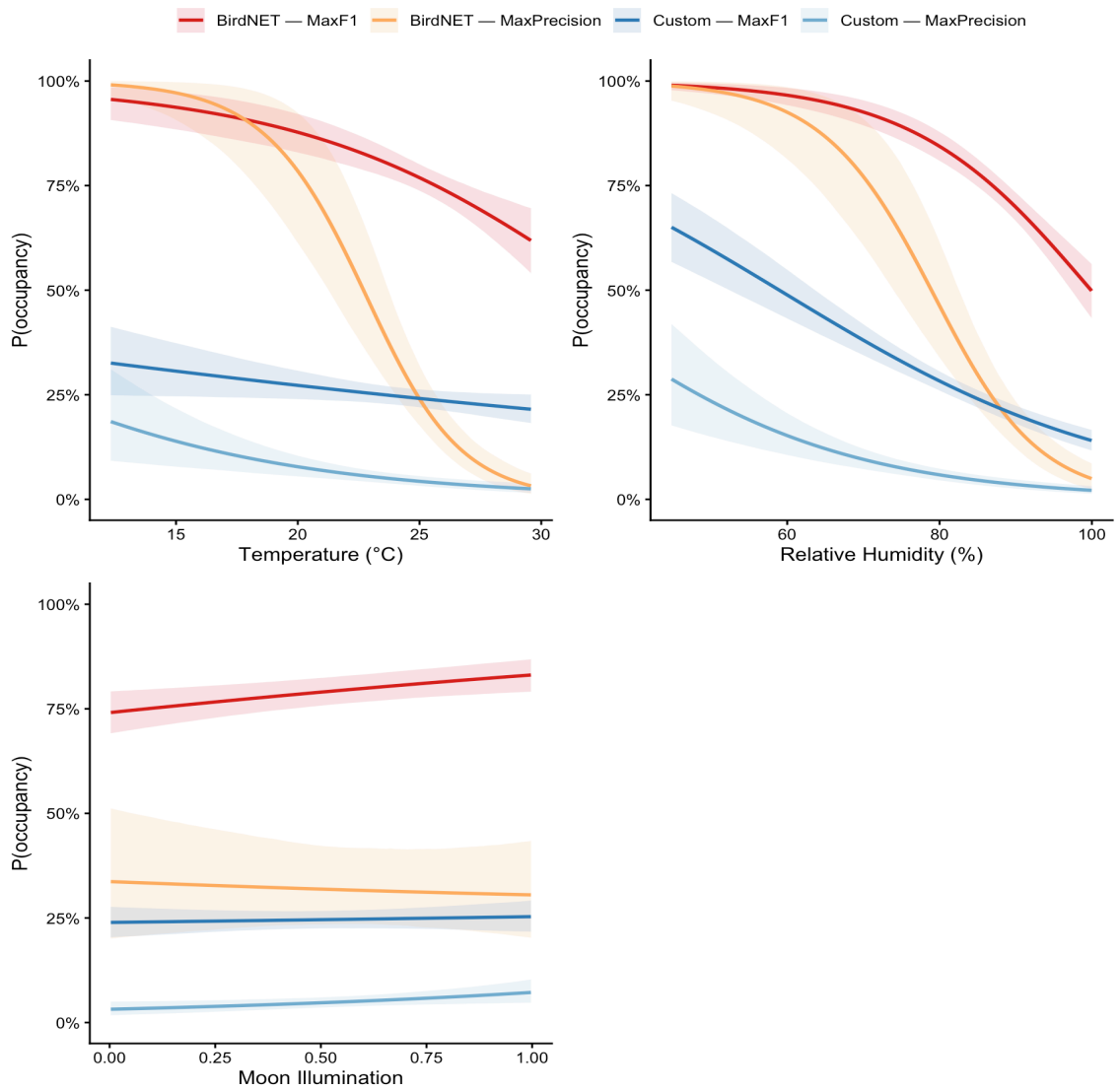


Figure 6. Posterior means ± 95% CI for detection (α) parameters. Positive BIO/ACI coefficients indicate higher detection probability in acoustically complex recordings.

The Acoustic Complexity Index (ACI) showed variable effects across models (Fig. 6). Under Custom — MaxF1, ACI had a strong negative association with detection probability, whereas under MaxPrecision models, the effect was near zero or weakly positive. This inconsistency suggests that the relationship between ACI and species-specific detection depends on the confidence threshold applied to classifier output, likely because lower thresholds admit more false positives in acoustically complex environments.

Marginal occupancy predictions

Marginal predicted occupancy probabilities, obtained by varying each covariate across its observed range while holding others at their mean, differed dramatically across models (Fig. 7). Under PreBuilt — MaxF1, predicted occupancy at low temperatures (~13 °C) exceeded 90%, declining to approximately 75% at the warmest observed nights (~30 °C). Under Custom — MaxPrecision, by contrast, predicted occupancy ranged from approximately 20% at cool temperatures to near 0% at warm temperatures. Similar divergence was observed across the relative humidity gradient: BirdNET — MaxF1 predicted near-complete occupancy at low humidity, whereas Custom — MaxPrecision predicted occupancy below 25% across most of the humidity range. These differences illustrate how the choice of automated detector directly propagates into the ecological quantities estimated by the occupancy model.



592

593 Figure 7. Marginal predicted occupancy probability of *N. albicollis* as a function of
594 each environmental predictor. Lines = posterior mean; shaded bands = 95% credible
595 interval. Covariates are on their original scales for interpretability.

596

597 *Comparison of approaches*

598 Coefficient estimates from the naïve GLMM and the occupancy model were
599 qualitatively consistent in the direction of environmental effects across all four
600 detector × threshold combinations, but differed in magnitude and uncertainty
601 structure in ways that depended on both the classifier and the thresholding strategy
602 (Fig. 8). Under the MaxF1 threshold, the two modeling frameworks produced broadly

similar estimates for both detectors. For the Custom classifier, GLMM and occupancy coefficients for temperature, relative humidity, and moon illumination were closely aligned, with confidence and credible intervals largely overlapping (Fig. 8, upper-left panel). For the BirdNET pre-built model, the occupancy posterior means were shifted away from zero relative to the GLMM point estimates for temperature and relative humidity (Fig. 8, lower-left panel), indicating that the occupancy framework estimated somewhat stronger environmental effects once imperfect detection was explicitly modelled.

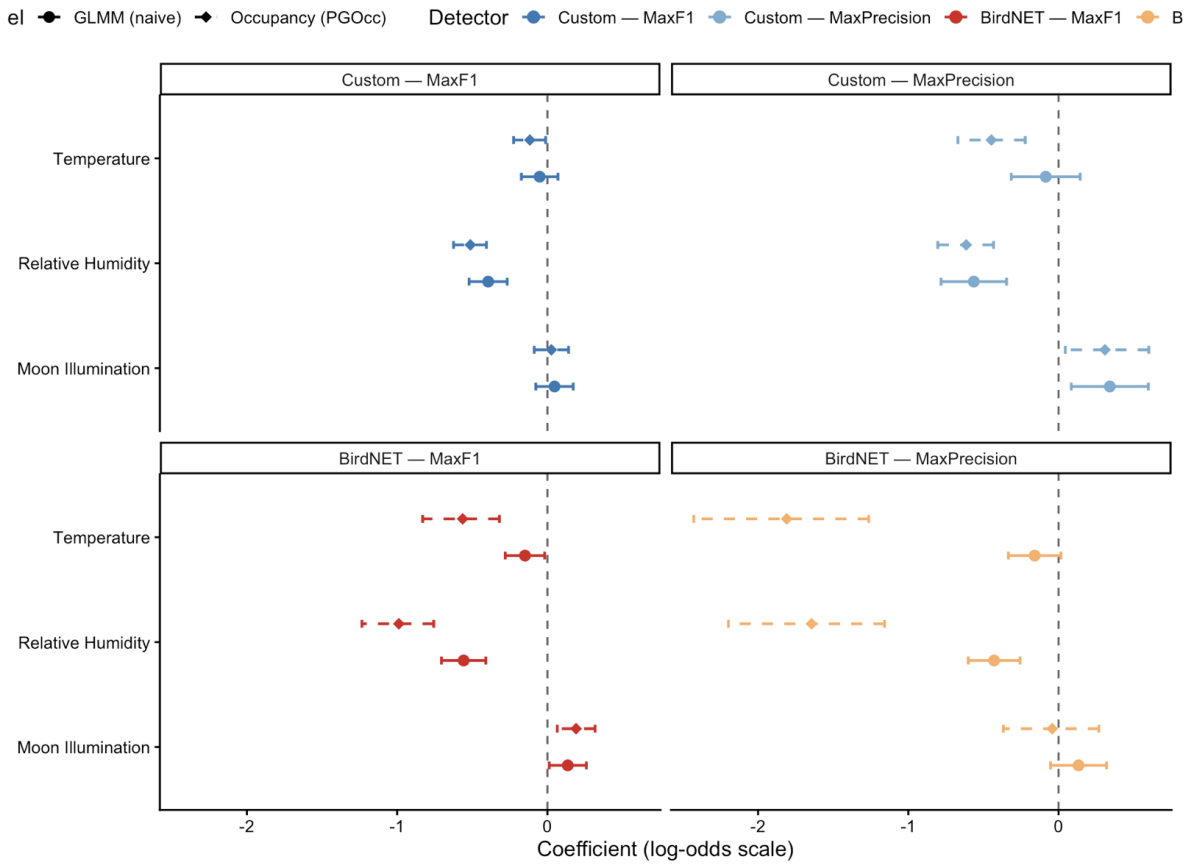


Figure 8. Coefficient comparison between the naive GLMM and the occupancy model across all four detector × threshold combinations. Circles/solid = GLMM; diamonds/dashed = Occupancy. Dashed vertical line at 0 (null effect).

Under the MaxPrecision threshold, the discrepancy between frameworks was substantially amplified (Fig. 8, right panels). In both detectors, the occupancy model

estimated environmental effect sizes that were considerably larger in absolute magnitude than the corresponding GLMM estimates. This pattern was most pronounced for PreBuilt — MaxPrecision, where the occupancy posterior means for temperature and relative humidity were approximately two to three times larger than the GLMM point estimates (Fig. 8, lower-right panel). The Custom — MaxPrecision combination showed a similar but more moderate pattern, with occupancy estimates for all three covariates shifted further from zero relative to the GLMM (Fig. 8, upper-right panel).

The full posterior distributions (Fig. 9) provided additional insight into how uncertainty structure differed between frameworks and analytical pathways. Under MaxF1, both the Custom and PreBuilt posteriors were relatively narrow, with GLMM point estimates falling within or near the central mass of the occupancy posteriors. Under MaxPrecision, the posteriors were markedly wider — particularly for PreBuilt — MaxPrecision, where the distributions for temperature and relative humidity spanned several log-odds units. In these cases, the GLMM point estimate (×) consistently fell near the upper tail of the occupancy posterior, indicating that the naïve approach produced attenuated effect sizes relative to the hierarchical model when applied to high-precision detection data. This attenuation is consistent with the expectation that naïve models underestimate environmental effects by conflating the ecological and observation processes: when false positives are largely eliminated by the MaxPrecision threshold, the remaining detections more closely approximate true occupancy events, and the occupancy model can recover larger environmental effects that the GLMM, which models raw detection probability, cannot distinguish from baseline noise.

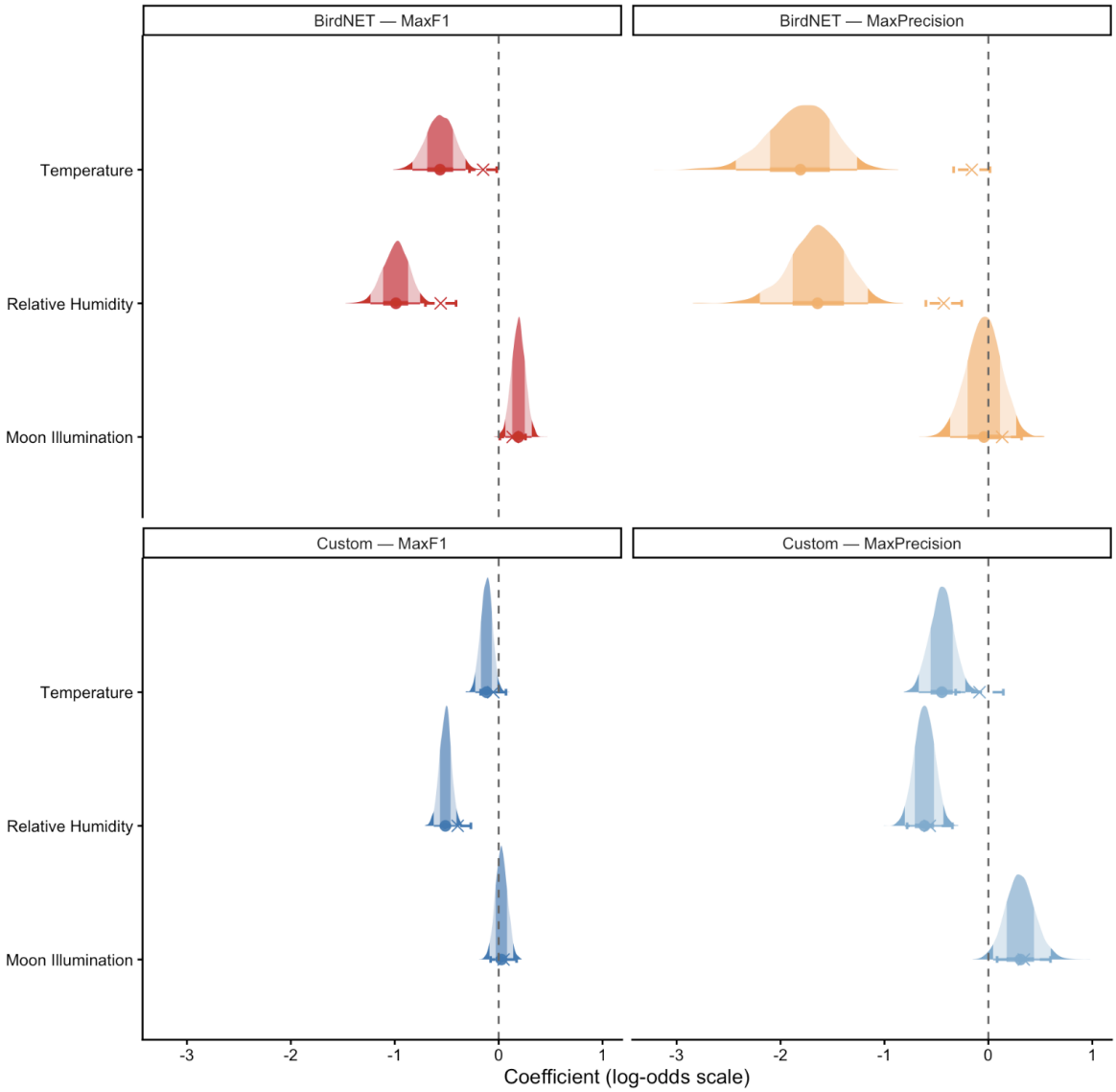


Figure 9. Full posterior distributions (occupancy model, filled halfeye) alongside GLMM point estimates ± 95% CI (× with dashed bars) across all four detector × threshold combinations.

A notable interaction emerged between detector quality and threshold choice. For the Custom classifier, switching from MaxF1 to MaxPrecision moderately increased the gap between GLMM and occupancy estimates, reflecting the reduction in sample size and the corresponding widening of posteriors. For the PreBuilt classifier, the same switch produced a much larger divergence, suggesting that MaxPrecision thresholds applied to the pre-built model not only reduced false

positives but also substantially changed the composition of the retained detection data — likely by removing condition-dependent false positives that had systematically biased environmental relationships in the MaxF1 dataset.

Together, these results demonstrate that the degree of discrepancy between naïve and hierarchical modeling approaches is not fixed, but depends on the upstream detection workflow. When detection data are of high quality — either through locally trained classifiers or stringent confidence thresholds — the GLMM provides a reasonable approximation of directional environmental effects, though with attenuated effect sizes and underestimated uncertainty. When detection data contain systematic noise, the occupancy model becomes essential for recovering unbiased environmental relationships, but the quality of its inference is bounded by the structure of the detection data it receives.

DISCUSSION

Our study demonstrates that methodological decisions made across the passive acoustic monitoring workflow — from detector choice and confidence threshold to statistical modeling framework — jointly shape ecological inference in ways that cannot be anticipated by evaluating any single step in isolation. Using acoustic data from *Nyctidromus albicollis* in the Pantanal, we showed that the direction of environmental effects on vocal activity was robust across all analytical pathways, but the magnitude of those effects, the predicted occupancy probabilities, and the calibration of fitted models varied dramatically depending on the combination of decisions applied to the same underlying recordings.

Methodological sensitivity: the garden of forking paths

The largest source of variation in estimated effect sizes was not the choice of statistical framework (GLMM *vs.* occupancy model), but the combination of automated detector and confidence threshold. Predicted occupancy at low temperatures differed by more than 65% between the most and least conservative analytical pathways (PreBuilt BirdNET — MaxF1 *vs* Custom — MaxPrecision), a discrepancy that would lead to fundamentally different ecological interpretations of the same data. This finding echoes the broader concern that analytical flexibility in ecology and evolutionary biology can produce substantial variation in conclusions drawn from a single dataset (Gould et al. 2025), and extends this concern to the upstream data-processing stages that are rarely subjected to the same scrutiny as model specification.

In passive acoustic monitoring, the sequential nature of analytical decisions compounds this issue (Gelman and Loken 2014, Steegen et al. 2016). The detector determines *which* recordings are labeled as detections; the confidence threshold determines *how many* of those labels are retained; and the modeling framework determines *how* those detections are related to environmental covariates. Our factorial design reveals that these decisions do not act independently: the difference between GLMM and occupancy estimates was small under the Custom classifier but substantially larger under BirdNET, indicating an interaction between detector quality and statistical framework. This interaction has practical implications: a well-calibrated detector can partially compensate for a simpler modeling approach, whereas a noisy detector amplifies the bias introduced by ignoring detection heterogeneity.

Precision versus recall in hierarchical models

A striking result was the dramatic difference in model calibration between confidence threshold strategies. MaxPrecision models yielded WAIC values that were one to two orders of magnitude lower than MaxF1 models, and posterior predictive checks confirmed that only MaxPrecision datasets were adequately described by the assumed data-generating process. The inflated effective number of parameters (pD) in MaxF1 models — particularly Custom — MaxF1 (pD = 342.5) — indicates that the occupancy model was forced to absorb systematic noise from false-positive detections by overfitting its random effects structure.

This result has important implications for the design of PAM workflows intended to feed into hierarchical models. The conventional wisdom in species detection is often to maximize recall (sensitivity) to avoid missing true presences, accepting some false positives as a tolerable cost (MacKenzie et al. 2018, Guilleta-Arroita 2017). However, our results show that when detection data are subsequently analyzed with occupancy models, false positives are not merely noise (Royle & Link 2006) — they represent a systematic violation of the model's closure assumption at the survey level (i.e. the assumption that a detection implies the species was truly present and available). Under these conditions, prioritizing precision, accepting fewer but more reliable detections, produces substantially better-calibrated models and less biased coefficient estimates, even at the cost of reduced statistical power from lower detection rates. This aligns with recent guidelines recommending careful threshold calibration for BirdNET and similar tools when scores are used in downstream statistical analyses (Wood & Kahl 2024).

The wider posterior distributions observed under MaxPrecision models reflect this tradeoff: greater data quality comes at the cost of greater statistical uncertainty. We argue that this honest uncertainty is preferable to the spurious precision of MaxF1

models, whose narrow posteriors are conditional on a poorly specified data-generating process. In practice, the choice between precision and recall should be guided by the downstream analytical framework: studies using simple summaries of detection frequency may benefit from higher recall, whereas those applying hierarchical models should prioritize precision. However, such decisions are rarely informed by dataset-specific evaluations of precision and recall, making the ecological consequences of threshold choice difficult to anticipate under local acoustic conditions.

The value of locally trained classifiers

The Custom classifier, trained on recordings from the study region, produced detection data that were more compatible with both modeling frameworks than the BirdNET pre-built model. Under the Custom classifier, GLMM and occupancy estimates were more closely aligned, suggesting that unmodeled detection heterogeneity was smaller — likely because the locally trained model produced fewer false positives and more consistent detection probabilities across sites and conditions. In contrast, the BirdNET model, trained on a global dataset, yielded higher baseline detection rates and stronger temporal clustering (transition probability of 0.84 versus 0.58 for consecutive detection nights), consistent with a higher rate of false-positive detections that inflated apparent occupancy.

This finding is consistent with recent global evaluations showing substantial variation in BirdNET performance across continents, biomes, and species (Funosas et al. 2026), and with evidence that fine-tuning classifiers to local acoustic conditions improves detection accuracy (Schiavo et al. 2025, Pérez-Granados et al. 2026). From the perspective of ecological inference, the advantage of local training goes beyond improved classification metrics: it produces detection data that better satisfy

751 the assumptions of hierarchical models, reducing the bias in estimated
752 environmental effects. However, training custom classifiers requires annotated
753 acoustic data, taxonomic expertise, and computational resources that may not
754 always be available. Our results suggest that when custom training is not feasible
755 and that a classifier performance is reasonable, applying stringent confidence
756 thresholds to pre-built models can partially compensate, as the BirdNET —
757 MaxPrecision model achieved the best overall calibration (lowest WAIC) among all
758 four models. This suggests that threshold optimization may be a more accessible
759 intervention than full model retraining for monitoring programs focused on species
760 well represented in prebuilt models. However, we caution against generalizing this
761 result to other taxa, regions, or soundscapes without local validation. The
762 transferability of globally trained classifiers to tropical ecosystems, particularly in the
763 Global South, where acoustic communities are often more diverse and
764 underrepresented in training datasets, remains uncertain. Under such conditions,
765 relying exclusively on pre-built models without evaluating local precision–recall
766 performance may propagate substantial ecological bias into downstream analyses.
767
768 *Environmental drivers of nightjar vocal activity*
769 Despite the substantial methodological variation documented above, the qualitative
770 ecological conclusions were remarkably consistent across all analytical pathways.
771 Occupancy probability of *N. albicollis* declined with increasing temperature and
772 relative humidity, and showed a weak positive association with lunar illumination.
773 The robustness of these directional effects provides confidence that the underlying
774 ecological signal is genuine, even though its precise magnitude remains sensitive to
775 analytical choices.

776 The negative effect of temperature on vocal activity is consistent with the
777 thermal ecology of caprimulgids, which are adapted to exploit cool nocturnal
778 conditions for aerial insectivory (McKechnie et al. 2023). This pattern is strongly
779 corroborated by recent findings in subtropical Yungas forests, where the vocal
780 activity of the rufous nightjar (*Antrostomus rufus*) was negatively predicted by
781 temperature (Schaaf et al. 2023). At higher temperatures, reduced thermal gradients
782 between the bird and its environment may decrease the energetic advantage of
783 nocturnal foraging, or prey availability may shift in ways that reduce vocal activity
784 (McKechnie & Ashdown 2008, O'Connor et al. 2018). The strong negative effect of
785 relative humidity likely reflects the suppressive influence of moist conditions, and
786 their association with precipitation events, on both insect flight activity and the
787 acoustic transmission environment (Taylor 1963, Grüebler et al. 2008). High humidity
788 dampens sound propagation and may also reduce the motivation to vocalize if
789 foraging conditions are poor (Marten et al. 1977, Richards and Wiley 1980,
790 Tsioutsiourigas et al. 2025). This pattern is consistent with the annual vocal cycle of
791 *N. albicollis* in the Brazilian Pantanal, where vocal activity drops sharply during the
792 rainy season (November–April), coinciding with peak humidity and flooding, and
793 recovers only as conditions dry out toward the end of the dry season (Pérez-
794 Granados and Schuchmann 2020).
795 The weak positive association with lunar illumination is consistent with the
796 well-documented lunar affinity of *N. albicollis* and other nightjars, which use
797 moonlight to enhance visual foraging (Mills 1986, Jetz et al. 2003, Pérez-Granados
798 et al. 2022). The modest statistical support for this effect in our occupancy models
799 may reflect the temporal resolution of our analysis: we modeled nightly presence
800 (whether the species vocalized at any point during the night), whereas lunar effects

may operate more strongly on the timing and intensity of vocal activity within a night rather than on the binary probability of vocalizing at least once. Future analyses examining within-night temporal patterns at finer resolution may reveal stronger lunar effects.

Acoustic environment and detection heterogeneity

In addition to environmental effects on vocal activity itself, detection probability was also influenced by the acoustic structure of the recordings. The two soundscape metrics included in the detection sub-model, BIO and ACI, revealed that automated detection performance depends not only on species behavior, but also on broader acoustic conditions.

BIO showed a consistent positive association with detection probability across all detector × threshold combinations, suggesting that recordings with greater overall biological acoustic activity were more likely to contain detectable vocalizations of *N. albicollis*. At first glance, this pattern may appear counterintuitive, since high levels of background acoustic activity could be expected to reduce detectability through masking effects. However, our recordings spanned broad seasonal and diel periods, including many relatively quiet nights outside the peak activity periods of other nocturnal taxa, particularly anurans. Under these conditions, increases in BIO may have been driven partly by the vocal activity of *N. albicollis* itself, whose calls contribute directly to acoustic energy within the frequency band used to calculate the index.

In contrast, ACI showed inconsistent effects across detector × threshold combinations, suggesting that it captures a different dimension of the soundscape. Whereas BIO primarily reflects accumulated acoustic energy, ACI is more sensitive

to rapid temporal fluctuations and overlapping acoustic events, making it more strongly associated with soundscape complexity. The negative association observed under permissive thresholds is consistent with the hypothesis that acoustically complex environments increase masking and classifier confusion, particularly when lower confidence detections are retained. The disappearance of this pattern under MaxPrecision thresholds suggests that stricter filtering removed many ambiguous detections associated with complex recordings.

Naïve vs. hierarchical approaches

The comparison between the GLMM and occupancy framework revealed that the naïve approach consistently underestimated uncertainty and produced attenuated effect sizes relative to models that accounted for imperfect detection. This attenuation was more pronounced under the BirdNET detector, where the higher false-positive rate inflated baseline detection probabilities and compressed the apparent environmental gradients in the GLMM. Under the Custom classifier, the two approaches produced more similar estimates, suggesting that the degree of bias in naïve models depends on the quality of the input detection data. This result reinforces the growing consensus that occupancy models are essential for robust inference from acoustic monitoring data (Chambert et al. 2018, Rhinehart et al. 2026, Martins et al. 2025). However, our findings add an important nuance: the benefit of hierarchical modeling is not uniform, but depends on the upstream detection workflow (Kéry and Royle 2015, 2021). When detection data are of high quality (low false-positive rate, consistent detection probabilities), naïve models may provide adequate approximations. When detection data are noisier, hierarchical models become essential, but their performance also degrades if the noise exceeds what the

model can accommodate, as demonstrated by the poor calibration of MaxF1 occupancy models.

We emphasize that this comparison does not imply that one framework is inherently superior, but that the two approaches address complementary questions (Tredennick et al. 2021). The GLMM asks 'what predicts whether we *detect* the species on a given night?', whereas the occupancy model asks 'what predicts whether the species is *vocally active*, after accounting for imperfect detection?' (Goldstein et al. 2026). The degree to which answers to these two questions diverge is itself informative about the role of unmodelled detection heterogeneity in shaping ecological inference.

Limitations and recommendations

Several limitations should be considered when interpreting our results. First, the single-season occupancy framework assumes closure within each site-night (the species is either present or absent throughout the night), which is reasonable for nocturnal vocal activity of a resident species, but does not model temporal dynamics between nights (MacKenzie et al. 2018). The GLMM partially addresses this through the lagged detection covariate, but the two frameworks estimate fundamentally different quantities (detection probability versus true occupancy), limiting direct numerical comparison. Second, posterior predictive checks were based on 250 posterior draws due to computational constraints; while sufficient for assessing model calibration, finer-grained diagnostics might benefit from larger posterior samples. Third, our study focused on a single, relatively common species; the sensitivity of ecological inference to methodological choices may differ for rare

875 species with lower detection rates, where the distinction between false negatives and
876 true absences becomes even more consequential (Guillera–Arroita 2017).
877 The single-season occupancy framework treats site-nights as independent,
878 which does not account for the strong first-order temporal dependence documented
879 in our GLMM analysis. A dynamic (multi-season) occupancy model (MacKenzie et al.
880 2003) could explicitly estimate colonization and extinction probabilities between
881 nights. However, implementing such models for >1,700 primary periods with the
882 current data structure would be computationally prohibitive. We note that the GLMM
883 analysis, which explicitly models temporal dependence via the lagged detection
884 covariate, complements the occupancy analysis by providing an alternative
885 perspective that accounts for this temporal structure. The consistency of directional
886 effects between the two frameworks suggests that temporal autocorrelation does not
887 qualitatively alter ecological conclusions, although it may influence the magnitude of
888 estimated effects.
889 We did not employ a formal false-positive occupancy model (Royle & Link
890 2006, Chambert et al. 2018, Miller et al. 2011, 2012, Kéry and Royle 2020) because
891 (i) our factorial design allows direct comparison of threshold strategies, revealing that
892 MaxPrecision thresholds effectively eliminate false positives as confirmed by
893 posterior predictive checks, and (ii) the sp0ccupancy framework does not currently
894 implement false-positive extensions. Nevertheless, we note that integrating false-
895 positive models with automated acoustic classifiers represents a promising avenue
896 for future research, but may also be very time consuming or even prohibitive,
897 depending on the scope of the sampling.
898 Based on our findings, we offer four recommendations for PAM studies aiming
899 to produce robust ecological inference. First, we recommend reporting results across

multiple analytical pathways — at minimum, two confidence thresholds and both naïve and hierarchical modeling frameworks — to characterize the sensitivity of conclusions to methodological choices. Second, when hierarchical models are used, we recommend prioritizing precision over recall in the detection workflow, as high-precision data produce substantially better-calibrated models. Third, when feasible, training classifiers on locally annotated data improves detection quality and reduces bias in downstream inference. Fourth, when local training is not feasible and performance of a pre-built model is relatively known for a focal species, applying stringent confidence thresholds to pre-built models can partially compensate for their higher false-positive rates.

In conclusion, our study illustrates that passive acoustic monitoring is not a single method but a workflow of interdependent decisions, each of which shapes the ecological signal that emerges from the data. By explicitly evaluating these decisions within a single empirical framework, we provide evidence that methodological transparency and multi-pathway reporting are essential for the credibility and reproducibility of ecological inference from acoustic data.

REFERENCES

Alho, C. J. R. and Sabino, J. 2012. Seasonal Pantanal flood pulse: implications for biodiversity conservation – a review. Oecologia Australis, Rio de Janeiro, v. 16, n. 4, p. 958–978.

Authors (2026). Methodological choices influence ecological inference in passive acoustic monitoring of a Neotropical nightjar. Journal of Avian Biology, Special Issue: From Chirps to Insights. https://doi.org/10.5281/zenodo.20402320

924 Bates, D., Mächler, M., Bolker, B. and Walker, S. 2015. Fitting Linear Mixed-Effects
925 Models Using lme4. – J. Stat. Soft. <https://doi.org/10.18637/jss.v067.i01>.
926 Boelman, N. T., Asner, G. P., Hart, P. J. and Martin, R. E. 2007. Multi-trophic
927 invasion resistance in Hawaii: bioacoustics, field surveys, and airborne remote
928 sensing. – Ecological Applications 17: 2137–2144.
929 Breznau, N., Rinke, E. M., Wuttke, A., Nguyen, H. H. V., Adem, M., Adriaans, J.,
930 Alvarez-Benjumea, A., Andersen, H. K., Auer, D., Azevedo, F., Bahnsen, O.,
931 Balzer, D., Bauer, G., Bauer, P. C., Baumann, M., Baute, S., Benoit, V.,
932 Bernauer, J., Berning, C., Berthold, A., Bethke, F. S., Biegert, T., Blinzler, K.,
933 Blumenberg, J. N., Bobzien, L., Bohman, A., Bol, T., Bostic, A., Brzozowska,
934 Z., Burgdorf, K., Burger, K., Busch, K. B., Carlos-Castillo, J., Chan, N.,
935 Christmann, P., Connelly, R., Czymara, C. S., Damian, E., Ecker, A.,
936 Edelmann, A., Eger, M. A., Ellerbrock, S., Forke, A., Forster, A., Gaasendam,
937 C., Gavras, K., Gayle, V., Gessler, T., Gnambs, T., Godefroidt, A., Grömping,
938 M., Groß, M., Gruber, S., Gummer, T., Hadjar, A., Heisig, J. P., Hellmeier, S.,
939 Heyne, S., Hirsch, M., Hjerm, M., Hochman, O., Hövermann, A., Hunger, S.,
940 Hunkler, C., Huth, N., Ignácz, Z. S., Jacobs, L., Jacobsen, J., Jaeger, B.,
941 Jungkunz, S., Jungmann, N., Kauff, M., Kleinert, M., Klinger, J., Kolb, J.-P.,
942 Kołczyńska, M., Kuk, J., Kunißen, K., Kurti Sinatra, D., Langenkamp, A.,
943 Lersch, P. M., Löbel, L.-M., Lutscher, P., Mader, M., Madia, J. E., Malancu, N.,
944 Maldonado, L., Marahrens, H., Martin, N., Martinez, P., Mayerl, J., Mayorga, O.
945 J., McManus, P., McWagner, K., Meeusen, C., Meierrieks, D., Mellon, J.,
946 Merhout, F., et al. 2022. Observing many researchers using the same data and
947 hypothesis reveals a hidden universe of uncertainty. – Proc. Natl. Acad. Sci.
948 U.S.A. 119: e2203150119.

Chambert, T., Waddle, J. H., Miller, D. A. W., Walls, S. C. and Nichols, J. D. 2018. A new framework for analysing automated acoustic species detection data: occupancy estimation and optimization of recordings post-processing. – Methods Ecol. Evol. 9: 560–570.

Chirino F, Willink B, Elizondo-Calvo J, Chaves-Acuña W, Salas-Solano D, Salazar-Zúñiga JA, Araya-Salas M. 2025 Environmental drivers of calling activity in the critically endangered lemur leaf frog, *Agalychnis lemur* (Hylidae: Phyllomedusinae). Phil. Trans. R. Soc. B, 380, 20240050. (doi:10.1098/rstb.2024.0050)

COPE Council. 2024. COPE position - Authorship and AI - English. https://doi.org/10.24318/cCVRZBms

Cornell Lab of Ornithology. 2026. All About Birds. Cornell Lab of Ornithology, Ithaca, New York. https://www.allaboutbirds.org Accessed on May 22 2026.

Diggle, P. J., Heagerty, P., Liang, K.-Y. & Zeger, S. L. (2002**).** *Analysis of Longitudinal Data*, 2nd ed. Oxford University Press.

Drobniak, S. M., Cendrowska-Pek, M., Gudowska, A., Janas, K., Podkowa, P., Skorb, K., Gronowska, M., Oleś, W., Bikmurzina, F., Boroń, N., Zagalska-Neubauer, M., Nakagawa, S., Lagisz, M., Arct, A. and Rutkowska, J. 2026. A systematic map of generative AI guidelines and reporting in ecology and evolutionary biology: towards the framework of AI disclosure for Improved Transparency (AIdIT). <https://doi.org/10.21203/rs.3.rs-9160721/v1>.

Doser, J. W., Finley, A. O., Kéry, M. and Zipkin, E. F. 2022. spOccupancy: An R package for single-species, multi-species, and integrated spatial occupancy models. – Methods in Ecology and Evolution 13: 1670–1678.

973 Funosas, D., Sebastián-González, E., Morant, J., Gómez, O. H. M., Mendoza, I.,
974 Mohedano-Muñoz, M. A., ... & Pérez-Granados, C. (2026). A global
975 assessment of BirdNET performance: Differences among continents, biomes,
976 and species. Ecological Indicators, 182, 114550.
977 Gelman, A. and Rubin, D. B. 1992. Inference from iterative simulation using multiple
978 sequences. – Statistical Science 7: 457–472.
979 Gelman, A., & Loken, E. (2014). The Statistical Crisis in Science. *American Scientist*,
980 *102*(6), 460–465. http://www.jstor.org/stable/43707868
981 Gould, E., Fraser, H. S., Parker, T. H., Nakagawa, S., Griffith, S. C., Vesk, P. A.,
982 Fidler, F., Hamilton, D. G., Abbey-Lee, R. N., Abbott, J. K., Aguirre, L. A.,
983 Alcaraz, C., Aloni, I., Altschul, D., Arekar, K., Atkins, J. W., Atkinson, J., Baker,
984 C. M., Barrett, M., Bell, K., Bello, S. K., Beltrán, I., Berauer, B. J., Bertram, M.
985 G., Billman, P. D., Blake, C. K., Blake, S., Bliard, L., Bonisoli-Alquati, A.,
986 Bonnet, T., Bordes, C. N. M., Bose, A. P. H., Botterill-James, T., Boyd, M. A.,
987 Boyle, S. A., Bradfer-Lawrence, T., Bradham, J., Brand, J. A., Brengdahl, M. I.,
988 Bulla, M., Bussière, L., Camerlenghi, E., Campbell, S. E., Campos, L. L. F.,
989 Caravaggi, A., Cardoso, P., Carroll, C. J. W., Catanach, T. A., Chen, X., Chik,
990 H. Y. J., Choy, E. S., Christie, A. P., Chuang, A., Chunco, A. J., Clark, B. L.,
991 Contina, A., Covernton, G. A., Cox, M. P., Cressman, K. A., Crotti, M., Crouch,
992 C. D., D'Amelio, P. B., de Sousa, A. A., Döbert, T. F., Dobler, R., Dobson, A.
993 J., Doherty, T. S., Drobniak, S. M., Duffy, A. G., Duncan, A. B., Dunn, R. P.,
994 Dunning, J., Dutta, T., Eberhart-Hertel, L., Elmore, J. A., Elsherif, M. M.,
995 English, H. M., Ensminger, D. C., Ernst, U. R., Ferguson, S. M., Fernandez-
996 Juricic, E., Ferreira-Arruda, T., Fieberg, J., Finch, E. A., Fiorenza, E. A., Fisher,
997 D. N., Fontaine, A., Forstmeier, W., Fourcade, Y., Frank, G. S., Freund, C. A.,

Fuentes-Lillo, E., Gandy, S. L., Gannon, D. G., García-Cervigón, A. I., Garretson, A. C., Ge, X., Geary, W. L., et al. 2025. Same data, different analysts: variation in effect sizes due to analytical decisions in ecology and evolutionary biology. – BMC Biol 23: 35.

Goldstein B. R., Keller A. G., Calhoun K. L., Barker K. J., Montealegre-Mora F., Serota M. W., Van Scoyoc A., Parker-Shames P., Andreozzi C. L. and de Valpine P. 2024. How do ecologists estimate occupancy in practice? – Ecography 2026: e07402.

Grüebler, M. U., Morand, M. and Naef-Daenzer, B. 2008. A predictive model of the density of airborne insects in agricultural environments. – Agriculture, Ecosystems & Environment 123: 75–80.

Guilherme, E. and Lima, J. (2020) Breeding biology and morphometrics of Common Pauraque *Nyctidromus a. albicollis* in south-west Amazonia and the species' breeding season and clutch size in Brazil. *Bulletin of the British Ornithologists' Club* 140: 344–352. https://doi.org/10.25226/bboc.v140i3.2020.a7

Guillera-Arroita, G. 2017. Modelling of species distributions, range dynamics and communities under imperfect detection: advances, challenges and opportunities. – Ecography 40: 281–295.

Hartig, F. 2024. DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models. doi:10.32614/CRAN.package.DHARMa <https://doi.org/10.32614/CRAN.package.DHARMa>, R package version 0.4.7, <https://CRAN.R-project.org/package=DHARMa>.

Heckman, C. H. 1999. Geographical and climatic factors as determinants of the biotic differences between the northern and southern parts of the Pantanal Mato-Grossense. In: II Simpósio sobre Recursos Naturais e Sócio-Econômicos

1023 do Pantanal: Manejo e Conservação. Corumbá: Embrapa Pantanal/UFMS. p.
1024 167–175.
1025 Jetz, W., Steffen, J. and Linsenmair, K. E. 2003. Effects of light and prey availability
1026 on nocturnal, lunar and seasonal activity of tropical nightjars. – Oikos 103:
1027 627–639.
1028 Junk, W. J., Piedade, M. T. F., Lourival, R., Wittmann, F., Kandus, P., Lacerda, L. D.,
1029 Bozelli, R. L., Esteves, F. A., Nunes da Cunha, C., Maltchik, L., Schöngart, J.,
1030 Schaeffer-Novelli, Y. and Agostinho, A. A. 2014. Brazilian wetlands: their
1031 definition, delineation, and classification for research, sustainable
1032 management, and protection. Aquatic Conserv: Mar. Freshw. Ecosyst., 24: 5-
1033 22. https://doi.org/10.1002/aqc.2386
1034 Kahl, S., Wood, C. M., Eibl, M. and Klinck, H. 2021. BirdNET: A deep learning
1035 solution for avian diversity monitoring. – Ecological Informatics 61: 101236.
1036 Kéry, M. and Royle, J. A. 2015. Applied Hierarchical Modeling in Ecology: Analysis
1037 of distribution, abundance and species richness in R and BUGS: Volume 1:
1038 Prelude and Static Models. – Academic Press.
1039 Kéry, M. and Royle, J. A. 2021. Applied hierarchical modeling in ecology: analysis of
1040 distribution, abundance and species richness in R and BUGS. Vol. 2: Dynamic
1041 and advanced models. – Academic Press.
1042 Knight, E. C., K. C. Hannah, G. Foley, C. Scott, R. Mark Brigham, and E. Bayne.
1043 2017. Recommendations for acoustic recognizer performance assessment
1044 with application to five common automated signal recognition programs. Avian
1045 Conservation and Ecology 12(2):14. https://doi.org/10.5751/ACE-01114-
1046 120214

1047 Lai, J., Zou, Y., Zhang, S., Zhang, X. and Mao, L. 2022. glmm.hp: an R package for
1048 computing individual effect of predictors in generalized linear mixed models. –
1049 Journal of Plant Ecology 15: 1302–1307.
1050 Latta, S. C. & Howell, S. N. G. (2020) Common Pauraque (*Nyctidromus albicollis*),
1051 version 1.0. In: Billerman SM, Keeney BK, Rodewald PG & Schulenberg TS
1052 (eds) *Birds of the World*. Cornell Lab of Ornithology, Ithaca, NY.
1053 https://doi.org/10.2173/bow.compau.01
1054 Lüdecke, D. 2018. ggeffects: Tidy Data Frames of Marginal Effects from Regression
1055 Models. – JOSS 3: 772.
1056 Lüdecke, D., Makowski, D., Ben-Shachar, M. S., Patil, I., Wiernik, B. M., Bacher, E.
1057 and Thériault, R. 2022. easystats: Framework for Easy Statistical Modeling,
1058 Visualization, and Reporting. CRAN. doi:10.32614/CRAN.package.easystats
1059 <https://doi.org/10.32614/CRAN.package.easystats>
1060 Marten, K., Quine, D. and Marler, P. 1977. Sound transmission and its significance
1061 for animal vocalization. – Behav Ecol Sociobiol 2: 291–302.
1062 Martins, F. C., Segurado, P., & Marques, J. T. (2025). Acoustic detection and
1063 occupancy models: A systematic review with insights for future monitoring
1064 programs. Ecological Indicators, 179, 114081.
1065 MacKenzie, D. I., Nichols, J. D., Hines, J. E., Knutson, M. G. and Franklin, A. B.
1066 2003. Estimating site occupancy, colonization, and local extinction probabilities
1067 when a species is detected imperfectly. – Ecology 84: 2200–2207.
1068 MacKenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L. L. and Hines,
1069 J. E. 2018. Occupancy Estimation and Modeling. 2nd edition. – Elsevier.
1070 McKechnie, A. E. & Ashdown, R. A. B. (2008). Environmental correlates of Freckled
1071 Nightjar (*Caprimulgus tristigma*) activity in a seasonal, subtropical habitat.

Journal of Ornithology, 149, 615–619. https://doi.org/10.1007/s10336-008-0309-7

McKechnie, A. E., Gerson, A. R., McWhorter, T. J., Smith, E. K., Talbot, W. A. and Wolf, B. O. 2023. Avian heterothermy: a review of patterns and processes. – Integrative and Comparative Biology 63: 1028–1043.

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

Miller, D. A. W., Weir, L. A., McClintock, B. T., Grant, E. H. C., Bailey, L. L. and Simons, T. R. 2012. Experimental investigation of false positive errors in auditory species occurrence surveys. – Ecological Applications 22: 1665–1674.

Moussy, C., Burfield, I.J., Stephenson, P.J., Newton, A.F.E., Butchart, S.H.M., Sutherland, W.J., Gregory, R.D., McRae, L., Bubb, P., Roesler, I., Ursino, C., Wu, Y., Retief, E.F., Udin, J.S., Urazaliyev, R., Sánchez-Clavijo, L.M., Lartey, E. & Donald, P.F. (2022). A quantitative global review of species population monitoring. *Conservation Biology*, 36, e13721. ttps://doi.org/10.1111/cobi.13721

Nakagawa, S. and Schielzeth, H. 2013. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. – Methods Ecol Evol 4: 133–142.

Nguyen Chi T, Vu TT, Nguyen HT, Clink DJ. 2025 Circadian rhythms and the use of transfer learning for critically endangered crested argus *Rheinardia ocellata* in

1096 the central highlands of Vietnam: the implications for conservation. Phil. Trans.
1097 R. Soc. B, 380, 20240056. (doi:10.1098/rstb.2024.0056)
1098 Nichols, J. D., Bailey, L. L., O'Connell, A. F., Talancy, N. W., Campbell Grant, E. H.,
1099 Gilbert, A. T., Annand, E. M., Husband, T. P. and Hines, J. E. 2008. Multi-scale
1100 occupancy estimation and modelling using multiple detection methods. –
1101 Journal of Applied Ecology 45: 1321–1329.
1102 O'Connor, R. S., Smit, B., Talbot, W. A., Gerson, A. R., Brigham, R. M., Wolf, B. O.
1103 and McKechnie, A. E. 2018. Avian thermoregulation in the heat: is evaporative
1104 cooling more economical in nocturnal birds? – Journal of Experimental Biology
1105 221: jeb181420.
1106 Pérez-Granados, C.; Schuchmann, K.-L. Nocturnal vocal behavior of the diurnal
1107 Undulated Tinamou *Crypturellus undulatus* is associated with temperature and
1108 moon phase**.** *Ibis*, 2020a, 163, 684–694.
1109 Pérez-Granados, C.; Schuchmann, K.-L. Illuminating the nightlife of two Neotropical
1110 nightjars: vocal behavior over a year and monitoring recommendations.
1111 Ethology Ecology & Evolution, 2020b, 32, 466–480.
1112 Pérez-Granados, C., Schuchmann, K.-L. and Marques, M. I. 2022. Addicted to the
1113 moon: vocal output and diel pattern of vocal activity in two Neotropical nightjars
1114 is related to moon phase. – Ethology Ecology & Evolution 34: 66–81.
1115 Pérez-Granados, C., Funosas, D., Morant, J., Marín Gómez, O.H., Mendoza, I.,
1116 Mohedano-Munoz, M.A., Santamaría, E., Bastianelli, G., Márquez-Rodríguez,
1117 A., Budka, M., Bota, G., De la Peña-Rubio, J.M., García De La Morena, E.,
1118 Santa-Cruz, M., De la Nava, P., Fernández-Tizón, M., Sánchez-Mateos, H.,
1119 Barrero, A., Traba, J., Osiejuk, T.S., Hart, P.J., Navine, A.K., Montoya Muñoz,
1120 A.F., De Araújo, C.B., Rosa, G.L.M., Torres, I.M.D., Catalano, A.L.C., Simões,

C.R., Llusia, D., Morales, M.B., Acebes, P., Medina, J.A., Brown, N., Astaras, C., Karmiris, I., Navarrete, E., Cauchoix, M., Barbaro, L., Arend, D., Müeller, S., González-García, F., González-Romero, A., Mammides, C., Pontikis, M., Jacuzzi, G., Olden, J.D., Bombaci, S.P., Marcacci, G., Jacot, A., Zurano, J.P., Gangenova, E., Varela, D., Di Sallo, F., Zurita, G.A., Atemasov, A., Tremblay, J.A., Hutschenreiter, A., Monroy-Ojeda, A., Díaz-Vallejo, M., Chaparro-Herrera, S., Briers, R.A., Sousa-Lima, R., Pinheiro, T., Da Silva, W.C., Calvente, A., Dal Molin, A., Antonelli, A., Gogoleva, S., Palko, I., Trong, H.V., Duarte, M.H.L., Saturnino, N.D.S., Silva, S.R., Rainho, A., Lopes, P., Schuchmann, K.-L., Marques, M.I., De Oliverira Tissiani, A.S., Littlewood, N.A., Tuanmu, M.-N., Cheng, Y.-R., Chao, H., Kepfer-Rojas, S., Aguilera, A.L., Brotons, L., Feldman, M.J., Imbeau, L., Panwar, P., Weed, A.S., Dehwal, A., Attisano, A., Theuerkauf, J., Oliveira-Júnior, D.D., Lima-Santos, C.S., Salustio-Gomes, C., Paz, R.V., Pichorim, M., Goodale, E. and Sebastián-González, E. (2026), Optimization of passive acoustic bird surveys: a global assessment of BirdNET settings. Ibis, 168: 785-798. https://doi.org/10.1111/ibi.70013

Pieretti, N., Farina, A. and Morri, D. 2011. A new methodology to infer the singing activity of an avian community: the Acoustic Complexity Index (ACI). – Ecological Indicators 11: 868–873.

Plummer, M., Best, N., Cowles, K. and Vines, K. 2006. CODA: Convergence Diagnosis and Output Analysis for MCMC. – R News 6: 7–11.

Polson, N. G., Scott, J. G. and Windle, J. 2013. Bayesian inference for logistic models using Pólya–Gamma latent variables. – Journal of the American Statistical Association 108: 1339–1349.

R Core Team (2025). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Rhinehart, T. A., Czarnecki, C., Lyon, R. P., Chronister, L. M., Lapp, S., Larkin, J. L., ... & Kitzes, J. (2026). Addressing widespread detection heterogeneity in avian occupancy modeling using passive acoustic surveys. Ornithological Applications, duag006.

Richards, D. G. and Wiley, R. H. 1980. Reverberations and Amplitude Fluctuations in the Propagation of Sound in a Forest: Implications for Animal Communication. – The American Naturalist 115: 381–399.

Riede K, Balakrishnan R. 2025 Acoustic monitoring for tropical insect conservation. Phil. Trans. R. Soc. B, 380, 20240046. (doi:10.1098/rstb.2024.0046)

Rolim F. G. and Theodorovicz A. 2012. Geoparque Bodoquena-Pantanal (MS): proposta. *In*: SCHOBBENHAUS, Carlos; SILVA, Cassio Roberto da (Org.). Geoparques do Brasil: propostas. Rio de Janeiro: CPRM. Cap. 8

Ross, S.-J., O'Connell, D. P., Deichmann, J. L., Desjonquères, C., Gasc, A., Phillips, J. N., Sethi, S. S., Wood, C. M., & Burivalova, Z. (2023). Passive acoustic monitoring provides a fresh perspective on fundamental ecological questions. Functional Ecology, 37, 959–975. https://doi.org/10.1111/1365-2435.14275

Royle, J. A. and Link, W. A. 2006. Generalized Site Occupancy Models Allowing For False Positive And False Negative Errors. – Ecology 87: 835–841.

Salustio-Gomes, C., Paz, R. V., Lima-Santos, C. S., Oliveira-Júnior, D. D., Cardoso-Oliveira, V. D., Venticinque, E. M., & Pichorim, M. (2026). Biotic and abiotic influences on vocal activity of tropical screech-owl (*Megascops choliba*) and

spectacled owl (*Pulsatrix perspicillata*). *Ornithology Research*, *34*(23), 1–13. https://doi.org/10.1007/s43388-026-00289-w

Schaaf, A. A., Boullhesen, M., Gandoy, F. A., & Rivera, L. O. (2023). Influence of environmental factors on the vocal activity of birds in a subtropical forest. *Ibis*, *165*(3), 912–925. https://doi.org/10.1111/ibi.13191

Schiavo, G., Portaccio, A., & Testolin, A. (2025). Fine-Tuning BirdNET for the Automatic Ecoacoustic Monitoring of Bird Species in the Italian Alpine Forests. Information, 16(8), 628. https://doi.org/10.3390/info16080628

Śmielak, M. K. 2023. Biologically meaningful moonlight measures and their application in ecological research. – Behav Ecol Sociobiol 77: 21.

Steegen, S., Tuerlinckx, F., Gelman, A. and Vanpaemel, W. 2016. Increasing transparency through a multiverse analysis. – Perspectives on Psychological Science 11: 702–712.

Studer, A. & Crozariol, M. A. (2025) New breeding information on Brazilian birds. 3: Nyctibiidae, Caprimulgidae, Apodidae and Trochilidae. *Bulletin of the British Ornithologists' Club* 145: 193–272. https://doi.org/10.25226/bboc.v145i3.2025.a2

Sugai, L. S. M., Silva, T. S. F., Ribeiro, J. W. and Llusia, D. 2019. Terrestrial Passive Acoustic Monitoring: Review and Perspectives. – BioScience 69: 15–25.

Sugai, L. S. M., Balantic, C., Clink, D. J., Ramesh, V., Kahl, S., Klinck, H., & Wood, C. M. (2026). Acoustic indices are not useful for biodiversity research. Methods in Ecology and Evolution, 00, 1–13. https://doi.org/10.1111/2041-210x.70285

Taylor, L. R. 1963. Analysis of the Effect of Temperature on Insects in Flight. – The Journal of Animal Ecology 32: 99.

1193 Teixeira D, Roe P, van Rensburg BJ, Linke S, McDonald PG, Tucker D, Fuller S.
1194 2024 Effective ecological monitoring using passive acoustic sensors:
1195 recommendations for conservation practitioners. Conservat. Sci. and Prac., 6,
1196 e13132. (doi:10.1111/csp2.13132)
1197 Tredennick, A. T., Hooker, G., Ellner, S. P. and Adler, P. B. 2021. A practical guide
1198 to selecting models for exploration, inference, and prediction in ecology. –
1199 Ecology 102: e03336.
1200 Tsioutsiourigas, D., Rodiou, A., Nalmpantis, E., Kyriakopoulou, A., Petrusková, T.
1201 and Astaras, C. 2025. By the Moonlight Shadow: Examining the Acoustic
1202 Ecology of the European Nightjar (*Caprimulgus europaeus*) in Northern
1203 Greece Using Passive Acoustic Monitoring.
1204 <https://doi.org/10.20944/preprints202503.0673.v1>.
1205 Valdez, J.W., Callaghan, C.T., Junker, J., Purvis, A., Hill, S.L.L. & Pereira, H.M.
1206 2023. The undetectability of global biodiversity trends using local species
1207 richness. *Ecography*, 2023, e06604. https://doi.org/10.1111/ecog.06604
1208 Villanueva-Rivera, L. J., & Pijanowski, B. C. (2018). soundecology: Soundscape
1209 Ecology (R package version 1.3.3).
1210 Watanabe, S. 2010. Asymptotic equivalence of Bayes cross validation and widely
1211 applicable information criterion in singular learning theory. – Journal of
1212 Machine Learning Research 11: 3571–3594.
1213 Wood, C. M. and Peery, M. Z. 2022. What does ‘occupancy’ mean in passive
1214 acoustic surveys? – Ibis 164: 1295–1300.
1215 Wood, C. M., Socolar, J., Kahl, S., Peery, M. Z., Chaon, P., Kelly, K., Koch, R. A.,
1216 Sawyer, S. C., & Klinck, H. (2024). A scalable and transferable approach to
1217 combining emerging conservation technologies to identify biodiversity change

after large disturbances. Journal of Applied Ecology, 61, 797–808. https://doi.org/10.1111/1365-2664.14579

Wood, C.M., Kahl, S. 2024. Guidelines for appropriate use of BirdNET scores and other detector outputs. J Ornithol 165, 777–782. https://doi.org/10.1007/s10336-024-02144-5

Zeger, S. L. & Qaqish, B. (1988). Markov regression models for time series: A quasi-likelihood approach. *Biometrics*, 44, 1019–1031.

Zuur, A. F., Ieno, E. N., Walker, N., Saveliev, A. A. & Smith, G. M. (2009). *Mixed Effects Models and Extensions in Ecology with R*. Springer.

AI USE STATEMENT

The authors used Claude (Opus 4.6, Anthropic, 2025; claude.ai) and ChatGPT (OpenAI, 2026; chat.openai.com) as AI assistants during the preparation of this manuscript. AI assistance was employed across the following stages of the research workflow, with human oversight maintained at each step:

Statistical analysis design. The AI tool was used to evaluate candidate model formulations for the binary detection response, identify conceptual and formal errors in the initial modelling approach (including an incorrect binomial specification), and advise on the inclusion of temporal autocorrelation structure. All modelling decisions — including the final choice of a Bernoulli GLMM with a random slope for the lagged detection covariate — were made by the authors following critical assessment of the AI's suggestions against the study design and ecological context.

Code development. AI tools were used to assist in creating R code used to generate the analytical workflow diagram and figures describing the study variables as well as for assisting in developing R workflows for processing BirdNET outputs and aggregating detections by sampling units (audio recordings and survey nights). R code for model fitting (`lme4`), residual diagnostics (`DHARMa`), autocorrelation function plots, hierarchical variance partitioning (`glmm.hp`), and marginal effects visualization (`ggeffects`) was also drafted with AI assistance. All code was reviewed, tested, and validated by the authors on the study dataset.

Interpretation of diagnostic outputs. The AI tool assisted in interpreting DHARMa residual plots, Durbin–Watson test results, ACF patterns, and AIC/BIC discordance. Ecological interpretation of results remained the sole responsibility of the authors.

Manuscript writing. The Statistical Analysis subsection of the Methods, draft Results subsections for both the regression-based and occupancy modeling approaches, and a draft Discussion were prepared with AI assistance, based on analytical decisions, ecological interpretations, and structural choices made by the authors. The Abstract was refined with AI assistance. All drafted text was critically reviewed, revised, and approved by all authors prior to submission.

Literature support. The AI tool was used to identify candidate references, verify bibliographic details, and format citations according to journal style.

No AI tool was used to generate or modify primary data, produce figures included in the manuscript, conduct the acoustic recordings or species identifications, or make final decisions on ecological interpretation. In accordance with COPE guidelines (COPE 2024), Wiley (https://www.wiley.com/en-us/publish/article/ai-guidelines/), and the AIdIT framework (Drobniak et al. 2026), no AI tool is listed as an author. The authors accept full responsibility for the integrity and accuracy of all content in this manuscript.

Methodological choices influence ecological inference in passive acoustic monitoring of a Neotropical nightjar

Supplementary Information

Table of Contents:

MATERIAL AND METHODS

Study area and Sampling design

Table S1. Monitoring sites, abbreviated codes, habitat descriptions, and geographic coordinates.

Site Code	Site name	Habitat/Ecoregion	Lat	Long	Elevation (m)
ARA	Caiman Ranch - Site 1	Miranda Pantanal	-19.961	-56.234	90 - 130
BAI	Caiman Ranch - Site 2	Miranda Pantanal	-19.778	-56.229	95 - 105
BEP	Pantanal Study Base - UFMS	Miranda Pantanal	-19.574	-57.017	85 - 95
MIM	Estância Mimosa	Cerrado	-20.970	-56.514	250 - 320
REF	Refúgio da Ilha	Cerrado-Pantanal Ecotone	-20.226	-56.578	140 - 180
UEM	UEMS Aquidauana	Aquidauana Pantanal	-20.440	-55.666	180 - 240
XAR	Xaraés Farm Lodge	Abobral Pantanal	-19.495	-56.956	110 - 130

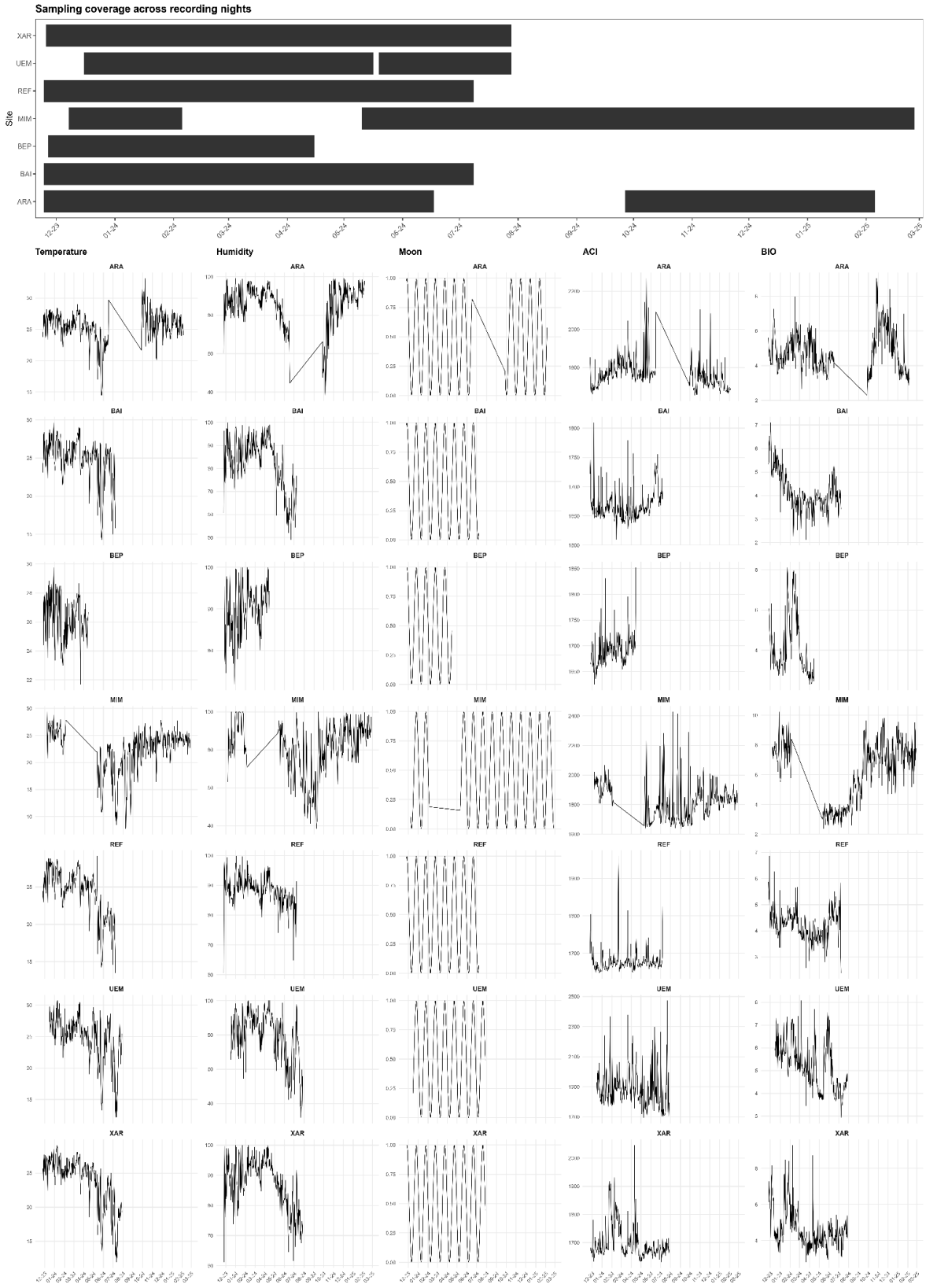


Figure S1. Temporal sampling coverage and nightly variation in environmental and acoustic variables across study sites. The upper panel shows sampling coverage for each recording night and site throughout the study period. Lower panels present nightly trends in temperature, relative humidity, moon illumination, Acoustic Complexity Index (ACI), and Bioacoustic Index (BIO) for each site.

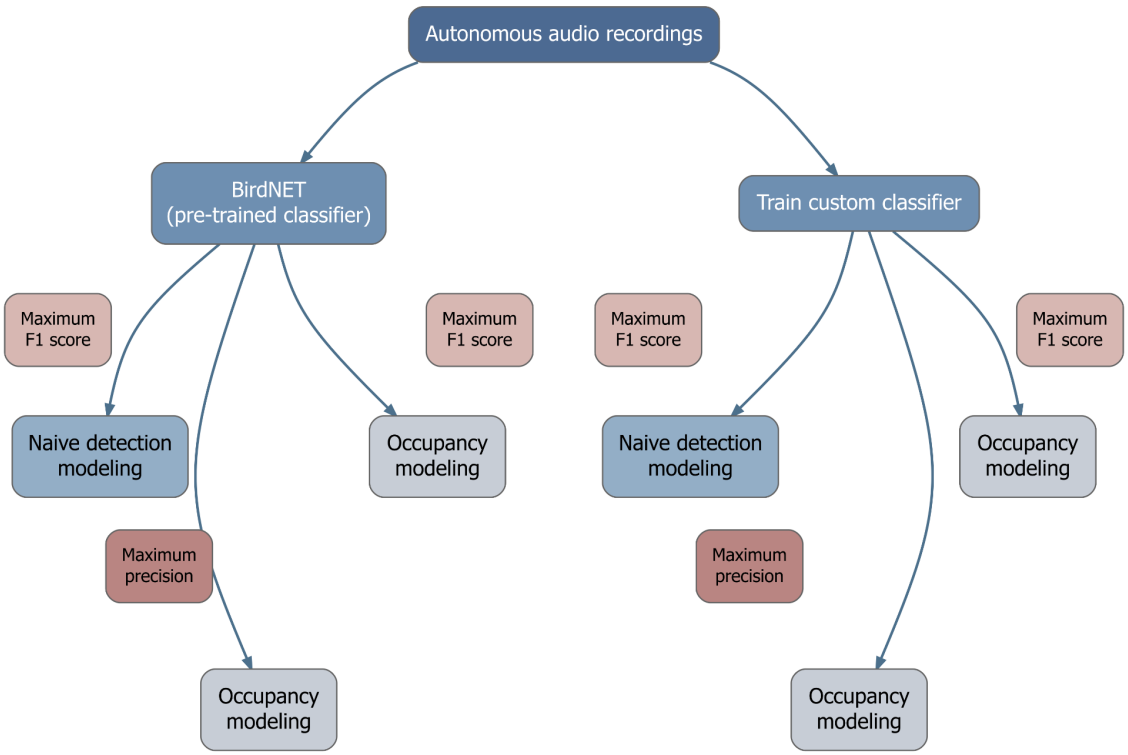


Figure S2**.** Workflow illustrating the analytical framework used to compare automated species detection approaches. Autonomous audio recordings were processed using either the pre-trained BirdNET classifier or a custom-trained classifier. Detection outputs were then filtered using two threshold selection criteria: maximum F1 score and maximum precision. Resulting detections were subsequently used in naïve detection analyses and occupancy modeling to evaluate species occurrence patterns and model performance.

Training of the custom classifier

To construct a custom classifier trained using region-specific acoustic data, we compiled training data from the same acoustic dataset used in this study to ensure compatibility with local soundscape conditions. Following BirdNET requirements, we constructed a labeled dataset of 3 s audio clips, including 377 clips containing vocalizations of the target species and 342 clips representing non-target sounds. The training dataset was assembled through random inspection of recordings from all study sites and designed to balance target and non-target samples. Target-class clips captured a range of acoustic conditions, including variation in signal intensity (likely reflecting distance to the recording unit) and overlapping calls from the same species, other species, and background noise. Non-target samples included a diversity of sounds present in the soundscape, including vocalizations of other species (particularly anurans and insects) as well as geophonic noise, such as rain and wind. This approach aimed to improve classification performance under local acoustic conditions and reduce biases associated with applying globally trained models to novel

soundscapes. Importantly, the audio files used to construct the training dataset were not included in the set of recordings used for subsequent analyses. Training was conducted using 1-4 kHz band filtering and the autotune and mixup option, which iterates across different hyperparameters to select the best performing ones. Final settings used were: i) hidden units 2048, ii) dropout 0.33, iii) batch size 64, iv) learning rate 0.001, and v) upsampline ratio 0.25, focal loss alpha 0.1.

Data Analysis

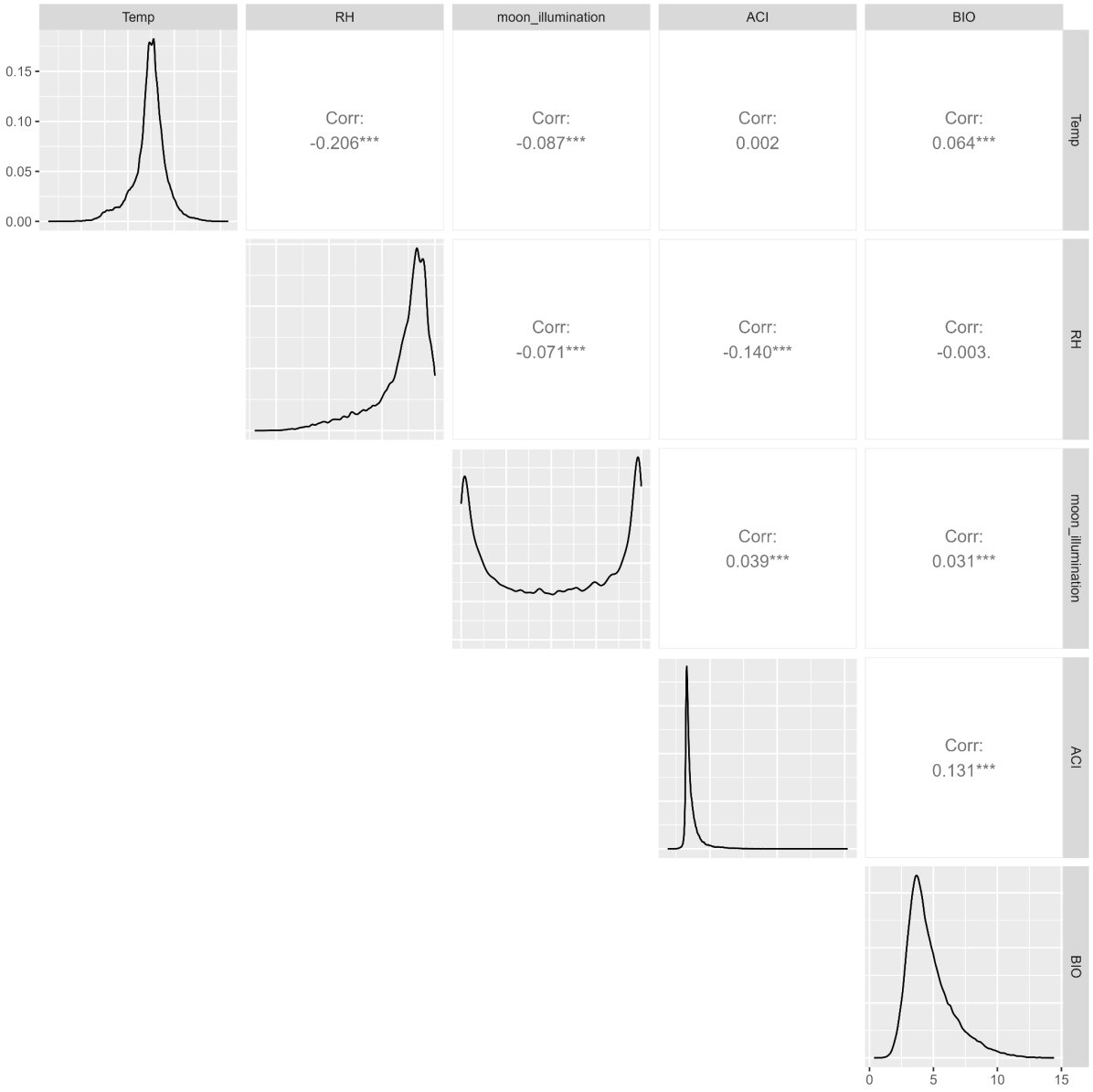


Figure S3. Pairwise correlation structure and variable distributions among environmental and acoustic variables. Pearson correlation coefficients are shown in the upper panels, while the diagonal panels display the distribution density of each variable.

Table S2. Empirical first-order transition probabilities.

Detector	transition	n	prob
Custom classifier	0 → 0	1117	0.859
Custom classifier	0 → 1	183	0.141
Custom classifier	1 → 0	184	0.416
Custom classifier	1 → 1	258	0.584
BirdNET pre-built	0 → 0	438	0.706
BirdNET pre-built	0 → 1	182	0.294
BirdNET pre-built	1 → 0	182	0.162
BirdNET pre-built	1 → 1	940	0.838

RESULTS

Table S3. Performance metrics of the pre-trained BirdNET model and the custom-trained classifier at the confidence thresholds selected for ecological analyses.

Classifier	Threshold strategy	Threshold	Precision	Recall	F1
BirdNET	Max F1	0.10	0.80	0.76	0.78
BirdNET	Max Precision	0.95	0.92	0.14	0.24
Custom	Max F1	0.40	0.82	0.86	0.84
Custom	Max Precision	0.95	0.98	0.16	0.28

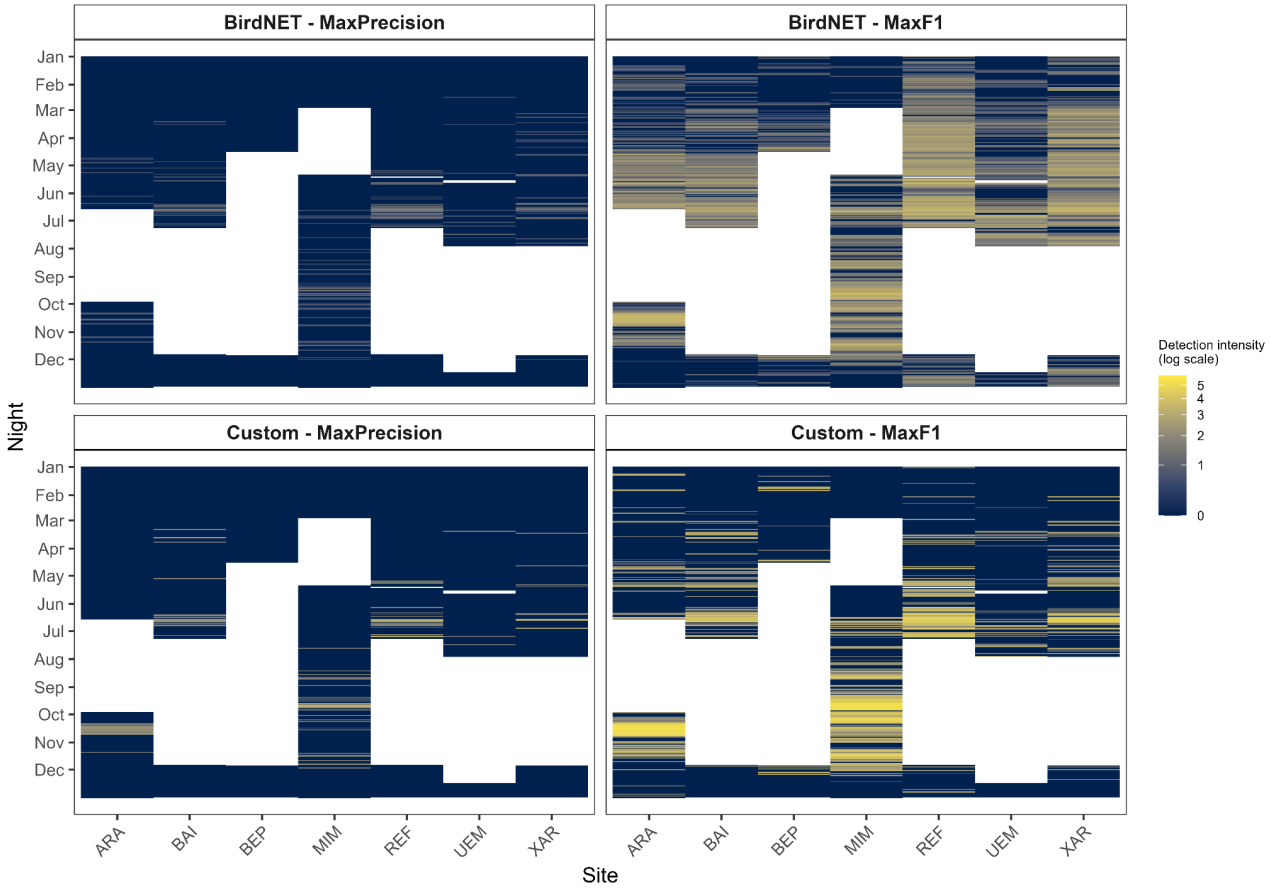


Figure S4. Nightly detection intensity of Nyctidromus albicollis across seven monitoring sites under four automated detection workflows combining two classifiers (BirdNET pre-built and Custom-trained) and two confidence-threshold strategies (MaxF1 and MaxPrecision). Each cell represents a site-night combination, with color intensity indicating the number of detections per night on a log-transformed scale.

Naïve approach (GLMM-based regression)

Analyses of temporal autocorrelation, random-effects structure, and model diagnostics

Autocorrelation diagnostics of DHARMa residuals from the random-intercept-only baseline model revealed heterogeneous temporal dependence among sites under both classifiers. In the Custom classifier, MIM showed persistent positive autocorrelation across all 14 lags examined, indicating sustained periods of consecutive detection or absence not explained by environmental predictors alone. BEP exhibited an oscillatory pattern with lag values crossing significance bands, and BAI showed borderline early-lag autocorrelation. The remaining four sites (ARA, REF, UEM, XAR) showed no residual autocorrelation. In the PreBuilt classifier, residual

autocorrelation was more widespread and of greater magnitude, reflecting the higher and more temporally clustered detection rates of the pre-built model. In both datasets, including lag_presence as a fixed effect with a random slope by site substantially improved model fit relative to the random-intercept-only model. For the Custom classifier: ΔAIC = −10.4, LRT $\chi^2_{(2)}$ = 14.45, p < 0.001. For the PreBuilt classifier: ΔAIC = 0.8, LRT $\chi^2_{(2)}$ = 3.18, p = 0.20. Although ΔBIC was marginally positive in the Custom classifier (+0.5), it increased substantially in the PreBuilt model (+11.7), we retained the random-slope structure for the reasons described in Methods.

Residual diagnostics indicated adequate fit for both final models. Under the Custom classifier, randomized quantile residuals showed no significant overdispersion (dispersion test: *P* = 0.99) and no influential outliers (outlier test: *P* = 0.59). The global KS test reached nominal significance (*P* = 0.026), but visual inspection of the QQ plot revealed no meaningful departure from uniformity; we attribute this to the high sensitivity of the KS test at n = 1,742 rather than to model misspecification. Under the PreBuilt classifier,residual diagnostics indicated no evidence of departure from model assumptions, with non-significant KS (*P* = 0.386), dispersion (*P* = 0.224), and outlier tests (*P* = 0.057).

Site-specific Durbin–Watson tests on the DHARMa residuals of both final models confirmed that temporal autocorrelation had been effectively resolved at all seven sites by the inclusion of the random slope. For the Custom classifier, all DW statistics were close to 2.0 (range: 1.85–2.15, all *P* > 0.15). For the PreBuilt classifier, DW statistics were likewise centered near 2.0 across most sites (range: 1.86–2.27), with non-significant tests in six of the seven sites (all *P* > 0.24). A marginally significant negative autocorrelation remained at site REF (DW = 2.268, *P* = 0.034), although its magnitude was small and no consistent residual temporal structure was evident across the remaining sites.

Table S4. Fixed-effect estimates (Odds Ratios) with 95% Wald CI for both final models. Predictors were z-standardized; OR represents the multiplicative change in detection odds per 1-SD increase.

Detector	Parameter	OR	SE	CI_low	CI_high	z	p
Custom classifier							
	(Intercept)	0.161	0.025	0.119	0.218	-11.814	0.000
	Temp_s	0.947	0.063	0.831	1.080	-0.811	0.417
	RH_s	0.659	0.045	0.577	0.753	-6.136	0.000
	moon_s	1.053	0.067	0.930	1.192	0.815	0.415
	lag_presence	5.583	1.549	3.241	9.617	6.197	0.000
BirdNET pre-built							
	(Intercept)	0.667	0.221	0.349	1.276	-1.224	0.221
	Temp_s	0.872	0.061	0.761	0.999	-1.971	0.049
	RH_s	0.550	0.043	0.471	0.642	-7.557	0.000
	moon_s	1.159	0.073	1.024	1.311	2.333	0.020
	lag_presence	7.399	1.297	5.248	10.434	11.415	0.000

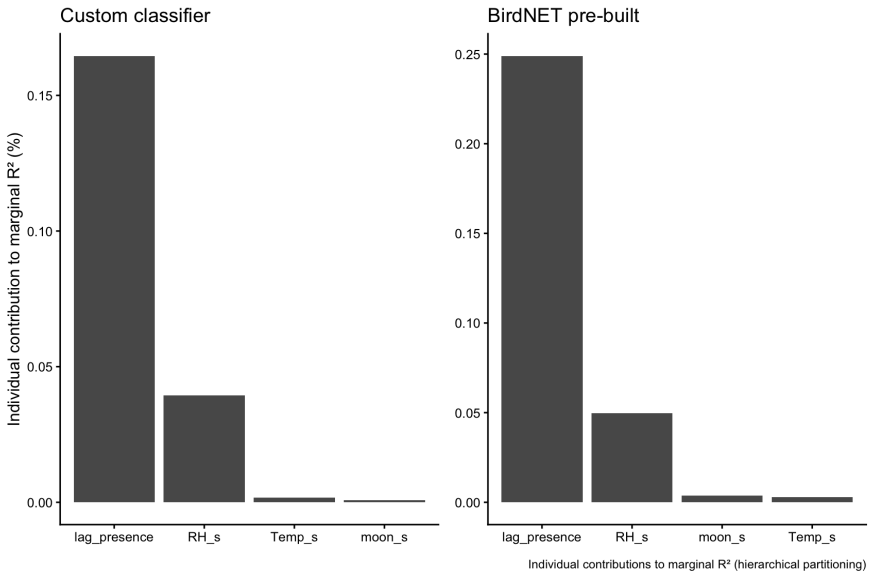


Figure S5**.** Hierarchical partitioning of marginal R² (all fixed effects). Uses the package's native plot method to avoid version-specific parsing issues.

Occupancy models

Analyses of of MCMC convergence, posterior predictive checks, and WAIC-based model comparison

All four occupancy models achieved adequate convergence. Across all occupancy (β) and detection (α) parameters, Gelman–Rubin statistics were below 1.05 and effective sample sizes exceeded 400. Traceplots showed well-mixed chains with no evidence of trend or drift (Fig. S6).

Posterior predictive checks revealed marked differences in model calibration across detector × threshold combinations (Tab. S5, Fig. S6). All four models showed poor fit at the site level (Bayesian *P*-value = 0 for all models), indicating that none of the models adequately reproduced the distribution of the Freeman-Tukey statistic across site-nights. At the survey level, models fitted to MaxF1 datasets showed extreme lack of fit (Bayesian *P*-value = 0.000 for both detectors), consistent with systematic false-positive contamination violating the model's assumptions. Models fitted to MaxPrecision datasets performed substantially better at the survey level, with BirdNET — MaxPrecision showing acceptable calibration (Bayesian *P*-value = 0.208). The Custom — MaxPrecision model yielded a survey-level p-value of 1.000, indicating that replicated Freeman-Tukey statistics consistently exceeded observed values — a pattern that may reflect the very low detection rates under this detector–threshold combination, which limit the model's ability to generate sufficient variation in replicated datasets. Together with the WAIC comparisons (Table S6), these results confirm that MaxPrecision thresholds produce substantially better-calibrated models than MaxF1 thresholds, although no single model achieved ideal calibration across both grouping levels. The site-level *P* = 0 for all models may reflect non-modelled heterogeneity between site-nights, which is expected given that the single-season model does not include any temporal structure.

WAIC comparisons reinforced this pattern (Table S6). The BirdNET — MaxPrecision model had the lowest WAIC, followed by Custom — MaxPrecision . Both MaxF1 models performed substantially worse: BirdNET — MaxF1 and Custom — MaxF1). The large pD values for the MaxF1 models, particularly Custom — MaxF1, indicate that the model requires an excessive number of effective parameters to accommodate the noise in the detection data, further evidence of poor model–data compatibility.

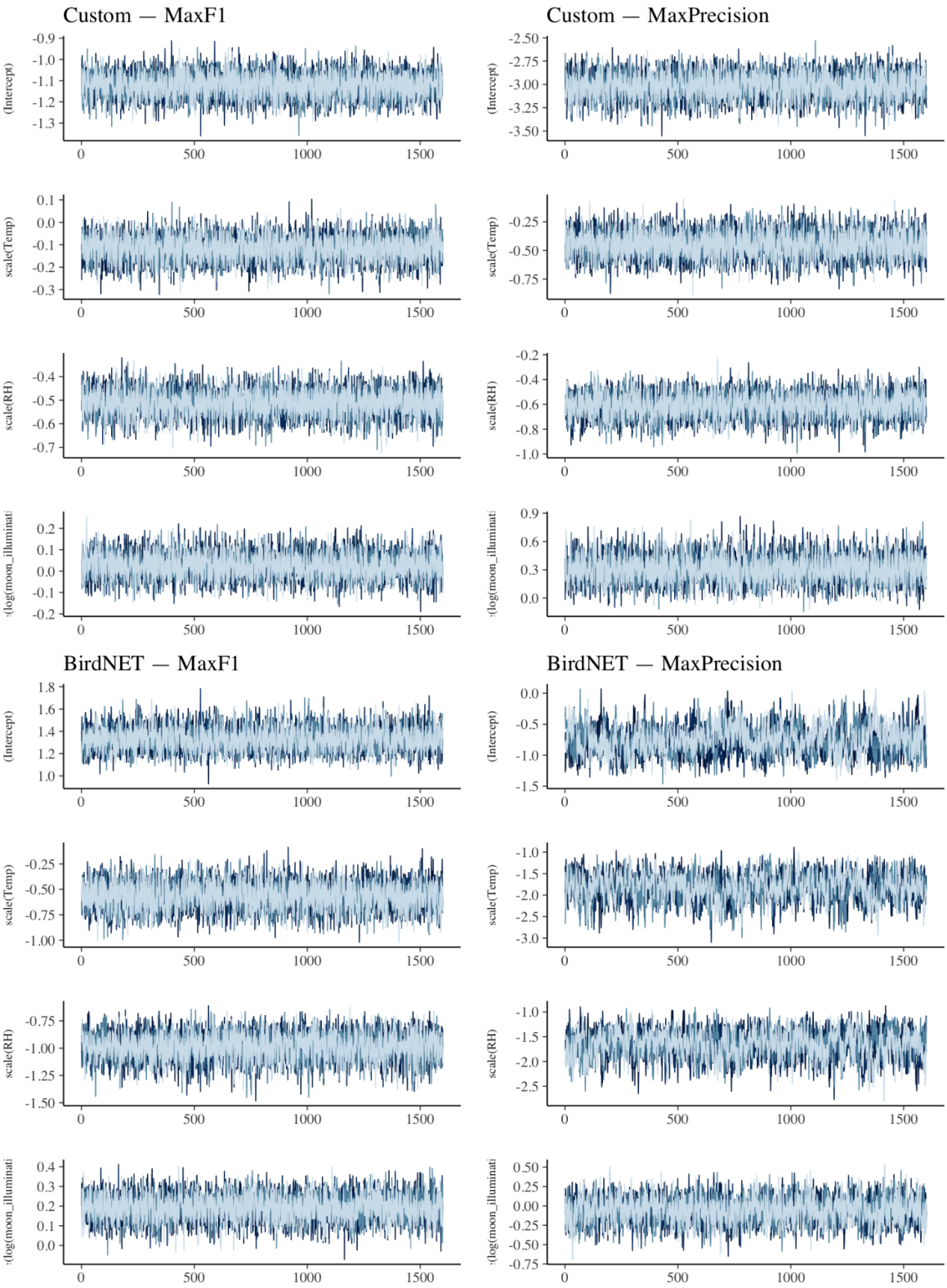


Figure S6. Traceplots for occupancy (β) parameters. Each colour = one MCMC chain. Chains should overlap without trend or drift.

Table S5. Bayesian *P*-values from Freeman-Tukey PPC. Values between 0.1 and 0.9 indicate adequate fit. Models with MaxF1 are worse than the ones with MaxPrecision.

Detector	Sites Bayesian p-values	Survey Bayesian p-values
Custom – MaxF1	0	0
Custom – MaxPrecision	0	1
BirdNET – MaxF1	0	0
BirdNET – MaxPrecision	0	0.208

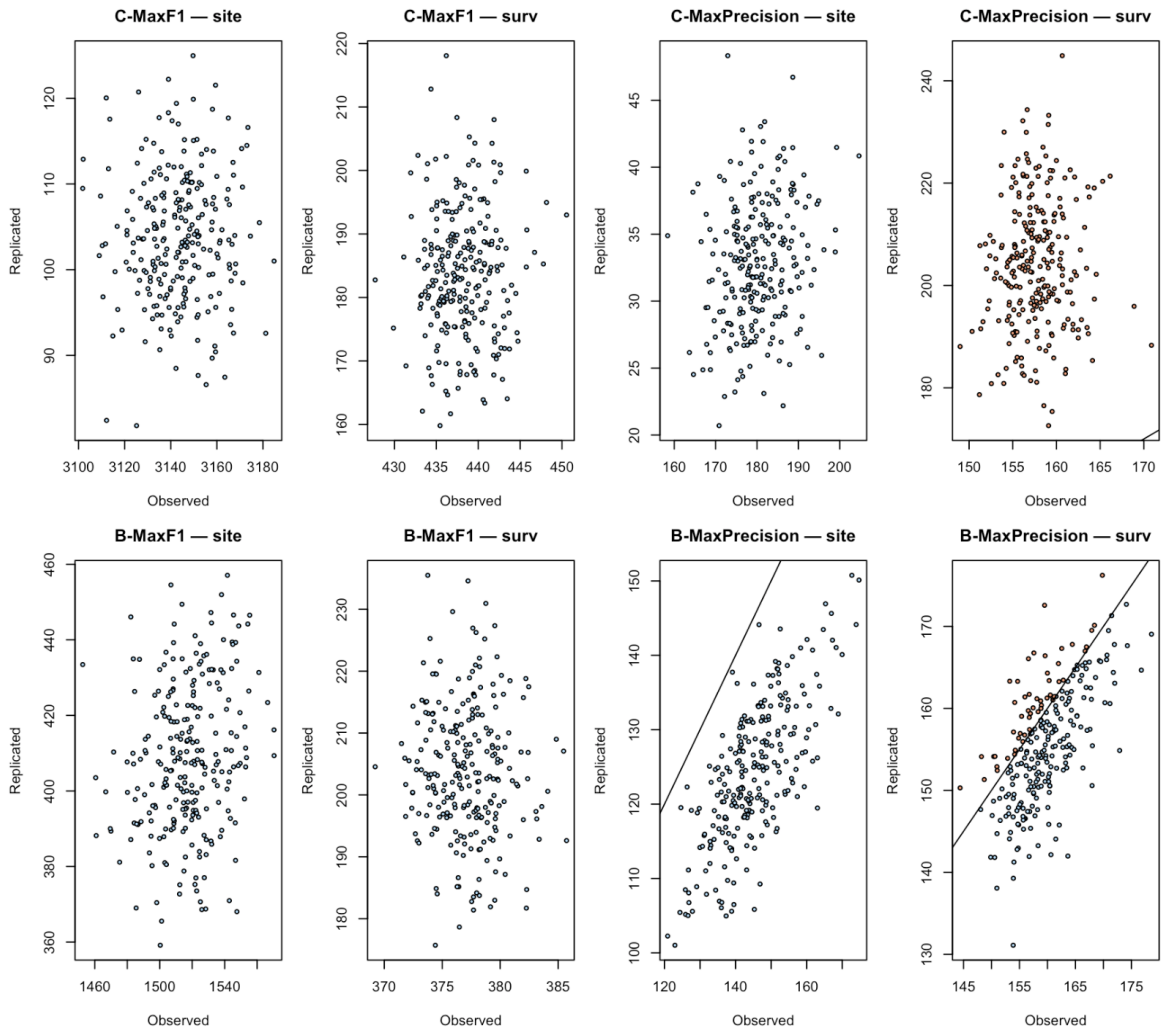


Figure S7. Posterior predictive checks (Freeman-Tukey statistic, grouped by site). Blue = replicated fit < observed fit; red = replicated > observed. Bayesian p-value near 0.5 indicates good calibration. C = Custom detector. B = BirdNet pre-built detector. The 1:1 line does not appear in all plots because the axes are at different scales. This means a lack of fit: the model cannot reproduce the observed patterns. When points are equally spread throughout the 1:1 line, it means the model is well calibrated. Models with MaxPrecision threshold seem to have a better fit than the ones with MaxF1.

Table S6. WAIC for all four occupancy models. Lower WAIC indicates better out-of-sample predictive accuracy. pD = effective number of parameters.

Detector	WAIC	pD	delta_WAIC
BirdNET — MaxPrecision	2984.7	16.0	0.0
Custom — MaxPrecision	6792.4	84.8	3807.7
BirdNET — MaxF1	71892.1	56.5	68907.4
Custom — MaxF1	92895.7	342.5	89911.0

Table S7. Posterior means (SD) and 95% credible intervals for occupancy (β) parameters across all four models. Credible intervals excluding zero indicate strong evidence of an effect.

Detector	Parameter	Mean	SD	Lower	Upper
Custom — MaxF1					
	Temperature	-0.117	0.056	-0.226	-0.009
	Relative Humidity	-0.512	0.055	-0.622	-0.406
	Moon Illumination (log)	0.026	0.058	-0.088	0.141
Custom — MaxPrecision					
	Temperature	-0.449	0.114	-0.670	-0.223
	Relative Humidity	-0.611	0.096	-0.803	-0.425
	Moon Illumination (log)	0.309	0.142	0.040	0.599
BirdNET — MaxF1					
	Temperature	-0.561	0.132	-0.828	-0.311
	Relative Humidity	-0.985	0.122	-1.237	-0.757
	Moon Illumination (log)	0.192	0.065	0.063	0.319
BirdNET — MaxPrecision					
	Temperature	-1.810	0.309	-2.464	-1.259
	Relative Humidity	-1.653	0.268	-2.221	-1.169
	Moon Illumination (log)	-0.050	0.163	-0.377	0.261