University of Cape Town

Passive Acoustic Monitoring: Estimating Population Density

Student:

Selaelo Kgafela

KGFSEL001

Supervisor:

Greg Distiller

Co-supervisor:

Matt Rogan

Honours research project

in the

Department of Statistical Sciences

November 19, 2025

Contents

Introduction **3**

Literature Review **6**

Ecological monitoring and birds as bio-indicators 6

Passive acoustic monitoring and bird population metrics 7

Automated recognisers, deep learning, and BirdNET 8

Thresholds, classifier uncertainty and the move toward threshold-free metrics . . . 10

Call density as a threshold-free indicator 11

Methodology **13**

Overview 13

Data Sources and Preprocessing 14

Data Structure 14

Data Readjustment 16

Species Information 17

Baglafecht weaver 18

White-browed coucal 20

African Gray Flycatcher 22

Abyssinian/Montane Nightjar 24

African Pipit 26

Quantile binning of classifier scores 28

Validation and Annotation 30

Study-Level Call Density Estimation 32

Site-Level (Strata-Level) Estimation under Distribution Shift 36

Strategy 1 36

Strategy 2 39

Strategy 3 42

Results **44**

Baglafecht weaver 46

White-Browed Coucal 54

African Gray Flycatcher 65

Abyssinian/Montane Nightjar 74

African Pipit 83

Discussion **92**

Conclusion **97**

Future Work **99**

Appendix **101**

References **111**

Introduction

Monitoring biodiversity at scale is essential for understanding ecosystem health and guiding conservation efforts. Population abundance and species diversity are examples of state variables representing ecosystem structure, while carbon fluxes and trophic interactions reflect ecosystem functioning (Yoccoz, 2012). Birds, in particular, have long been recognized as valuable bio-indicators due to their sensitivity to environmental change and distribution; the role they play is that of a litmus for ecosystem stability and health (Pérez-Granados, Traba, et al., 2021).

Traditionally, the monitoring of birds was done through survey methods such as point counts, mist-netting, and transect walks. These have contributed extensively in ornithology and long-term datasets for monitoring bird population changes (Eastern Ecological Science Center, 2025). However, these techniques have introduced significant limitations: they are labour-intensive, geographically constrained, and subject to observer bias, and they struggle to detect rare or cryptic species. Such constraints can hinder inference about population parameters and introduce uncertainty into conservation decision-making (Gregory et al., 2004; Morelli et al., 2022)

Passive Acoustic Monitoring (PAM) is revolutionizing the field of biodiversity monitoring. Through the use of affordable, autonomous recording units (ARUs), scientists can unobtrusively gather extensive acoustic datasets from multiple bird habitats over long durations (Sugai et al., 2019). These datasets capture continuous information on vocalizing species and can reveal species presence, activity patterns, and habitat use without the constant presence of human observers and when paired with machine learning classifiers such as BirdNET, PAM provides an efficient means to process tera-bytes of recordings and detect target species with minimal manual effort (Wood and Kahl, 2024).

While PAM offers significant potential, it presents distinct challenges. Classifier-based methods often simplify detections to binary outcomes of detection or non-detection by applying confidence thresholds. Although this approach is straightforward, it introduces biases that manifest in the form of false positives that lead to overestimated species presence or false negatives that can obscure actual occurrences. Overly lenient thresholds may classify background noise or non-target calls as detections, inflating apparent species presence (false positives). Conversely, overly strict thresholds can exclude genuine detections with lower confidence scores, leading to false negatives and the underestimation of true occurrence. Moreover, the optimal threshold may differ across sites, seasons, or species, reflecting distribution shifts in the classifier's error structure (Knight et al., 2017). As a result, threshold-dependent estimates risk misrepresenting ecological phenomena and may perform inconsistently when applied across heterogeneous datasets (Navine et al., 2024).

A growing body of research suggests that the focus should move away from threshold-dependent detections toward directly estimating biologically meaningful quantities such as call density; the proportion of fixed-duration audio windows (e.g., 3 seconds each) that contain at least one confirmed vocalization of the target species. It provides a threshold-free estimate of vocal activity, which can be measured in percentage, ranging from 0% to 100%, and can serve as a proxy for species presence, behaviour, or abundance when monitored over time. Call density is particularly compelling because it links occupancy (whether a species is present) with abundance and activity (quantity of birds in a habitat and how frequently they call), providing a more robust ecological signal than binary detection counts (Navine et al., 2024). Once occupancy is established, changes in call density may also reflect shifts in population abundance, behaviour, or habitat quality, making it a versatile tool for monitoring population dynamics.

Threshold-free approaches to call density estimation are timely for two reasons. First, they allow researchers to harness the full distribution of classifier scores rather than discarding information that falls below a certain threshold. Secondly, they explicitly address the problem of distribution shifts across sites and conditions, which are a pervasive challenge in ecological monitoring (Navine et al., 2024). By integrating limited human validation effort with probabilistic modelling, these approaches generate reliable estimates of call density without requiring "perfect" classifiers; a critical advantage for real-world monitoring programs.

This study builds upon the threshold-free framework of Navine et al. (2024) to evaluate its applicability in an African context, acknowledging that many existing bird classifiers are not well trained on the vocalizations of African bird species. Using ARU data provided by a conservation organisation called Natural State that was collected from conservancies in and around Mount Kenya, it investigated whether direct estimation of call density can produce robust ecological insights for vocal bird species. Specifically, it assesses whether call density can be reliably estimated without reliance on arbitrary thresholds, whether alternative strategies can account for distribution shifts across strata and still give robust estimates of call density, and whether the resulting estimates align with ecological expectations of species distribution and activity. By focusing on call density rather than binary detection counts, this study contributes to the growing effort to standardize PAM analysis pipelines and enhance the utility of bio-acoustics for biodiversity monitoring and conservation decision-making.

This study's investigation of the threshold-free method will advocate for threshold-free bioacoustic monitoring techniques. If successful, as a start, it could lead to further research and possible implementation in conservancies surrounding Mount Kenya by Natural State, enabling them to estimate bird density at various altitudes and habitats.

Literature Review

Ecological monitoring and birds as bio-indicators

Ecological monitoring is fundamentally concerned with tracking changes in ecosystem state variables over space and time and linking these changes to their potential drivers (Yoccoz, 2012). Well-designed monitoring programmes rely on clearly articulated questions, explicit conceptual models, and carefully selected indicators that capture both ecosystem structure (e.g. species richness, population abundance) and ecosystem functioning (e.g. carbon fluxes, trophic interaction strength). Birds are particularly well suited to this role as they are taxonomically well studied, respond rapidly to environmental change, and are comparatively easy to detect across a broad range of environments. Their abundance, diversity, and community composition therefore serve as widely used proxies for ecosystem integrity and for ecosystem services such as recreation, aesthetic value, and cultural significance (Yoccoz, 2012; Pérez-Granados, Traba, et al., 2021).

Traditional bird monitoring methods such as point counts, line transects, spot mapping, and mist-netting have long underpinned large-scale population indices and trend analyses. These approaches remain critical for understanding long-term population change under land-use and climate pressures (Eastern Ecological Science Center, 2025). However, they are labour-intensive, constrained by accessibility, and sensitive to observer related variation in detection, fatigue, and identification skills. They also perform poorly for rare, nocturnal, or cryptic species and are difficult to scale to the spatial and temporal resolutions needed to track rapid biodiversity change (Pérez-Granados, Traba, et al., 2021). These limitations have motivated the development of complementary approaches that maintain ecological relevance while improving scalability and standardisation.

Passive acoustic monitoring and bird population metrics

Passive Acoustic Monitoring (PAM) with autonomous recording units (ARUs) has emerged as a transformative method for surveying vocal wildlife, enabling efficient monitoring of birds, amphibians, and acoustically active mammals across large spatial and temporal scales. ARUs can be deployed for extended periods, record continuously or on programmed schedules, and operate effectively in remote or inaccessible environments (Shonfield and Bayne, 2017; Pérez-Granados, Traba, et al., 2021). Compared with human-based surveys, PAM reduces observer bias, enables archival re-analysis, and allows researchers to revisit historical recordings as analytical methods improve. However, the large volume of data generated by ARUs necessitates automated or semi-automated processing pipelines in which automated recognisers and machine learning classifiers now play a central role.

The application of PAM to estimate bird population metrics has diversified into several methodological strands. Pérez-Granados, Traba, et al. (2021) identify at least four major classes of approaches: (i) localisation methods, which use microphone arrays to triangulate individual callers and estimate density; (ii) stereo or multi-channel analyses, which infer distance via signal amplitude or arrival direction; (iii) single-recorder distance sampling, which combines call counts with statistical models of detection and signal attenuation; and (iv) vocal-activity or soundscape indices, such as the Acoustic Complexity Index. These approaches differ in their logistical demands, assumptions regarding detectability and calling behaviour, and the type and volume of data required. For example, array-based localisation offers high spatial resolution but requires substantial field infrastructure, while simpler soundscape indices are computationally efficient yet may conflate variation in calling rates, detection probabilities, and acoustic background noise (Pérez-Granados, Traba, et al., 2021)

Common challenges across PAM-based density estimation approaches include heterogeneous

calling rates, imperfect detection, and the influence of environmental covariates such as wind, rain, vegetation structure and hardware differences such as microphone sensitivity on signal propagation. Cue-counting methods, which infer density from call rates, require reliable estimates of *cue rate* (calls per individual per unit time). However, cue rates are often species-specific, context-dependent, and poorly quantified. As highlighted in recent reviews, these limitations underscore the need for well-designed calibration studies that pair PAM data with traditional surveys or experimental protocols to estimate detection functions, cue rates, and other parameters with sufficient precision (Pérez-Granados, Traba, et al., 2021).

Automated recognisers, deep learning, and BirdNET

Advances in PAM have paralleled rapid improvements in automated signal recognition. Early recognisers relied on spectrogram cross-correlation or template matching, which perform well for a limited set of species with stereotyped calls but degrade quickly in complex soundscapes. Modern systems increasingly use supervised machine learning methods such as random forests, support vector machines, and deep convolutional neural networks (CNNs) to learn discriminative features directly from audio (Priyadarshani et al., 2018). Reviews emphasise that model performance is sensitive to background noise, overlapping calls, recording quality, and the taxonomic and geographic coverage of the training data, underscoring the need for rigorous validation under realistic field conditions (Priyadarshani et al., 2018; Shonfield and Bayne, 2017).

BirdNET is one of the most widely adopted CNN-based classifiers for birds. It was originally trained primarily on recordings from roughly 1 000 European and North American species, and although its v2.4 release now includes more than 6 000 species worldwide (Wood and Kahl, 2024), the training data remain geographically uneven. African species, dialects, and

soundscapes are comparatively underrepresented, which can lead to reduced classification performance for taxa in this region. BirdNET processes audio in fixed 3-second windows and assigns a continuous *score* for each species-window pair. This score output by *BirdNET* as a unitless non-probabilistic score that is transformed via sigmoid function to lie in the interval 0 and 1. As emphasised by Wood and Kahl (2024), BirdNET confidence scores are *not* calibrated probabilities. A given score does not necessarily correspond to the same likelihood of a true detection across species, recording environments, or hardware configurations, meaning that identical numerical outputs may reflect very different underlying levels of precision and recall [1]. Instead, the scores provide a monotonic ranking of classifier confidence whose relationship with true positive and false positive rates must be empirically validated.

Despite these limitations, BirdNET-based pipelines have demonstrated strong potential for ecological inference. Wood et al. (2019) introduced a bioacoustic site-occupancy framework capable of detecting subtle population changes, while Wood and Peery (2022) clarified how acoustic occupancy relates to biological presence and calling behaviour. BirdNET has also become an important tool for citizen science through the BirdNET app, which provides open access to classifier outputs for research (Wood et al., 2022). In addition, BirdNET-derived feature embeddings have been shown to contain ecologically meaningful information beyond simple detections (McGinn et al., 2023). However, as reviews note, species-specific validation and alternatives to ad hoc thresholding remain critical (Pérez-Granados, 2023; Wood and Kahl, 2024).

[1] In classification, *precision* measures the proportion of detected calls that are true positives, while *recall* measures the proportion of actual vocalizations that the model successfully detects.

Thresholds, classifier uncertainty and the move toward threshold-free metrics

Most PAM workflows reduce continuous classifier scores to binary detection/non-detection values using fixed thresholds, typically derived from heuristic rules or validation-based trade-offs between precision and recall. Knight et al. (2017) provide best-practice guidelines for recogniser evaluation, emphasising ROC curves, independent test sets, and transparent reporting. Knight et al. (2020) extend this with a "validation–prediction" framework designed to improve the efficiency of automated pipelines. Despite this, threshold-based approaches still remain limited because thresholds rarely transfer across species or sites, they introduce strong biases toward false positives or false negatives, and discard informative sub-threshold variation in score distributions (Pérez-Granados, 2023; Wood and Kahl, 2024).

These challenges are exacerbated under distribution shifts, where score distributions at new sites differ from those observed during training or calibration. Such shifts may arise from differences in background noise, habitat structure, species assemblages, or recording hardware, including changes in sampling rate or microphone sensitivity (Wood and Kahl, 2024). Under these conditions, thresholds optimised on one dataset may perform poorly elsewhere, leading to inconsistent or biased inference. This has motivated the development of methods that (i) use the full distribution of classifier scores, (ii) explicitly model uncertainty, and (iii) produce ecologically interpretable metrics suitable for comparison across space and time (Pérez-Granados, Traba, et al., 2021). Pérez-Granados, Traba, et al. (2021) also highlight the potential of calibrated call-rate or vocal-activity metrics to minimise dependence on arbitrary thresholds. However, many existing approaches still rely on some form of binary detection step.

Call density as a threshold-free indicator

Navine et al. (2024) propose *call density* as a threshold-free, PAM-based indicator of population status. In their formulation, *call density is defined as the proportion of fixed-length audio windows (e.g. 3 s segments) that contain a true vocalisation of the target species.* Conceptually, a call density of 0.4 means that 40% of all recorded windows include at least one true vocalisation, providing a direct measure of acoustic activity that is interpretable without converting classifier scores to binary detections. By measuring the proportion of recorded audio that contains verified vocalizations, call density provides a quantitative index of vocal output. This metric has been shown to correlate with local abundance, presence, or behaviour in many vocal species.

Using fully annotated datasets and simulated validation designs, Navine et al. (2024) demonstrate that call density can be estimated from classifier outputs by calibrating the relationship between score and truth within quantile-based bins. This design avoids reliance on a single decision threshold and is robust to variation in classifier performance, species detectability, and underlying abundances. Importantly, call density can also be linked to traditional ecological metrics such as occupancy or distance-sampling estimates allowing PAM-based indicators to align with established monitoring frameworks.

In their Hawaiian case study, Navine et al. (2024) show that call density tracks population changes inferred from point counts while circumventing the limitations of threshold-dependent workflows. Their approach models the full distribution of classifier scores using Beta–Binomial bin-level detection models and mixture components to accommodate differences between study-level and site-level score distributions. This structure resolves several persistent challenges in automated acoustic monitoring, particularly those related to threshold sensitivity and distributional mismatch between different strata.

However, most applications to date remain concentrated in temperate regions or among species well represented in classifier training data (Pérez-Granados, Traba, et al., 2021; Scarpelli et al., 2019). Few studies have explored threshold-free methods in African bird communities, where calling behaviour, soundscape complexity, and field conditions can differ substantially from datasets used to train global models. These challenges are compounded by variation in recording hardware and sampling rates, which is known to affect score distributions and introduce further sources of distribution shift (Wood and Kahl, 2024).

In response, the present study applies the threshold-free call-density framework of Navine et al. (2024) to BirdNET outputs collected across the Mount Kenya landscape. By analysing species-specific logit-score distributions and modelling for call density across distributional shifts between the 32 kHz and 48 kHz audio sampling rates as strata-level shifts. The study also examines how hardware-driven variation affects the calibration of score-to-truth relationships. Through the evaluation of three complementary strategies for strata-level call density estimation, this work provides an empirical assessment of threshold-free call-density estimation within an African acoustic environment, an important geographic and methodological gap in the current literature.

Methodology

Overview

At a high level, this study estimates call density by assigning human-generated labels of bird species detection and non-detection (human annotations) to 3-second audio examples expressed as logit confidence scores for a given bird species of interest (focal species). These audio examples are grouped into bins [2] based on quantiles that contain a specified amount of data. Following this, representative samples are then selected per bin for human validation and annotation. Per-bin *call-density* estimates are then calculated using methods/techniques that are described in greater detail later in this *Methodology* Section. These methods are specifically, the Study-level approach which makes use of the full dataset of logit confidence scores (All of the data available in the study) to calculate the call density and three complementary strata-level strategies (Strategy 1, Strategy 2 and Strategy 3), with each designed to estimate call density across two strata comprising a splits of the study level full dataset into a 32 kHz audio sampling rate representing Stratum 1 and the 48 kHz rate representing Stratum 2. These strata capture distributional shifts in the logit score distributions associated with the two sampling rates. Together, these procedures yielded final call-density estimates at both the study level and the strata level

The variability in the estimates was assessed through the standard deviations (SD) of the estimated call densities. At the study level, both empirical and bootstrap-based standard deviations were computed to quantify two complementary sources of variation. The empirical standard deviation reflected the variability observed across repeated simulation[3] runs, and the bootstrap standard deviation captured the internal sampling uncertainty within the

[2] A bin refers to an interval of the logit confidence scores defined by adjacent quantile boundaries. In this study, bins partition the range of logit confidence scores into discrete segments

[3] *Simulations* in this study referred to the generation of empirical results across multiple independent model runs (study-level bins having new random human-validated populations of within-bin samples).

fitted Beta–Binomial model. Bootstrapping was therefore used to re-estimate the call density and its associated standard error, producing a distribution that represented uncertainty under finite sampling. These two standard deviations were implemented only at the study level, since Strategies 1 and 2 directly built on the study-level calibration. Strategy 3, an algebraic ensemble of Strategies 1 and 2 formed through a geometric mean, had no re-sampling bootstrapping stage and thus no explicit estimation error was calculated for it.

This framework follows the threshold-free methodology for bio-acoustic monitoring proposed by Navine et al. (2024), which integrates study-level validation, logarithmic or quantile binning, and Beta-binomial distribution modelling with small priors[4] to support robust strata-level extrapolation. In this study, the analysis was applied to five focal species. Namely, the Baglafecht weaver, white-browed Coucal, African Gray Flycatcher, Abyssinian Nightjar and African Pipit.

Data Sources and Data Preprocessing

Data Structure

The acoustic data used in this study were provided by the organisation called *Natural State* and were collected across multiple conservancies in and around the Mount Kenya landscape in Kenya. Natural State's Research Centre, situated at the base of Mount Kenya, operates within a heterogeneous ecological matrix of open grasslands, shrublands, montane forests, and wetland habitats that support a high diversity of bird species.

Autonomous Recording Units (ARUs) deployed by Natural State's monitoring teams collected extensive audio data, ranging from tens to hundreds of hours per site, over multiple

[4]Small constant value added to both parameters of the beta-binomal model to remedy situations where the model is undefined because 0 is a parameter.

recording phases. During field deployment, ARUs continuously captured soundscapes at two distinct sampling rates, 48 kHz during the initial phase of data collection and 32 kHz in a subsequent phase. Once retrieved from the field, recordings were spliced into 3-second segments during pre-processing, producing a large corpus of short, standardized audio examples that were fed to a convolutional neural network called *BirdNET* that classifies bird species vocalisations.

Each audio segment was analyzed using the BirdNET classifier and assigned a confidence score between 0 and 1, representing the classifier's certainty that a segment contains vocalizations from the focal bird species. The confidence scores were transformed into the logit domain to stabilize skewness and emphasize their distribution, facilitating more effective quantile-based binning and model calibration. The resulting datasets with accompanying logit confidence scores attached to corresponding audio examples from both sampling rates (32 kHz and 48 kHz) were merged to create the study-level dataset, while pure copies of the 32 kHz and 48 kHz subsets were still retained for strata-level analysis. Each dataset contained detailed metadata across several variables. These were, *Selection*, *View*, *Channel*, *Begin Time (s)*, *End Time (s)*, *Low Freq (Hz)*, *High Freq (Hz)*, *Species Code*, *Common Name*, *Confidence*, *logit_Confidence* and *source_file*. Together, these variables describe the temporal and spectral boundaries of detected calls, the associated species identity, the classifier's confidence score (non-probabilistic confidence of detection), and the file-level source of each observation.

From the entire data collected by *Natural State*, data on only five species was selected for analysis based on the species' relative abundance and detectability in the Mount Kenya landscape. As stated in the *Overview* section, the species chosen were, the *Baglafecht weaver*, *White-browed Coucal*, *Abyssinian/Montane Nightjar*, *African Gray Flycatcher* and *African Pipit*. These species collectively represent a gradient of vocal behaviours, acoustic detectabil-

ity, and call structure; from the highly tonal songs of the Baglafecht weaver to the low, booming calls of the White-browed coucal(Kirwan et al., 2024), making them suitable for evaluating the performance of threshold-free call density estimation across distinct acoustic profiles.

For each species, the BirdNET classifier produced a distribution of confidence scores across all 3-second audio segments, reflecting the model's degree of certainty in identifying vocalizations. These distributions, when annotated (see *Validation and Annotation* section in subsequent pages), form the basis for estimating call density and assessing classifier reliability across strata.

Data Readjustment

The raw output from the *BirdNET* classifier comprised confidence scores ranging from 0 to 1, reflecting *BirdNET's* confidence that each 3-second audio window contained a vocalization from the target species. During pre-processing, all scores between 0 and 0.1 were discarded by *Natural State's* research team as these primarily represented background noise or other non-informative acoustic events. While this filtering improved data quality, it also eliminated a substantial portion of potential non-detection samples, which are crucial for unbiased estimation of call density.

To account for missing low-confidence audio windows, an adjustment procedure was performed for each focal species. The earliest and latest ARU recording times were identified across the 32 kHz and 48 kHz datasets to define the full temporal extent of acoustic sampling. Each audio segment reflected a three-second recording window. Additionally insights into the recording schedule were gained from both datasets. For the 32 kHz dataset, record-

ings were taken continuously at regular hourly intervals over several weeks. By converting these timestamps into three-second windows and comparing the expected total number of windows across the full monitoring period with the number of audio examples available, the number of omitted low-confidence windows could be estimated. 48 kHz audio recording data showed structured sampling design where one hour audio recordings were captured daily at approximately 03h00 and 15h00, targeting early-morning and afternoon vocal activity. The 32 kHz dataset recorded acoustic data every hour across the entire 24-hour cycle, providing continuous coverage of variation in vocal behaviour. Together, these datasets enabled both fine-scale and broad-scale temporal analysis of vocal activity, while the adjustment procedure ensured that the missing low-confidence segments were accounted for in downstream density estimation.

To restore temporal completeness, these missing examples were imputed by generating new confidence scores randomly drawn from a uniform distribution over $[0, 0.1)$. These synthetic values approximate the distribution of discarded background segments, ensuring that the empirical distribution of logit-transformed scores more accurately represented the full soundscape present during data collection. This adjustment preserved the original sampling effort and maintained the balance between detections and non-detections necessary for robust call density estimation.

Species Information

This subsection provides an overview of the data about each of the focal species analysed in this study. Outlining its typical behaviour, vocal characteristics, and natural habitat. Alongside their ecological background, the corresponding distribution of BirdNET logit confidence scores and their five number summary for each species is also presented to illustrate how classifier's certainty varied across acoustic environments and species call types.

Baglafecht weaver

Below is information relating to the Baglafecht weaver bird species.

Figure 1: Picture of a Baglafecht weaver.

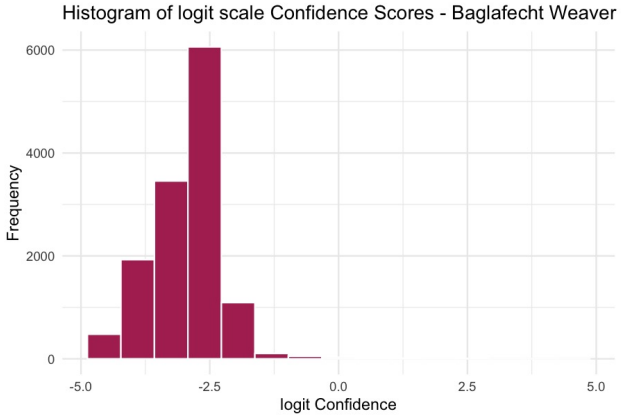

Figure 2: Confidence histogram for Baglafecht weaver.

Level	**Min**	**Q1 (25%)**	**Med**	**Q3 (75%)**	**Max**
Study Level (All)	-4.498	-3.366	-2.821	-2.446	4.575
Strata 1 (32 kHz)	-4.498	-3.360	-2.816	-2.447	3.867
Strata 2 (48 kHz)	-4.494	-3.386	-2.836	-2.446	4.575

Table 1: Five-number summary of logit confidence scores at the study level and by sampling rate.

Craig (2020) described the call of the Baglafecht weaver as follows, "Song is a chattering mixed with musical notes, including high-pitched glissando and swizzling sounds, usually a two-part or three-part series. Common call in Kenya is a chatter, "swii chee chee cheechit", and pair-members exchange regular "pseet" or "shreeep" contact". According to Oschadleus

(2024), the Baglafecth weaver is a territorial species and is never found in large flocks. It is often spotted singly, in pairs, or in small parties, typically inhabiting open vegetation and forest clearings, and is well distributed across eastern and central Africa.

As illustrated in Figure 2, the Baglafecht weaver's logit-scale confidence scores are strongly left-skewed, with most classifier outputs falling in the low-confidence region. The study-level scores range from approximately –4.50 to 4.58, with the majority of values concentrated between –3.5 and –2.0. This pattern indicates that BirdNET assigned relatively low confidence to most audio segments, consistent with the species' soft, unobtrusive vocalisations and its generally low detectability in the field (Craig, 2020).

Table 1 summarises the five-number statistics for the study-level dataset and for each sampling-rate stratum separately. The 32 kHz and 48 kHz subsets displayed similar central tendencies, though the 48 kHz stratum showed a slightly wider range and slightly lower median confidence. These patterns reflected mild distributional shifts between sampling rates, which were further examined later in the *Results* section of this study.

White-browed coucal

Below is information relating to the White-browed coucalbird species.

Figure 3: Picture of a white-browed coucal.

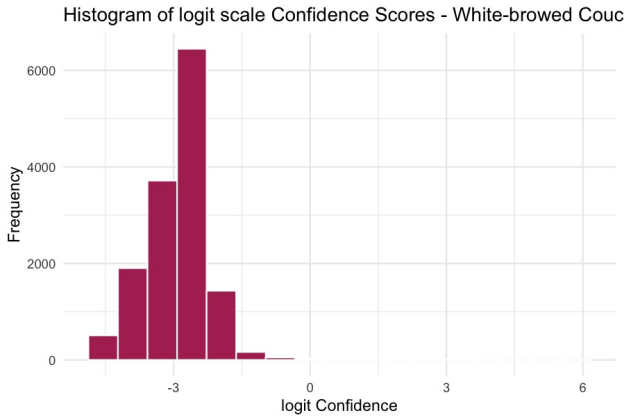

Figure 4: Logit confidence histogram for white-browed coucal.

Level	Min	Q1 (25%)	Med	Q3 (75%)	Max
Study Level (All)	-4.499	-3.340	-2.788	-2.424	6.030
Strata 1 (32 kHz)	-4.499	-3.341	-2.791	-2.426	6.030
Strata 2 (48 kHz)	-4.496	-3.333	-2.767	-2.416	2.216

Table 2: Five-number summary of logit confidence scores for white-browed coucal.

The White-browed coucalis a bird species commonly found in wetland and riparian[5] habitats throughout East Africa, extending southward into the northern regions of Southern Africa. Its vocalizations are described as low, bubbling calls that resemble the sound of water being poured into a bottle. These calls are consistent across members of the coucal family, maintaining a similar acoustic structure among related species (Kirwan et al., 2024). The

[5]Riparian zones are transitional areas between land and rivers, forming semi-terrestrial habitats that both affect and are shaped by the flow and dynamics of freshwater systems.(Urbanič et al., 2022)

White-browed coucalis also noted for its territorial and monogamous behaviour, particularly during the breeding season.

As illustrated in Figure 4, the distribution of *logit-transformed* confidence scores for the White-browed coucalwas strongly right-skewed, with the vast majority of observations concentrated at low logit confidence values. Only a small proportion of windows fell into higher logit-score regions, indicating that the classifier rarely assigned high-confidence predictions for this species. This pattern was consistent with the acoustic properties of the white-browed coucal's vocalisations, which are low-frequency, bubbling, and often difficult for automated classifiers to distinguish from background sound.

The White-browed coucalexhibited a predominantly low logit confidence distribution, with the interquartile range tightly clustered between approximately -3.34 and -2.42 across all levels of analysis (Table 2). This indicated that the classifier assigned similarly low confidence to most audio segments regardless of sampling rate. The 48 kHz subset showed a noticeably reduced maximum logit confidence score (2.216 compared to 6.030 at both the study level and 32 kHz strata), suggesting that high-confidence detections were substantially less frequent at this sampling rate. However, the median and quartile values remained nearly identical across strata, implying that the bulk of the score distribution was stable and that the primary difference lay in the suppression of extreme high-confidence values at 48 kHz.

African Gray Flycatcher

Below is information relating to the African Gray Flycatcher.

Figure 5: Picture of a African Gray Flycatcher.

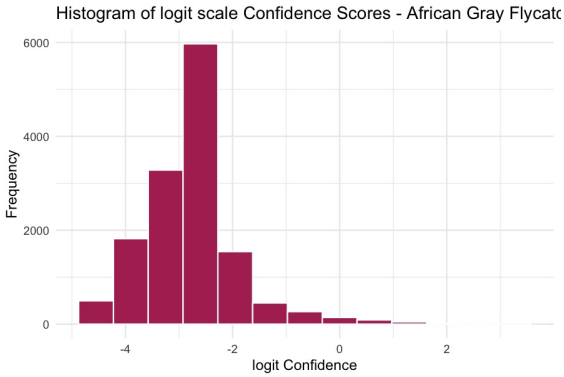

Figure 6: Logit confidence histogram for African Gray Flycatcher.

Level	**Min**	**Q1 (25%)**	**Med**	**Q3 (75%)**	**Max**
Study Level (All)	-4.498	-3.277	-2.734	-2.360	3.240
Strata 1 (32 kHz)	-4.498	-3.278	-2.733	-2.360	3.240
Strata 2 (48 kHz)	-4.472	-3.261	-2.777	-2.383	0.924

Table 3: Five-number summary of logit confidence scores for African Gray Flycatcher.

The African Gray Flycatcher was found in savanna ecosystems, woodland, orchard country, bushland, and wooded grassland, particularly in clearings and secondary growth. In Kenya, where the data for this study were collected, it occurred in the lowlands and hills up to an altitude of 2000 metres (Taylor, 2020). Taylor (2020) also described the species' vocalisations as complex, with harsh and scratchy notes. Its distribution, noisiness, and detectability were considered to be at average levels.

The logit confidence distribution for the African Gray Flycatcher (Figure 6) showed a characteristic left-skewed shape, with most classifier outputs concentrated between approximately -3.5 and -2 on the logit confidence scale. A smaller number of detections appeared in the higher-confidence range (logit > -1), indicating that the classifier occasionally assigned strong confidence to true vocalizations but did so infrequently. Compared to the white-browed coucal, the African Gray Flycatcher displayed a broader spread of mid-range logit confidence scores and a more gradual decline toward the upper tail, suggesting that the classifier distinguished this species with slightly greater confidence overall, even though high-certainty detections remained sparse.

The African Gray Flycatcher exhibited a tightly clustered range of low logit confidence values across all strata, with first and third quartiles consistently falling between approximately -3.28 and -2.36, and median scores near -2.74. The 48 kHz subset showed a noticeably reduced maximum logit score (0.924) relative to the 32 kHz recordings (3.240), indicating that high-confidence detections were substantially rarer at this sampling rate. Apart from this ceiling effect, the lower quantiles and central tendency were nearly identical across strata, suggesting that the classifier assigned similarly low baseline confidence regardless of sampling rate, but was less able to produce high-certainty predictions under 48 kHz conditions.

Abyssinian/Montane Nightjar

Below is information relating to the Montane/Abyssinian Nightjar bird species.

Figure 7: Picture of an Abyssinian/Montane Nightjar

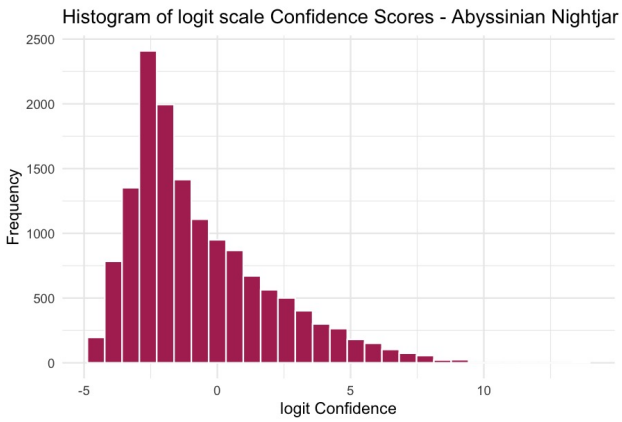

Figure 8: Logit confidence score histogram for Abyssinian/Montane Nightjar

Level	**Min**	**Q1 (25%)**	**Med**	**Q3 (75%)**	**Max**
Study Level (All)	-4.498	-2.517	-1.438	0.754	13.816
Strata 1 (32 kHz)	-4.498	-2.512	-1.433	0.774	13.816
Strata 2 (48 kHz)	-4.494	-2.543	-1.464	0.676	9.210

Table 4: Five-number summary of logit confidence scores for the Abyssinian/Montane Nightjar.

The Abyssinian/Montane Nightjar is a nocturnal bird species found throughout the upland regions of eastern and central Africa, including Ethiopia, Kenya, Uganda, Tanzania, and Zambia. It inhabits high-altitude areas, typically between 1000 meters and 3 000 meters above sea level, where it frequents moist montane forests, open grasslands, forest clearings, and shrublands near forest edges. During the day, the species roosts on shaded ground or

low branches, relying on its cryptic plumage for camouflage, and becomes active at dusk to feed on flying insects captured in mid-air. Its vocalisation is distinctive, consisting of a repeated nasal "ank-ank-ank" followed by a high-pitched whistle rendered as "piiiyu-pirrr," the first note descending and rising, while the second is tremulous and falling in tone. These calls are often delivered at night and serve both territorial and mate-attraction functions (Cleere et al., 2022; Oiseaux.net, 2025).

The logit confidence histogram shown in Figure 8 exhibited a strongly right-skewed distribution for the Abyssinian/Montane Nightjar. Most 3-second audio windows received low to moderate logit values between approximately -3 and 1, with a clear concentration near the lower end of the scale. However, unlike the other focal species, the distribution extended into a long positive tail, reaching logit values above 8, indicating that the classifier occasionally assigned extremely high confidence to presumed detections. Mid-range values were comparatively sparse, suggesting that BirdNET tended to classify windows either as low-confidence non-detections or as high-confidence detections, with relatively few ambiguous scores in between. Overall, this pattern reflected a relatively strong separation between likely calls and background noise for this species.

The Abyssinian/Montane Nightjar exhibited a broad range of logit confidence values across all recordings, with quartiles spanning from roughly -2.5 to 0.75, and a median near -1.44. The lower quantiles were nearly identical between the 32 kHz and 48 kHz strata, indicating consistent classifier behaviour in the low-confidence region. However, the maximum logit score dropped from 13.816 at 32 kHz to 9.210 at 48 kHz, suggesting that the classifier produced fewer extreme high-confidence detections in the 48 kHz subset. Apart from this ceiling difference, the overall distribution remained structurally similar between strata, reflecting the species' generally high separability in the classifier's score space.

African Pipit

Below is information relating to the African pipit bird species.

Figure 9: Picture of a African Pipit.

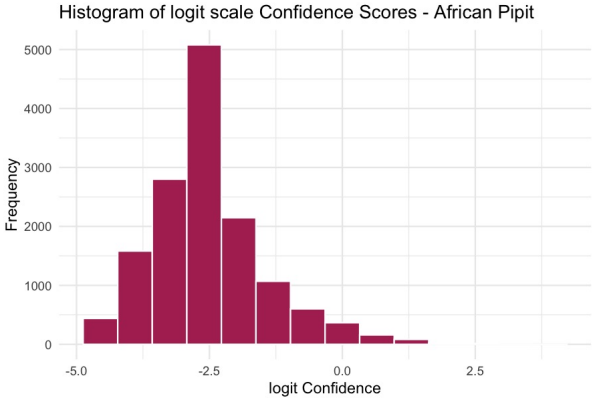

Figure 10: Logit confidence score histogram for the African Pipit.

Level	**Min**	**Q1 (25%)**	**Med**	**Q3 (75%)**	**Max**
Study Level (All)	-4.498	-3.157	-2.572	-2.158	3.771
Strata 1 (32 kHz)	-4.498	-3.157	-2.572	-2.158	3.771
Strata 2 (48 kHz)	-4.449	-3.122	-2.592	-2.136	1.606

Table 5: Five-number summary of logit confidence scores for African Pipit at the study level and by sampling rates.

The African Pipit is a widespread species found throughout Eastern and Southern Africa, migrating southward during the non-breeding season. Its calls are short, typically lasting between 0.011 and 0.18 seconds. During flight, vocalizations are usually produced as single *chup* notes, whereas take-off calls consist of a rapid series of repeated phrases. When on the

ground, the call closely resembles the flight call (Teichmann, 2019).

As shown in Figure 10, the logit confidence scores for the African Pipit were strongly left-skewed, with the highest concentration of values clustered between approximately -3.5 and -2.0. Only a small proportion of segments received logit scores above 0, and scores greater than 2 were extremely rare. This pattern indicated that BirdNET most often assigned low confidence to African Pipit detections, with few instances in which the classifier expressed high certainty. The histogram therefore reflected a score distribution dominated by low-confidence predictions, consistent with classifier uncertainty rather than strong positive detection signals.

The African Pipit exhibited a tightly constrained range of low logit confidence values across all levels of aggregation. As shown in Table 5, both the study-level and 32 kHz strata shared identical quartiles and maximum values, with scores spanning from approximately -4.50 to 3.77. The 48 kHz stratum displayed a slightly narrower upper range, with a reduced maximum logit score of 1.61, although the median and interquartile bounds remained highly similar to the 32 kHz data. These results indicated that, while the overall location and spread of confidence scores were nearly identical across strata, the classifier produced fewer high-confidence outputs at 48 kHz compared to 32 kHz.

With the species-level data and score distributions now characterised, the analysis proceeds by organising the logit-transformed BirdNET confidence scores into a structured form suitable for call-density estimation. The next section, *Quantile binning of classifier scores*, outlines how these scores were partitioned into quantile-based bins, creating the framework from which per-bin detection probabilities and the subsequent call-density estimates were derived.

Quantile binning of classifier scores

This section describes one of the key preprocessing techniques used to prepare the dataset for call-density estimation, the quantile-based binning of logit-transformed confidence scores. The process involved partitioning the score distributions for each focal species into quantile-defined bins, forming the basis for subsequent modelling of per-bin call densities.

Let z denote the (logit-transformed) BirdNET confidence score for each 3 second audio window. Each z was allocated into one of B segments, $\{b_1, \ldots, b_B\}$ using quantile cut-points that were obtained by first examining the logit-confidence score distribution for each species and identifying the quantiles that divided the data into the desired proportions. These quantile thresholds produced 3, 4, 5, or 6 segment partitions, referred to as bins, throughout this study. The quantiles segmented the data by establishing boundaries, with each interval between two consecutive quantiles defining a bin. All observations falling within the interval between a quantile and the previous quantile were assigned to the same bin. This process continued until the entire set of logit confidence scores was segmented or binned.[6] By using binning, a higher resolution on data points could be achieved across different regions of the logit confidence score distribution during call density estimation, enabling more accurate call density estimation, as also demonstrated by (Navine et al., 2024). In this context, $P(b_i)$ represented the relative probability or weight of bin b_i, expressing the proportion of all annotated audio examples, both positive and negative, that fall within that bin. It thus quantified each bin's contribution to the overall call density estimate, ensuring that bins containing more data exerted proportionally greater influence in the aggregation of per-bin call density estimates.

Several binning schemes were used and as aluded to in the previous paragraph, these included

[6]The first and last quantiles used in any binning scheme correspond to the minimum and maximum values of the complete distribution of logit confidence scores, respectively.

quantile schemes that sectioned off the data into 3, 4, 5 or 6 segments/bins. Each binning scheme, except for the three-bin scheme, were designed to optimize binning of the distributional patterns of logit confidence scores for a higher resolution in regions where scores were most densely concentrated. The three-bin scheme served as a control to assess call density estimation when all bins contained equal proportions of logit confidence scores and it made use of equal quantile splits of (0.33, 0.33, 0.33) as a configuration, enabling assessment of model behaviour under coarse segmentation. The four-bin scheme of quantiles (0.25, 0.25, 0.30, 0.20) introduced finer resolution in the lower-to-mid logit score regions, where *BirdNET* confidence scores typically cluster due to background noise and ambiguous detections. The five-bin configuration of quantiles (0.20, 0.25, 0.10, 0.25, 0.20) was tailored to bi-modal or logit confidence score distributions with a relatively high proportion of high *BirdNET* scores as seen for the Montane/Abyssian Nightjar species in the *Species Information* section; where both low and high confidence scores dominate while mid-range scores are relatively sparse. This structure enhanced the Beta–Binomial model's sensitivity at the extremes of the logit distribution. Finally, the six-bin scheme of quantiles (0.20, 0.18, 0.18, 0.18, 0.16, 0.10) was designed for long-tailed or exponentially skewed distributions, a pattern often observed when the majority of audio windows yield low classifier confidence and only a small subset produce high-confidence detections. By progressively refining bin granularity and concentrating resolution where the score distribution was most informative, these schemes collectively aimed to provide a robust framework for assessing how bin structure influenced both the accuracy and interpretation of threshold-free call-density estimates.

With the quantile-binning framework in place and the logit-score distributions organised into species-specific segments, the next step was to link these binned classifier outputs to biological truth. This required reliably annotated data against which detection probabilities could be calibrated. The following section, *Validation and Annotation*, outlines how the quantile-binned audio windows were assigned expert-derived detection labels, providing the

empirical foundation for all per-bin probability estimates and subsequent call-density modelling.

Validation and Annotation

This section details the procedure used to assign detection $+1$ and non-detection -1 labels to the species of interest, and describes the human validation process that informed the annotation of the unlabelled *BirdNET* output introduced earlier in the *data* and the *species information sections*.

Following classification by the *BirdNET* model and transformation of confidence scores into the logit domain and being binned according to quantiles, the resulting dataset lacked explicit ground-truth labels indicating whether a 3-second audio segment contained a true vocalization by the target species or not. The purpose of the annotation process was to generate these labels, where *detections* $(+1)$ corresponded to confirmed species vocalizations and *non-detections* (-1) confirmed absences. These annotations collectively formed the empirical foundation for evaluating model accuracy and estimating call density.

The annotations for each 3-second audio example are assigned manually by trained ornithologists who reviews individual 3-second audio clips and assigned one of three possible outcomes, $+1$ for a confirmed vocalization by the focal species, -1 for a confirmed non-target species vocalization, and 0 for uncertain or ambiguous examples. However, the large volume of data generated by autonomous recording units (ARUs), often encompassing hundreds of hours of audio data per strata, rendered manual annotation of the full dataset by a human effort infeasible. Consequently, following the approach outlined by Navine et al. (2024), only a random subset of audio clips within each bin was selected for expert validation. This subset was treated as a representative sample of the full population of 3-second audio windows within its

respective bin, enabling the reliable estimation of per-bin detection rates while keeping the manual annotation effort manageable. Given the limited availability of manually validated annotations, a simulation-based approach was used to assign detection labels across the full dataset. Specifically, for each species, the same quantile-based binning thresholds that were applied to the unlabelled dataset of *BirdNET* outputs were super imposed onto *Natural State*'s existing validated datasets. Within each bin of the *Natural State's* validated data, the proportion of +1 (detection) labels was calculated. These per-bin detection proportions were then used to simulate the assignment of +1 (detection) labels onto the corresponding bins in the unlabelled dataset. In each bin, the remaining non-annotated scores after having assigned +1s were assigned −1 (non detection). 0 (unknown) detections were group with −1 and consequently assigned −1s in this study. This approach preserved the relative distribution of detections across the score space, while enabling comprehensive annotation of the dataset for downstream call-density estimation. 3 second audio windows of the dataset that formed part of the confidence scores between 0 and 0.1 were all assigned labels of -1.

With the validation step completed and each audio segment assigned a detection or non-detection label, the dataset now contained the essential annotated information required for modelling. These annotated examples form the empirical basis for estimating how classifier scores relate to true species activity. The following section, *Study-Level Call Density Estimation*, outlines how these validated data are used to quantify call density. Forming the core metric for the threshold-free estimation framework.

Study-Level Call Density Estimation

Notation and bin design

Let $X = \{x\}$ denote a 3 second audio window's confidence score and let x be mapped to a score $z \in \mathbb{R}$ (z being logit scale confidence score). Let $P(\oplus)$ be the *call density*: the probability that a randomly sampled window contains a true vocalization of the focal species, and let complement of $P(\oplus)$ be $P(\ominus) = 1 - P(\oplus)$. By the law of total probability, the study-level logit confidence score distribution decomposed as:

$$P(z) \;=\; P(z \mid \oplus)\, P(\oplus) \;+\; P(z \mid \ominus)\, \{1 - P(\oplus)\}$$

Equivalently, expanding the probability of the target focal species over $P(z)$ yielded

$$P(\oplus) \;=\; \int P(\oplus \mid z)\, P(z)\, dz$$

For efficient validation, the z were converted into B discrete bins $\{b_1, \ldots, b_B\}$ using the quantile schemes explained in the previous section (*Quantile binning of classifier scores*). Writing b for a generic bin (and b_i when explicit indexing is needed), the marginal probability of landing in bin b expanded over $P(\oplus)$ by the law of total probability was:

$$P(b) \;=\; P(b \mid \oplus)\, P(\oplus) \;+\; P(b \mid \ominus)\, \{1 - P(\oplus)\} \qquad \text{(eq.3)}$$

With the discrete nature of binning, the integral representation of $P(\oplus)$ above became a finite mixture, as shown below:

$$P(\oplus) \;=\; \sum_{b=1}^{B} P(\oplus \mid b)\, P(b)$$

$P(b)$ was computed from the study-level score histogram and the conditional probability of

a positive z given a bin b_i ($P(\oplus \mid b)$) was then estimated from the per-bin annotations, as described in the subsequent subsections.

Following Navine et al. (2024), the bin-level probability of a positive is modelled with a Beta-binomial model parametrized as below:

$$P(\oplus \mid b) \;\approx\; \mathrm{Beta}\big(k_{b,\oplus} + c,\; k_{b,\ominus} + c\big),$$

where c is a small uninformative prior to handle zero counts ($c = 0.1$ used throughout as it gave good coverage across densities in Navine et al. (2024)). The $k_{b,\oplus}$ and $k_{b,\ominus}$ where respectively the counts of positive and negatively annotated audio examples in bin b_i

Estimator for study-level call density

Using the bin probabilities $P(b)$, which were derived from the logarithmic binning scheme and the law of total probability (eq.3), the study-level call density was defined as the expected overall probability of a positive detection, computed as a weighted sum of the per-bin expected probabilities of a positive outcome. Each bin's contribution to the total density was proportional to its weight $P(b)$ and its corresponding Beta mean, reflecting the expected probability of a positive annotation within that bin. Formally, the study-level call density was expressed as a *summation of Beta means* as follows:

$$P(\oplus) \;=\; \sum_{b=1}^{B} \mathbb{E}[P(\oplus \mid b)]\; P(b) \;\approx\; \sum_{b=1}^{B} \frac{k_{b,\oplus} + c}{k_{b,\oplus} + k_{b,\ominus} + 2c}\, P(b),$$

where $k_{b,\oplus}$ and $k_{b,\ominus}$ denote the number of positively and negatively annotated audio segments in bin b, respectively, and c is a small regularization constant ensuring numerical stability for bins with few annotations.

This formulation provided a single point estimate of the study-level call density, derived as the weighted sum of expected per-bin detection probabilities. However, as with any estimator built from finite validation samples and probabilistic estimates, quantifying the associated uncertainty was critical for robust inference.

Uncertainty quantification and variance decomposition

To assess the reliability of the study-level call density estimate $P(\oplus)$, two complementary measures of variability were computed. One reflecting the *uncertainty within a single model run* (bootstrap variance), and the other reflecting *variability across repeated validation samples* (empirical variance). These captured distinct but related sources of uncertainty in the estimation process.

Bootstrap variance

It quantified uncertainty arising from the call density estimation using a finite number of validation examples within each bin. It was derived by repeatedly drawing random samples from each per-bin Beta-Binomial model and getting a bin weighted $P(b)$ sum of all of them. Specifically, for each repetition r, one sample $\theta_b^{(r)}$ was drawn from each Beta distribution $\text{Beta}(k_{b,\oplus} + c, k_{b,\ominus} + c)$ and combined as

$$\tilde{P}^{(r)}(\oplus) \;=\; \sum_{b=1}^{B} \theta_b^{(r)}\, P(b),$$

where $r = 1, \ldots, R$ and $R = 1000$ in this study. The collection $\{\tilde{P}^{(r)}(\oplus)\}_{r=1}^{R}$ formed a bootstrap distribution whose standard deviation reflected the model-based uncertainty (uncertainty per given set of validated audio examples).

Empirical variance

By contrast, the empirical variance represented the uncertainty that arises from the stochastic design of the validation process itself. To estimate it, the entire pipeline comprising validation sampling, and the 1000-fold bootstrap estimation was repeated across multiple independent simulation runs. In each run, a new subset of annotated examples was drawn within each bin, and the mean of the beta-binomial distribution was recorded as the study-level call density estimate. The variability (standard deviation) of these mean estimates across all runs represented the empirical standard deviation, capturing design-level uncertainty due to differences in the validation samples.

Variance reporting

Therefore two complementary measures of uncertainty were reported. The the bootstrap standard deviation of $\{\tilde{P}^{(r)}(\oplus)\}_{r=1}^{R}$, representing within-run (model-based) uncertainty derived from the Beta posteriors, and the empirical standard deviation of study-level call density means across repeated runs, representing across-run (design-based) uncertainty.

Strategy-level estimates inherited the Beta–Binomial model parameters established at the study level and therefore did not require a simulation stage in which different validation sets were used. Rather than re-estimating the model, these strategy-level analyses extend the study-level calibration to strata-specific datasets, allowing the evaluation of how well the threshold-free estimation framework generalizes under varying acoustic conditions and recording parameters. In particular, the strategy-level estimation procedure adapted the study-level Beta–Binomial framework to estimated call density in the presence of shifts in the distribution of classifier scores between the 32 kHz (strata 1) and 48 kHz (strata 2)

strata, thereby providing more localized and context-sensitive estimates of call density under distributional shift.

Site-Level (Strata-Level) Estimation under Distribution Shift

In practice, $P_s(z)$, the distribution of classifier scores at the site or stratum level, often differs from the overall study-level distribution $P(z)$ due to factors such as ambient noise, local species composition, and recording variability. This mismatch, known as a *distribution shift*, can cause bias if a global detection threshold is applied uniformly across all sites. In this study, the 32 kHz and 48 kHz sampling rates served as a practical example of how $P_s(z)$ may differ between strata. Accordingly, site or stratum level density estimates were derived through three complementary strategies (Strategy 1, Strategy 2, and Strategy 3) designed to robustly estimate call density in the face of potential distributional shifts.

Strategy 1: Fixed Study-Level Calibration

Strategy 1 assumes that the mapping of positively annotated z to the probability that the 3 second audio window truly contains the target species is the same at both the site and study levels. Formally, this implied that:

$$P_s(\oplus \mid z) \;=\; P(\oplus \mid z),$$

so that the site-level model leveraged the study-level calibration without further adjustment. Under this assumption, similar to the study level, the expansion of $P(\oplus)$ across site-level call density followed from the law of total probability as follows:

$$P_s(\oplus) \;=\; \int_z P_s(\oplus \mid z)\, P_s(z)\, dz \;\approx\; \int_z P(\oplus \mid z)\, P_s(z)\, dz.$$

In accordance with the discrete nature of the binning scheme used in this study, as described

in the subsubsection *Study-Level Call Density Estimation*, the integral expanded over the B bins as:

$$P_s(\oplus) \approx \sum_{b=1}^{B} P(\oplus \mid b)\, P_s(b).$$

Here, $P(\oplus \mid b)$ represented the probability of a true detection conditional on bin b, estimated from the study-level calibration, while $P_s(b)$ is the weight proportion of site-level audio examples whose classifier scores fall within bin b set in accordance with the bin boundaries found at the study level. The resulting estimator was therefore a mixture of the study-level Beta means weighted by the site-level bin weights and is formally expressed as:

$$P_s(\oplus) \approx \sum_{b=1}^{B} \left[\frac{k_{b,\oplus} + c}{k_{b,\oplus} + k_{b,\ominus} + 2c}\right] P_s(b),$$

where $k_{b,\oplus}$ and $k_{b,\ominus}$ denote the numbers of positively and negatively validated samples per bin at the study level, and c is the same small prior used to stabilize sparse bins.

Intuitively, this strategy reused the global mapping between *BirdNET* confidence scores and detection probability while re-weighting it according to the local (site-specific) bin weights $P_s(b)$. It was computationally straightforward and performed well when environmental and acoustic conditions are similar between sites and the global dataset.

Uncertainty estimation

Uncertainty around the site-level density estimate in Strategy 1 was quantified through a bootstrap resampling procedure analogous to that used for the study-level estimator, adapted to reflect the strata-specific mixture formulation. The goal of this procedure was to approximate the sampling distribution of the estimator $P_s(\oplus)$ by repeatedly drawing random

realizations of the per-bin Beta-Binomial models and aggregating them according to the site-level bin weights $P_s(b)$.

For each bootstrap iteration $r = 1, \ldots, R$ (with $R = 1000$ replicates in this study), one random value $\theta_b^{(r)}$ was drawn from each per-bin Beta distribution.

$$\theta_b^{(r)} \sim \text{Beta}(k_{b,\oplus} + c,\ k_{b,\ominus} + c),$$

and combined as a weighted sum to yield a single bootstrap realization of the site-level call density:

$$\tilde{P}_s^{(r)}(\oplus) = \sum_{b=1}^{B} \theta_b^{(r)}\, P_s(b).$$

The resulting ensemble of bootstrap replicates $\{\tilde{P}_s^{(r)}(\oplus)\}_{r=1}^{R}$ formed a bootstrap sampling distribution whose mean represented a smoothed point estimate of the site-level density, while its standard deviation reflected model-based uncertainty attributable to finite validation counts within bins.

The bootstrap-based standard error was represented as:

$$\widehat{\text{SE}}\big(P_s(\oplus)\big) = \text{SD}\Big(\{\tilde{P}_s^{(r)}(\oplus)\}_{r=1}^{R}\Big),$$

This procedure provided a non-parametric, data-driven measure of estimator variability that captured the propagation of bin-level uncertainty into the overall site-level call density estimate.

Bias under data sparsity

However, when bins contain few or no positively validated samples ($k_{b,\oplus} = 0$), the pseudo-

count c provided only minimal regularization, resulting in a downward bias. In such cases,

$$P(\oplus \mid b) \;=\; \frac{k_{b,\oplus} + c}{k_{b,\oplus} + k_{b,\ominus} + 2c} \;\approx\; \frac{c}{k_{b,\ominus} + 2c}, \qquad \text{(eq.15)}$$

so as $k_{b,\ominus}$ increases and $k_{b,\oplus}$ decreases, the per-bin estimate rapidly approaches zero. This behaviour caused $P_s(\oplus)$ to underestimate true activity when certain bins lacked positively annotated 3 second audio examples. An effect that compounds across bins in sparse or low-prevalence settings.

This strategy is analogous to applying a fixed-threshold binary classifier to a new subset of data. It assumes the same true-positive rate applies across all strata, even when the underlying score distributions differ. Consequently, while Strategy 1 was efficient and unbiased under stable conditions, it became conservative when the distribution shift between $P(z)$ and $P_s(z)$ were substantial. This motivates the use of Strategies 2 and 3, which aimed to more accurately estimate call-density under distributional drift.

Strategy 2: KL–mixture matching of bin shapes

Unlike Strategy 1 (which assumed site transferability of per–bin calibration, $P_s(\oplus \mid b) \approx P(\oplus \mid b)$), Strategy 2 assumes that the distribution of positively annotated z is the same at the study level and at the site level:

$$P_s(b \mid \oplus) \;\approx\; P(b \mid \oplus), \qquad P_s(b \mid \ominus) \;\approx\; P(b \mid \ominus),$$

where $P(b \mid \oplus)$ and $P(b \mid \ominus)$ learned at the study level from the validated data (*Study-Level Call Density Estimation*) Section.

With this pattern invariance assumption, the site (or stratum) histogram over bins, $P_s(b)$, decomposed as:

$$P_s(b) \;\approx\; P(b \mid \oplus)\, P_s(\oplus) \;+\; P(b \mid \ominus)\, (1 - P_s(\oplus)).$$

Let q be a mixture weight that lies in the interval $[0, 1]$.

$$Q_q(b) \;=\; q\, P(b \mid \oplus) \;+\; (1 - q)\, P(b \mid \ominus).$$

In this case, q can be made to vary between a weight allocation of between 100% or 0% to $P(b|\oplus)$ and $P(b|\ominus)$. The site density is then chosen as the mixture weight (q) whose study mixture Q_q is closest (in KL divergence[7]) to the observed site distribution $P_s(b)$:

$$P_s(\oplus) \;\equiv\; \hat{q} \;=\; \arg\min_{q \in [0,1]} \mathrm{KL}\big(P_s \,\big\|\, Q_q\big) \;=\; \arg\min_{q \in [0,1]} \sum_b P_s(b)\, \log \frac{P_s(b)}{Q_q(b)}$$

Quantities used were $P_s(b)$, which was the proportion of site-level audio examples whose classifier scores fell within bin b_i for $i \in \{1, 2, ..., B\}$, set in accordance with the bin boundaries found at the study level. $P(b \mid \oplus)$ and $P(b \mid \ominus)$ were also used and were the *study–level* conditional distributions of positive and negative annotations . From the study level validation counts $k_{b,\oplus}, k_{b,\ominus}$, the following smoothed estimates were used.

$$P(b \mid \oplus) \;\approx\; \frac{k_{b,\oplus} + c}{\sum_u (k_{u,\oplus} + c)}, \qquad P(b \mid \ominus) \;\approx\; \frac{k_{b,\ominus} + c}{\sum_u (k_{u,\ominus} + c)},$$

with the same small $c = 0.1$ pseudo count that was used in the study-level Beta model.

Interpretation

If the site is mostly positively annotated, (its bin weights, $P_s(b)$ place more mass where $P(b \mid \oplus)$ is large), the minimizer $\hat{q}$ moves toward 1. If it is mostly negatively annotated, $\hat{q}$

[7]Kullback-Leibler divergence is a measure of the difference between two probability distributions (Nawa and Nadarajah, 2024).

moves toward 0. This avoids the downwards bias that Strategy 1 can suffer when some bins have $k_{b,\oplus} = 0$, because strategy 2 fits the distribution of positively annotated z according to bin weights $P_s(b)$ rather than relying solely on per bin positive rates at the site.

Uncertainty estimation

Uncertainty around the site-level density estimates derived through the KL-divergence framework was quantified using a non-parametric bootstrap resampling procedure. Rather than relying on an analytical expression for the estimator's variance, this approach empirically approximated the sampling distribution of the site-level density estimates through repeated resampling and re-estimation.

For each bootstrap replicate $r = 1, \ldots, R$ (with $R = 1000$ in this study), the following steps were performed:

1. A bootstrap sample of the site-level binned logit scores was drawn with replacement from the original strata dataset, producing a pseudo-site distribution $P_s^{(r)}(b)$.

2. Using this resampled distribution, a mixture distribution was constructed as

$$Q_q^{(r)}(b) = q\, P(b \mid \oplus) + (1 - q)\, P(b \mid \ominus),$$

   where $P(b \mid \oplus)$ and $P(b \mid \ominus)$ denote the study-level reference distributions for positive and negative detections, respectively.

3. The divergence between $P_s^{(r)}(b)$ and $Q_q^{(r)}(b)$ was then computed using the Kullback–Leibler measure:

$$\mathrm{KL}\big(P_s^{(r)}(b) \,\|\, Q_q^{(r)}(b)\big) = \sum_b P_s^{(r)}(b)\, \log \frac{P_s^{(r)}(b)}{Q_q^{(r)}(b)}.$$

The minimizing mixture weight was subsequently identified as

$$P_s^{(r)}(\oplus) = \arg\min_{q\in(0,1)} \mathrm{KL}\big(P_s^{(r)}(b) \,\|\, Q_q^{(r)}(b)\big),$$

with the q that minimized the mixture, yielding a bootstrap estimate of the site-level call density estimate.

The ensemble of bootstrap replicates, $\{P_s^{(r)}(\oplus)\}_{r=1}^{R}$, formed a measure upon which the bootstrap standard deviation was be derived as follows:

$$\widehat{\mathrm{SE}}\big(P_s(\oplus)\big) = \mathrm{SD}\big(\{P_s^{(r)}(\oplus)\}_{r=1}^{R}\big)\,.$$

This procedure provided a data-driven measure of estimator uncertainty that captured both sampling variability within strata and the stochastic behaviour of the KL-divergence optimization process.

Strategy 3: Ensemble Synthesis of Study- and Site-Level Calibrations

Strategy 3 combined the outputs of Strategies 1 and 2 into a unified ensemble estimator of site-level call density. As proposed by Navine et al. (2024), the two preceding strategies employ contrasting assumptions. Strategy 1 transfers the conditional probability $P(\oplus \mid z)$ from the study-level calibration to each site, generally yielding a conservative (downward-biased) estimate under distributional shift. Strategy 2, by contrast, assumes stability in the positive score distribution $P(b \mid \oplus)$ across sites, often producing inflated estimates when the site's empirical distribution resembles the study-level positive reference. To mitigate these opposing biases, the two estimates were combined multiplicatively as a geometric mean as follows:

$$P_s^{(3)}(\oplus) \;=\; \sqrt{P_s^{(1)}(\oplus)\; P_s^{(2)}(\oplus)}\,.$$

Where $P_s^{(1)}(\oplus)$ is the call density estimate from strategy 1 and $P_s^{(2)}(\oplus)$ is the call density estimate from strategy 2.

This construction empirically balanced the tendencies of the first two strategies, stabilizing predictions across heterogeneous recording environments. It preserved the property that if either Strategy 1 or Strategy 2 yields a null density for an unoccupied site ($P_s(\oplus) = 0$), the ensemble also collapses to zero, preventing false-positive detection of occupancy. Conversely, when both strategies indicate occupancy, the geometric mean provides a moderated estimate between the conservative and optimistic extremes.

No additional re-sampling or bootstrapping is required for this strategy, as it relied directly on the point estimates $P_s^{(1)}(\oplus)$ and $P_s^{(2)}(\oplus)$ obtained under their respective assumptions. The ensemble thus inherits the uncertainty structure of the underlying methods, serving as a bias–variance compromise under heterogeneous or shifting acoustic conditions, as demonstrated in the Hawaiian case studies of Navine et al. (2024).

Results

This section presents the results obtained from the analysis of the data introduced in the *Data* section and analysed using the statistical framework described in the *Methodology*. The results address the three central questions stated in the introduction. Which were assessing whether call density can be reliably estimated without reliance on arbitrary thresholds, whether alternative estimation strategies can accommodate distributional shifts across strata while still producing robust and stable call-density estimates and lastly, whether the resulting estimates align with ecological expectations regarding species distribution and activity. In doing so, the results summarise the behaviour of the threshold-free call-density estimator across species, sampling-rate strata (32 kHz and 48 kHz), binning schemes (3, 4, 5, and 6 bins), and estimation strategies (Study Level and Strategies 1, 2 and 3), providing a comprehensive assessment of how classifier score distributions, bin structures, and distributional shifts influence the accuracy, robustness, and ecological validity of the estimator.

For each species, call-density performance across all binning schemes and both sampling-rate strata is presented. Central to the evaluation are the relative bias plots, which compare estimated call density to ground-truth density across all strategies and binning configurations. Relative bias formally defined as;

$$\text{Relative Bias} = \frac{\widehat{D} - D}{D},$$

provided a scale-free measure of estimator accuracy by quantifying the proportional deviation of the estimated call density from the true value. Positive values indicate overestimation and negative values indicate underestimation, while values close to zero reflect accurate estimation. In addition to the direction of error, the magnitude of relative bias conveys the severity of misestimation. A relative bias value of 0.10 represents a 10% overestimate, whereas a value of 1.00 implies that the call density estimation model used overestimated the true call density by 100%. The inverse analogy goes for negative relative bias values. With the rela-

tive bias metric expressed as the difference between the estimate and the true call density, scaled relative to the true call density, it becomes particularly informative for species with low call densities. In such cases, even small absolute errors can lead to large proportional deviations due to the low underlying density. Examining patterns in relative bias across binning levels, strata, sample sizes, and estimation strategies therefore allows assessment not only of systematic tendencies toward overestimation or underestimation but also of the stability and consistency of estimator performance across different modelling configurations.

Results presentation

Across all analyses, call-density estimates and their corresponding ground-truth values were examined under each binning scheme (3, 4, 5, and 6 bins) at the study level and for Strategies 1, 2, and 3 for each species. For every strategy, three key quantities were reported. The estimated call density , The ground-truth density derived from manual annotation and finally the relative bias computed with respect to the ground-truth estimate

This structure enabled direct comparison between estimated and true call-density values and facilitated evaluation of how calibration choices and strata-specific adjustments influenced estimator performance. To assess the effect of validation effort, bias plot results were generated for all three validation-sample sizes considered in this study, 25, 50, and 75 validated samples per bin. Note that the results were jittered to aid visualization as results may overlap. Providing a complete overview of how estimation accuracy changed as additional annotated data became available. The bootstrap standard deviation associated with the strategies 1, 2 and 3 and the study-level call density estimation, and the empirical standard deviation associated with the study level call density estimation are presented in the *Appendix*

Together, these outputs offered a coherent and systematic view of estimator behaviour across species, sampling rates, binning schemes, and validation-sample sizes. They revealed how patterns in classifier score distributions propagated into call-density estimates and highlighted the conditions under which the estimator remained stable versus those in which it became more sensitive to distributional differences or limited validation data. In the subsections that follow, these dynamics were examined individually for each species, with specific attention to the role of acoustic characteristics, detectability profiles, and empirical score distributions in shaping estimation performance.

Baglafecht weaver

25 validated samples per bin:

The results presented below are of the Baglafecht weaver at 25 validated samples per bin.

Table 6: Estimated call densities across binning levels and sampling rates for Baglafecht weaver (25 validated samples per bin).

Level	3 bins		4 bins		5 bins		6 bins	
	32 kHz	48 kHz	32 kHz	48 kHz	32 kHz	48 kHz	32 kHz	48 kHz
Study Level	0.028	0.028	0.029	0.029	0.029	0.029	0.028	0.028
Strategy 1	0.033	0.034	0.031	0.029	0.031	0.031	0.033	0.034
Strategy 2	0.033	0.033	0.030	0.029	0.030	0.030	0.033	0.034
Strategy 3	0.033	0.034	0.031	0.029	0.031	0.031	0.033	0.034

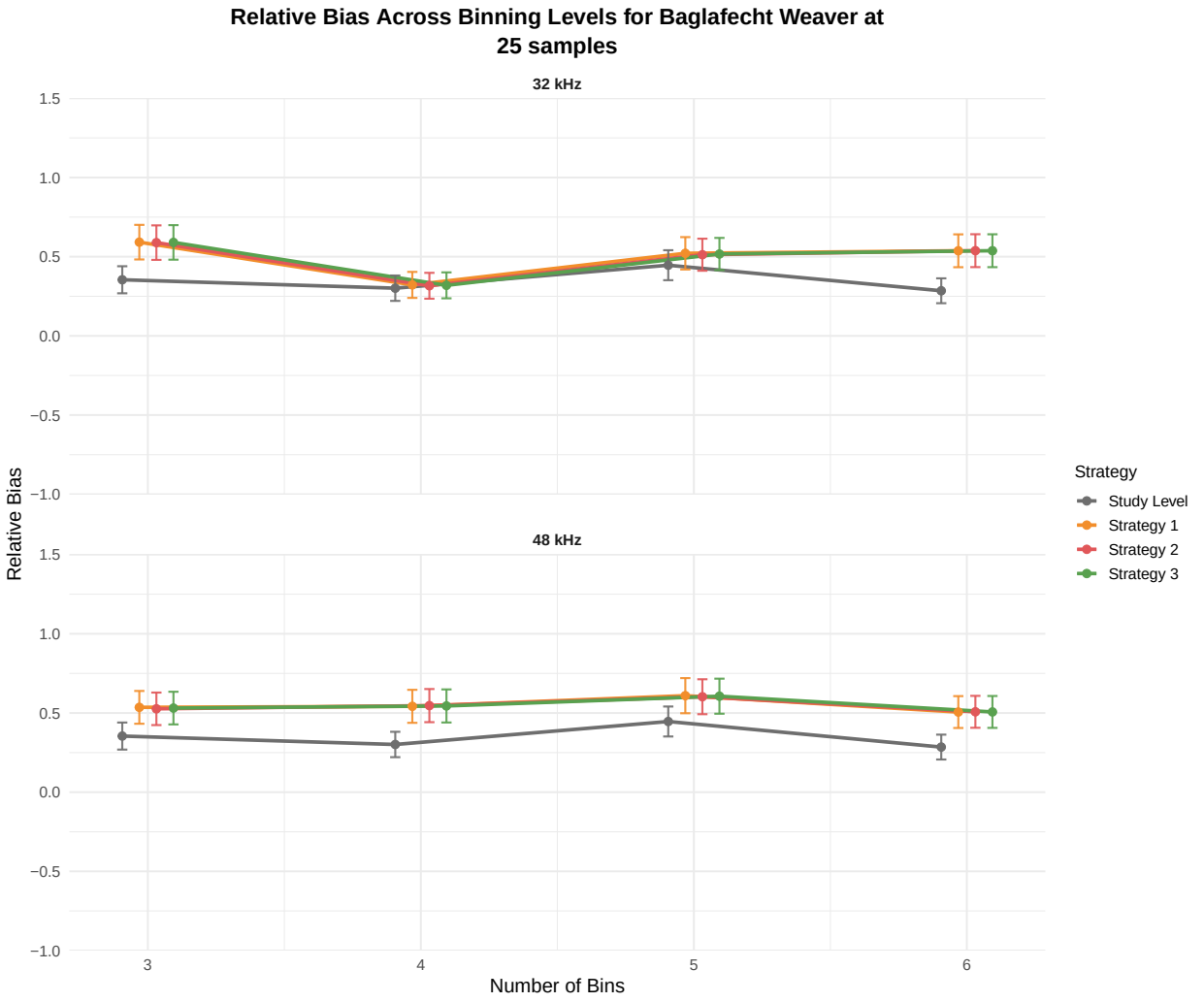

Figure 11: Bias plot of Baglafetch weaver at 25 validated samples per bin

Across all binning levels and sampling rates, the estimated call densities for the Baglafecht weaver remained consistently low, with study-level values ranging from approximately 0.028 to 0.029 at 25 validated samples per bin. Strategy-level estimates fell within a similar range, typically between 0.030 and 0.034, and followed the same pattern across both the 32 kHz and 48 kHz strata. Estimates also remained stable across the 3-, 4-, 5-, and 6-bin configurations, with no substantial differences attributable to binning scheme or estimation strategy. Call density estimates varied only modestly across configurations while remaining within a narrow numerical range.

Across all binning levels, estimation strategies, and sampling-rate strata, the Baglafecht weaver displayed uniformly positive relative bias. Relative bias values ranged between 0.25 and 0.65, indicating that estimated call density was consistently higher than the ground-truth density derived from manual annotation.

Differences between sampling-rate strata were evident. The 32 kHz recordings showed higher relative bias and greater variability across binning levels than the 48 kHz recordings. Standard deviations ranged from 0.11 to 0.12 for the 32 kHz stratum and from 0.04 to 0.07 for the 48 kHz stratum, indicating more stable estimates at the higher sampling rate. For comparison, the combined empirical standard deviation at the study level was substantially smaller (0.012; Table 22), reflecting the stabilising effect of pooling all annotated data prior to calibration.

Across strategies, relative bias values were broadly similar, with no estimation method showing consistently superior or inferior performance. All strategies exhibited comparable levels of bias and variability across binning levels and sampling configurations.

The Baglafecht weaver tended to be systematically higher than ground-truth densities and that estimation performance differed more strongly by audio sampling rate than by binning scheme or strategy. No meaningful differences were observed between the study-level and strata-specific estimators, which produced similar patterns of bias and variability across all configurations.

50 validated samples per bin:

The results presented below are of the Baglafecht weaver at 50 validated samples per bin.

Table 7: Estimated call densities across binning levels and sampling rates for Baglafecht weaver (50 validated samples per bin).

Level	3 bins		4 bins		5 bins		6 bins	
	32 kHz	48 kHz	32 kHz	48 kHz	32 kHz	48 kHz	32 kHz	48 kHz
Study Level	0.025	0.025	0.025	0.025	0.025	0.025	0.024	0.024
Strategy 1	0.025	0.025	0.030	0.030	0.020	0.020	0.044	0.043
Strategy 2	0.024	0.025	0.030	0.030	0.020	0.020	0.044	0.042
Strategy 3	0.024	0.025	0.030	0.030	0.020	0.020	0.044	0.043

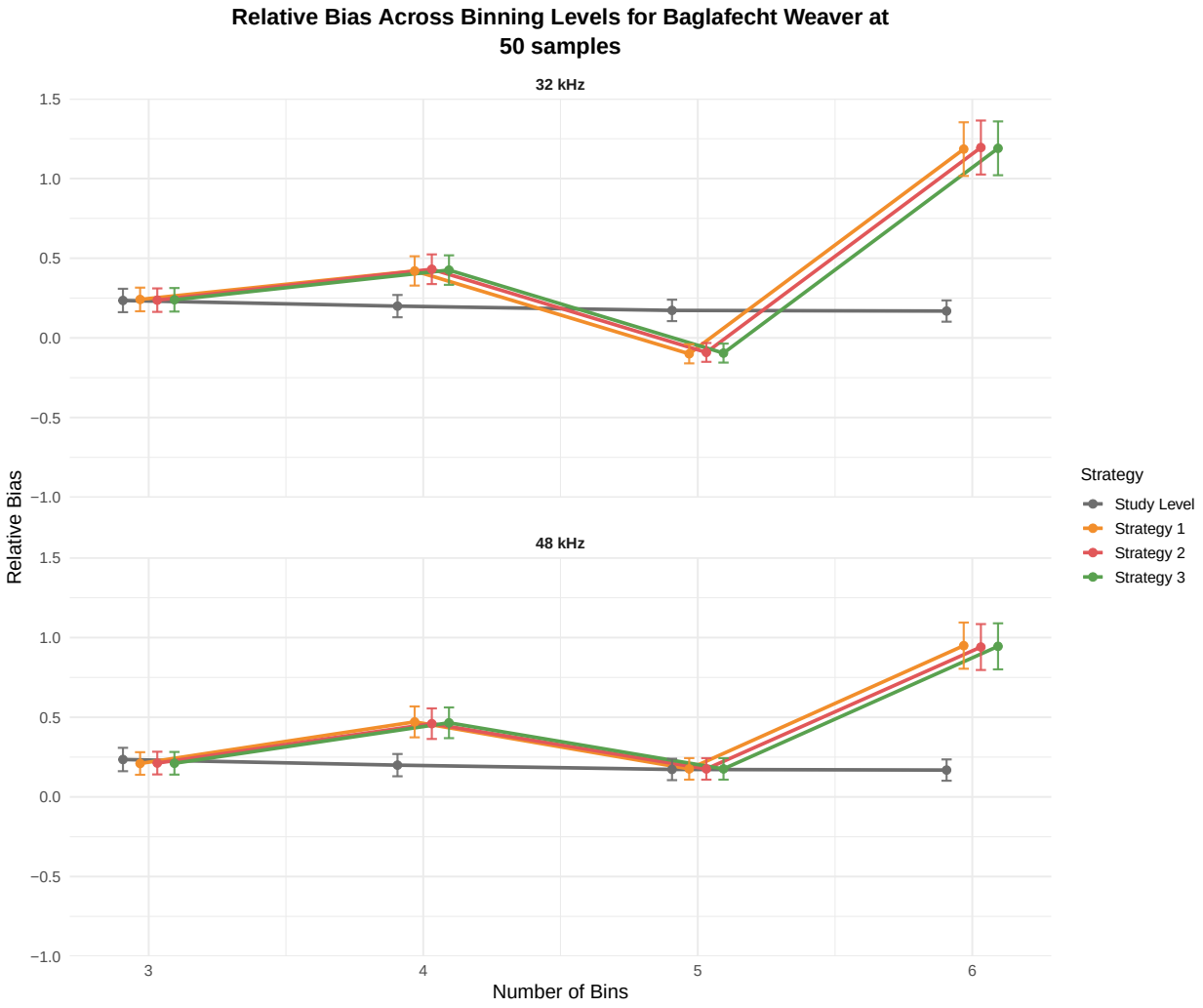

Figure 12: Bias plot of Baglafetch weaver at 50 samples per bin

The call density estimates for the Baglafecht weaver at 50 validated samples per bin remained uniformly low across all binning levels and sampling rates (Table 7). Study-level estimates ranged narrowly from approximately 0.024 to 0.025, while strategy-level estimates spanned a slightly wider range of roughly 0.020 to 0.044, depending on binning scheme and sampling-rate stratum. Estimates in both the 32 kHz and 48 kHz strata followed similar patterns, and no major shifts in the magnitude of call density were observed across the 3, 4, 5 and 6-bin configurations. The estimator produced consistent outputs across modelling configurations, with density values remaining within a compressed numerical range.

Relative bias values for this species at 50 samples per bin remained consistently positive across all strategies, binning levels, and sampling-rate strata (Figure 12). Estimated call density exceeded the ground-truth density in every configuration. Relative bias generally ranged from approximately 0.20 to 1.10, indicating overestimation from 20% up to more than twice the true density. Bias tended to increase between the three and four bin configurations, decline slightly at five bins, and reach its highest values under the six-bin scheme at just below 100% overestimation for the 48 kHz stratum and just above 100% for the 32 kHz stratum.

Differences between sampling-rate strata were again evident. The 32 kHz recordings consistently produced larger and more variable relative bias estimates, whereas the 48 kHz stratum yielded lower and more tightly clustered values. This pattern mirrors the call-density outputs, where estimates in the 48 kHz stratum were slightly lower and more stable than those derived from the 32 kHz data. These differences reflect the greater amount of validated data available at 32 kHz, which produced higher detection probabilities but also amplified estimation error due to distributional sparsity in the high-score region.

Although all estimation strategies produced broadly similar patterns of bias, the study-level estimator displayed the lowest variability across both binning levels and sampling rates. As shown in Table 23, the study-level standard deviations were approximately 0.031 for both strata, whereas Strategies 1 through 3 exhibited substantially higher variability, particularly in the 32 kHz stratum, where SD values exceeded 0.54. This contrast highlights the stabilising effect of pooling data at the study level, where bins are better populated and the Beta–Binomial model is more robustly informed. The empirical combined standard deviation at the study level, was markedly smaller (0.010; Table **??**) than any of the strategy-level standard deviations or bootstrap study level SDs. The empirical SD quantifies the variability

across repeated simulated calibrations.

Taken together, the 50-sample results show that estimated call densities for the Baglafecht weaver remained systematically higher than the ground-truth densities across all modelling configurations. Relative bias remained strongly positive throughout, driven more by sampling-rate differences than by binning scheme or estimation strategy. While all strategies produced similar bias patterns, the study-level estimator consistently produced the most stable and tightly clustered estimates. The low empirical standard deviation further demonstrates the reliability of the study-level calibration relative to strata-level approaches, which showed inflated variability due to data sparsity and distributional mismatch.

75 validated samples per bin:

The results presented below are of the Baglafecht weaver at 75 validated samples per bin.

Table 8: Estimated call densities across binning levels and sampling rates for Baglafecht weaver (75 validated samples per bin).

Level	3 bins		4 bins		5 bins		6 bins	
	32 kHz	48 kHz	32 kHz	48 kHz	32 kHz	48 kHz	32 kHz	48 kHz
Study Level	0.023	0.023	0.024	0.024	0.023	0.023	0.024	0.024
Strategy 1	0.023	0.025	0.016	0.015	0.016	0.016	0.012	0.011
Strategy 2	0.023	0.025	0.016	0.015	0.016	0.016	0.011	0.011
Strategy 3	0.023	0.025	0.016	0.015	0.016	0.016	0.012	0.011

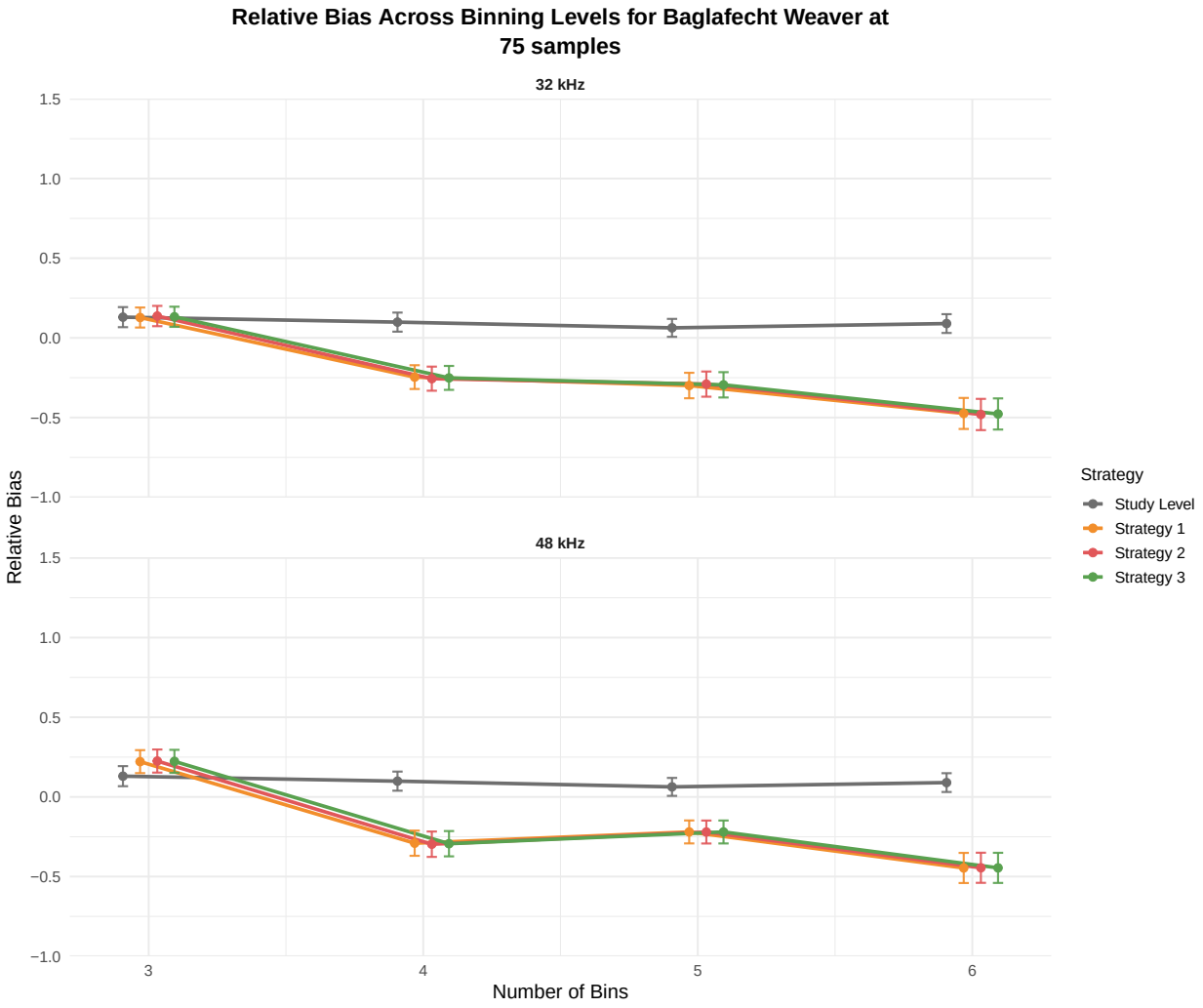

Figure 13: Bias plot of Baglafetch weaver at 75 samples per bin

For the 75 sample configuration, estimated call densities for the Baglafecht weaver remained uniformly low across all binning schemes and sampling rates. Study-level estimates ranged from 0.023 to 0.024, while strategy-level values occupied a slightly wider range (0.011 to 0.025), although these differences were numerically small. Estimates in the 32 kHz and 48 kHz strata followed the same overall pattern, and no major shifts in density magnitude were observed across the 3, 4, 5, and 6 bin configurations.

Relative bias values at this validation depth were concentrated near zero across all strategies,

strata, and binning levels, indicating that estimation error was substantially reduced at this larger sample size. Mild positive bias appeared under the three bin scheme in the 32 kHz stratum, while small negative values emerged for the four, five, and six bin configurations. The same trend occurred in the 48 kHz recordings, where relative bias remained close to zero at three bins and shifted slightly negative as bin counts increased.

Variability in relative bias also declined relative to the lower sample size conditions. The study-level estimator again showed the lowest variability, with standard deviations of approximately 0.028 across strata, while Strategies 1 through 3 exhibited slightly higher but still comparable dispersion, with standard deviations between 0.25 and 0.29. Variability was similar between the 32 kHz and 48 kHz strata, indicating consistent performance across sampling rates at this higher annotation effort.

Overall, the results for 75 validated samples per bin showed that call density estimates converged toward the ground truth values, with relative bias centred close to zero and reduced variability across strategies, binning levels, and sampling rate conditions. Although the study-level estimator remained the most stable, all estimation strategies produced near identical results at this sample depth, and no meaningful performance differences were observable between them.

White-browed Coucal

25 validated samples per bin:

The results presented below are of the White-browed coucalat 25 validated samples per bin.

Table 9: Estimated call densities across binning levels and sampling rates for White-browed coucal(25 validated samples per bin).

Level	3 bins		4 bins		5 bins		6 bins	
	32 kHz	48 kHz	32 kHz	48 kHz	32 kHz	48 kHz	32 kHz	48 kHz
Study Level	0.041	0.041	0.041	0.041	0.040	0.040	0.041	0.041
Strategy 1	0.039	0.041	0.030	0.029	0.031	0.030	0.076	0.077
Strategy 2	0.038	0.041	0.030	0.029	0.031	0.030	0.076	0.076
Strategy 3	0.039	0.041	0.030	0.029	0.031	0.030	0.076	0.076

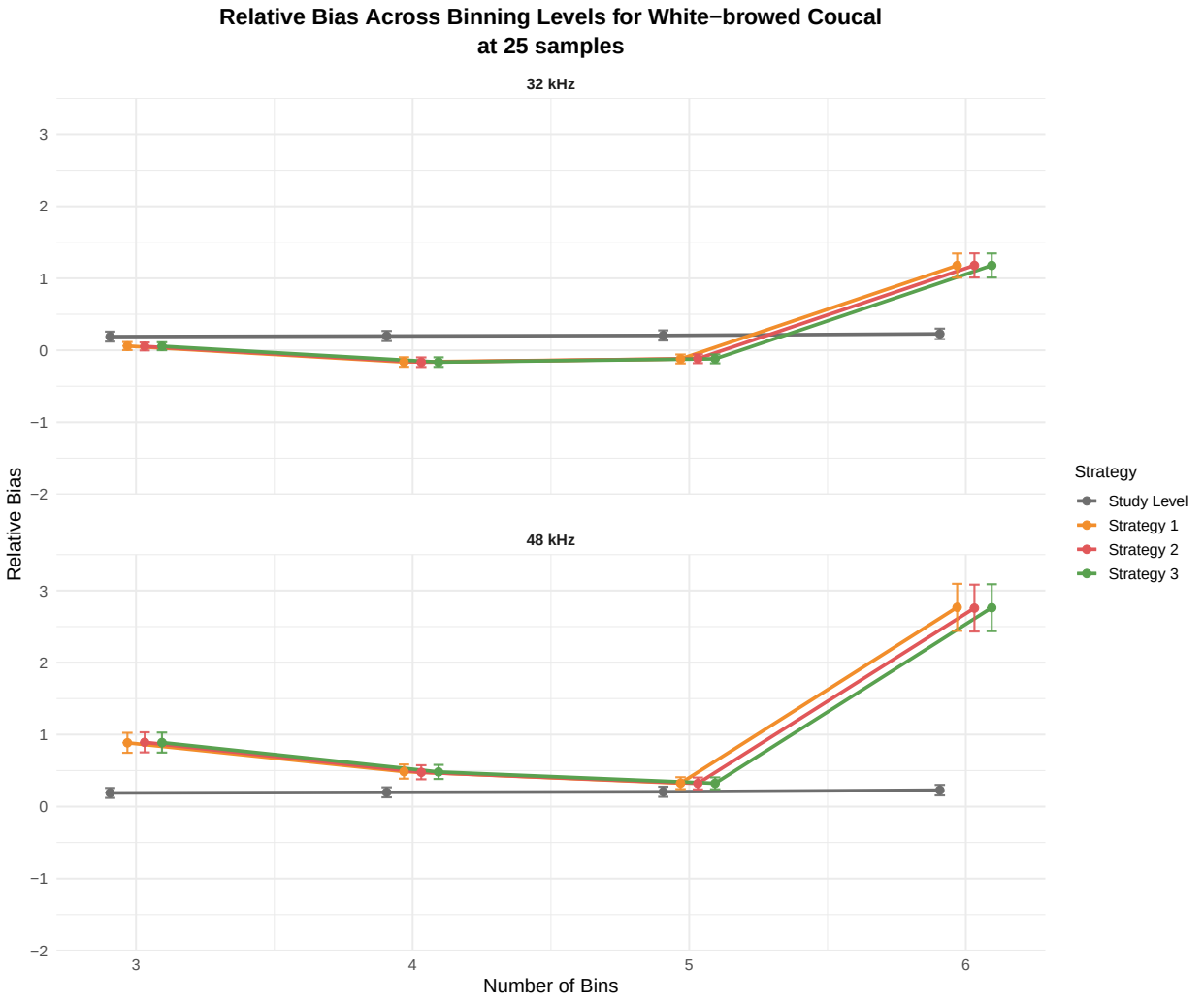

Figure 14: Bias plot of White-browed coucalat 25 samples per bin

Across all binning levels and sampling rates, the estimated call densities for the White-browed coucalat 25 validated samples per bin remained low and numerically consistent. Study-level estimates ranged narrowly between 0.040 and 0.041 across all binning schemes and both sampling-rate strata, while strategy-specific estimates occupied a slightly wider range, reaching approximately 0.076 to 0.077 under the six-bin configuration. Although these differences were somewhat larger at higher bin counts, all estimates remained within a relatively constrained numerical range, indicating broadly similar behaviour across estimation strategies.

Relative bias results followed the same general pattern. Bias values were consistently *positive* across all binning schemes and strategies, meaning that estimated call density exceeded the manually derived ground-truth density in every configuration. However, the magnitude of this overestimation depended strongly on sampling-rate stratum and binning level.

In the 32 kHz recordings, relative bias values remained close to zero for the three, four, and five bin schemes across all strategies, indicating only slight underestimation at lower bin counts. Larger positive bias values appeared under the six-bin scheme for all strategies, ranging from roughly 1.0 to 1.2. The Study-level estimator showed only a modest increase under this configuration and remained near zero across all binning levels, demonstrating greater numerical stability.

In the 48 kHz recordings, a similar overall pattern was observed, but with a substantially significant magnitude of overestimation. Bias values for all strategies approached 1.0 under the three-bin scheme (Figure 14), decreased slightly at four and five bins, and then rose sharply under the six-bin segmentation, reaching values of approximately 200% to 260% overestimations. In contrast, the Study-level estimator again showed only a small upward shift and remained close to zero across all configurations.

The variability results reinforce this behaviour. As shown in Table , variability for the Study-level estimator was extremely low (SD $\approx$ 0.016 for both the 32 kHz and 48 kHz strata), whereas the strategy-based estimators exhibited far greater dispersion, particularly in the 48 kHz recordings where SD values exceeded 1.12. This contrast aligns closely with the empirical combined standard deviation of 0.015 reported in Table 28, which reflects the run-to-run variability of the estimator under repeated calibration. The near-identical values

between the empirical SD and the Study-level SDs indicate that the Study-level estimator is highly stable under finite sampling. In contrast, the substantially higher bootstrap SDs for Strategies 1–3 reflect the sensitivity of strata-specific calibration to sparse validated positives and distributional shifts, particularly in the higher sampling-rate stratum.

In summary, the White-browed coucalresults at 25 validated samples per bin showed that all estimation strategies produced low call-density estimates with consistently positive relative bias. Most estimates remained close to the ground-truth density for three- to five-bin schemes, while marked overestimation appeared under the six-bin segmentation for all strategies. The Study-level estimator displayed the lowest variability and the greatest numerical stability across all modelling configurations, while the strata-specific strategies exhibited substantially greater dispersion driven by sampling-rate differences and limited annotated data at 48 kHz. Collectively, these results indicate that the Study-level method is the most reliable estimator in this setting, whereas strata-level strategies remain highly sensitive to distributional mismatch and data sparsity.

50 validated samples per bin:

The results presented below are of the White-browed coucalat 50 validated samples per bin.

Table 10: Estimated call densities across binning levels and sampling rates for White-browed coucal(50 validated samples per bin).

Level	3 bins		4 bins		5 bins		6 bins	
	32 kHz	48 kHz	32 kHz	48 kHz	32 kHz	48 kHz	32 kHz	48 kHz
Study Level	0.037	0.037	0.037	0.037	0.037	0.037	0.037	0.037
Strategy 1	0.030	0.029	0.049	0.048	0.037	0.036	0.031	0.032
Strategy 2	0.030	0.029	0.048	0.048	0.037	0.035	0.030	0.032
Strategy 3	0.030	0.029	0.048	0.048	0.037	0.036	0.030	0.032

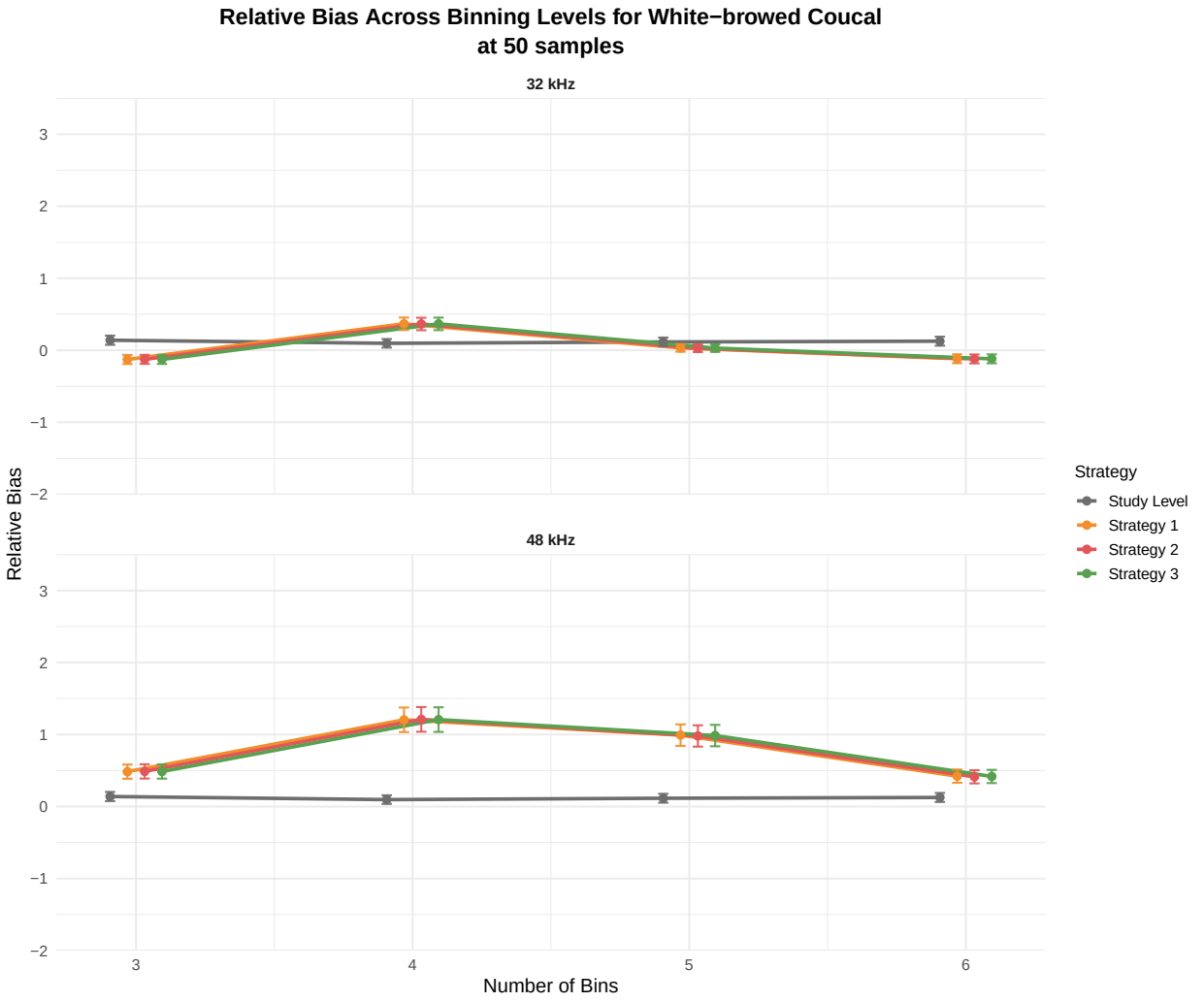

Figure 15: Bias plot of White-browed coucal at 50 samples per bin

When 50 validated samples per bin were used, the White-browed Coucal produced call-density estimates that remained low and numerically consistent across all binning schemes and sampling rates. Study-level estimates clustered tightly around 0.037 across configurations, while Strategies 1–3 generated values ranging from approximately 0.029 to 0.049 in the 32 kHz recordings and from 0.029 to 0.048 in the 48 kHz stratum. Although the strategy-based estimates exhibited slightly wider variation than the Study-level estimates, all values remained within a narrow numerical range, indicating no meaningful differences in estimator behaviour.

Relative-bias patterns followed the same structure. In the 32 kHz recordings, all estimation strategies including the Study-level estimator produced bias values oscillating only slightly between small negative and positive values across binning levels (see Figure 15). In the 48 kHz recordings, Strategies 1–3 showed large positive bias values at an overestimate of approximately 120% at 4 bins. The Study-level estimator again remained centred near zero, demonstrating numerical stability across strata.

The variability results reinforce these trends. As shown in Table 29, standard deviations for the Study-level estimator were extremely small and nearly identical across strata (SD ≈ 0.018 for 32 kHz and 48 kHz). In contrast, Strategies 1–3 exhibited substantially greater variability (SD ≈ 0.23 in the 32 kHz stratum and ≈ 0.38 in the 48 kHz stratum). Despite these numerical differences, the broader pattern of estimator behaviour remained similar across strata.

A comparison of the Study-level bootstrap SD with the empirical SD provides further insight. The Study-level combined bootstrap SD was very small (SD ≈ 0.017), indicating that internal estimation uncertainty within the Beta–Binomial model was minimal. The empirical standard deviation derived from repeated simulations (Table 24) was also low (0.011), closely matching the bootstrap result. The agreement between these two uncertainty measures shows that the Study-level estimator was highly stable both internally (bootstrap) and externally (simulation-based empirical variation). In contrast, the strategy-based estimators displayed substantially higher combined SDs (SD ≈ 0.491), suggesting that strata-specific calibration introduced considerable additional noise without improving estimator accuracy.

Overall, the White-browed Coucal results at 50 validated samples per bin demonstrate that

call-density estimates, bias values, and variability metrics remained stable across all estimation strategies and binning configurations. Differences in performance were driven primarily by sampling rate rather than estimator choice. The strong agreement between the empirical and bootstrap SDs further confirms that the Study-level estimator was the most robust and reliable approach for this species at this validation level.

75 validated samples per bin:

The results presented below are of the White-browed coucal at 75 validated samples per bin.

Table 11: Estimated call densities across binning levels and sampling rates for White-browed coucal (75 validated samples per bin).

Level	3 bins		4 bins		5 bins		6 bins	
	32 kHz	48 kHz	32 kHz	48 kHz	32 kHz	48 kHz	32 kHz	48 kHz
Study Level	0.036	0.036	0.036	0.036	0.036	0.036	0.035	0.035
Strategy 1	0.027	0.027	0.027	0.025	0.036	0.034	0.048	0.050
Strategy 2	0.027	0.027	0.026	0.024	0.036	0.033	0.048	0.050
Strategy 3	0.027	0.027	0.026	0.025	0.036	0.034	0.048	0.050

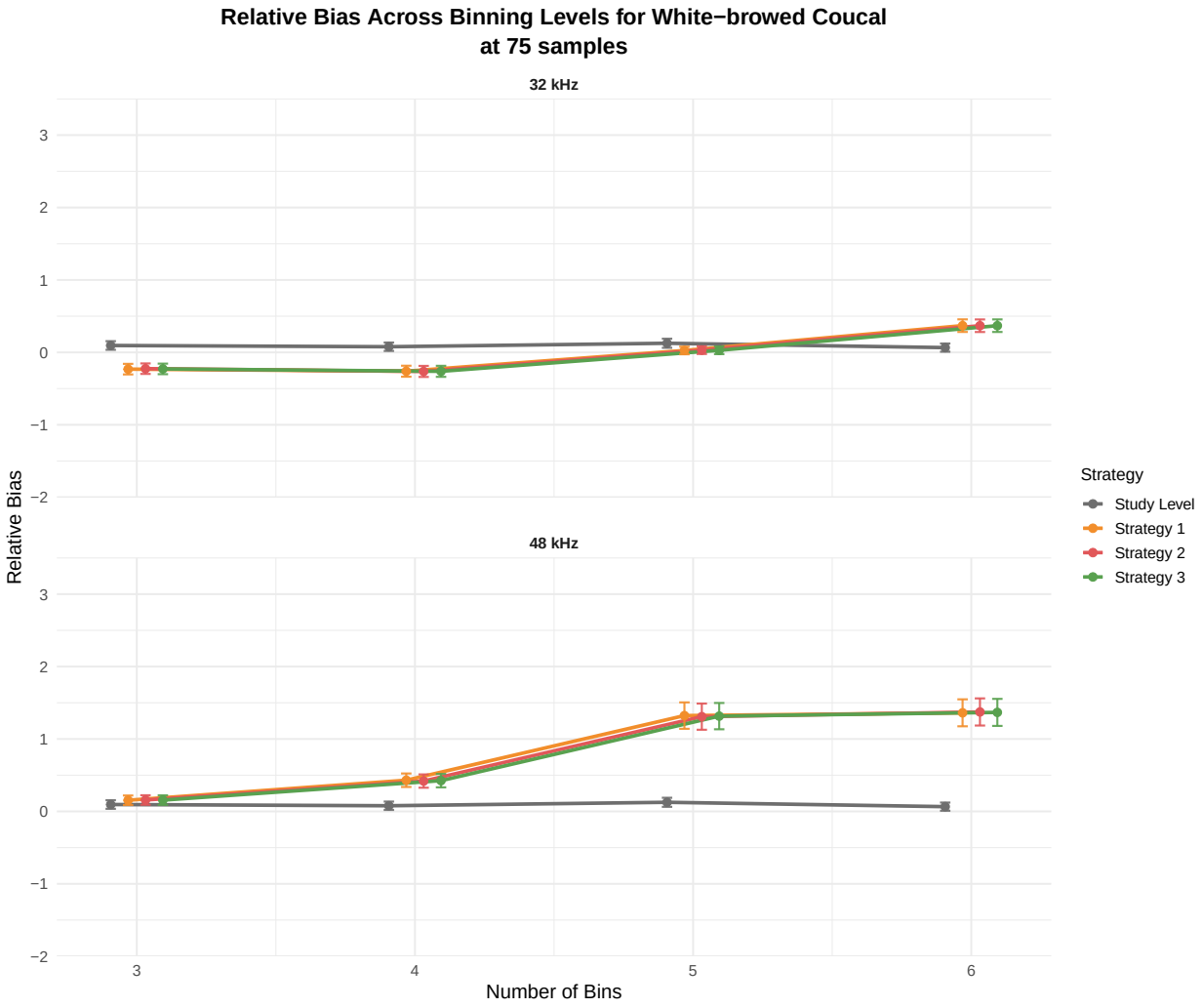

Figure 16: Bias plot of White-browed coucal at 75 samples per bin

Across all binning configurations, the White-browed Coucal showed a consistent pattern in both call-density estimates and relative bias when 75 validated samples per bin were used. Study-level call-density estimates remained tightly concentrated between 0.035 and 0.036 across all binning schemes and both sampling-rate strata, indicating a stable overall baseline. Strategy-level estimates occupied a slightly wider numerical range, spanning approximately 0.026 to 0.048 in the 32 kHz recordings and approximately 0.024 to 0.050 in the 48 kHz recordings. However, these differences were small in magnitude, and all estimates remained within a narrow overall range, indicating no meaningful performance distinctions

between the estimation strategies.

Relative-bias behaviour followed the same general structure. In the 32 kHz stratum, both the Study-level estimator and Strategies 1–3 produced bias values that hovered near zero across all binning levels, with only minor fluctuations between slightly negative and slightly positive values. In the 48 kHz recordings, bias values remained positive and increased gradually with bin count for all approaches, reaching as high as 145% overestimation in the six-bin configuration. Despite this increase, the overall pattern of behaviour remained consistent, and the Study-level estimator continued to show the smallest deviations.

As shown in Table 31, the Study-level estimator had very low dispersion (SD $\approx$ 0.026 for each stratum), while Strategy-based estimators displayed larger SDs (ranging from approximately 0.291 at 32 kHz to 0.617 at 48 kHz). The combined-strata variability followed the same pattern, with the Study-level estimator again showing the smallest SD (0.024), whereas Strategies 1–3 produced substantially larger combined SDs (0.632–0.634).

A comparison between the empirical SD and the Study-level bootstrap SD further highlights the stability of the Study-level estimator. The empirical SD derived from repeated simulations (0.009; Table 32) was even smaller than the Study-level combined bootstrap SD (0.024). This close agreement suggests that the Study-level estimator was highly stable both in terms of model-based uncertainty and empirical performance.

Taken together, the results for 75 validated samples per bin demonstrate that call-density estimates and relative-bias patterns were stable across all estimation strategies, binning schemes, and sampling rates. While the Study-level estimator continued to show the smallest numerical variability and strong agreement between empirical and bootstrap SDs the

differences between it and Strategies 1–3 remained minor in practical terms.

African Gray Flycatcher

25 validated samples per bin:

The results presented below are of the African Gray Flycatcher at 25 validated samples per bin.

Table 12: Estimated call densities across binning levels and sampling rates for African Gray Flycatcher (25 validated samples per bin).

Level	3 bins		4 bins		5 bins		6 bins	
	32 kHz	48 kHz	32 kHz	48 kHz	32 kHz	48 kHz	32 kHz	48 kHz
Study Level	0.103	0.103	0.094	0.094	0.095	0.095	0.094	0.094
Strategy 1	0.138	0.125	0.085	0.087	0.081	0.078	0.141	0.159
Strategy 2	0.138	0.125	0.085	0.087	0.081	0.078	0.141	0.159
Strategy 3	0.138	0.125	0.085	0.087	0.081	0.078	0.141	0.159

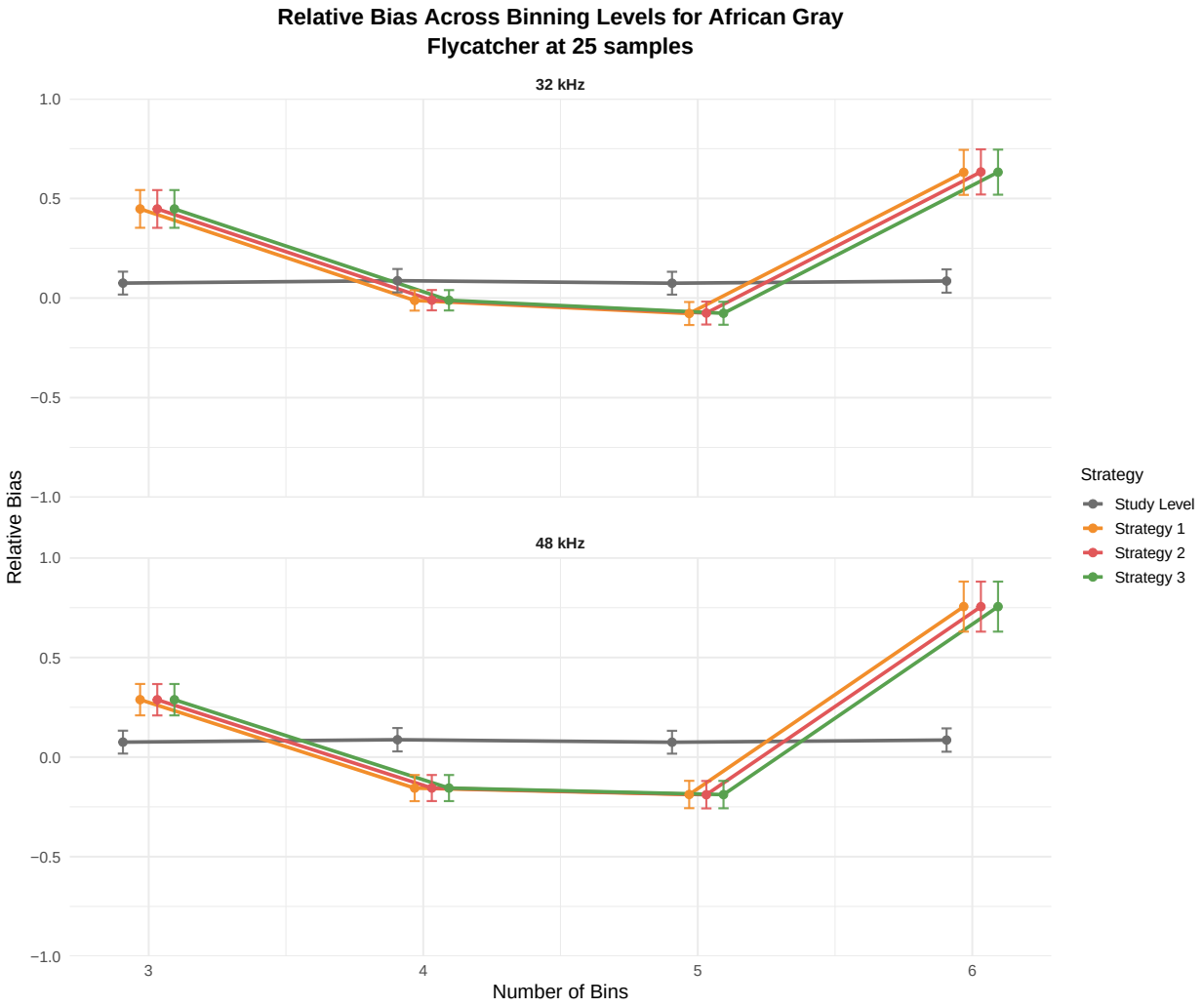

Figure 17: Bias plot of African Gray Flycatcher at 25 samples per bin

The call-density estimates for the African Grey Flycatcher at 25 validated samples per bin remained within a relatively narrow numerical range across binning levels and sampling rates. Study-level estimates ranged from 0.094 to 0.103 across both strata, while Strategies 1–3 produced estimates between approximately 0.081 and 0.141 at 32 kHz and between 0.078 and 0.159 at 48 kHz. Although the strategy-based estimators displayed a wider numerical spread than the Study-level approach, all estimates still fell within a broadly coherent range of call-density values.

Relative-bias patterns showed the same general structure. In the 32 kHz stratum, bias values for Strategies 1–3 were slightly positive at three bins, close to zero at four and five bins, and moderately positive at six bins with an overestimation of about 60%. Results for the 48 kHz recordings mirrored this pattern. Small positive bias at three bins, slight negative bias at four and five bins, and moderate positive bias at six bins with an overestimation of approximately 75%. Across both sampling-rate strata, the Study-level estimator exhibited consistently low bias, indicating stable behaviour even as binning increased.

The variability results highlighted clear differences between the Study-level and strategy-based estimators. As shown in Table 33, the Study-level estimator displayed very low bootstrap standard deviations (SD ≈ 0.007 in each stratum and SD ≈ 0.006 when combined). In contrast, Strategies 1–3 showed substantially higher dispersion, with SD values of approximately 0.346 at 32 kHz and 0.444 at 48 kHz, and combined SD values of roughly 0.371. The empirical standard deviation, derived from the full simulation output (Table 34), was also small (0.020), closely aligned with the Study-level bootstrap SD but far smaller than the strategy-specific SDs.

Taken together, the 25-sample results showed that all estimation approaches produced broadly similar call-density and relative-bias patterns for the African Grey Flycatcher. However, the Study-level estimator was markedly more stable than any of the strategy-based approaches, with both empirical and bootstrap SDs indicating exceptionally low variability. The strata-specific strategies, although following the same qualitative trends, consistently exhibited much larger dispersion and showing no intra-strategy differences, reinforcing the conclusion that the Study-level estimator was the most stable.

50 validated samples per bin:

The results presented below are of the African Gray Flycatcher at 50 validated samples per bin.

Table 13: Estimated call densities across binning levels and sampling rates for African Gray Flycatcher (50 validated samples per bin).

Level	3 bins		4 bins		5 bins		6 bins	
	32 kHz	48 kHz	32 kHz	48 kHz	32 kHz	48 kHz	32 kHz	48 kHz
Study Level	0.100	0.100	0.091	0.091	0.091	0.091	0.092	0.092
Strategy 1	0.092	0.079	0.111	0.104	0.124	0.128	0.097	0.099
Strategy 2	0.092	0.079	0.111	0.104	0.124	0.128	0.097	0.099
Strategy 3	0.092	0.079	0.111	0.104	0.124	0.128	0.097	0.099

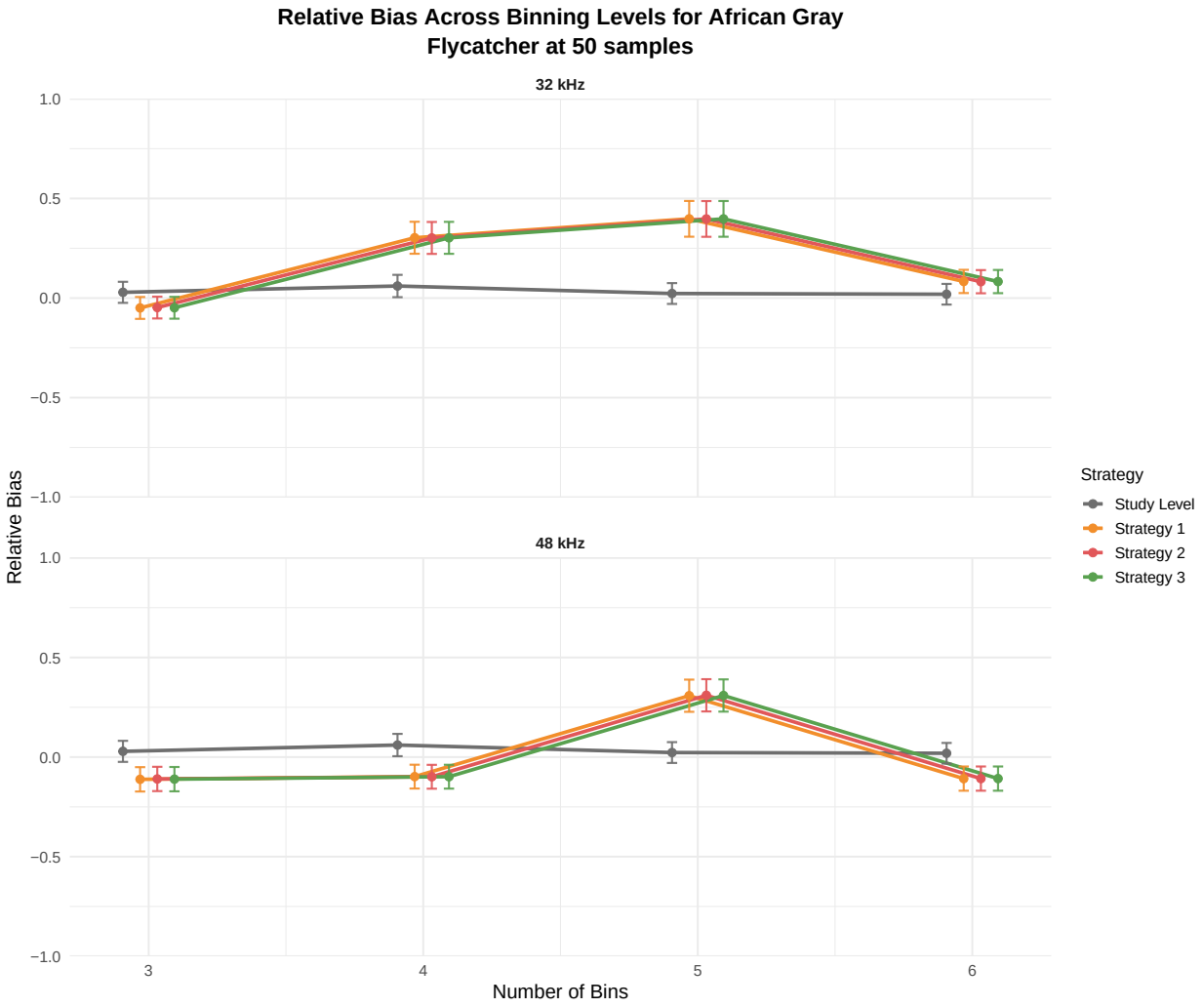

Figure 18: Bias plot of African Gray Flycatcher at 50 samples per bin

The call-density estimates for the Abyssinian Nightjar at 50 validated samples per bin remained remarkably consistent across all binning schemes, sampling-rate strata, and estimation strategies. Study-level estimates occupied an extremely narrow range between 0.600 and 0.602, while strategy-based estimates differed only slightly, spanning approximately 0.595 to 0.610 in the 32 kHz recordings and 0.596 to 0.616 in the 48 kHz stratum. These minimal differences indicate that all calibration approaches produced nearly identical call-density values and that estimator choice had no practical influence on the magnitude of the estimates.

Relative-bias results reinforced this uniformity. Across all configurations, bias values for every strategy remained very close to zero, typically between about -0.05 and 0.05, with no consistent directional trend across binning levels or sampling-rate strata. The Study-level estimator showed the same centred behaviour as Strategies 1–3, indicating that none of the methods produced systematic over- or underestimation of the true call density.

Variability estimates reflected the same pattern of stability. As shown in Table 42, the Study-level estimator produced extremely small standard deviations in both strata (SD ≈ 0.002), representing near-zero variability across Monte Carlo replicates. Strategies 1–3 exhibited slightly higher variability, but the magnitude remained small: SD values of approximately 0.013 in the 32 kHz stratum and 0.039–0.040 in the 48 kHz recordings. Combined-strata standard deviations followed the same structure, with the Study-level SD remaining minimal (0.002) and the strategy-based estimators clustering around 0.131.

Comparison with the empirical combined standard deviation provides additional insight into model stability. As shown in Table 43, the empirical SD at 50 samples per bin was 0.017—substantially larger than the Study-level bootstrap SD of 0.002, but still very small in absolute terms. The discrepancy reflects the fact that bootstrap SD quantifies Monte Carlo variability of the fitted model, whereas the empirical SD captures broader sampling uncertainty across simulated ground-truth populations. Importantly, even the empirical SD remains low, indicating that model uncertainty for this species is minimal under real-world sampling fluctuation.

Taken together, the results at 50 validated samples per bin demonstrate that the Abyssinian Nightjar represents one of the most stable cases in the study. Call-density estimates were tightly grouped, relative bias remained close to zero, and both bootstrap-derived and em-

pirical measures of variability were small. No estimator exhibited meaningful deviation in accuracy or stability, confirming that all strategies performed equivalently across strata and binning configurations.

75 validated samples per bin:

The results presented below are for the African Gray Flycatcher at 75 validated samples per bin.

Table 14: Estimated call densities across binning levels and sampling rates for African Gray Flycatcher (75 validated samples per bin).

Level	3 bins		4 bins		5 bins		6 bins	
	32 kHz	48 kHz	32 kHz	48 kHz	32 kHz	48 kHz	32 kHz	48 kHz
Study Level	0.098	0.098	0.090	0.090	0.089	0.089	0.089	0.089
Strategy 1	0.100	0.092	0.102	0.096	0.086	0.082	0.070	0.064
Strategy 2	0.099	0.092	0.102	0.096	0.086	0.082	0.069	0.063
Strategy 3	0.099	0.092	0.102	0.096	0.086	0.082	0.069	0.063

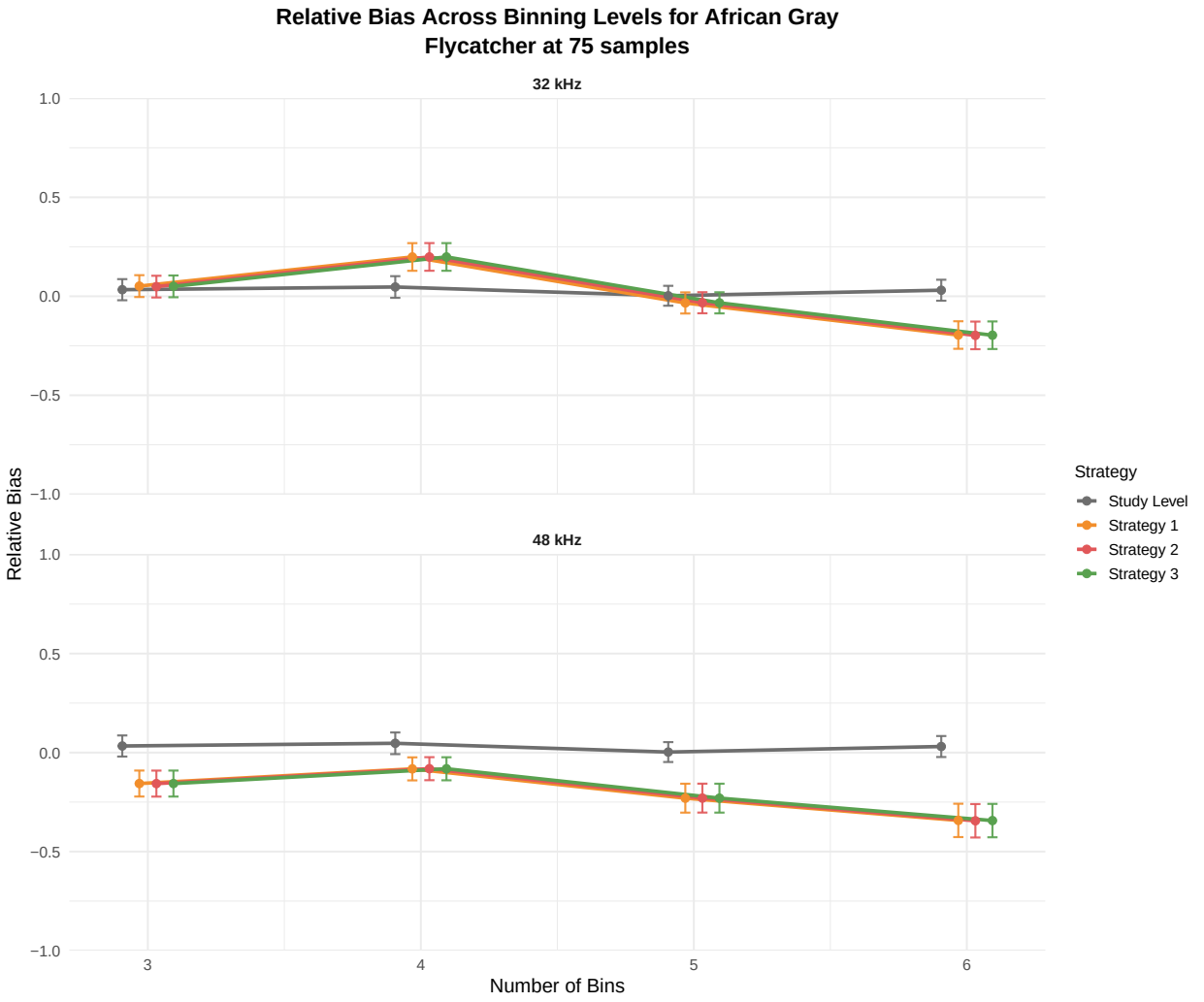

Figure 19: Bias plot of African Gray Flycatcher at 75 samples per bin

The call-density estimates for the African Grey Flycatcher at 75 validated samples per bin formed a tightly constrained set of values across both sampling-rate strata. Study-level estimates ranged narrowly from 0.089 to 0.098 across all binning configurations, indicating a high degree of numerical stability. Strategies 1–3 produced only slightly wider ranges, with densities between approximately 0.070 and 0.102 at 32 kHz and between 0.064 and 0.096 at 48 kHz. Although these values showed marginally greater spread than the Study-level estimates, they remained well within the same overall density band, confirming that estimator choice had minimal influence on call-density magnitude at this higher validation depth.

Relative-bias patterns reinforced this overall stability. In the 32 kHz recordings, Strategies 1–3 displayed small positive bias at three and four bins, transitioning to mild negative values at six bins. A similar pattern appeared in the 48 kHz stratum, where bias values remained close to zero and became only slightly negative at higher bin counts. Across all configurations, the Study-level estimator remained centred near zero, and none of the estimation approaches produced large systematic deviations from the ground truth except for the 48 kHz that underestimated by more than 20% at 6 bins.

The variability results were consistent with these observations. As shown in Table 38, the Study-level estimator exhibited very low dispersion, with standard deviations of approximately 0.019 in both sampling-rate strata and a combined SD of 0.017. Strategies 1–3 showed higher but still moderate variability, with SD values of roughly 0.165 at 32 kHz and 0.111–0.112 at 48 kHz, and combined SDs of approximately 0.171. Importantly, the empirical combined SD computed independently from the simulation results (Table **??**) was also very small (0.013). The close correspondence between the Study-level bootstrap SD (0.017) and the empirical SD (0.013) demonstrates that the Study-level estimator captured the sampling uncertainty well and remained highly stable at this validation depth. In contrast, although the strategy-based estimators showed limited variability relative to lower sample sizes, their SDs remained an order of magnitude larger than those of the Study-level method.

Taken together, the 75-sample results showed that call-density estimation for the African Grey Flycatcher was highly stable across binning levels, sampling rates, and calibration strategies. Relative bias remained tightly centred around zero, and variability decreased further compared with lower validation sample sizes. While all strategies behaved similarly, the Study-level estimator continued to exhibit the lowest variability, with empirical and bootstrap SDs confirming its superior numerical stability. The strategy-based estimators

produced consistent results but retained modestly higher variability, reinforcing the advantages of the Study-level approach under high data availability.

Abyssinian/Montane Nightjar

25 validated samples per bin:

The results presented below are for the Abyssinian/Montane Nightjar at 25 validated samples per bin.

Table 15: Estimated call densities across binning levels and sampling rates for Abyssinian/Montane Nightjar (25 validated samples per bin).

Level	3 bins		4 bins		5 bins		6 bins	
	32 kHz	48 kHz	32 kHz	48 kHz	32 kHz	48 kHz	32 kHz	48 kHz
Study Level	0.600	0.600	0.601	0.601	0.603	0.603	0.600	0.600
Strategy 1	0.582	0.586	0.585	0.593	0.551	0.548	0.586	0.585
Strategy 2	0.582	0.585	0.585	0.593	0.550	0.548	0.586	0.585
Strategy 3	0.582	0.585	0.585	0.593	0.550	0.548	0.586	0.585

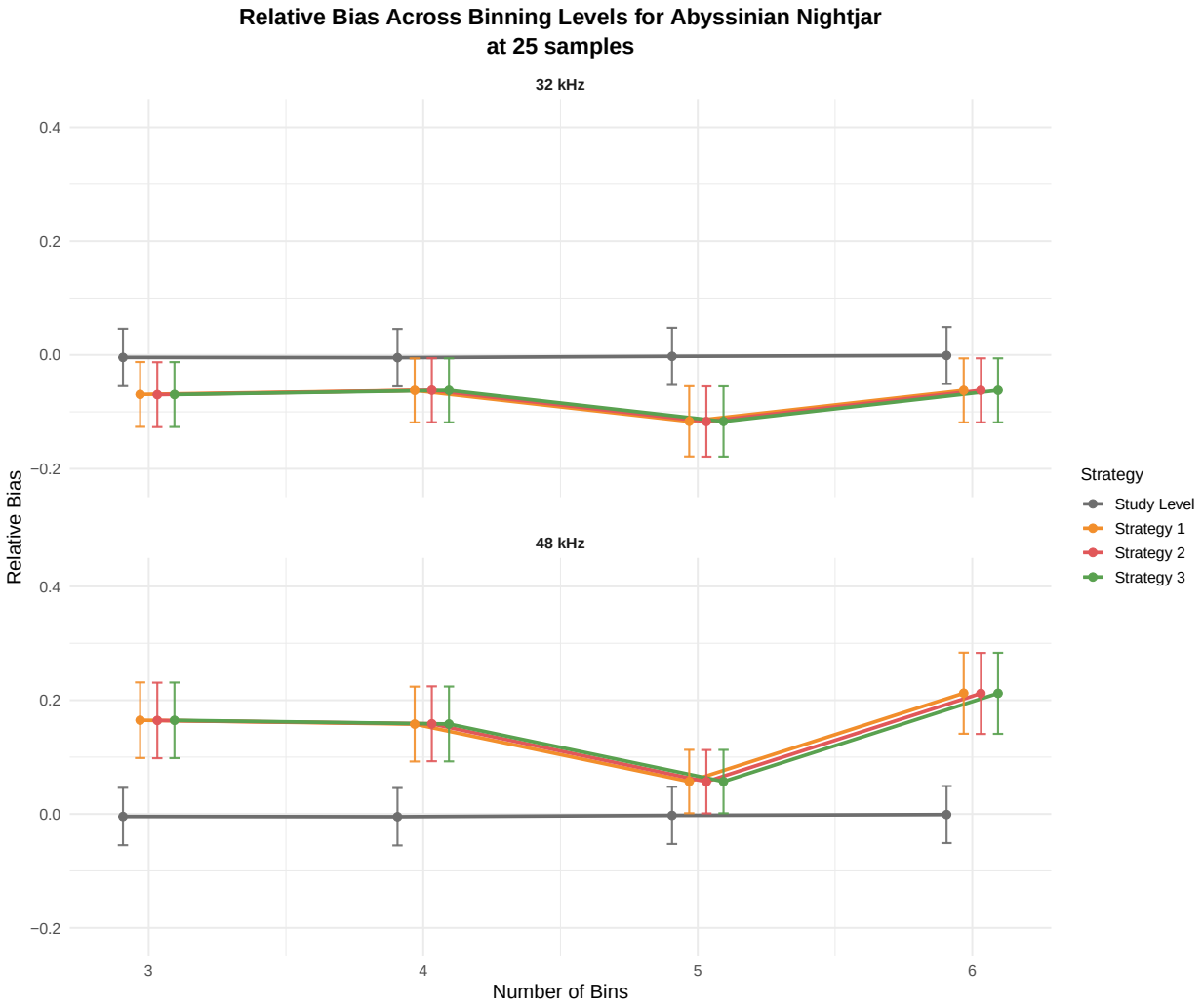

Figure 20: Bias plot of Abyssinian Nightjar at 25 samples per bin

The call-density estimates for the Abyssinian Nightjar at 25 validated samples per bin showed a tightly constrained set of values across all binning levels and sampling rates. Study-level estimates ranged only from 0.600 to 0.603 across the 32 kHz and 48 kHz strata, indicating almost complete insensitivity to binning configuration. Strategies 1–3 produced slightly wider but still highly concentrated estimates, ranging from approximately 0.550 to 0.593 in both strata. These results demonstrate that even under strategy-based calibration, estimated densities remained numerically very close to the Study-level outputs, with no meaningful divergence attributable to estimator choice.

Relative-bias results reflected the same degree of stability. Across all binning schemes, Strategies 1–3 produced bias values clustered tightly around zero, with only mild negative deviations at five bins and small positive deviations at three and six bins. This structure was present in both sampling-rate strata, although the 48 kHz recordings showed slightly more positive bias on average. The Study-level estimator remained centred almost exactly at zero across all configurations, producing negligible systematic deviation from the ground truth. Overall, all estimation approaches yielded bias patterns that were extremely close to zero and showed no evidence of directional distortion at deviations of less than 20% from the ground truth for both strata.

The variability results reinforced this interpretation. As shown in Table 40, the Study-level estimator exhibited extremely low dispersion, with standard deviations of approximately 0.002 in both strata and a combined SD of 0.002. Strategies 1–3 displayed higher variability, but the increase was modest. SDs of about 0.026 in the 32 kHz stratum and 0.065 in the 48 kHz recordings. Combined-strata variability followed the same pattern, with Study-level variability near zero and strategy-based variability remaining well bounded.

A comparison to the empirical combined standard deviation (Table 41) provides additional insight. The empirical SD of 0.021 is substantially larger than the Study-level bootstrap SD (0.002), suggesting that although the bootstrap estimator reflects extremely low variability, the real-world dispersion in score–truth relationships for this species is higher. Importantly, however, even the empirical SD remains small in absolute terms, indicating that the estimator is still performing with high stability, particularly given the low validation effort of only 25 samples per bin.

Taken together, these results show that the Abyssinian Nightjar exhibited one of the most stable estimation profiles observed across all species. Call-density estimates were tightly grouped across strata and calibration strategies, relative bias remained extremely close to zero, and variability stayed low in both bootstrap-derived and empirical assessments. At this validation depth, all strategies behaved similarly and produced accurate, robust, and internally consistent estimates of call density.

50 validated samples per bin:

The results presented below are for the Abyssinian/Montane Nightjar at 50 validated samples per bin.

Table 16: Estimated call densities across binning levels and sampling rates for Abyssinian/Montane Nightjar (50 validated samples per bin).

Level	3 bins		4 bins		5 bins		6 bins	
	32 kHz	48 kHz	32 kHz	48 kHz	32 kHz	48 kHz	32 kHz	48 kHz
Study Level	0.600	0.600	0.601	0.601	0.602	0.602	0.600	0.600
Strategy 1	0.610	0.616	0.607	0.613	0.595	0.596	0.609	0.600
Strategy 2	0.610	0.616	0.607	0.613	0.595	0.596	0.609	0.600
Strategy 3	0.610	0.616	0.607	0.613	0.595	0.596	0.609	0.600

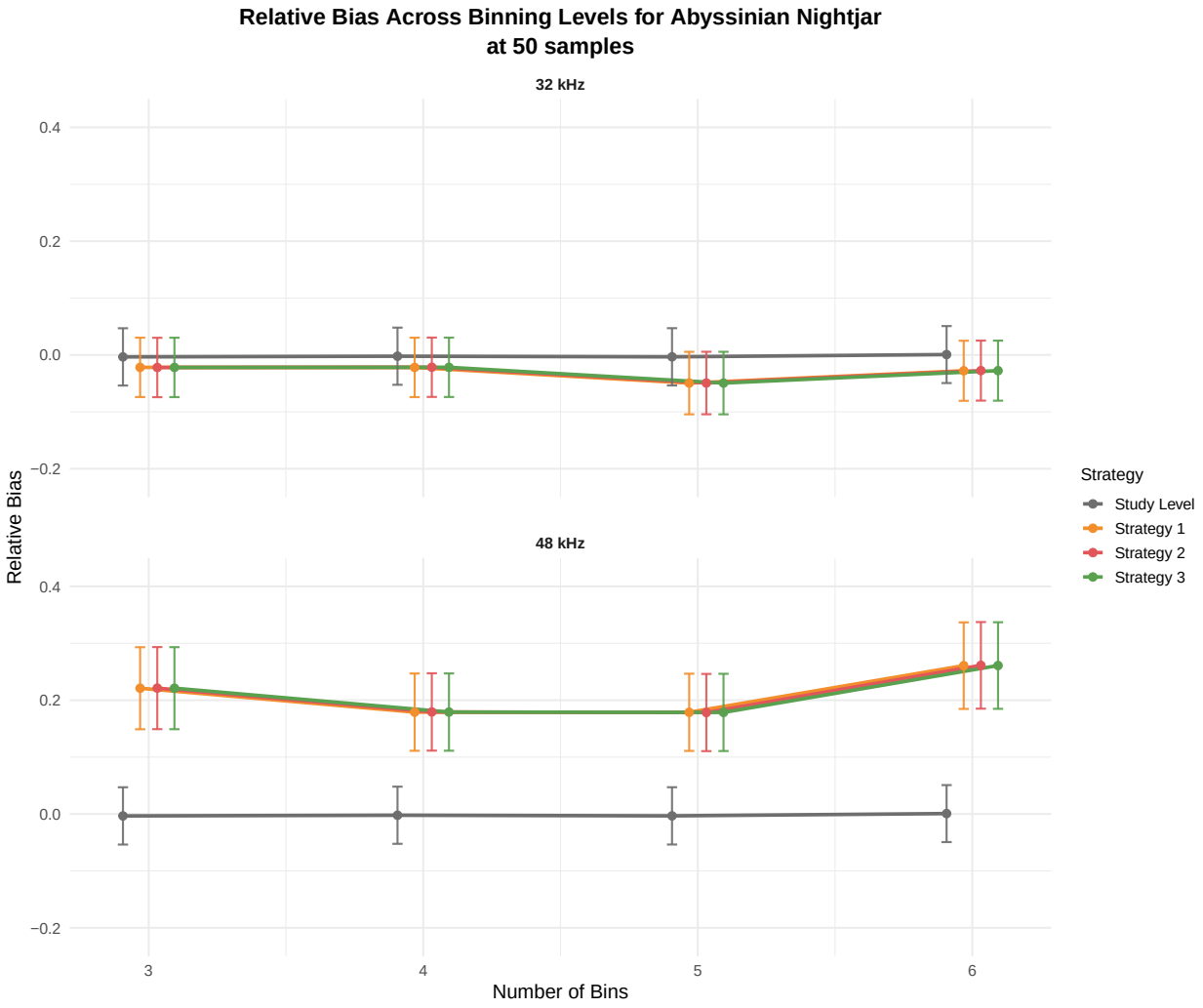

Figure 21: Bias plot of Abyssinian/Montane Nightjar at 50 samples per bin

The call-density estimates for the Abyssinian Nightjar at 50 validated samples per bin remained remarkably consistent across all binning schemes, sampling-rate strata, and estimation strategies. Study-level estimates occupied an extremely narrow range between 0.600 and 0.602, while strategy-based estimates differed only slightly, spanning approximately 0.595 to 0.610 in the 32 kHz recordings and 0.596 to 0.616 in the 48 kHz stratum. These minimal differences indicate that all calibration approaches produced nearly identical call-density values and that estimator choice had no practical influence on the magnitude of the estimates.

Relative-bias results reinforced this uniformity. Across all configurations, bias values for every strategy remained very close to zero, typically between about -0.05 and 0.05, with no consistent directional trend across binning levels or sampling-rate strata. The Study-level estimator showed the same centred behaviour as Strategies 1–3, indicating that none of the methods produced systematic over- or underestimation of the true call density.

Variability estimates reflected the same pattern of stability. As shown in Table 42, the Study-level estimator produced extremely small standard deviations in both strata (SD ≈ 0.002), representing near-zero variability across replicates. Strategies 1–3 exhibited slightly higher variability, but the magnitude remained small. SD values of approximately 0.013 in the 32 kHz stratum and 0.039–0.040 in the 48 kHz recordings. Combined-strata standard deviations followed the same structure, with the Study-level SD remaining minimal (0.002) and the strategy-based estimators clustering around 0.131.

Comparison with the empirical combined standard deviation provides additional insight into model stability. As shown in Table 43, the empirical SD at 50 samples per bin was 0.017, substantially larger than the Study-level bootstrap SD of 0.002, but still very small in absolute terms. The discrepancy reflects the fact that bootstrap SD quantifies variability of the fitted model, whereas the empirical SD captures broader sampling uncertainty across simulated ground-truth populations. Importantly, even the empirical SD remains low, indicating that model uncertainty for this species is minimal under real-world sampling fluctuation.

Taken together, the results at 50 validated samples per bin demonstrate that the Abyssinian Nightjar represents one of the most stable cases in the study. Call-density estimates were tightly grouped, relative bias remained close to zero, and both bootstrap-derived and empirical measures of variability were small. No estimator exhibited meaningful deviation in

accuracy or stability except the 48 kHz at 3 and 6 bins. Confirming that generally, strategies performed equivalently across strata and binning configurations.

75 validated samples per bin:

The results presented below are for the Abyssinian/Montane Nightjar at 75 validated samples per bin.

Table 17: Estimated call densities across binning levels and sampling rates for Abyssinian Nightjar (75 validated samples per bin).

Level	3 bins		4 bins		5 bins		6 bins	
	32 kHz	48 kHz	32 kHz	48 kHz	32 kHz	48 kHz	32 kHz	48 kHz
Study Level	0.601	0.601	0.602	0.602	0.603	0.603	0.601	0.601
Strategy 1	0.605	0.594	0.618	0.613	0.622	0.624	0.609	0.611
Strategy 2	0.605	0.594	0.618	0.613	0.622	0.624	0.609	0.611
Strategy 3	0.605	0.594	0.618	0.613	0.622	0.624	0.609	0.611

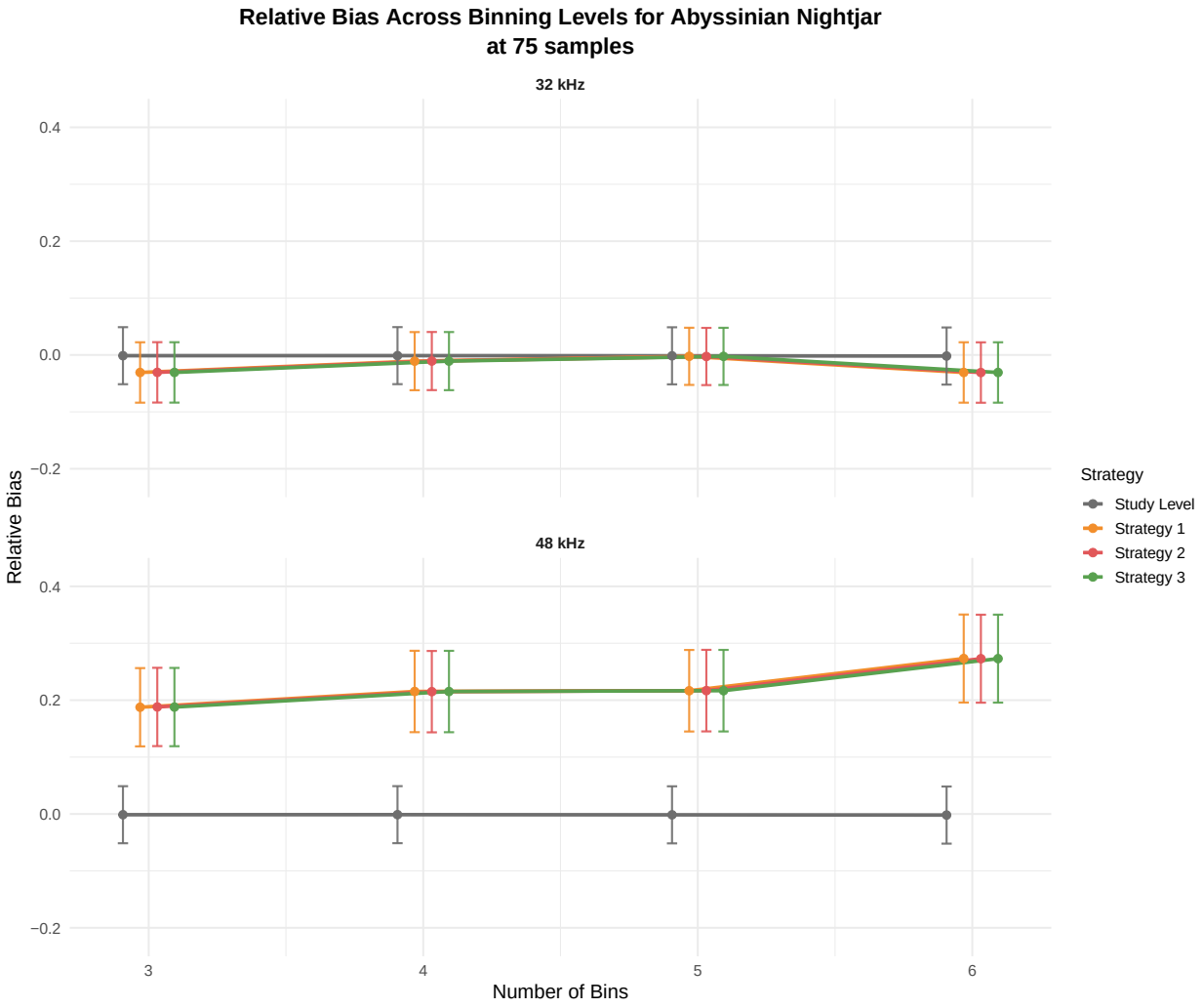

Figure 22: Bias plot of Abyssinian/Montane Nightjar at 75 samples per bin

The call-density estimates for the Abyssinian Nightjar at 75 validated samples per bin remained highly consistent across all binning levels, sampling-rate strata, and estimation strategies. Study-level estimates were tightly concentrated between 0.601 and 0.603 across all configurations, indicating complete insensitivity to bin structure at this validation depth. Strategy-based estimates were similarly stable, ranging from approximately 0.605 to 0.622 in the 32 kHz recordings and from 0.594 to 0.624 in the 48 kHz stratum. These values remained centred around the same density level as the Study-level estimator, demonstrating that all approaches produced effectively identical call-density outputs when sufficient validation data

were available.

Relative-bias patterns reflected the same stability. In the 32 kHz stratum, all estimators including the Study-level approach produced bias values that were almost zero across every binning configuration, indicating exact or near-exact recovery of the true density. In the 48 kHz recordings, bias values showed positive deviations (over estimations of approximately 20–28%), but these shifts were consistent across strategies and did not depend on bin count. Across all configurations, relative-bias values remained within a narrow neighbourhood around zero.

Variability results further reinforced this pattern. As shown in Table 44, the Study-level estimator exhibited zero variability across strata (SD = 0.000), reflecting identical bias estimates across all replicates. Strategy-based estimators showed only modest dispersion, with SD $\approx$ 0.014 at 32 kHz and SD $\approx$ 0.036 at 48 kHz. When strata were combined, the resulting strategy-level standard deviations remained limited (SD $\approx$ 0.132), and none of the estimation approaches produced variability patterns large enough to indicate any meaningful behavioural differences.

A comparison with the empirical standard deviation provides additional insight into estimator stability. Table 45 reports an empirical SD of 0.013, which, while relatively larger than the Study-level bootstrap SD of 0.000 is still very small in absolute terms. Importantly, both remain extremely low, demonstrating that call-density estimation for this species may be highly robust under realistic sampling uncertainty.

Taken together, the results at 75 validated samples per bin demonstrate that call-density estimation for the Abyssinian Nightjar was stable, unbiased, and invariant to binning scheme,

sampling rate, and estimation strategy. All estimators converged near a density of 0.60, relative bias remained centred at zero, and both bootstrap-derived and empirical variability measures were minimal. With sufficient validation effort, all strategies performed equivalently, confirming that this species represents a highly reliable estimation case with no meaningful performance differences across modelling approaches.

African Pipit

25 validated samples per bin:

The results presented below are for the African Pipit at 25 validated samples per bin.

Table 18: Estimated call densities across binning levels and sampling rates for African Pipit (25 validated samples per bin).

Level	3 bins		4 bins		5 bins		6 bins	
	32 kHz	48 kHz	32 kHz	48 kHz	32 kHz	48 kHz	32 kHz	48 kHz
Study Level	0.115	0.115	0.117	0.117	0.116	0.116	0.106	0.106
Strategy 1	0.121	0.116	0.118	0.104	0.071	0.062	0.101	0.107
Strategy 2	0.121	0.116	0.118	0.103	0.071	0.062	0.101	0.107
Strategy 3	0.121	0.116	0.118	0.103	0.071	0.062	0.101	0.107

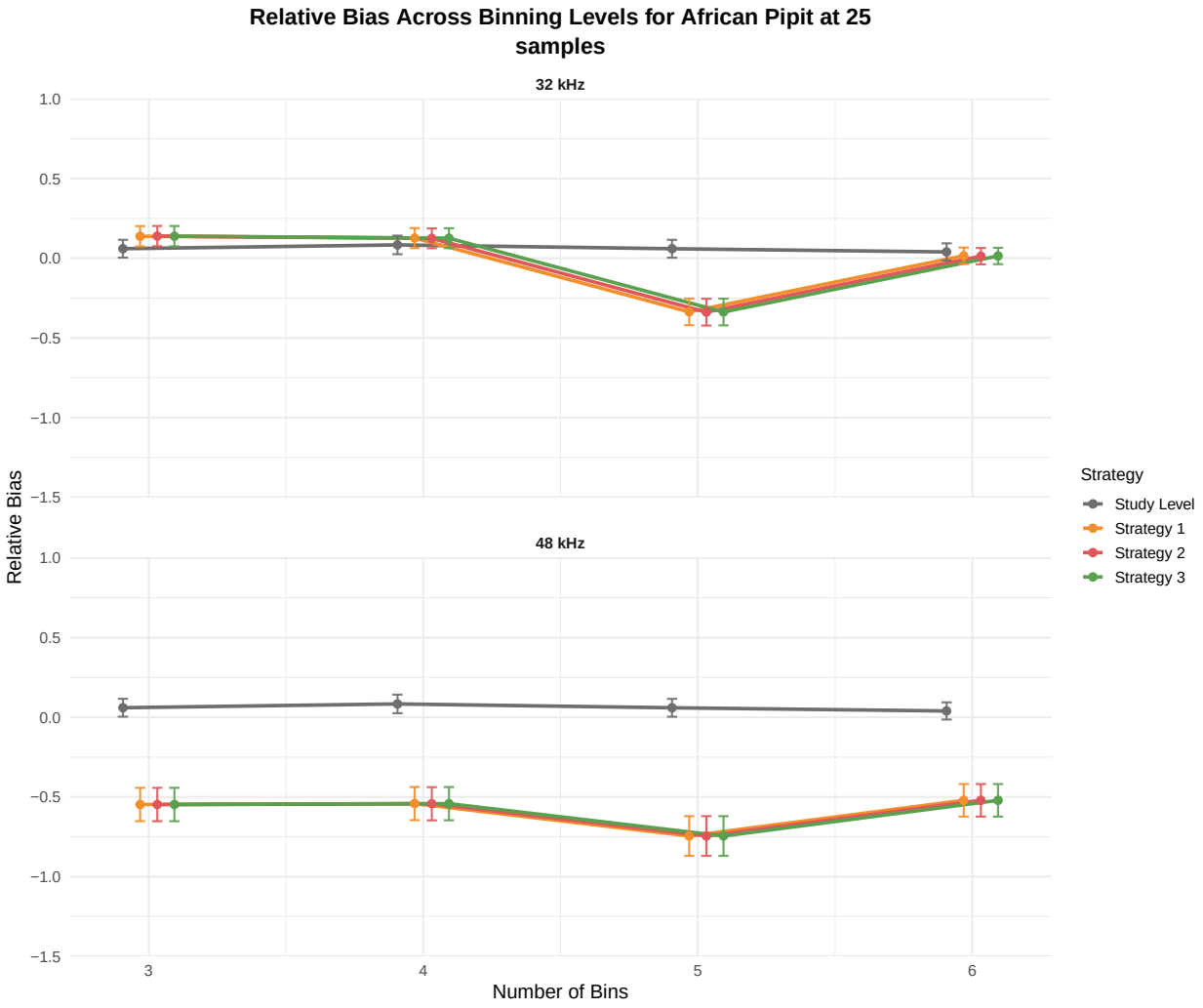

Figure 23: Bias plot of African Pipit at 25 samples per bin

The call-density estimates for the African Pipit at 25 validated samples per bin remained low across all binning configurations and sampling-rate strata. Study-level estimates were tightly grouped, ranging from 0.115 to 0.117 for the 3–5 bin schemes before decreasing slightly to 0.106 at six bins. Strategy-based estimates followed the same overall pattern, with values ranging from approximately 0.071 to 0.121 at 32 kHz and from roughly 0.116 down to 0.062 at 48 kHz before increasing again at six bins. These fluctuations were consistent across all strategies, indicating that the observed shifts in estimated density were driven by bin count and sampling rate rather than estimator choice.

Relative-bias behaviour mirrored these density patterns. In the 32 kHz recordings, Strategies 1–3 exhibited mild underestimation that began near zero at three bins, became increasingly negative and reached its strongest magnitude at five bins (around -0.4), before returning toward zero at six bins. The same progression occurred in the 48 kHz stratum but with deeper underestimation, reaching values near -0.8 (80% underestimation) at five bins. The Study-level estimator followed the same directional trend but with reduced magnitude, remaining closer to zero across all configurations. No meaningful differences emerged among Strategies 1–3, which produced nearly identical bias curves at both sampling rates.

Estimator variability followed a consistent pattern. As shown in Table 46, the Study-level estimator exhibited low dispersion around zero (SD ≈ 0.018 across both strata), whereas Strategies 1–3 displayed higher but nearly identical variability, with SD values of approximately 0.222–0.223 at 32 kHz and about 0.105 at 48 kHz. Combined-strata variability for the strategies (SD ≈ 0.346–0.347) reflected the joint effects of underestimation at intermediate binning levels and differences between sampling-rate strata, rather than differences attributable to strategy design.

A comparison between the bootstrap-based Study-level variability and the empirical variability provides further insight into estimator performance. Table 47 reports an empirical SD of 0.022, which is slightly larger than the Study-level combined bootstrap SD of 0.017. This difference arises because the empirical SD captures variation across independent simulated validation populations, while the bootstrap SD reflects internal uncertainty associated with the fitted Beta–Binomial model. Despite this discrepancy, both values remain small in absolute terms, confirming that the Study-level estimator maintained low uncertainty relative to the strategy-based approaches.

Taken together, the results for the African Pipit at 25 validated samples per bin demonstrate that all estimation approaches produced similar call-density behaviour characterised by systematic underestimation at intermediate binning levels, particularly in the 48 kHz recordings. Strategy-based estimators showed equivalent performance with indistinguishable bias and variability patterns. Although the Study-level estimator remained closest to the ground truth and showed the lowest variability.

50 validated samples per bin:

The results presented below are for the African Pipit at 50 validated samples per bin.

Table 19: Estimated call densities across binning levels and sampling rates for African Pipit (50 validated samples per bin).

Level	3 bins		4 bins		5 bins		6 bins	
	32 kHz	48 kHz	32 kHz	48 kHz	32 kHz	48 kHz	32 kHz	48 kHz
Study Level	0.113	0.113	0.113	0.113	0.114	0.114	0.103	0.103
Strategy 1	0.092	0.078	0.089	0.079	0.103	0.090	0.081	0.079
Strategy 2	0.092	0.078	0.089	0.079	0.103	0.090	0.081	0.079
Strategy 3	0.092	0.078	0.089	0.079	0.103	0.090	0.081	0.079

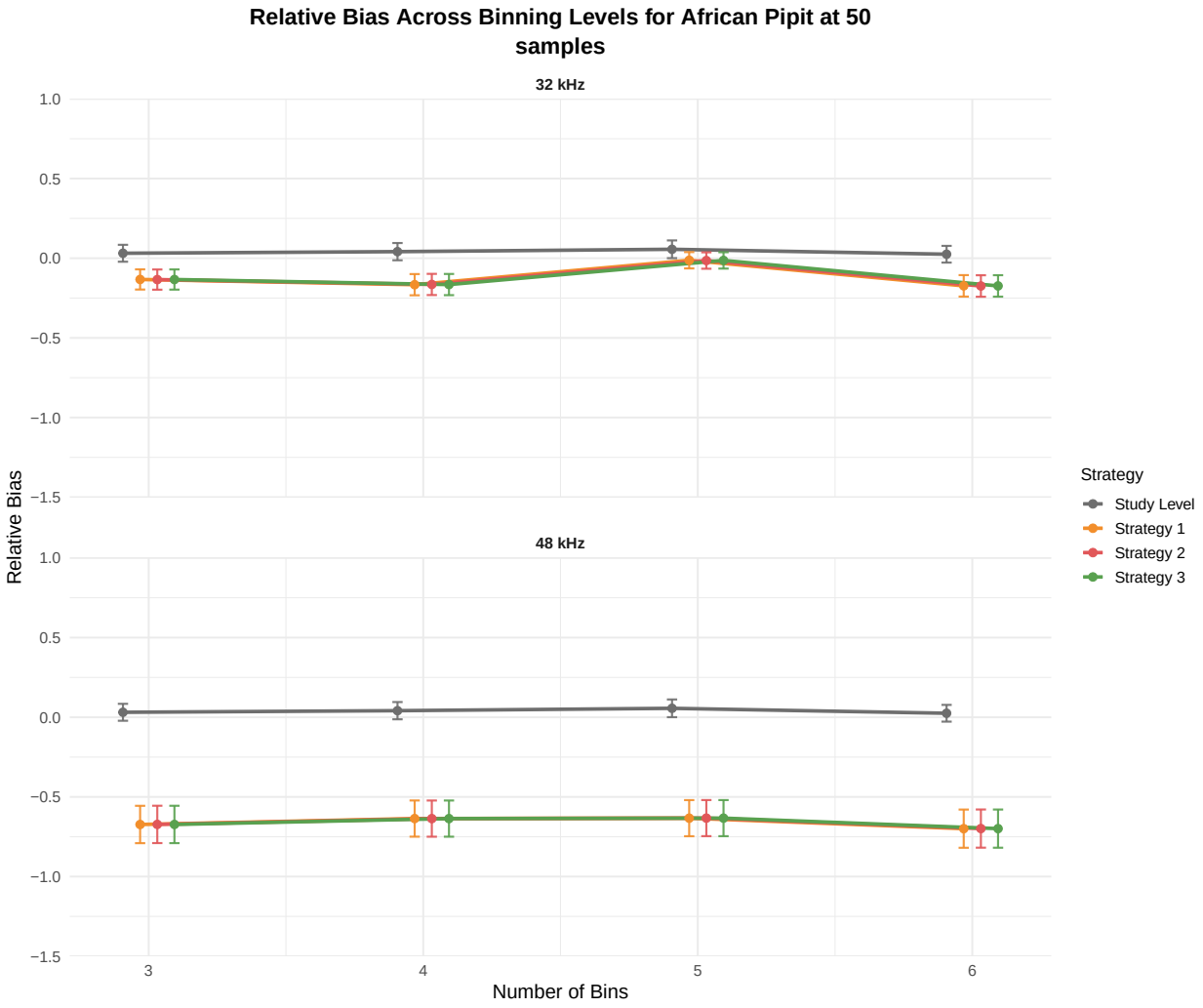

Figure 24: Bias plot of African Pipit at 50 samples per bin

The call-density estimates for the African Pipit at 50 validated samples per bin showed consistent behaviour across all three strategy-based estimators, with differences driven primarily by sampling rate rather than by estimator design. Study-level estimates were tightly grouped, ranging from 0.113 to 0.114 across the 3–5 bin configurations and decreasing slightly to 0.103 at six bins, indicating stable performance across binning levels. Strategy-based estimates followed the same overall pattern but at lower numerical values. In the 32 kHz recordings, densities ranged from approximately 0.089 to 0.103, while in the 48 kHz stratum values remained between roughly 0.078 and 0.090. These reductions were uniform across

Strategies 1–3, indicating that the slight downward shift in density reflected sampling-rate effects rather than any estimator-specific behaviour.

Relative-bias trends mirrored these density patterns. In the 32 kHz stratum, all strategy-based estimators produced mild underestimation at every binning level, beginning near -0.1 at three bins, moving slightly toward zero at four and five bins, and declining again at six bins. In the 48 kHz recordings, the same structure emerged but with consistently stronger negative bias (approximately -0.8 across bins 3–6). Strategies 1–3 produced nearly identical bias curves, confirming that the pattern originated from the per-stratum calibration rather than from estimator differences. By contrast, the Study-level estimator remained centred near zero across all binning levels, demonstrating reduced sensitivity to strata-specific deviations.

Estimator variability exhibited the same uniformity. As shown in Table 48, the Study-level estimator displayed low dispersion (SD ≈ 0.014 in both strata; combined SD ≈ 0.013), while Strategies 1–3 showed higher but similar variability, with SD values of approximately 0.074 at 32 kHz and 0.031–0.032 at 48 kHz. The combined-strata standard deviations (SD ≈ 0.292–0.293) were dominated by the consistently stronger underestimation in the 48 kHz recordings rather than by differences between the strategy-based estimators.

Additional insight comes from comparing the Study-level bootstrap SD to the empirical SD calculated across fully simulated datasets. Table 49 reports an empirical standard deviation of 0.016, slightly larger than the Study-level combined bootstrap SD of 0.013.

Overall, the African Pipit results at 50 validated samples per bin show that all strategy-based estimators behaved similarly, producing systematic underestimation, most notably in

the 48 kHz recordings and exhibiting moderate variability. The Study-level estimator remained the only configuration consistently yielding near-unbiased density estimates across all binning schemes and sampling rates.

75 validated samples per bin:

The results presented below are for the African Pipit at 75 validated samples per bin.

Table 20: Estimated call densities across binning levels and sampling rates for African Pipit (75 validated samples per bin).

Level	3 bins		4 bins		5 bins		6 bins	
	32 kHz	48 kHz	32 kHz	48 kHz	32 kHz	48 kHz	32 kHz	48 kHz
Study Level	0.111	0.111	0.112	0.112	0.113	0.113	0.100	0.100
Strategy 1	0.116	0.116	0.128	0.113	0.117	0.109	0.119	0.115
Strategy 2	0.116	0.116	0.128	0.113	0.117	0.109	0.119	0.115
Strategy 3	0.116	0.116	0.128	0.113	0.117	0.109	0.119	0.115

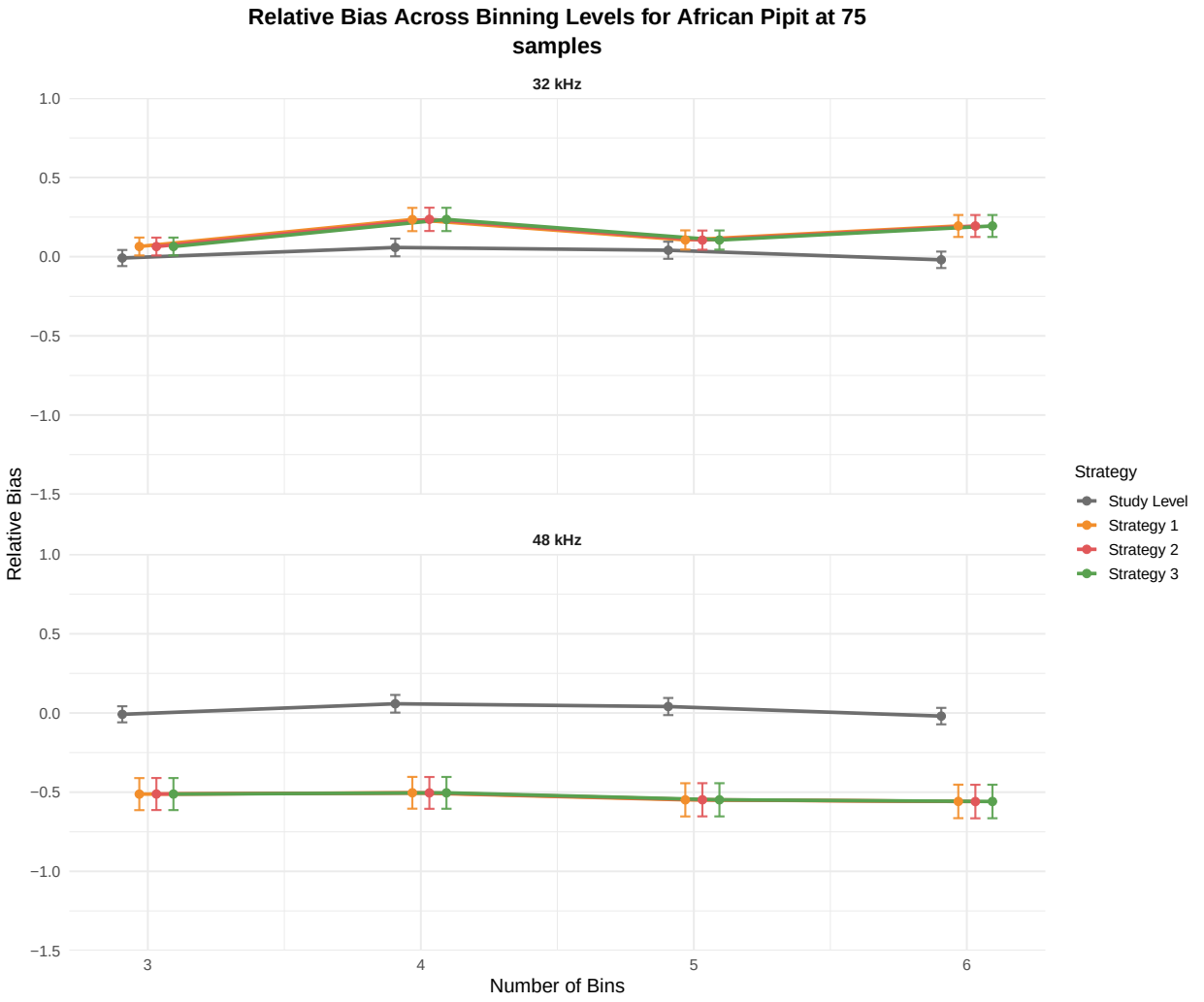

Figure 25: Bias plot of African Pipit at 75 samples per bin

The call-density estimates for the African Pipit at 75 validated samples per bin showed a pronounced improvement in numerical stability, with clear separation between the study-level estimator and the strategy-based approaches while retaining the overall structure seen at lower sample sizes. Study-level estimates remained tightly aligned across all binning levels, ranging from 0.111 to 0.113 in the 3–5 bin configurations and declining modestly to 0.100 at six bins. This narrow spread indicates that the pooled calibration continued to behave in a highly regular and predictable way as validation effort increased. Strategy-based estimates exhibited the same qualitative pattern across all calibration strategies. In the 32 kHz

recordings, densities rose from roughly 0.116 at three bins to approximately 0.128 at four bins before returning toward 0.117–0.119 at higher bin counts. In contrast, estimates from the 48 kHz stratum followed the same shape but lay at a slightly lower level (approximately 0.109–0.115).

Bias patterns closely tracked these density shifts. In the 32 kHz recordings, the strategy-based estimators produced uniformly positive bias, typically between 0.10 and 0.20, indicating mild overestimation of the true density. In the 48 kHz stratum, the sign reversed: all strategies exhibited consistent underestimation, with bias values centred between -0.6 and -0.8. The opposing directions of bias across strata therefore reflected systematic differences in the distribution of BirdNET scores rather than methodological differences among the strategies. Throughout all binning levels and in both strata, the study-level estimator remained effectively unbiased, with its bias curve anchored at zero.

Patterns of estimator variation reinforced this contrast. As shown in Table 50, the study-level bootstrap standard deviations were small and nearly identical across strata (SD ≈ 0.038), with a combined SD of 0.035. This compact spread indicates that the pooled estimator behaved very consistently across replicates. Strategy-based variability displayed a tiered structure. In the 32 kHz stratum, SDs were moderately larger (SD $\approx$ 0.078–0.079), while in the 48 kHz recordings variability dropped to SD ≈ 0.027, reflecting the very stable but systematically biased behaviour of those estimates.

A broader view of uncertainty emerges by comparing the study-level bootstrap SDs with the empirical standard deviation obtained across independently generated datasets. Table 51 reports an empirical SD of 0.013, which is substantially smaller than the study-level combined bootstrap SD of 0.035. Empirical SD being considerably lower may suggests that the

estimator's end-to-end performance is even more consistent than the bootstrap-based bias variation would imply.

Taken together, the African Pipit results at 75 validated samples per bin show that increasing the validation effort substantially stabilised all estimation approaches. The strategy-based estimators followed the same systematic pattern. Overestimation in the 32 kHz recordings and consistent underestimation in the 48 kHz stratum with no meaningful differences between strategies. The study-level estimator continued to outperform the stratified approaches in both accuracy and consistency, producing near-zero bias and relatively small bootstrap variability. Although the strategy-based estimators remained sensitive to sampling-rate differences, all methods benefited from the larger sample size, and the tight empirical SD confirms that overall density estimation was highly stable at this validation depth.

Discussion

This study evaluated a threshold-free framework for estimating call density from BirdNET outputs under realistic field conditions, using a combination of study-level model calibration and three strata-level strategies applied to two sampling-rate strata (32 kHz and 48 kHz). By embedding the estimator within a quantile-binning and Beta–Binomial modelling framework, and by explicitly representing distribution shifts between strata, the analysis provides insight not only into how well call density can be recovered, but also why estimator performance varies across species, strata, binning schemes, and validation-sample sizes. The overall results demonstrate that estimator performance is driven far more by the availability and distribution of validated information than by the theoretical differences between the calibration strategies themselves.

Study level

A central finding of this study is the consistently strong and stable performance of the study-level estimator. Across all five species and all binning schemes, its estimates of relative bias remained close to zero under all but the most limited validation conditions. The primary exception was the Baglafecht weaver, which showed substantial overestimation at 25 and 50 validation samples per bin, with bias reaching almost 50% (see Figure 11) in the 25-sample configuration. This behaviour reflects the species' low detectability and sparse high-confidence positive examples in the validation set, which resulted in many poorly populated bins. Under such sparse conditions, the Beta–Binomial posterior is weakly informed by the empirical counts and more sensitive to sampling noise or prior influence, causing large deviations in estimated detection probability. Once validation effort increased to 75 samples per bin, this overestimation diminished sharply, matching the expectation that additional validation support stabilises the model.

A similar, though more subtle, pattern appeared in species with moderate detectability. Small upward biases at low validation effort disappeared almost completely at 75 samples per bin, illustrating how the study-level framework naturally converges toward unbiased estimation when bins are adequately supported. A notable exception was the Abyssinian/Montane Nightjar, which showed almost no bias across all configurations. This behaviour is consistent with the species' densely populated distribution of high-confidence logit scores (see Figure 8), which ensured stable bin populations even at 25 samples per bin (see Figures 20, 21, and 22).

These findings collectively reinforce the strength of the study-level calibration architecture. By pooling all recordings across both sampling-rate strata before constructing bins and estimating per-bin detection probabilities, the study-level estimator benefits from larger effective sample sizes and more complete representations of the score distribution. As a result, the estimator is less sensitive to local irregularities in confidence scores, less affected by sparsely

populated bins, and more resilient to the distributional heterogeneity that arises when strata differ in recording conditions or annotation density. The study-level estimator therefore provides a consistent and reliable summary of call density that is robust to both ecological and methodological variation across recordings.

Strategies

Strategy 1 and Strategy 2 were designed to respond differently to distributional shifts across strata, yet the empirical results demonstrated that these distinctions did not materialise in practice. Across all species and sampling rates, the two strategies produced estimates that were either indistinguishable or differed only trivially from one another. This indicates that the mechanisms intended to detect and correct for strata-specific variation, such as reweighting and mixture-matching, did not meaningfully influence the final density estimates. In other words, although the strategies were theoretically constructed to adapt to changes in strata-level score distributions, the data did not contain enough structure or separation across strata for these mechanisms to activate in a meaningful way.

More importantly, the strategies failed to deliver accurate strata-level density estimates. Even in cases with clear differences between the 32 kHz and 48 kHz score distributions, both Strategy 1 and Strategy 2 produced estimates that either mirrored the study-level estimator without gaining sensitivity or deviated in ways that did not correspond to recognisable ecological or acoustic patterns. The lack of operational differentiation indicates that the strategies were not sufficiently informed by the data once the validation set was partitioned by sampling rate. The resulting sparsity appeared to overwhelm the strategy mechanisms, preventing them from isolating genuine strata-specific patterns.

Strategy 3, defined as the geometric mean of the other two strategies, simply produced a

midpoint between their outputs. Because Strategies 1 and 2 shared similar errors in direction and magnitude, Strategy 3 inherited these tendencies rather than correcting them. While its outputs sometimes appeared smoother or more moderate, this characteristic was not indicative of better performance but rather of averaging two similarly misaligned estimates. As such, Strategy 3 did not provide a meaningful improvement over either of its parent strategies.

Across all strategies, the choice of validation sample size (25, 50, or 75) did not lead to systematic differences in performance. Unlike the study-level estimator, which improved predictably as validation effort increased, the strategy-based estimators showed no consistent pattern linking sample size to performance. This suggests that, once the validation dataset was subdivided by sampling rate, the remaining annotated examples were insufficient to support stable per-bin estimation, regardless of the initial validation effort. A contributing factor is the influx of negative annotations introduced during the data reconstruction step described in the *Data Re-adjustment* section. The resulting imbalance between positive and negative examples likely overwhelmed the strategy mechanisms, further reducing their ability to differentiate strata.

Strata levels

Sampling-rate effects were most pronounced in the strategy-level estimates. Across species and binning schemes, the 32 kHz stratum consistently produced lower bias than the 48 kHz stratum. This difference is largely attributable to annotation density. The 48 kHz recordings contained far fewer validated examples, meaning that once the study-level bins were split by strata, many bins in the 48 kHz subset became sparsely populated. These sparsely supported bins produced unstable per-bin detection estimates and inflated bias values, leading to systematic underestimation for some species and overestimation for others. This pattern is clearly illustrated in Tables 1 to 5.

The sparsity of high-confidence bins in the 48 kHz stratum is particularly influential. Species such as the Baglafecht weaver and White-browed coucal exhibited extreme levels of misspecification at the strata level, with relative bias reaching 115% for the Baglafecht weaver and over 250% for the White-browed coucal under certain configurations. These effects reflect a fundamental mismatch between the resolution of the quantile-binning procedure and the available annotation density within each stratum. Without sufficient positive annotations in the high-confidence regions of the score distribution, the strategies were unable to reliably estimate detection probabilities, leading to extreme systematic errors.

The Abyssinian/Montane Nightjar illustrates this dynamic in a more controlled setting. Its strata-level bias patterns consistently diverged in opposite directions, with the 48 kHz stratum tending toward overestimation and the 32 kHz stratum tending toward underestimation. As validation effort increased, both effects diminished but did not disappear entirely. Similar patterns appeared for the African Pipit, with the 48 kHz strata-level estimates consistently falling below the study-level values even as sample size increased. These divergences likely reflect sampling variability combined with the broader distribution of informative examples in the 32 kHz data, which better supported stable estimation than the more sparsely annotated 48 kHz data.

Ecological expectations

Using the study-level estimator, which consistently delivered the most accurate and stable results, to evaluate ecological plausibility, it becomes clear that call density estimates for all species were substantially lower than the values reported by Navine et al. (2024). This discrepancy stems from the extensive data recovery procedure described in the *Data Readjustment* section. Large portions of the audio dataset were missing annotated labels and had to be reconstructed statistically. As these reconstructions were based on probabilistic

approximations rather than direct observations, the resulting dataset likely underestimated true vocal activity. Additionally, the recordings originally removed by *Natural State* could not be validated independently and may have contained important vocalisations that the reconstruction process failed to recover. The assignment of potentially unknown validation labels through the simulation procedure further amplified this problem by inflating the number of negative annotations, making reliable ecological inference even more difficult. As a result, the reconstructed dataset may not fully represent the true acoustic landscape of the study sites, leading to structurally deflated density estimates. These constraints limit the ecological interpretability of the results, even though the modelling framework itself performed well when sufficient validation support was available.

Conclusion

This study evaluated the applicability of the threshold-free call density framework proposed by Navine et al. (2024) within an African context, leveraging passive acoustic data from five focal bird species in the Mount Kenya region. By implementing quantile-based binning, Beta–Binomial modelling, and multiple strata-level estimation strategies, it addressed three primary questions which were, to assess whether call density can be reliably estimated without reliance on arbitrary score thresholds, to assess whether strata-level strategies can accurately estimate call density in the presence of distribution shifts due to differing acoustic conditions and sampling rates, and to assess whether the resulting call density estimates align with ecological expectations for vocal bird species being assessed.

The results provide a clear and consistent answer to the first question. Study-level call density estimation is robust and accurate across species, binning schemes, and validation sample sizes. By pooling all recordings prior to calibration, bins remained well-supported by anno-

tated data, and the Beta–Binomial models operated within their intended statistical regime. Despite some overestimation under sparse validation conditions for certain species, such as the Baglafecht weaver, increasing per-bin sample sizes helped correct for these discrepancies. This demonstrates the method's sensitivity to data sufficiency rather than structural instability.

In contrast, the second question yielded negative results. None of the strata-level strategies reliably adjusted for distribution shifts, nor did they behave in line with theoretical expectations. Strategies 1 and 2 produced estimates that were nearly indistinguishable from one another and often diverged substantially from both study-level estimates and ecological plausibility. Strategy 3, designed as a geometric-mean compromise between the two, simply inherited their shared biases. Across species and strata, especially at 48 kHz, call density estimates exhibited large and inconsistent biases, suggesting that neither framework assumptions nor data conditions supported effective strata-level inference. These shortcomings were likely exacerbated by the dominance of synthetic negative labels introduced during the data recovery process, which diluted the ability of strategies to engage with meaningful variation in strata-specific call activity.

Ecological inference: The study-level estimates, where the model is best supported, provide reasonable insight into relative differences in acoustic activity among the five focal species. For example, the Abyssinian/Montane Nightjar exhibited consistently higher call density estimates than the African Pipit, reflecting known differences in vocal behaviour and detectability. However, the absolute densities inferred in this study were markedly lower than those reported by Navine et al. (2024), and likely do not reflect true ecological conditions. This discrepancy was driven by the limitations of the reconstructed dataset, including the re-introduction of negative labels. As a result, while relative comparisons among species may retain ecological signal, absolute estimates cannot be interpreted as reliable indicators

of biological activity. Strata-level estimates, in turn, did not correspond with any plausible ecological patterns such as nocturnal vs. diurnal activity differences or habitat-derived shifts. Underscoring that the current strategies failed to capture spatial or acoustic heterogeneity. These findings collectively highlight that threshold-free call density estimation can support ecological inference when data coverage and score structure are adequate, but that strata-level approaches require significant methodological refinement before they can be reliably used for partitioned ecological insight in the African landscape.

Taken together, the results demonstrate that threshold-free call density estimation is viable and informative at the study level, but current strategy-based approaches are inadequate for strata-level inference under distribution shifts, especially in regions where bio-acoustic classifiers remain under-trained. For passive acoustic monitoring programmes aiming to produce robust, partitioned estimates of acoustic activity across space, time, or sensor types, it is therefore critical to either ensure sufficient calibration data within each stratum or develop more flexible modelling strategies such as utilising data-driven adaptive binning and only then may possible improvements in the methods/ techniques used come about.

Future Work

This study highlights key limitations in strata-level call density estimation under distribution shifts from the study level, suggesting several areas for improvement. Future work should explore hierarchical extensions to the Beta–Binomial model that allow partial pooling across strata, enabling more stable estimates even under sparse validation data. The use of different strata such as geographical site could also yield improvement in the analysis and prove to be a valid stratification technique and is therefore an area to explore in future work. Additionally, developing adaptive or data-driven binning strategies could mitigate the sensitivity to skewed score distributions observed in this dataset.

Finally, enhancing annotation efficiency through active learning or selectively increasing validation effort in high-uncertainty regions would strengthen the calibration underpinning estimator performance.

Appendix

Outlined in this appendix is the standard errors associated with the error bars on the relative bias plots in the *Results* Section. The are organised in tables of validation sample size for every species analysed in this study.

Baglafecht weaver

25 samples per bin

Level	**SD (32 kHz)**	**SD (48 kHz)**	**SD (Combined)**
Study Level	0.073	0.073	0.067
Strategy 1	0.118	0.044	0.083
Strategy 2	0.119	0.041	0.088
Strategy 3	0.119	0.042	0.089

Table 21: Standard deviations of relative bias for the Baglafecht weaver at 25 validated samples per bin.

Baglafecht weaver (25 samples)	0.012

Table 22: Combined empirical standard deviation for Baglafecht weaver using 25 validation samples.

50 samples per bin

Level	**SD (32 kHz)**	**SD (48 kHz)**	**SD (Combined)**
Study Level	0.0306	0.0306	0.0283
Strategy 1	0.5435	0.3571	0.4258
Strategy 2	0.5457	0.3524	0.4252
Strategy 3	0.5446	0.3548	0.4255

Table 23: Standard deviations of relative bias for the Baglafecht weaver at 50 validated samples per bin.

Baglafecht weaver (50 samples)	0.010

Table 24: Combined empirical standard deviation for Baglafecht weaver using 50 validation samples.

75 samples per bin

Table 25: Standard deviations of relative bias for the Baglafecht weaver at 75 validated samples per bin.

Level	**SD (32 kHz)**	**SD (48 kHz)**	**SD (Combined)**
Study Level	0.028	0.028	0.026
Strategy 1	0.253	0.286	0.251
Strategy 2	0.259	0.288	0.255
Strategy 3	0.256	0.287	0.253

Baglafecht weaver (75 samples)	0.008

Table 26: Combined empirical standard deviation for Baglafecht weaver using 75 validation samples.

White browed coucal

25 samples per bin

Level	SD (32 kHz)	SD (48 kHz)	SD (Combined)
Study Level	0.016	0.016	0.015
Strategy 1	0.634	1.126	0.967
Strategy 2	0.635	1.123	0.965
Strategy 3	0.634	1.125	0.966

Table 27: Standard deviations of relative bias for the White-browed coucal at 25 validated samples per bin.

White-browed coucal (25 samples)	0.015

Table 28: Combined empirical standard deviation for White-browed coucal using 25 validation samples.

50 samples per bin

Level	SD (32 kHz)	SD (48 kHz)	SD (Combined)
Study Level	0.018	0.018	0.017
Strategy 1	0.231	0.383	0.491
Strategy 2	0.230	0.385	0.491
Strategy 3	0.231	0.384	0.491

Table 29: Standard deviations of relative bias for the White-browed coucal at 50 validated samples per bin.

White-browed coucal (50 samples)	0.011

Table 30: Combined empirical standard deviation for White-browed coucal using 50 validation samples.

75 samples per bin

Level	**SD (32 kHz)**	**SD (48 kHz)**	**SD (Combined)**
Study Level	0.026	0.026	0.024
Strategy 1	0.292	0.617	0.634
Strategy 2	0.291	0.617	0.632
Strategy 3	0.292	0.617	0.633

Table 31: Standard deviations of relative bias for the White-browed coucal at 75 validated samples per bin.

White-browed coucal (75 samples)	0.009

Table 32: Combined empirical standard deviation for White-browed coucal using 75 validation samples.

African Gray Fly catcher

25 samples per bin

Level	SD (32 kHz)	SD (48 kHz)	SD (Combined)
Study Level	0.007	0.007	0.006
Strategy 1	0.346	0.444	0.371
Strategy 2	0.346	0.444	0.371
Strategy 3	0.346	0.444	0.371

Table 33: Standard deviations of relative bias for the African Gray Flycatcher at 25 validated samples per bin.

African Gray Flycatcher (25 samples)	0.020

Table 34: Combined empirical standard deviation for African Gray Flycatcher using 25 validation samples.

50 samples per bin

Table 35: Standard deviations of relative bias for the African Gray Flycatcher at 50 validated samples per bin.

Level	SD (32 kHz)	SD (48 kHz)	SD (Combined)
Study Level	0.019	0.019	0.017
Strategy 1	0.204	0.207	0.215
Strategy 2	0.203	0.208	0.214
Strategy 3	0.203	0.208	0.215

Table 36: Standard deviations of relative bias for the African Gray Flycatcher at 50 validated samples per bin.

African Gray Flycatcher (50 samples)	0.014

Table 37: Combined empirical standard deviation for African Gray Flycatcher using 50 validation samples.

75 samples per bin

Level	**SD (32 kHz)**	**SD (48 kHz)**	**SD (Combined)**
Study Level	0.019	0.019	0.017
Strategy 1	0.165	0.111	0.171
Strategy 2	0.165	0.112	0.171
Strategy 3	0.165	0.112	0.171

Table 38: Standard deviations of relative bias for the African Gray Flycatcher 75 samples.

African Gray Flycatcher (75 samples)	0.013

Table 39: Combined empirical standard deviation for African Gray Flycatcher using 75 validation samples.

Abyssinian/Montane Nightjar

25 samples per bin

Level	SD (32 kHz)	SD (48 kHz)	SD (Combined)
Study Level	0.002	0.002	0.002
Strategy 1	0.026	0.065	0.129
Strategy 2	0.026	0.065	0.129
Strategy 3	0.026	0.065	0.129

Table 40: Standard deviations of relative bias for the Abyssinian Nightjar at 25 validated samples per bin.

Abyssinian Nightjar (25 samples)	0.021

Table 41: Combined empirical standard deviation for Abyssinian Nightjar using 25 validation samples.

50 samples per bin

Level	SD (32 kHz)	SD (48 kHz)	SD (Combined)
Study Level	0.002	0.002	0.002
Strategy 1	0.013	0.039	0.131
Strategy 2	0.013	0.040	0.131
Strategy 3	0.013	0.039	0.131

Table 42: Standard deviations of relative bias for the Abyssinian Nightjar at 50 validated samples per bin.

Abyssinian Nightjar (50 samples)	0.017

Table 43: Combined empirical standard deviation for Abyssinian Nightjar using 50 validation samples.

75 samples per bin

Level	**SD (32 kHz)**	**SD (48 kHz)**	**SD (Combined)**
Study Level	0.000	0.000	0.000
Strategy 1	0.014	0.036	0.132
Strategy 2	0.014	0.036	0.132
Strategy 3	0.014	0.036	0.132

Table 44: Standard deviations of relative bias for the Abyssinian Nightjar at 75 validated samples per bin.

Abyssinian Nightjar (75 samples)	0.013

Table 45: Combined empirical standard deviation for Abyssinian Nightjar using 75 validation samples.

African Pipit

25 samples per bin

Level	SD (32 kHz)	SD (48 kHz)	SD (Combined)
Study Level	0.018	0.018	0.017
Strategy 1	0.222	0.105	0.347
Strategy 2	0.223	0.105	0.346
Strategy 3	0.222	0.105	0.347

Table 46: Standard deviations of relative bias for the African Pipit at 25 validated samples per bin.

African Pipit (25 samples)	0.022

Table 47: Combined empirical standard deviation for African Pipit using 25 validation samples.

50 samples per bin

Level	SD (32 kHz)	SD (48 kHz)	SD (Combined)
Study Level	0.014	0.014	0.013
Strategy 1	0.074	0.032	0.293
Strategy 2	0.074	0.031	0.292
Strategy 3	0.074	0.031	0.293

Table 48: Standard deviations of relative bias for the African Pipit at 50 validated samples per bin.

African Pipit (50 samples)	0.016

Table 49: Combined empirical standard deviation for African Pipit using 50 validation samples.

75 samples per bin

Level	**SD (32 kHz)**	**SD (48 kHz)**	**SD (Combined)**
Study Level	0.038	0.038	0.035
Strategy 1	0.078	0.027	0.368
Strategy 2	0.079	0.027	0.368
Strategy 3	0.079	0.027	0.368

Table 50: Standard deviations of relative bias for the African Pipit at 75 validated samples per bin.

African Pipit (75 samples)	0.013

Table 51: Combined empirical standard deviation for African Pipit using 75 validation samples.

References

Cleere, Nigel, Guy M. Kirwan, and Peter F. D. Boesman (2022). *Montane Nightjar (Caprimulgus poliocephalus)*. Version 1.0. In *Birds of the World* (B. K. Keeney, Editor). Cornell Lab of Ornithology, Ithaca, NY, USA. DOI: 10.2173/bow.monnig1.01. URL: https://birdsoftheworld.org/bow/species/monnig1/1.0/introduction (visited on 10/31/2025).

Craig, Adrian J.F. (2020). *Baglafecht Weaver (*Ploceus baglafecht*)*. BirdLife International factsheet, based on the IUCN Red List of Threatened Species. Status: Least Concern. URL: https://datazone.birdlife.org/species/factsheet/baglafecht-weaver-ploceus-baglafecht (visited on 10/11/2025).

Dunn, Erica H. and C. John Ralph (2002). "Use of mist nets as a tool for bird population monitoring". In: *Studies in Avian Biology* 29, pp. 1–6.

Eastern Ecological Science Center (2025). *Training: BBS Methodology - Point Counts*. Training manual. U.S. Geological Survey, Eastern Ecological Science Center. Laurel, Maryland, United States of America.

Gregory, Richard D., David W. Gibbons, and Paul F. Donald (2004). "Bird census and survey techniques". In: *Bird Ecology and Conservation: A Handbook of Techniques*. Ed. by William J. Sutherland, Ian Newton, and Rhys E. Green. Oxford, UK: Oxford University Press, pp. 17–56. ISBN: 9780198520863.

Kgafela, Selaelo (2025). *GitHub repository of MrSelaeloK*. https://github.com/MrSelaeloK. Accessed on 2025-11-19.

Kirwan, Guy M. et al. (2024). *White-browed Coucal (*Centropus superciliosus*)*. In *Birds of the World*, eds. J. del Hoyo, A. Elliott, J. Sargatal, D. A. Christie, and E. de Juana. Cornell Lab of Ornithology, Ithaca, NY. Version 1.2 published October 22, 2024. URL: https://birdsoftheworld.org/bow/species/whbcou2/1.2/introduction (visited on 10/11/2025).

Knight, E. C. et al. (2017). "Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs". In: *Avian Conservation and Ecology* 12.2, p. 14. DOI: 10.5751/ACE-01114-120214. URL: https://www.ace-eco.org/vol12/iss2/art14/.

Knight, Emily C. et al. (2020). "Validation prediction: a flexible protocol to increase efficiency of automated acoustic processing for wildlife research". In: *Ecological Applications* 30.4, e02140. DOI: 10.1002/eap.2140.

Krebs, Charles J. (1989). *Ecological Methodology*. New York, USA: Harper & Row.

McGinn, Kate et al. (2023). "Feature embeddings from the BirdNET algorithm provide insights into avian ecology". In: *Ecological Informatics* 74, p. 101995. DOI: 10.1016/j.ecoinf.2022.101995. URL: https://www.sciencedirect.com/science/article/abs/pii/S1574954123000249.

Morelli, Federico et al. (2022). "Detection Rate of Bird Species and What It Depends on: Tips for Field Surveys". In: *Frontiers in Ecology and Evolution* 9, p. 671492. DOI: 10.3389/fevo.2021.671492. URL: https://www.frontiersin.org/articles/10.3389/fevo.2021.671492/full.

Navine, Carl et al. (2024). "All thresholds barred: Direct estimation of call density in bioacoustic data". In: *Frontiers in Bird Science* 3. Advocates direct estimation of call density as a biologically meaningful measure linking occupancy, abundance, and activity in passive acoustic monitoring., p. 1380636. DOI: 10.3389/fbirs.2024.1380636. URL: https://www.frontiersin.org/articles/10.3389/fbirs.2024.1380636/full.

Nawa, Victor and Saralees Nadarajah (2024). "Exact Expressions for Kullback–Leibler Divergence for Multivariate and Matrix-Variate Distributions". In: *Entropy* 26.8, p. 663. DOI: 10.3390/e26080663. URL: https://doi.org/10.3390/e26080663.

Oiseaux.net (2025). *Montane Nightjar – Caprimulgus poliocephalus*. Accessed 31 October 2025. URL: https://www.oiseaux.net/birds/montane.nightjar.html.

Oschadleus, Dieter (Apr. 2024). *Baglafecht Weaver (*Ploceus baglafecht*)*. Macaulay Library, Cornell Lab of Ornithology. Male, nominate race. Addis Ababa, Ethiopia. ML617262232, eBird S168209128. URL: https://birds4africa.org/weaver-research/weavers/476a/.

Pérez-Granados, Carlos, Juan Traba, et al. (2021). "Review on estimating bird density using acoustic monitoring". In: *Ecological Indicators* 121, p. 106977. DOI: 10.1016/j.ecolind.2020.106977.

Pérez-Granados, Cristian (2023). "BirdNET: applications, performance, pitfalls and future opportunities". In: *Ibis* 165.3, pp. 1068–1075. DOI: 10.1111/ibi.13193. URL: https://onlinelibrary.wiley.com/doi/full/10.1111/ibi.13193?utm_source=chatgpt.com.

Priyadarshani, Nisansala, Stephen Marsland, and Isabel Castro (2018). "Automated birdsong recognition in complex acoustic environments: a review". In: *Journal of Avian Biology* 49.5, jav–01447. DOI: 10.1111/jav.01447. URL: https://onlinelibrary.wiley.com/doi/full/10.1111/jav.01447.

Scarpelli, Mariana D. et al. (2019). "Gaps in terrestrial soundscape research: it's time to focus on tropical wildlife". In: *Science of the Total Environment* 707, p. 135403. DOI: 10.1016/j.scitotenv.2019.135403. URL: https://www.sciencedirect.com/science/article/abs/pii/S0048969719353963.

Shonfield, John and Erin M. Bayne (2017). "Autonomous recording units in avian ecological research: current use and future applications". In: *Avian Conservation and Ecology* 12.1, p. 14. DOI: 10.5751/ACE-00974-120114.

Sugai, Larissa Sayuri Moreira et al. (2019). "Terrestrial passive acoustic monitoring: Review and perspectives". In: *BioScience* 69.1, pp. 15–25. DOI: 10.1093/biosci/biy147. URL: https://academic.oup.com/bioscience/article/69/1/15/5193506.

Taylor, Barry (2020). *African Gray Flycatcher (*Bradornis microrhynchus*)*. Version 1.0 — Published March 4, 2020. Text last updated January 24, 2013. Status: Least Concern. The

IUCN Red List of Threatened Species. URL: https://www.iucnredlist.org/species (visited on 10/11/2025).

Teichmann, Holger (Jan. 16, 2019). *African Pipit (African) (*Anthus cinnamomeus *[cinnamomeus Group])*. Photographed in Yabello area, Oromia, Ethiopia. eBird Checklist S64974455. Accessed 2025-10-11. Macaulay Library, Cornell Lab of Ornithology. URL: https://macaulaylibrary.org/asset/204308481.

Urbanič, Gorazd et al. (2022). "Riparian Zones—From Policy Neglected to Policy Integrated". In: *Frontiers in Environmental Science* 10, p. 868527. DOI: 10.3389/fenvs.2022.868527. URL: https://www.frontiersin.org/articles/10.3389/fenvs.2022.868527/full.

Wood, C. M. and S. Kahl (2024). "Guidelines for interpreting BirdNET scores in ecological studies". In: *Ecological Informatics* 80, p. 102387. DOI: 10.1016/j.ecoinf.2024.102387. URL: https://connormwood.com/wp-content/uploads/2024/02/wood-kahl-2024-guidelines-for-birdnet-scores.pdf.

Wood, Connor M. and M. Zachariah Peery (2022). "What does 'occupancy' mean in passive acoustic surveys?" In: *Ibis* 164.4, pp. 1295–1300. DOI: 10.1111/ibi.13097.

Wood, Connor M. et al. (2019). "Detecting small changes in populations at landscape scales: a bioacoustic site-occupancy framework". In: *Ecological Indicators* 98, pp. 492–507. DOI: 10.1016/j.ecolind.2018.10.027. URL: https://connormwood.com/wp-content/uploads/2021/02/wood.etal-2019-detecting-small-pop-changes-at-landscape-scales.pdf.

Wood, Connor M. et al. (2022). "The machine learning–powered BirdNET App reduces barriers to global bird research by enabling citizen science participation". In: *PLOS Biology* 20.6, e3001670. DOI: 10.1371/journal.pbio.3001670. URL: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3001670.

Yoccoz, Nigel G. (2012). *Ecological Monitoring*. Tromsø, Norway: University of Tromsø.

Zhang, Yuzhe et al. (2021). "On the Properties of Kullback–Leibler Divergence Between Multivariate Gaussian Distributions". In: *arXiv preprint arXiv:2102.05485*. URL: https://arxiv.org/abs/2102.05485.