8-1-2024

Passive acoustic monitoring: Considerations for recording units, BirdNET settings, and filtering methods for long-term avian population monitoring

Shasta Corvus
Southern Illinois University Carbondale, shastacorvus@gmail.com

PASSIVE ACOUSTIC MONITORING: CONSIDERATIONS FOR RECORDING UNITS, BIRDNET SETTINGS, AND FILTERING METHODS FOR LONG-TERM AVIAN POPULATION MONITORING

by

Shasta S. W. Corvus

B.S., Shawnee State University, 2019

A Thesis
Submitted in Partial Fulfillment of the Requirements for the
Master of Science Degree

School of Forestry and Horticulture
in the Graduate School
Southern Illinois University Carbondale
August 2024

THESIS APPROVAL

PASSIVE ACOUSTIC MONITORING: CONSIDERATIONS FOR RECORDING UNITS, BIRDNET SETTINGS, AND FILTERING METHODS FOR LONG-TERM AVIAN POPULATION MONITORING

by

Shasta S. W. Corvus

A Thesis Submitted in Partial

Fulfillment of the Requirements

for the Degree of

Master of Science

in the field of Forestry

Approved by:

Dr. Brent Pease

Dr. Eric Holzmueller

Dr. Charles Ruffner

Graduate School
Southern Illinois University Carbondale
April 9, 2024

AN ABSTRACT OF THE THESIS OF

Shasta S. W. Corvus, for the Master of Science degree in Forestry, presented on April 9, 2024, at Southern Illinois University Carbondale.

TITLE: PASSIVE ACOUSTIC MONITORING: CONSIDERATIONS FOR RECORDING UNITS, BIRDNET SETTINGS, AND FILTERING METHODS FOR LONG-TERM AVIAN POPULATION MONITORING

MAJOR PROFESSOR: Dr. Brent Pease

This research investigated several aspects of passive acoustic monitoring (PAM) which were previously unexplored or understudied. A comparison of autonomous recording units (ARUs) for use with BirdNET for the purpose of bird monitoring was conducted. Four ARUs were compared, including AudioMoth, SM4, SMMicro, and SwiftOne. We found that, of the performance metrics for which ARU choice made a statistically significant difference ($P>0.01$), which included sensitivity, specificity, F1 harmonic mean, and Matthews Correlation Coefficient, (but not precision: $P = 0.94$), AudioMoth had the best performance for all statistically significant performance metrics except for specificity, for which SMMicro had the highest. The same audio was then processed using 18 combinations of Overlap and Sensitivity, including default settings. We found that Overlap and Sensitivity values were highly significant ($P>0.001$) for all performance metrics: precision, sensitivity, specificity, F1 harmonic mean, and Matthews Correlation Coefficient. No individual Overlap-Sensitivity setting combination performed outperformed others in most of the performance metrics; however, in general, as Overlap or Sensitivity increased, the number of true and false positive species reports increased while the number of false negatives decreased. Four confidence-based threshold types were then used to filter BirdNET output to compare threshold performances, comparing two arbitrary thresholds and two species-specific thresholds which were calculated using manual validation data. Of the thresholds tested, one of the arbitrary threshold types and one of the species-specific threshold types achieved a precision ≥ 0.95. We hope this research will help guide PAM decisions regarding ARU choice, BirdNET settings, and threshold type choice.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Dr. Brent Pease, who has given me this amazing opportunity and has been such a patient and supportive mentor. I would like to thank my committee members, Dr. Charles Ruffner and Dr. Eric Holzmueller, for their time and expertise. I would also like to give thanks to my undergraduate professors, Dr. Sarah Ivers, the late Dr. Jon Bedick, and Dr. Logan Minter, for inspiring me to pursue my passion for ecology.

I would like to express my sincere gratitude towards the landowners who allowed us access to their property for our bird monitoring efforts, as this work would have been impossible without them. I would like to thank my lab-mate Elaine Metz, as well as our technicians, Kelsee Dodd, Zach Skubiszewski, and Madison Johnson, for their help in the field and in the lab. I would like to thank all the volunteers of both the Illinois Breeding Bird Atlas and the North American Breeding Bird Survey, as their hard work allowed us to pull from external sources to better understand long-term avian population trends.

Finally, I would like to give thanks to my grandpa and foster father, Ray Radabaugh Sr., who not only gave me a chance at life, but who has served as my role model for hard work and perseverance. I would not be here today if not for you.

TABLE OF CONTENTS

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

CHAPTER 1

THE EFFECTS OF AUTONOMOUS RECORDING UNIT CHOICE AND BIRDNET-ANALYZER SETTINGS ON BIRDNET PERFORMANCE

INTRODUCTION

Point counts have historically been the conventional field method for estimating avian population abundance and distribution, assessing species-environmental relationships, and understanding drivers of population change (Bibby et al., 2000). However, with the advent of autonomous recording units (ARUs) and rapidly developing automatic species classifiers, such as BirdNET Analyzer, many researchers have switched to using Passive Acoustic Monitoring (PAM) as an alternative approach to traditional point count methods (Kahl et al., 2021; Sugai et al., 2021; Shonfield et al., 2017). Recent research comparing traditional avian monitoring techniques with ARUs has documented that ARUs can be more cost-effective, allow for greater duration of observation, and reduce observer bias, among other advantages (Alquezar & Machado, 2015; Klingbeil & Willig, 2015; Hobson et al. 2002; Tegeler, Morrison & Szewczak, 2012; Shonfield & Bayne, 2017). Despite these benefits, there are also potential drawbacks to PAM including limited detection distance of ARUs and data processing challenges (Hutto & Stutzman, 2009; Yip et al., 2017).

Automatic species classifiers offer a promising path forward to alleviate the data processing bottleneck of acoustic recordings, as terabytes of data can be collected during a sampling period. Though promising, even the most reliable open-access bird sound recognizer currently available for general use, BirdNET Analyzer (herein, BirdNET), has an average precision of only 72—85% (percent of detections correctly classified) and a recall rate usually ranging from 33—84% (percent of target species vocalizations detected), highlighting the continued issue of false positive and false negative errors (Pérez-Granados, 2023; Kahl et al., 2021; Wood et al., 2022). This problem is further exacerbated in that precision may vary widely depending on the species being examined and the environmental conditions at the sampling location (Sethi et al., 2023). Although BirdNET has many advantages including its open-accessibility, its capability of identifying more bird species than any other open-access classifier, as well as its ability to exceed other classifiers in average precision, it is nevertheless prone to producing both false positive species reports and false negative species errors (Lauha et al., 2021; LeBien et al., 2020; Ruff et al., 2020; Kahl et al., 2021; Sethi et al., 2023). Consequently, identifying streamlined ways to minimize these error rates is an active area of research (e.g., Cole et al. 2022). Further complicating the issue of the measurement error in species classifiers is the growing range of sound recording hardware (i.e., ARUs) available for ecological studies (Toenies & Rich, 2021).

Currently, many ARU devices are available for ecological studies, yet there is no consensus as to which ARU device produces the highest quality recordings for the purpose of using the audio obtained for analysis via BirdNET. Commonly used ARUs for avian PAM include the AudioMoth (Hill et al., 2018), SwiftOne (Cornell Lab or Ornithology, 2023a), and a range of products from Wildlife Acoustics (Wildlife Acoustics, 2023). Although an array of available ARU devices exist, quantitative, empirical comparisons of the impact that ARU choice has on BirdNET output are lacking across a range of ecosystems and recording conditions. Using BirdNET, Toenies and Rich (2021) documented variation in the mean number of species reported by the Swift Recorder, AudioMoth, SMMini, and SM3BAT. In addition, many of the recorders evaluated by Toenies and Rich (2021) are now “outdated,” as newer models have replaced most of the tested units. Furthermore, the SMMicro, SM4, and SwiftOne—all newer versions of the units examined by Toenies and Rich—have yet to be assessed. Furthermore, given that Toenies and Rich (2021) performed their test in Monterey County, CA—a location which has a unique avian species composition that differs vastly from most of the United States and North America—it is important that additional research be conducted to ascertain whether their results hold true for different species and environmental conditions.

Along with hardware comparison, little guidance on optimal BirdNET settings exists in the literature. Although research on mitigating false positive errors is ongoing, finding ways to mitigate false negative errors when using BirdNET has received less attention, but could potentially be remediated using features currently available in BirdNET. Specifically, BirdNET offers two fine-tuning parameters in the classifier—Overlap and Sensitivity—that could minimize false negative errors. BirdNET's Overlap setting controls the temporal overlap of prediction segments, which defaults to 0s but could be adjusted up to 2.9s (Kahl et al., 2021). Modifying the Overlap setting could increase detection resolution—particularly in cases in which there are multiple species vocalizing simultaneously (e.g., during a dawn chorus; Kahl et al., 2021). The Sensitivity setting aims to control *detection sensitivity,* which can be thought of as the prediction sensitivity, meaning that vocalizations (or potentially noise in general) captured on the recording will be more likely to lead to a species prediction than it otherwise would be with a lower Sensitivity setting (Kahl et al., 2021a). The impacts of adjusting the Overlap and Sensitivity settings are yet to be explored, as there currently exists no publications which investigate the effects of altering the Overlap and Sensitivity on the performance of BirdNET (Pérez-Granados, 2023). It is possible that adjustments made to the Overlap and Sensitivity settings may reduce false negative species rates, although the affect this may also have on the number of false positive species reported is likewise underexplored, highlighting an important yet uninvestigated aspect of BirdNET.

Collectively, two outstanding albeit understudied issues regarding the use of ARUs and BirdNET exist: 1.) uncertainty as to how ARU device quality and functionality impacts BirdNET performance, and 2.) the ability of BirdNET's Overlap and Sensitivity settings to increase the number of true positive species reported, thereby decreasing the number of false negative species. Here, I describe a field trial comparing the output of BirdNET across 4 leading ARU devices, and a systematic analysis of various Overlap and Sensitivity setting combinations in the BirdNET software. Specifically, I 1.) compared the mean number of true positive, false positive, and false negative species between 4 ARUs: SwiftOne, AudioMoth, SM4, and SMMicro; and 2.) compared the effects of 18 Overlap and Sensitivity settings on BirdNET's species report, comparing the average numbers of true positive and false positive species reported between setting combinations using audio collected from a single device type (AudioMoth), further comparing unfiltered output versus output filtered using a confidence-based threshold.

METHODS

To test the effect of ARU quality on BirdNET performance, I deployed 4 ARU devices across six sampling locations for six consecutive days, each set to record 30-min sampling periods per visit. A single site visit consisted of deploying all 4 ARUs (SwiftOne, AudioMoth, SMMicro, and SM4) and concurrently conducting a point count for the duration of the ARU recording period. The point counts consisted of recording all bird species detected during the 30-minute period and served as "truth" in the experiment. Although sight-only observations (e.g., flyovers) were noted, they were not included in the analysis. The selection of these 4 ARUs was guided by specific criteria: product versioning, targeted use and functionality, and cost effectiveness. The SM4 was chosen due to its status as the latest iteration of the SM3, a device previously examined by Toenies and Rich (2021). Similarly, the SwiftOne, being the updated version of the Swift Recorder, was also tested. The selection of the SM Micro over the SM Mini was based on cost-effectiveness, being the most economical option available from Wildlife Acoustics. Furthermore, the SM Micro's comparable size to the AudioMoth positioned it as a nominal competitor in both functionality and price point (Wildlife Acoustics, 2023). Finally, I included the AudioMoth due to its significance as the most economical unit at the time of writing; this choice aims to provide valuable insights into how the AudioMoth, with its competitive pricing, performs in comparison to other contemporary ARUs (Hill et al., 2018).

For deployment, the 4 ARUs were distributed equally around a center point every 90 degrees, resulting in approximately 0.5 m from neighboring units. Each ARU was fastened 1 m above ground to a 1.5 m t-post using a zip tie. The units were arranged facing outward from one another but were close enough such that distance and direction from sound had minimal to no impact on species detection. Because all ARUs tested in this study except for the AudioMoth had an external, weatherproof casing and could be zip tied directly to the posts, the AudioMoth needed to be placed in sealable plastic bags alongside a silica desiccant packet prior to deployment; previous research suggests there is little to no loss in performance in this housing (Lapp et al., 2023). Each site was visited six times for 36 total site visits, producing 144 audio recordings (6 sites x 6 days = 36 site visits x 4 recorders = 144 audio recordings). Each recording occurred between the hours of 0500 and 0830, with all recordings beginning on the hour (e.g., 0500) and recording for a total of 30 minutes.

To compare ARU performance, all recordings were processed using BirdNET with default classifier settings except for the use of a custom species list. The custom species list was generated using the R package *rebird* (package version 1.3.0; Maia et al., 2023) in Program R (Version 4.2.3) with the restrictions of using species reported on eBird during the past 10 years from June 21 to July 4, 2023 (the date range of the sampling period) in Jackson County, Illinois, USA (in which all sites were located;). Audio from all 4 units at each site and visit (144 files in total) was processed using this methodology for the ARU comparison, generating 144 unique BirdNET outputs. Point count data was used to verify whether a species on a given BirdNET species output was truly present during the recording period (true positive), not truly present (false positive), or truly present but not reported on BirdNET's output (false negative).

The number of false negative, false positive, and true positive species reported were calculated for each visit individually and then averaged by unit for all 36 visits. The number of true negative species reported were calculated for each visit individually by adding the number of false negative, false positive, and true positive species for that visit and subtracting the sum from the total number of species contained on the custom species list (121), and the true negative values were then averaged by unit for all 36 visits. For each unit, I calculated precision, sensitivity, specificity, F1 harmonic mean, and Matthews correlation coefficient (MCC) using data from all 36 site visits. A mixed model nested ANOVA tested the effect of ARU choice on performance metrics (precision, sensitivity, specificity, F1 harmonic mean, and MCC), with date nested within site surveyed; this model determined if ARU choice affects BirdNET performance metrics while accounting for random variation due to site and date differences.

To assess the impact of adjusting Overlap and Sensitivity settings on BirdNET performance, I isolated recordings from the top-performing ARU in the comparison test (6 sites * 6 visits = 36 recordings) then used 18 combinations of these settings to cover the entire range of setting values. Overlap controls the temporal overlap of prediction segments (0.0 to 2.9, default

0.0; Kahl et al., 2021) and Sensitivity influences whether BirdNET will generate a prediction for an image present on the spectrogram (0.5 to 1.5, default: 1.0; Kahl et al., 2021). I tested Overlap settings 0.0, 0.5, and 1.5 with all Sensitivity settings (0.5, 1.0, 1.5, 2.0, and 2.5) resulting in 18 combinations. Running all 18 combinations on the 36 recordings generated 648 species outputs (18 settings * 36 recordings). The same custom species list used for the ARU comparison tests was also used in this analysis.

The average number of false positive, true positive, false negative, and true negative species from the BirdNET outputs collected from all visits using the AudioMoth were compared between all 18 combinations to generate a confusion matrix heatmap to show the effects of changing these settings on BirdNET output. A three-way ANOVA was used to discern 1.) whether Overlap and Sensitivity settings have a significant effect on the number of true positive species reported per visit, and 2.) whether Overlap and Sensitivity have a significant effect on the number of false positive species reported per visit. The three-way ANOVA method was conducted twice, once for true positive species reports and again for false positive species reports, as changing the Overlap and Sensitivity settings might differently affect the two. Data was aggregated the mean number of false positive or true positive species by visit across all visits. The BirdNET outputs from all site visits obtained using the AudioMoth were used for the three-way ANOVAs testing the effects of Overlap and Sensitivity.

RESULTS

ARU Comparison

A total of 75 species were heard across all site visits during the point counts (Table A.1). Although no species was detected during all 36 points counts, the Common Yellowthroat (*Geothlypis trichas*, n = 35) was detected during more point counts than any other species (Table

A.1). A few species were detected by the observer during the point counts but were not present on any of the BirdNET outputs, including the American Robin (*Turdus migratorius*, n = 3), Barred Owl (*Strix varia*, n = 1), Chimney Swift (*Chaetura pelagica*, n = 1), and Wild Turkey (*Meleagris gallopavo*, n = 1). On the contrary, a few species were detected on all ARU outputs each time they were detected during points counts, including the American Kestrel (*Falco sparverius*, n = 1), Hooded Warbler (*Setophaga citrina*, n = 3), Northern Bobwhite (*Colinus virginianus*, n = 3), and Prothonotary Warbler (*Protonotaria citrea*, n = 1). Additionally, several species were detected as many times as the point counts by at least one ARU, including the American Redstart (*Setophaga ruticilla*, n = 2), Baltimore Oriole (*Icterus galbula*, n = 2), Chipping Sparrow (*Spizella passerina*, n = 2), Green Heron (*Butorides virescens*, n = 4), Hairy Woodpecker (*Leuconotopicus villosus*, n = 2), Kentucky Warbler (*Geothlypis formosa*, n = 15), Orchard Oriole (*Icterus spurius*, n = 14), Ovenbird (*Seiurus aurocapilla*, n = 4), Pileated Woodpecker (*Dryocopus pileatus*, n = 1), Prairie Warbler (*Setophaga discolor*, n = 7), Purple Martin (*Progne subis*, n = 7), Red-headed Woodpecker (*Melanerpes erythrocephalus*, n = 4), Scarlet Tanager (*Piranga olivacea*, n = 2), Warbling Vireo (*Vireo gilvus*, n=9), Yellow-throated Vireo (*Vireo flavifrons*, n=5), and Yellow-throated Warbler (*Setophaga dominica*, n=6). For 59 out of the 75 total species, AudioMoth performed better than or equal to all other ARUs based solely on the number of detections per species for the 36 site visits, using the point count data as truth (Table A.1). Further details regarding the exact species which were heard during points counts, how many visits each species was detected via point counts and for each ARU tested, can be found in Table A.1.

I conducted a comparison between the number of true positive, false positive, and false negative species reported on the unfiltered BirdNET output for each unit for each visit, which revealed the AudioMoth to have the best performance in terms of the highest number of true positive species reported and the lowest number of false negative species errors, but AudioMoth also had the highest number of false positive species reports (Figure 1.1). As for the detection type means—averaged across all site visits—AudioMoth had the greatest mean number of true positives (14.92 species) and the least mean number of false negatives (8.56), though it also had the greatest number of false positives (11.97; Figure 1.1; Table A.2). The SwiftOne consistently ranked second across all detection types while the Wildlife Acoustic devices had the poorest performance across all detection types (Figure 1.1; Table A.2).

The mixed effects nested ANOVA indicated that ARU model choice had a highly significant effect on sensitivity, specificity, F1 harmonic mean, and MCC ($P<0.001$; Table 1). The only performance metric for which ARU choice did not have a significant effect was precision (P-value = 0.97; Table 1). I conducted a comparison between the precision, sensitivity, specificity, F1 harmonic mean, and MCC values for each of the ARUs, examining each performance metric calculated for each visit per ARU (Figure 1.2), as well as examining the means for each ARU—averaged across all site visits (Table A.2). For the performance metric means—which were calculated for each ARU using data for each site visit—the AudioMoth achieved the highest mean sensitivity (0.64), F1 harmonic mean (0.59), and MCC (0.49; Table A.2). The SM4, SwiftOne, and SMMicro were tied for the highest precision (0.57; Table A.2). The SM4 also had the highest specificity (0.92; Table A.2).

Overlap and Sensitivity in BirdNET

Examining the effects of Overlap and Sensitivity on false and true positive species reported per visit, both a three-way randomized block ANOVA for true positives and a separate ANOVA for false positives indicated highly significant effects of Overlap and Sensitivity on the mean number of true positives and false positives, respectively (P<0.001; Table 2). Additionally, the results from both ANOVAs indicated a significant interactive effect between Overlap and Sensitivity (P<0.01; Table 2). Furthermore, the results from the Overlap and Sensitivity tests indicated that increasing the Overlap or Sensitivity settings causes an increase in the total number of reports and the total number of species reported, resulting in an increase in both false positive and true positive species (Figure 1.3).

Although both Overlap and Sensitivity both have highly significant effects on the number of both false positive and true positive reports based on the results of the three-way ANOVA (Table 2), Sensitivity had the greatest impact on the number of species reported (Figure 1.3; Table A.3). Increasing BirdNET's Sensitivity caused the number of false negatives and true negatives to decrease, while causing both the number of false positives and true positives to increase (Figure 1.3; Table A.3). The effects of adjusting the Sensitivity setting were especially prominent from 1.0 (the default setting) to 1.5 (the maximum sensitivity possible; Figure 1.3). Adjustments to the Overlap setting had a lesser effect on species output than Sensitivity, but increases in the Overlap setting (from its default value of 0.0) had a stronger effect at the maximum Sensitivity value (Figure 1.3). Like with Sensitivity, increasing the Overlap setting led to a decrease in the number of false negatives and true negatives, while leading to an increase in the number of false positives and true positives (Figure 1.3).

Both the highest mean number of true positive species reported (20.47 species) and the highest mean number of false positive species reported (38.08 species) were obtained at Overlap, Sensitivity setting combination (2.5, 1.5; Table A.4). The lowest mean number of false negative species errors (0.67) was achieved at the setting combination (2.5, 1.5), whereas the highest mean number of false negative species errors (11.53) was obtained at the setting combination

(0.5, 0.5; Table A.4). The lowest mean number of true positive species reported per visit (9.61) was obtained at the setting combination (0.5, 0.5), while the lowest mean number of false positive species reported (3.22) was achieved at the setting combination (0, 0.5; Table A.4).

The results comparing the mean values of precision, sensitivity, specificity, F1, and MCC between the 18 Overlap, Sensitivity setting combinations tested showed the highest mean precision (0.76), the highest mean specificity (0.93), and the highest mean MCC (0.46) were achieved at Overlap, Sensitivity setting pair (0, 0.5; Table A.3); however, the highest mean MCC (0.46) was also achieved at another Overlap, Sensitivity setting combination (1.5, 0.5; Table A.3). The highest sensitivity (0.97) was obtained using the setting combination (2.5, 1.5; Table A.3). The highest mean F1 score (0.60) was reached at the setting combination (2, 1; Table A.3). The lowest mean precision (0.35), specificity (0.22), F1 harmonic mean (0.51), and MCC (0.23) were obtained at Overlap, Sensitivity combination (2.5, 1.5; Table A.3). The lowest mean sensitivity (0.46) was achieved at setting combinations (0, 0.5) and (0.5, 0.5; Table A.3).

DISCUSSION

ARU Model Comparison Analysis

Despite PAM becoming increasingly used in avian monitoring, and the reliance of PAM efforts on ARUs for obtaining audio data, little research has been conducted to discern the impact that ARU device choice has on BirdNET species output (Sugai et al., 2019; Toenies & Rich, 2021). AudioMoths outperformed other units tested in the number of false negative and true positive species reported, but AudioMoths generally reported higher overall species richness, resulting in the units additionally having the highest number of false positives. On behalf of unfiltered BirdNET outputs having an unacceptable level of false positive reports, which was highly evident in this study as well as previous studies which found that BirdNET has an average unfiltered precision between 72—85%, it is thus imperative to implement some method of filtering to mitigate the false positives in BirdNET outputs, such as by using a confidence-based threshold or by manually validating BirdNET reports (Pérez-Granados, 2023; Kahl et al., 2021; Wood et al., 2022). Additionally, of the four performance metrics for which ARU choice differed (sensitivity, specificity, F1, and MCC), AudioMoth performed the best for three of the four metrics—failing to outperform other units only for specificity. However, given the importance of implementing a filtering method when using automatic species classifier data, the lower mean specificity, as well as the higher number of false positive species reported by the AudioMoth, could be remedied by implementing a filtering method, such as confidence-score filtering (e.g., Cole et al., 2022; Bota et al., 2023; Wood et al., 2023).

The SwiftOne ranked second best in the average number of true positive and false negative species reports, but it had the second highest number of false positive species reported; the same pattern then holds true for the SM4 and SM-Micro. This pattern of scoring well in true positive and false positive but also doing poorly with false positives intuitively makes sense: as more species are reported, there is a greater chance that both the number of true positive species and false positive will increase, and as the number of true positive species increases, the number of true negatives decrease. However, an increased number of total species reported among units using the same audio processing methods (BirdNET using the same settings) might suggest that more distant vocalizations were recorded in general—or perhaps were recorded more clearly, as an increase in either the quantity or quality of the vocalizations captured could lead to more overall BirdNET reports overall. In other words, although the number of both true positive and false positive species reported per visit were higher for the AudioMoth, this might be due to it capturing an increased quantity or quality of bird vocalizations compared to the other units, which would thereby result in more overall detections to be made by BirdNET.

This study further emphasizes the importance of filtering BirdNET reports (or other automatic species classifiers), as each unit reported a greater number of false positive species than it did true positive species, granted, many of the false positive species had few reports per recording whereas the true positive species typically had multiple reports per recording. Given this, some filtering method should be used to mitigate the number of false positive species reports being accepted into any data set to be used for population monitoring to safeguard against considering species which are truly absent to be present. Ways to mitigate the number of false positives include changing the minimum confidence threshold setting in BirdNET (Kahl et al., 2021; Pérez-Granados, 2023), calculating species-specific thresholds using results from a subset of validated species reports for each species (e.g., Cole et al., 2022; Wood et al., 2023; Bota et al., 2023), or perhaps by requiring a minimum number of vocalizations per species per recording. Despite some form of false positive mitigation being necessary, it is also important to note that such measures will also increase the number of false negatives (e.g., Cole et al., 2022). Nevertheless, false negatives can be dealt with, to some degree, using well-established occupancy modeling methods (MacKenzie et al., 2017).

Overlap and Sensitivity Analysis

Automatic species classifiers, such as BirdNET, are essential to large-scale PAM efforts on account of the large quantity of audio data that can feasibly be collected using PAM techniques (Cole et al., 2022; Lauha et al., 2021; LeBien et al., 2020; Ruff et al., 2020). Despite this, the consequences of adjusting BirdNET's Overlap and Sensitivity settings on the performance have, as of writing this, yet to be investigated (Pérez-Granados, 2023). My analyses indicated that Overlap and Sensitivity have highly significant effects on both the number of true positive species reported and the number of false positive species reported, and a significant interactive effect was documented between Overlap and Sensitivity. Additionally, Sensitivity had a greater impact on the number of true positive, false positive, and false negative species reported than did Overlap.

As Sensitivity increased from its default value (1) to its maximum value (1.5) and Overlap increased from its default value (0) to the maximum value I tested (2.5), the average number of false negative species decreased to an average of almost 0 for the *unfiltered* BirdNET output. Although this seems promising given that one could potentially curtail the increased number of false positives by simply using a threshold or thresholds to filter the output, it is quite possible that the use of confidence-based thresholds might remove any net gain in true positives, as the detections which were not made prior to increasing the Overlap and Sensitivity values might not have confidence scores high enough to be considered true positives using such a filtering method. At any rate, it is quite possible that adjusting either or both the Overlap and Sensitivity settings might result in a different confidence score to be assigned to the same detection, which means the species-specific threshold would therefore need to be calculated for whichever Overlap-Sensitivity setting combination a person decides to use.

Limitations

This study took place during the later portion of the breeding season, and as some species may discontinue singing earlier in the breeding season than others, it is possible that the acoustic density during dawn chorus may not have been as dense as it otherwise would have had I sampled earlier in the breeding season. Collectively, my results may be a reflection of the community diversity and bird behavior occurring during this study. Additionally, the lack of some species vocalizing means that it is possible that a lower species composition (and thus a lower diversity in avian vocalizations) was tested. It is also possible that the vocalization of juvenile birds during my sampling period may have slightly biased the results for BirdNET's species output given that juvenile birds often do not have the same vocalizations as adults. For example, juvenile American Crows (*Corvus brachyrhynchos*), were often mistaken for Fish Crows (*Corvus ossifragus*) by BirdNET, which has an impact on the classification metrics; BirdNET uses embedded features to capture age and sex variation in vocalizations, but it is still an area of development (Kahl et al., 2021). Despite the elevated possibility of miscategorizations on account of the presence of juvenile birds, the amount of bias that this introduces should overall be minimal—especially given that BirdNET was trained using species recordings on Xeno-canto and Macaulay Library—both of which include juvenile vocalizations (Kahl et al., 2021; Xeno-Canto, 2023; Cornell Lab of Ornithology, 2023b).

It is also important to note that this study was only able to compare the number of species known to have truly been present during each of the recording periods versus the number that were absent, meaning that I do not know the actual number of true positive or false positive reports for each BirdNET output. Furthermore, it is possible, albeit unlikely, that some of the species labeled as true positive species on each of the BirdNET outputs for a given visit (species truly present during the time of recording) could have actually been the call of a different species during an individual recording, meaning that there may have only been false positive reports of a species for a given BirdNET species output despite the species truly being heard during the concurrent point count. I state this as being unlikely, or at the very least likely seldom occurred and therefore minimally impactful, on behalf of the fact that most species that were truly present during a given point count vocalized more than once during the recording period, and due to the fact that most species that were present during a point count and therefore labeled as a "true positive" species report if present on a BirdNET output for that particular site visit were often reported more than once on the species output—making it likely that at least one of those reports was a true positive.

CONCLUSION

The best performing unit was the AudioMoth, followed by the SwiftOne, SM4, and finally the SMMicro. Despite this, and though several of the performance metrics were statistically different, the difference in performance between the 4 different units was not significant enough to warrant a caution against the use of any specific unit; however, I do hope that this information can be used by researchers to assist in deciding between which ARU model they should use for their avian population monitoring work. There is a significant cost difference between the 4 different ARUs tested, with the AudioMoth costing $79.99 per unit, the SwiftOne costing $349.00 per unit, the SM4 costing $899.00 per unit, and the SMMicro costing $249.00 per unit (as of August 2023). Acknowledging these stark differences in cost, it seems that for most researchers, opting for the cheap yet reliable AudioMoth may be the best bet; however, other considerations—such as durability of a given ARU model or maximum possible recording time for a single deployment—factors which I did not examine, may also be important to consider when deciding which ARU model to use for a given project. Nonetheless, my findings suggest that, regardless of which ARU model is purchased, the biggest challenge researchers will face is the false positive rate generated by the automatic classifier itself rather than the quality of the audio captured by the ARU.

Based on my findings of the Overlap and Sensitivity settings, whether it is worthwhile to run BirdNET with Overlap/Sensitivity settings deviating from the defaults depends on the goals of the researcher. If the goal is to examine whether a specific species is occupying a given area, and the researcher is particularly concerned with limiting false negative errors, then running BirdNET higher Sensitivity and Overlap setting values might be worthwhile, assuming that the researcher is able to invest additional time towards manual validation to contend with the increase in false positives (and overall decrease in precision) that comes with using higher Sensitivity and Overlap settings. However, if the goal is to examine species richness of a multitude of sites—or if there is limited funding to hire technicians for extensive manual validation efforts—then increasing these settings from their default values might be a poor decision given that, without further manual validation, making such an adjustment to the settings will increase the number of false positives and thus artificially bloat the "species richness." For most users who likely do not have the time or resources to invest in extensive manual validation efforts—especially for those who are aspire to establish a long-term population monitoring program and seek to maximize the amount of sites surveyed each year, then using the default Overlap and Sensitivity settings along with some streamlined filtering method seems to be the best choice, as doing so was able to achieve a precision value of 0.95 and a specificity of 0.99.

Given that this is the only paper thus far comparing the performance between these 4 ARU models, as well as the only study investigating the effects of different Overlap and Sensitivity setting combinations on BirdNET species output, it is critical that additional investigations be carried out to support my findings. I hope to see further research conducted to discern whether my ARU model comparison results hold up in different contexts. For example, future research regarding the comparison of different ARU models might examine whether certain ARU models pick up specific species better than others on behalf of vocalization differences, making it worthwhile to recreate this study using a different avian community composition; or examine the effects of anthropogenic noise on ARU model performance. Similarly, I hope to see further investigations conducted to determine whether my findings regarding the Overlap and Sensitivity settings hold true under various conditions. For example, future research regarding the effects of Overlap and Sensitivity could focus on discerning whether varying amounts of anthropogenic noise affects the impact had by different Overlap and Sensitivity setting combinations; or examine the effects of different Sensitivity settings on BirdNET output in a controlled environment in which the distance of the vocalization from the ARU is known; or examine the impacts of different Overlap settings in different environments—especially in locations which have exceedingly high species richness such as the neotropics.

Since my research focused only on whether a species was present or absent—marking species as "true positives" if they were present during the associated point count and marking them as "false positives" if they were not heard during the associated point count—I believe future studies should also investigate whether my findings for both the ARU model comparison and Overlap/Sensitivity work differs much from the number of true positive and false positive observations made by BirdNET. Granted, doing so would be a time-demanding task, as it would require a person to annotate either every second of an audio file or entire clips from audio files (assuming one would also like to examine false negative reports), making this an unfeasible task for most research teams. Despite this, having such data would allow for us to better understand the effects that ARU model choice might have on average confidence scores for true and false positive species reports, as well as the effects that adjusting the Overlap and Sensitivity settings have on average confidence scores for true and false positive reports. This would especially be helpful in discerning whether my findings pertaining to the ability of using a filtering method alongside higher Overlap and Sensitivity values is helpful in not only increasing the number of true positive species reported while minimizing the number of false positives, but also if it is as effective at minimizing individual false positive reports while maximizing true positive reports.

TABLES AND FIGURES

Tables

Table 1.1: Single factor (Unit) mixed model nested ANOVA between all ARUs for a.) sensitivity, b.) precision, c.) specificity, d.) F1 harmonic mean, and e.) Matthews correlation coefficient. Date (of visit) is nested within site (of visit). Type III ANOVA with Satterthwaite's method was used.

Single Factor, Mixed Model Nested ANOVA

a. Precision

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Unit	6	6.2e^{-3}	1.03e^{-3}	0.22	0.97

b. Sensitivity

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Unit	6	1.59	0.27	36.29	2.2e^{-16}

c. Specificity

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Unit	6	0.07	0.01	17.11	4.55e^{-16}

d. F1 Harmonic Mean

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Unit	6	0.37	0.06	15.98	4.09e^{-15}

e. Matthews Correlation Coefficient

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Unit	6	0.35	0.06	10.76	1.99e^{-10}

Table 1.2: Three-way, randomized block design ANOVA test results testing the effect of the Sensitivity and Overlap settings on a.), the number of false positive species reported per site visit across all 36 site visits and b.), the number of true positive species reported per site visit across all 36 site visits.

Three-Way, Randomized Block Design ANOVA

a. False Positives

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Sensitivity setting	2	310524	155262	6285.476	< 2e^{-16}

Overlap setting	3	6735	1347	54.528	$< 2e^{-16}$
Visit	35	19137	547	22.135	$< 2e^{-16}$
Sensitivity setting—Overlap setting	10	644	64	2.605	0.00421
Residuals	595	14698	25		

b. True Positives

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Sensitivity setting	2	13294	6647	2096.565	$< 2e^{-16}$
Overlap setting	3	394	79	24.881	$< 2e^{-16}$
Visit	35	8068	231	72.710	$< 2e^{-16}$
Sensitivity setting—Overlap setting	10	87	9	2.731	0.00272
Residuals	595	1886	3		

Figures

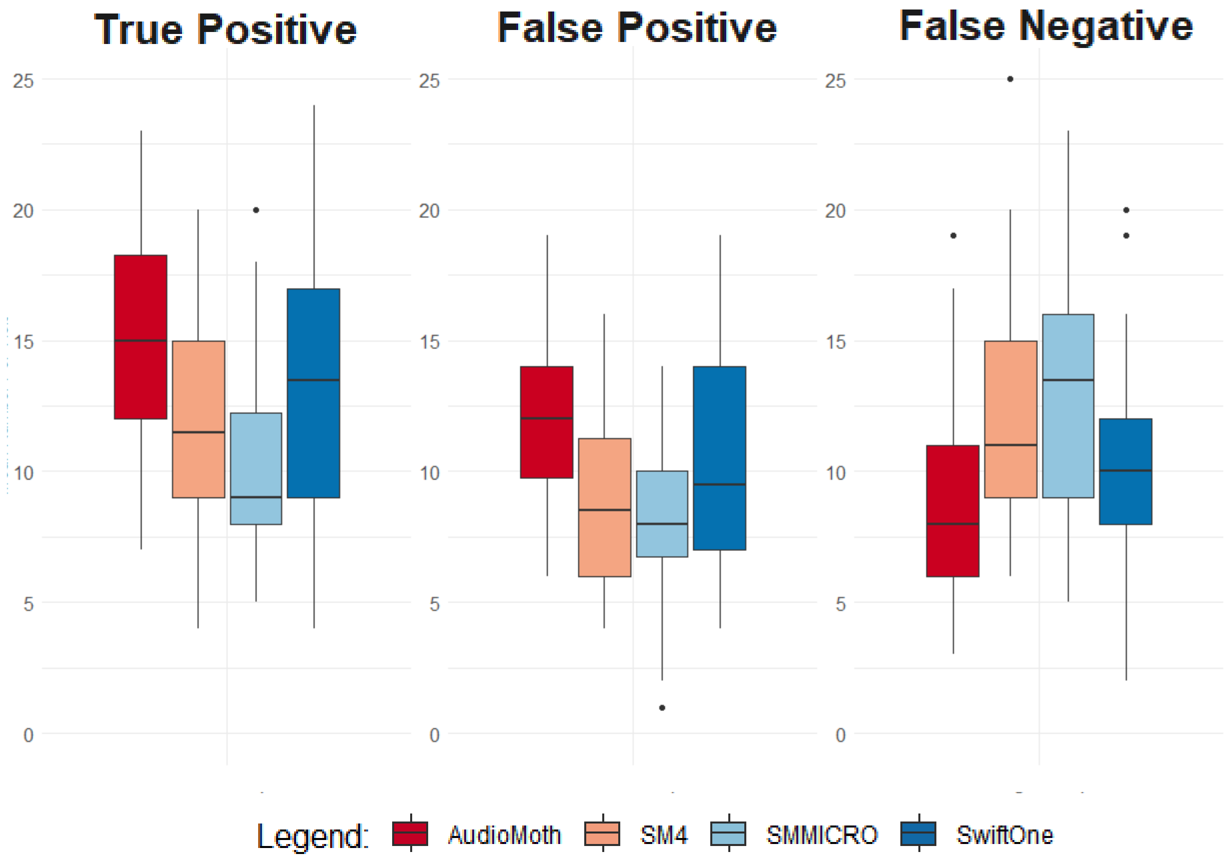

Figure 1.1: False negative (FN), false positive (FP), and true positive (TP) averages among AudioMoth (leftmost; red), SM4 (second to leftmost; orange), SMMicro (third to leftmost; light blue), and SwiftOne (rightmost; dark blue). Data represents average number per detection type per visit for 36 visits conducted from June 21—July 4, 2023 in Jackson County, Illinois, USA. Detections were produced using audio from each of the four ARUs and using BirdNET-Analyzer at default settings.

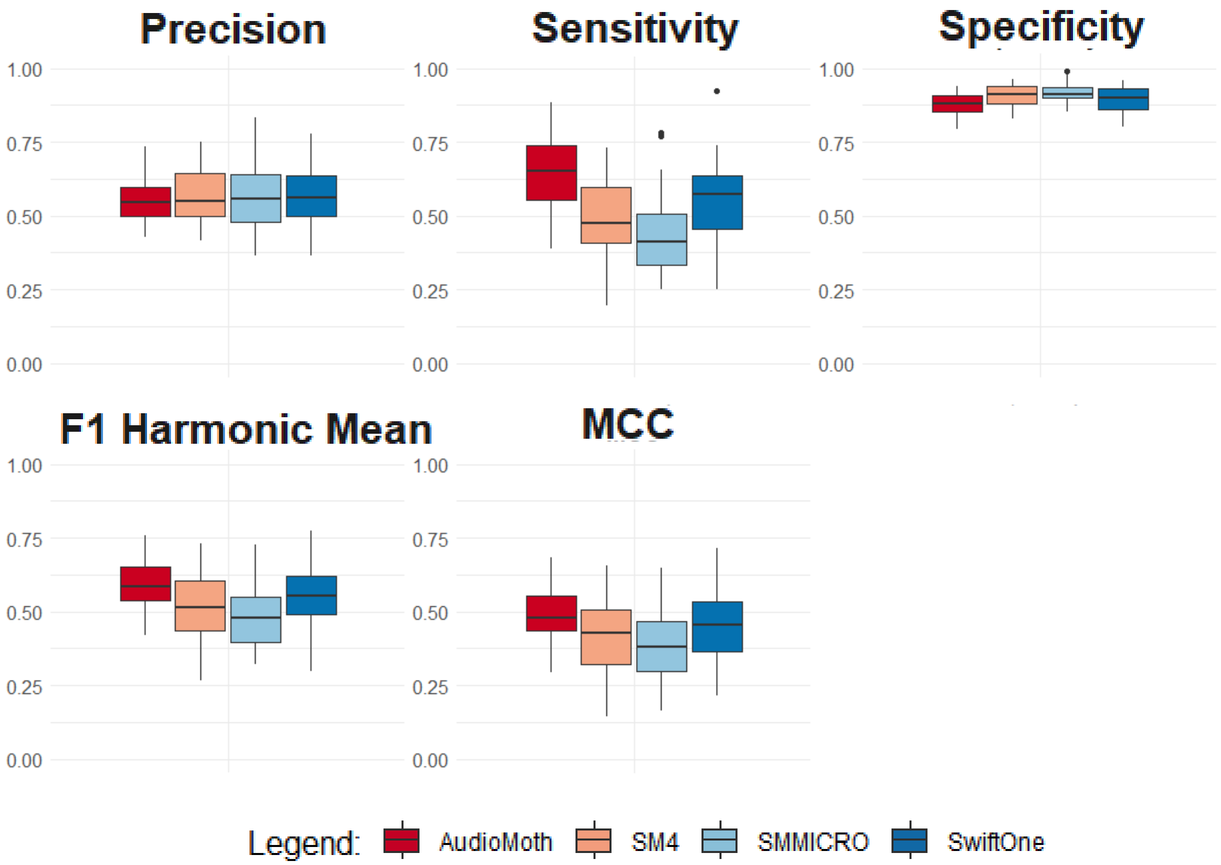

Figure 1.2: Precision, Sensitivity, Specificity, F1 harmonic mean, and MCC (Matthews Correlation Coefficient) averages among AudioMoth (leftmost; red), SM4 (second to leftmost; orange), SMMicro (third to leftmost; light blue), and SwiftOne (rightmost; dark blue). Data represents average value per performance metric per visit for 36 visits conducted from June 21—July 4, 2023 in Jackson County, Illinois, USA. Detections were produced using audio from each of the four ARUs and using BirdNET-Analyzer at default settings.

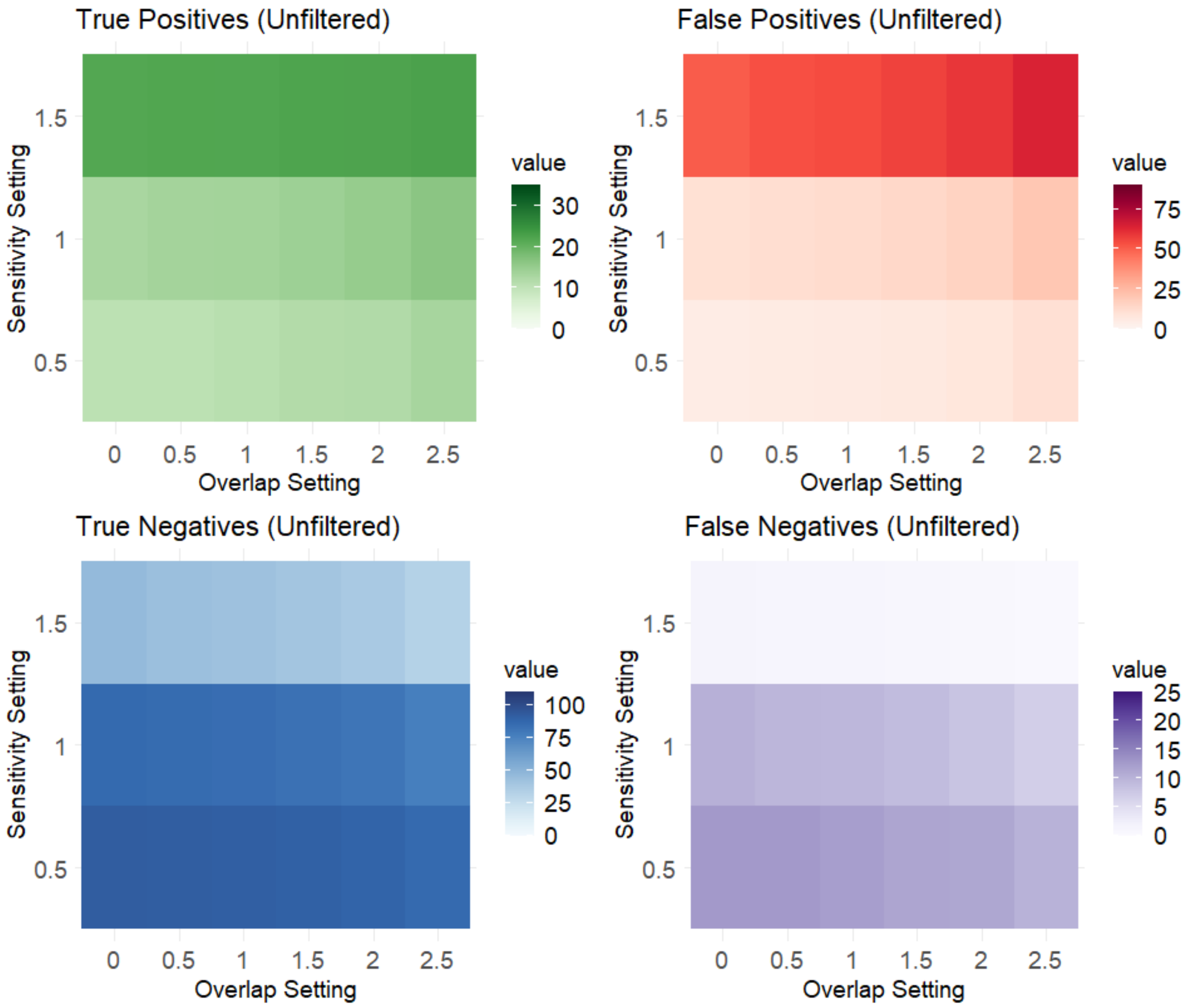

Figure 1.3: Confusion matrix heatmaps showing the effects of different Sensitivity and Overlap setting combinations on the mean number of true positive species reported (TP), false positive species reported (FP), true negative species reported (TN), and false negative species reported (FN). Darker colors indicate greater values. Values depicted were averaged across all 36 site visits. Only audio collected via AudioMoth was used for this analysis, and audio from each of the 36 visits were processed with all 18 combinations of Overlap and Sensitivity tested. Audio was collected during 36 visits between June 21—July 4, 2023 in Jackson County, Illinois, USA.

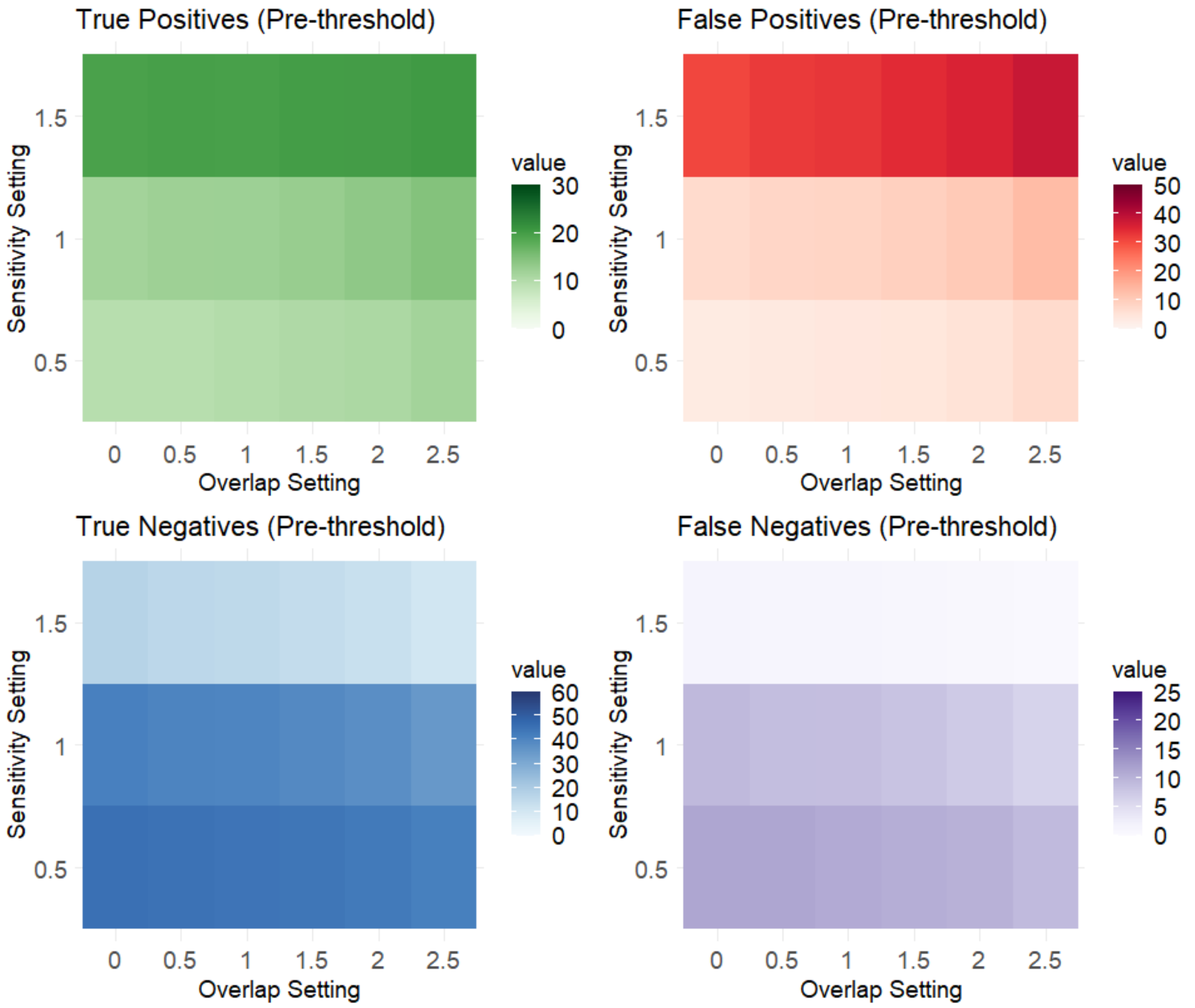

Figure 1.4: Confusion matrix heatmaps showing the effects of different Sensitivity and Overlap setting combinations on the mean number of true positive species reported (TP), false positive species reported (FP), true negative species reported (TN), and false negative species reported (FN) for the Pre-threshold dataset. To accurately compare the effects of implementing a species-specific threshold, the pre-threshold dataset only included species which had species-specific validations and were present during the point counts were included in the making of these heat maps. The data shown was not filtered using species-specific thresholds. Darker colors indicate greater values. Values depicted are means across all 36 site visits.

CHAPTER 2

COMPARING METHODS OF STREAMLINING BIRDNET-ANALYZER VALIDATION FOR LONG-TERM AVIAN POPULATION MONITORING

INTRODUCTION

Since 1970, there has been an estimated net-loss of approximately 3 billion birds in North America—a loss equal to 29% of 1970 abundance estimates (Rosenberg et al., 2019). Successful avian conservation is hinged on effective monitoring of population dynamics and trends over time. Effective monitoring enables researchers to prioritize their efforts towards the species and ecosystems most imperiled, identify population trends, and better gauge the effectiveness of various conservation practices (Nichols & Williams, 2006; Marsh & Trenham, 2008; Jones, 2011; Jones et al., 2013; Stowell et al., 2018). Point counts have historically been the conventional field method for estimating avian population abundance and distribution, assessing species-environmental relationships, and understanding drivers of population change (Bibby et al., 2000). However, with the advent of autonomous recording units (ARUs) and rapidly developing automatic species classifiers, such as BirdNET Analyzer (herein, BirdNET), many researchers have switched to using Passive Acoustic Monitoring (PAM) as an alternative approach to traditional point count methods (Kahl et al., 2021; Shonfield et al., 2017). Recent research comparing traditional avian monitoring techniques with PAM has documented that PAM can be more cost-effective, allow for greater observation windows than feasible with point counts, and reduce observer bias, among other advantages (Alquezar & Machado, 2015; Klingbeil & Willig, 2015; Hobson et al. 2002; Tegeler, Morrison & Szewczak, 2012; Shonfield & Bayne, 2017). Despite these benefits, the quantity of audio data that can conceivably be collected versus the quantity that can feasibly be manually validated by humans differs substantially.

Automatic species classifiers, which produce confidence-based species predictions based on audio data, offer a promising path forward to alleviate the data processing bottleneck of acoustic recordings, as terabytes of data can be easily collected during a sampling period. Even the most reliable automatic bird species classifier currently available for general use, BirdNET, has an average precision of 72—85% (percent of detections correctly classified), thus highlighting the continued issue of false positive and false negative errors with current species recognition software (Pérez-Granados, 2023; Kahl et al., 2021; Wood et al., 2021). This problem is further exacerbated by the fact that precision varies widely depending on the species being examined (Sethi et al., 2023). Although BirdNET is arguably the best available classifier for birds given it is free, capable of identifying more species than any other classifier, and exceeds other classifiers in terms of mean average precision, it too is prone to producing both false positive and false negative species reports (Lauha et al., 2021; LeBien et al., 2020; Ruff et al., 2020; Kahl et al., 2021; Sethi et al., 2023). As such, one of the most pressing issues regarding PAM is how to process massive quantities of audio data, especially in a way that minimizes false positive and false negative errors (Pérez-Granados, 2023; Lauha et al., 2021).

Many researchers have attempted to minimize the number of false positive reports generated by BirdNET by filtering BirdNET output using confidence-based thresholds. When using confidence-based thresholds, the user either chooses an arbitrary threshold value for all species (e.g. Kahl et al., 2021; Wood et al., 2021) or calculates species-specific thresholds using data gathered via manual validation of BirdNET reports (e.g. Cole et al., 2022; Bota et al., 2023). The confidence-based thresholds filter BirdNET outputs such that all detections with a confidence value that is less than the selected threshold value are removed from the dataset. Fixed albeit arbitrary thresholds that have been applied to BirdNET in previous studies include 0.5 (e.g., Wood et al., 2021), but these thresholds are usually accompanied with little justification or exploration of the implications. Alternatively, recent approaches have validated a subset of the PAM recordings and then used species-specific calculated or modeled thresholds, avoiding a one-size-fits-all approach, ideally minimizing the number of discarded valid detections (Cole et al., 2022; Wood et al., 2023). Although filtering BirdNET output using confidence-based thresholds has become common in PAM efforts in the last few years (e.g. Kahl et al., 2021; Wood et al., 2021; Cole et al., 2022; Bota et al., 2023; Perez-Granados, 2023), few have compared the performance of different confidence-based threshold filtering methods.

Here, I compare four confidence-based thresholds for filtering BirdNET output: fixed 0.5 and/or 0.75 thresholds for all species (e.g., Kahl et al, 2021 and Wood et al., 2021) and two species-specific thresholds which are calculated through manual validation of BirdNET output (e.g., Cole et al., 2022, Bota et al., 2023, and Wood et al., 2023). Coupled with concurrent point counts during ARU recording periods, I then evaluated which filtering method best achieves a mean precision value of at least 0.95 while still maximizing sensitivity, thus excluding most false positive species errors while also minimizing the number of false negative errors created by the filtering process. Following this, I applied the best performing threshold approach based on my requirements to filter a PAM dataset collected during April – June 2022 covering the southern 11 counties of Illinois, USA, aimed at resurveying the 1986-1991 Illinois Breeding Bird Atlas to assess changes in species composition, occurrence, and richness of the warbler and vireo community. By resampling primary blocks sampled by the Illinois Breeding Bird Atlas, which occurred during 1986—1991, I can gain insight to changes in bird species distribution in the

Southern Illinois region—a region which has not been probabilistically sampled at this scale since the Illinois Breeding Bird Atlas (IBBA) concluded over three decades ago (Kleen et al., 2004).

METHODS

Assessing the performance of various BirdNET filtering approaches requires us to know "truth" regarding either the validity of individual detections or the validity of individual species present per recording. I established "truth" by conducting point counts alongside ARU recordings, allowing us to identify species present at each site. Using point count data as a reference, I calculated multiple performance metrics (e.g., precision, sensitivity) for BirdNET classification. Then, to establish species-specific filtering thresholds, I manually validated a subset of BirdNET detections. These validations, sourced from a large-scale PAM survey effort in 2022, provided the necessary information for filtering threshold calculations that were then applied to study changes in the regional warbler and vireo community over a 30-year period.

2023 Point Counts and ARUs

From June 21 to July 4, 2023, six field sites in Jackson County, IL, USA with differing land use histories, physical attributes, plant communities, and avian community compositions were sampled using AudioMoths and simultaneous point counts. ARUs were placed in small, grip-sealed plastic bags along with a desiccant silica packet to ensure the units remained dry and were fastened to a 1.5-meter t-post with a fixed location at each site. This method of waterproofing AudioMoths was chosen as it appears to be the most cost-efficient method without losing a significant amount of sound (Lapp et al., 2023). All six sites were visited 6 times each for a total of 36 site visits. Each recording period (site visit) took place between the hours of 0500 and 0830, with all recordings beginning on the hour (ex: 5:00:00 AM) and recording for a total of 30 minutes while a point count was simultaneously conducted for the duration of the ARU recording period. Site visitation order was randomized during each field day such that all sites would be sampled during each of the possible start times.

BirdNET Analyzer was then used, with default Overlap and Sensitivity settings, the default confidence value minimum of 0.1, and a custom species list, to process all audio files from the 2023 PAM recordings. The custom species list used to process the 2023 point count and ARUs dataset was generated using the R package "rebird," and the list consisted solely of species reported on eBird during the past 10 years from June 21 to July 4 (the date range of the sampling period) in Jackson County, IL (in which all sites were located) (Maia et al., 2023).

Point counts were conducted during the entire duration of each 30-minute recording period for all 36 site visits for the 2023 point counts and ARUs dataset. For each point count, the observer stood approximately 5 meters away from the ARU and recorded all bird species which were heard. All bird species were marked only once but were not time-stamped. Although bird species which were seen but not heard during point counts were noted, they were not incorporated into the analysis due to my focus on vocalizing species.

2022 Breeding Bird Surveys

From May 1 to July 1, 2022, I surveyed 45 blocks—or 135 sites—across the southernmost 11 Illinois counties using AudioMoth ARUs (version 1.2.0, Hill et al., 2018, 2019) and BirdNET Analyzer. ARUs were placed in small, grip-sealed plastic bags along with a desiccant silica packet to ensure the units remained dry and were fastened to a 1.5-meter t-post with a fixed location at each site, as this method of waterproofing AudioMoths appears to be the most cost-efficient method without losing a significant amount of sound (Lapp et al., 2023). I further divided the primary census blocks (approximately 25.9 square kilometers) into nine, uniform sub-blocks (approximately 2.88 square kilometers) and randomly selected three to account for avian species composition variation within the census blocks (Kleen et al., 2004; Montgomery et al., 1987). Each of the 135 sites sampled represents a different sub-block. Given that many of the census blocks consist of exclusively private land, in some situations the initially selected sub-blocks had to be revised due to lack of access to private land. Due to these limitations, only 45 primary blocks were fully sampled (3 surveyed sub-blocks per block), and to remain consistent with my sampling efforts, for the purpose of this research, I analyzed the 45 primary blocks for which I was able to sample 3 sub-blocks.

During the 2022 breeding bird surveys, AudioMoths were used for recording bird vocalizations, as they are currently the lowest-cost ARU and have shown a comparable performance to higher-cost units (Toenies & Rich, 2021). ARUs were programmed to record for four hours beginning 30 minutes before local sunrise to capture the dawn chorus and peak hours of diurnal bird vocalization. For deployment, ARUs were placed in small, grip-sealed plastic bags along with a desiccant silica packet to ensure the units remained dry and were then secured to a tree at breast height (approximately 1.5 meters above ground) using a zip tie. The ARUs remained in the field to record for a minimum of five days—producing a minimum total of 5 4-hour recordings per site, thus resulting in a minimum of 20 observation hours per sub-block and a minimum of 60 observation hours per primary block. In some instances, ARUs were left out for more than 5 days due to ARU retrieval becoming infeasible due to inclement weather making dirt roads impassable or due to the need to prioritize deploying additional units rather than retrieving units.

To process ~3500 hours of audio data collected during the 2022 breeding bird surveys, the use of an automatic species classifier was necessary. As such, BirdNET Analyzer was selected to process the collected audio files using default settings and a custom species list. The R package 'rebird' was used to create a custom species list containing all species reported to eBird from 2021-2022 in Illinois (Maia et al., 2023). I ran BirdNET using default Overlap and Sensitivity settings, the default minimum confidence as 0.1, and with the use of the custom species list to process all audio files, resulting in a minimum of 5 daily BirdNET outputs per sub-block. Each BirdNET output corresponds with a single "visit" (4-hour recording for an individual site on a given day), with each row containing information regarding the species predicted by BirdNET, the recording time of the detection, and a confidence value, which is a rating of how confident the BirdNET algorithm is that the individual detection is a true positive, among other information.

Manual Validation & Threshold Calculations

Following audio analysis with BirdNET, 100 random BirdNET detections per species were selected using all outputs from the 2022 breeding bird surveys. The only species for which I validated 100 random samples were species present during the point counts (75 species) and all warbler and vireo species which breed in Southern Illinois (23 species), with many of the warbler and vireo species appearing in the point count dataset. All random samples were manually validated using audio playback, with each individual detection being marked as 1 (detected) or 0 (not detected). In addition to calculating species-specific thresholds, the results from the 100 manual validations allowed for species-specific calculations of: 1) a maximum confidence score for a false positive detection; 2) a maximum confidence score for a true positive detection; and 3) standard deviation and standard error of confidence scores.

The first set of species-specific thresholds calculated, herein referred to as the FP-based thresholds, uses methods adapted from Cole et al. (2022). The FP-based thresholds are calculated by summing the species-specific maximum confidence score of false positives, the standard error of the confidence scores for the species, and—if the sum of the first two values did not equal or exceed 0.95—then adding 0.05. Any species which either had a maximum confidence value amongst all false positives that was greater than the maximum confidence score amongst all true positives or which had a calculated FP-based threshold value exceeding 1.0 were excluded from further analyses.

The second set of species-specific threshold type tested is herein referred to as the Modeled thresholds (Wood et al. 2023; Bota et al., 2023). By this method, the same 100 manual validations used for the FP-based threshold were also used for the derivation of this set of species-specific thresholds to ensure consistency. To calculate the Modeled thresholds, first the BirdNET's confidence scores were back-transformed into their original logit scale using the following equation:

$$\text{Logit score} = \ln(1/(1 - \text{confidence score}))$$

After back-transforming the confidence scores, a logistic regression was fitted to establish the relationship between the independent variable (BirdNET logit-scale prediction score) and dependent variable (the probability that the detection is a true positive). For each species, the equation considering a probability of correct detection (true positive) used was as follows:

$$\text{logit}(P) = \ln(\text{intercept}) + 0.95 \times \ln(\text{logit-score})$$

Then, the minimum confidence score at which a 0.95 probability of a detection being correct was reached determined the threshold for the focal species. As in the case of the FP-based thresholds, any species for which I was unable to calculate a Modeled threshold score were excluded from further analyses. The 100 manual validations per species were used to derive the Modeled threshold values for each species.

Finally, I also included two arbitrary thresholds: confidence scores of 0.5 and 0.75. When comparing the four thresholds, I only included species that I was able to calculate a threshold for across each approach. These species were excluded from all analyses pertaining to the comparison of performance between the four thresholds, as well as comparisons between the four filtered outputs (which were created using the thresholds) and the unfiltered output.

Performance Comparisons

Following the threshold calculations, I compared the number of true positive species, false positive species, and false negative species in each of the four filtered outputs and unfiltered output. The number of true negative species was calculated for the unfiltered dataset and each filtered dataset by adding the number of true positive, false positive, and false negative species and subtracting the total from the number of species present on the custom species list. A confusion matrix heatmap was created to visually compare the average number of true positive, false positive, false negative, and true negative species per visit for each output version. Precision, sensitivity, specificity, F1 harmonic mean (herein F1), and Matthews Correlation Coefficient (herein, MCC) were calculated individually for all five versions of the output (including the unfiltered version), using the average number of true positive, false positive, false negative, and true negative species per visit for each output version, and the results were graphed using a boxplot. A single-factor, mixed-effects ANOVA in which date was nested within site to account for variation due to site and date, was conducted for precision, sensitivity, specificity, f1, and MCC separately to determine whether the observed differences between the various performance metrics between threshold types were significant.

Comparing Species Composition & Richness

Using the best performing threshold set, I then filtered the BirdNET output from the 2022 surveys to evaluate changes in species richness and composition in the warbler and vireo species guild over the 30-year period since the last IBBA survey. Both the number of total target species detected per block per survey year as well as the species composition per block per survey year were examined. To examine block-level changes in species composition and richness between the last IBBA (1986—1991) to 2022, I calculated the Jaccard's Similarity Index for individual blocks using the equation:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

wherein $A \cap B$ is equal to the number of species detected by both surveys (IBBA and the 2022 survey) within the same block, while $A \cup B$ equals the number of unique species detected per block across both survey years (in other words, the sum of the number of species detected in both surveys minus the number of shared species between the two surveys) (Jaccard, 1901). To further examine block-level changes in species composition and richness, the Sørensen's Similarity Index was calculated for each block using the equation:

$$SSI = 2a/(2a + b + c)$$

wherein a = the number of species shared by both the IBBA and 2022 surveys in a particular block, b = the number of species present in the IBBA survey but absent from the 2022 survey for a particular block, and c = the number of species present in the 2022 survey but absent from the IBBA survey for a particular block (Sørensen, 1948). Both the Jaccard's Similarity Index and Sørensen's Similarity Index were used as the two indices provide different means of assessing species composition change, with the key difference being that the Sørensen's emphasizes the importance of shared species between surveys for individual blocks. Both indices represent ways to examine changes in β-diversity over time, and both indices may be used in tandem for such applications (e.g., Dong et al., 2023). To further differentiate the two, the Sørensen's Similarity Index places a greater emphasis on the number of shared species and on changes in species composition between survey years whereas the Jaccard's Similarity Index places greater importance on the number of species detected and thereby changes in species richness. For both the Jaccard's Similarity Index and the Sørensen's Similarity Index, possible values range from 0 (indicating no similarity, or that no species are shared between the two survey years for an individual block) to 1 (indicating complete similarity, or that all species present in the first survey year are present in the second and vice versa). I then calculated the percent species turnover using the equation:

$$\% \text{ Species Turnover} = 100 \times (b+c)/(d)$$

Wherein b = the number of species present in the IBBA survey but absent from the 2022 survey for a particular block, c = the number of species present in the 2022 survey but absent from the IBBA survey for a particular block, and d = the total number of species found between both survey years for a particular block. In addition to these metrics, the change in species richness was also calculated for each block individually by subtracting the total number of species for a given block from the IBBA survey from the total number of species for the same block.

To provide further context for any observed changes in warbler and vireo species richness and composition, I compared my findings with data from the USGS North American Breeding Bird Survey (BBS) for the years 1991 and 2022 (Ziolkowski et al., 2022). The BBS also takes place during the breeding bird season, and theoretically should capture similar changes in the warbler and vireo guild. However, the BBS sampling design differs from mine. The BBS does not sample grid blocks in the same way as the IBBA and my survey efforts, and BBS routes are scarce in southern Illinois. Therefore, I compared state-level changes in warbler and vireo counts between 1991 and 2022, rather than relying solely on data from the southern Illinois region. Since most of my target species are forest specialists primarily found in southern Illinois (Crocker, 2018), and several have their entire Illinois breeding range located within (e.g., Black-and-white Warbler (*Mniotilta varia*), Northern Parula (*Setophaga americana*)), I compared state-level BBS data. This approach captures detections likely originating from southern Illinois, particularly compared to generalist species like the Common Yellowthroat (*Geothlypis trichas*) and Red-eyed Vireo (*Vireo olivaceus*). For each of the 21 target warbler and vireo species, the total count per species for the years 1991 and 2022 were examined to discern whether a positive, negative, or neutral change in the number of detections per species took place during the time between the last IBBA and my survey efforts. To further understand the effects of observation hours on species richness variation, I plotted the relationship between observation hours per block (independent variable) and observed species richness (dependent variable)—separately assessing the point count and PAM data.

RESULTS

Threshold Comparison

A total of 75 species were detected by the observer during the 36 point counts conducted during the concurrent 2023 PAM recordings. Of these 75 species—all of which had 100 random BirdNET detections sourced from the 2022 breeding bird surveys and manually validated—10 species had to be excluded as I was unable to calculate a species-specific threshold. Specifically, I was unable to calculate species-specific thresholds for 6 species using the Modeled threshold approach, and I was unable to calculate species-specific thresholds for 5 species using the FP-based threshold methodology; 1 species failed across both approaches. Using the 65 remaining species, averaging all species-specific thresholds, I documented a mean Modeled threshold value of 0.40 ± 0.20 SD and a mean FP-based threshold value of 0.48 ± 0.17 SD.

I was unable to derive thresholds for six species using the methodology used for deriving the Modeled thresholds, including the Barred Owl (*Strix varia*), Belted Kingfisher (*Megaceryle alcyon*), Eastern Towhee (*Pipilo erythrophthalmus*), Kentucky Warbler (*Geothlypis formosa*), and Green Heron (*Butorides virescens*) (Table B.1). I was unable to calculate species-specific thresholds using the FP-based threshold calculation methods for a total of five species, including the Summer Tanager (*Piranga rubra*), Eastern Towhee (*P. erythrophthalmus*), Eastern Wood-Pewee (*Contopus virens*), Great Crested Flycatcher (*Myiarchus crinitus*), and Song Sparrow (*Melospiza melodia*) (Table B.1). I was unable to calculate the FP-based threshold or Modeled threshold for the Eastern Towhee (*P. erythrophthalamus*). In total, I was unable to calculate one or both of the species-specific thresholds for 10 species (Table B.1), and thus these 10 species were removed from further analyses.

The single-factor (threshold type) mixed-effects nested ANOVA, in which "Date" was nested within "Site" to account for the variation due to site and date sampled, was conducted for the following metrics: a. true positive species (per visit), b. false positive species (per visit), c. false negative species (per visit), d. precision, e. sensitivity, f. specificity, g. F1 harmonic mean, and h. MCC (Matthews correlation coefficient) (Table 2.1). A unique ANOVA was fit to all 5 versions of the output: unfiltered, an arbitrary 0.5 confidence threshold for all species, an arbitrary 0.75 confidence threshold for all species, the species-specific Modeled thresholds, and the species-specific FP-based thresholds. The results of the ANOVA showed that, for each metric listed above, the threshold approach was found to significantly affect the classification metrics ($P < 0.001$; Table 2.1). To further understand the differences between individual output versions (the difference between unfiltered vs. different threshold filters, or the difference between different threshold types), I used a Tukey's post-hoc HSD test to examine pairwise comparisons for each of the metrics.

Tukey's post-hoc HSD test documented all pairwise comparisons between the unfiltered data and any of the outputs filtered using confidence-based thresholds were highly significantly different (P < 0.001) except for a few of the MCC pairwise comparisons (Table 2.2). Notably, there were no significant differences in true positive, false positive, or specificity metrics between the filtered outputs themselves (0.5, 0.75, FP-based, Modeled thresholds), except when compared to the unfiltered data (all p < 0.001). Interestingly, several comparisons involving false negative rates and MCC showed no significant differences between the 0.5, FP-based, and Modeled thresholds (Table 2.2).

The unfiltered data differed significantly from all filtering approaches in the number of true positive, false positive, and false negative species (Figure 2.1; Table 2.3). The unfiltered data had the highest number of true positive species reported per visit (11.31 ± 3.41 SD), followed by the output filtered using the Modeled thresholds (7.08 ± 2.2 SD; Figure 2.1; Table 2.3). The unfiltered data also had the greatest mean number of false positives (6.58 ± 2.96 SD), followed again by the Modeled threshold (0.69 ± 0.86 SD; Figure 2.1; Table 2.3). The 0.75-filtered data had the most false negative species (16.94 ± 4.24 SD) followed by the 0.5 threshold (14.86 ± 4.09 SD; Table 3).

Analysis of five performance metrics (precision, sensitivity, specificity, F1 harmonic mean, and Matthews Correlation Coefficient [MCC]) revealed key trends in BirdNET species identification with varying confidence thresholds. Threshold filtering significantly improved precision compared to unfiltered data. The 0.75 confidence threshold achieved the highest precision (0.99 ± 0.08 SD), followed by the FP-based threshold (0.95 ± 0.08 SD) and the 0.5 threshold (0.94 ± 0.09 SD; Table 3). Conversely, unfiltered data exhibited the highest sensitivity (0.55 ± 0.13 SD) but also the lowest precision (0.64 ± 0.10 SD; Table 3). Specificity, indicating the ability to correctly identify absent species, generally increased with threshold filtering. The 0.75 threshold achieved the highest specificity (1.00 ± 0.00 SD), followed by the FP-based and 0.5 thresholds (both 0.99 ± 0.01 SD; Table 3). The unfiltered data had the lowest specificity (0.85 ± 0.07 SD). F1 harmonic mean and MCC, which combine precision and sensitivity, showed a trade-off (Table 3). Unfiltered data had the highest F1 score (0.58 ± 0.09 SD) but a lower MCC (0.42 ± 0.11 SD). The Modeled threshold achieved a mid-range F1 score (0.50 ± 0.10 SD) and the highest MCC (0.47 ± 0.10 SD; Table 3).

Changes in Species Composition and Richness

For the 45 primary blocks sampled during both the IBBA and the 2022 data collection efforts, the average number of species detected per block was 9.71 ± 4.34 SD for the IBBA (1986—1991) data and 13.24 ± 2.01 SD for my 2022 data (Table 2.3). The average number of warbler/vireo species detected per hour of observation for the IBBA was 0.45 ± 0.30 SD, while the average number of warbler/vireo species detected per hour of observation for the 2022 dataset was 0.18 ± 0.033 SD. Across all blocks, there was an average 36.38% increase across all resampled blocks in warbler and species richness in the Southern Illinois region over the 30-year period (Table2. 3; Figure 3). The IBBA averaged 47.80 hours of observation per block while the 2022 study averaged 76.71 hours of observation per block (Table 2.3). A positive relationship was found to exist between the number of observation hours per block and the observed species richness per block for both the point count data from the IBBA and for the PAM data from 2022 (Figure 2.4).

We resampled a total of 45 primary blocks originally sampled by the IBBA. Of these 45 blocks, 26 blocks were considered public blocks (meaning either 2 or 3 of the 3 sites sampled within a given block were located on publicly owned land), whereas 19 blocks were considered private blocks (either 2 or 3 of the 3 sites sampled within a given block were located on privately owned land). Because the IBBA examined primary blocks rather than subblocks, our data was aggregated at the block-level rather than the sub-block level. Among the private blocks, the mean change in warbler/vireo species richness per block was 5.42 ± 0.86 SD, whereas for public blocks, it was 2.15 ± 0.83 SD (Table B.5). The IBBA mean warbler/vireo species richness among the private blocks was 7.00 ± 0.80 SD and for public blocks was 11.69 ± 0.75 SD, whereas the 2022 mean species richness among private blocks was 12.42 ± 0.44 SD and for the public blocks was 13.85 ± 0.37 SD (Table B.5). The average hours of survey effort for the private and public blocks during the IBBA were 29 ± 6.24 SD and 61.76 ± 24.27 SD, respectively, while the average hours of survey effort for the private and public blocks during the 2022 survey were 74.95 ± 1.70 SD and 78.46 ± 2.60 SD, respectively (Table B.5).

Of the 21 target warbler and vireo species, the species which were detected in the least number of blocks in the IBBA data was the Golden-winged Warbler (*Vermivora chrysoptera*), and for my 2022 dataset was both the Golden-winged Warbler (*V. chrysoptera*) and Yellow Warbler (*Setophaga petechia*) (Table 4). The target species detected in the most blocks in both the IBBA and 2022 data was the Common Yellowthroat (*G. trichas*), which was detected in all 45 blocks in both datasets (Table 4). Species which experienced an increase in the number of blocks detected from the IBBA to my dataset, ordered from greatest increase to least amount of increase: American Redstart (*Setophaga ruticilla*), Worm-eating Warbler (*H. vermivorum*), Ovenbird (*Seiurus aurocapilla*), Yellow-throated Warbler (*Setophaga dominica*), Prothonotary

Warbler (*Protonotaria citrea*), Louisiana Waterthrush (*Parkesia motacilla*), Black-and-white Warbler (*M. varia*), Hooded Warbler (*Setophaga citrina*), Yellow-throated Vireo (*Vireo flavifrons*), Pine Warbler (*S. pinus*), Bell's Vireo (*Vireo bellii*), Golden-winged Warbler (*V. chrysoptera*), Northern Parula (*S. americana*), and Red-eyed Vireo (*V. olivaceus*) (Table 4). The target species which saw a decrease in the number of blocks detected from the IBBA to my dataset, ordered from greatest amount of decrease to least amount of decrease: Yellow Warbler (*S. petechia*), Prairie Warbler (*S. discolor*), Blue-winged Warbler (*Vermivora cyanoptera*), Warbling Vireo (*Vireo gilvus*), and Kentucky Warbler (*G. formosa*) (Table 4).

The block-level Jaccard's Similarity Index values ranged from 0.07 to 0.89, with an average of 0.49 ± 0.21 SD across all 45 blocks (Table 5). The block-level Sørensen's Similarity Index values ranged from 0.14 to 0.94, with an average of 0.63 ± 0.20 SD (Table 5). The block-level percent species turnover ranged from 11.11% to 92.31%, with an average of 50.82% ± 20.7 SD (Table 5).

The Illinois state-wide North American Breeding Bird Survey count data for my target species showed an increase in counts for 16 of the 21 species between 1991 and 2022 (Table 6; Table B.2). A total of 3 of the 21 species decreased in count data between 1991 and 2022 for the Illinois Breeding Bird Survey (Table 6; Table B.2). Another 2 species experienced no change in the total number of counts between the 1991 and 2022 Breeding Bird Survey years (Table 6; Table B.2).

DISCUSSION

Threshold Comparison

Despite the increasing popularity in both PAM, BirdNET, and confidence-based threshold filtering over the last few years (e.g., Wood et al., 2021; Cole et al., 2022; Wood et al., 2023;

Bota et al., 2023), a lack of attention has been paid towards comparing the performance of various species-specific confidence thresholds, or even in comparing species-specific confidence thresholds with arbitrary confidence thresholds. Given the potential that such methods have to equip land managers and landowners with the tools needed to conduct long-term avian monitoring, I deemed it necessary to further investigate these topics. Although I found that the FP-based threshold had the highest sensitivity out of the only two thresholds which met my minimum mean precision requirement of 0.95, the use of such a threshold requires manual audio validation of at least 100 detections per species potentially present. As such, it is a time-consuming task that requires a skilled observer to discern the calls of all species present on the BirdNET output. Because of this, and due to the arbitrary threshold 0.75 likewise meeting the minimum requirement of a mean precision greater than or equal to 0.95, for those who may be unable to validate 100 detections of all possibly present species, it may be easiest to opt with an arbitrary threshold of 0.75. However, for those who are both skilled enough to discern species vocalizations and are able to dedicate enough time towards manual validation, a higher sensitivity can be reached by using the FP-based thresholds, and therefore less false negative errors.

A few different issues prevented us from being able to calculate Modeled thresholds for 6 species. For four of these 6 species (Barred Owl (*Strix varia*), Belted Kingfisher (*Megaceryle alcyon*), Eastern Towhee (*Pipilo erythrophthalmus*), and Kentucky Warbler (*Geothlypis formosa*)), the probability of a detection being a true positive detection never reached a value of 0.95 (95%) at any confidence score (e.g., Figure B.1), thereby rendering it impossible to derive Modeled thresholds for these species. One of the six species (American Robin (*Turdus migratorius*)) was found to have a negative relationship between the probability of a BirdNET detection being correct and the confidence score of a detection, meaning that the results from the logistic regression analysis showed that as the confidence score increased, the probability of a detection being a true positive decreased—the opposite relationship as expected (Figure B.2). I was also unable to derive a Modeled threshold for one species (Green Heron (*Butorides virescens*)) using this method due to the model not converging. I encountered additional issues which prevented us from calculating FP-based thresholds for 5 species (Table B.1). Of these five species, 2 (Summer Tanager (*Piranga rubra*) and Eastern Towhee (*P. erythrophthalmus*)) could not have valid thresholds calculated using the FP-based threshold methods due to their calculated thresholds exceeding a value of 1.0, which is the maximum possible confidence value for BirdNET detections. The other 3 species had incalculable thresholds using the FP-based threshold methods due to all 100 random BirdNET reports for each of these three species (Eastern Wood-Pewee (*Contopus virens*), Great Crested Flycatcher (*Myiarchus crinitus*), and Song Sparrow (*Melospiza melodia*)) being true positive reports, meaning there was no possible way to calculate species-specific threshold scores using this methodology. I was unable to calculate either a FP-based threshold or derive a Modeled threshold for the Eastern Towhee (*P. erythrophthalamus*).

Although there were several species for which I was unable to calculate one or both species-specific thresholds for, it is possible that this could potentially be due to an insufficient range of confidence scores represented by the 100 random samples validated for some species. On behalf of my interest vested in streamlining the process of filtering BirdNET outputs, I elected to validate 100 random BirdNET reports per species as was practiced by Cole et al. (2022), rather than the higher number of reports per species used by Bota et al. (2023). Due to the majority of BirdNET detections having confidence scores at the lower end of the range of possible values, coupled with my randomized sampling scheme for extracting individual species detections, a majority proportion of the manually validated detections were those with lower confidence scores. By intentionally stratifying the sampling scheme to select BirdNET detections representing the full range of possible confidence scores, it is perhaps possible to combat this issue. Despite this, only 2 of the 10 species for which I was unable to calculate one or both species-specific thresholds did not have the full range of possible confidence scores represented, meaning that this did not pose a problem for most species in my dataset. Nevertheless, given that I was also unable to calculate FP-based thresholds for 3 species due to them having 100 true positive detections, it is probable that by manually validating an additional number of random samples, at least one false positive could be found which would enable an FP-based threshold to be calculated; however, under such circumstances, it may be sufficient to conclude that in the geographic location and season in which the recording period for the validated dataset took place, BirdNET can accurately identify such a species with a high enough precision to perhaps lessen the need for any confidence-based filtering beyond BirdNET's default 0.1 cut-off for that species.

By filtering the unfiltered dataset with any of the threshold types, many statistically significant improvements were achieved, including a decrease in the number of false positive species reported per site visit, as well as an increase in precision, specificity, and in the case of the Modeled thresholds only—a statistically significant increase in MCC. Unfortunately, although the use of confidence-based thresholds proved to be quite proficient at excluding false positive species reports to reach an acceptable level of precision of 0.95 or greater for two of the threshold types (FP-based thresholds and the arbitrary 0.75 threshold), such methods also resulted in a statistically significant increase in the number of false negative species errors per visit, and subsequently a decrease in sensitivity and F1 harmonic mean when applying any of the threshold types tested. As mentioned previously, regardless of the value of the value of a confidence-based threshold used to filter out false positives, as long as an overlap in confidence score values exists between true and false positive detections for a species, no confidence-based threshold filter would be able to exclude all false positives without also excluding at least some true positive values, thus increasing the number of false negative errors. This increase in false negative errors resulting from the use of confidence-based thresholds was also observed by Cole et al. (2022), to which they suggested increasing the sampling duration to combat the increased false negative errors—a recommendation which they backed by demonstrating that the likelihood of obtaining at least one detection with a confidence score meeting the threshold requirements increases as the sampling duration increases (Cole et al., 2022).

The finding that warbler and vireo species richness increased more than double in privately owned blocks (5.42) compared to publicly owned blocks (2.15); however, this seems to be primarily since the mean richness for private blocks for the IBBA (7) was lower than the mean richness among the public blocks for the IBBA (11.69) (Table B.5). Additionally, the lower species richness detected among the privately owned blocks during the IBBA may be more likely due to the discrepancy between survey efforts between private versus public blocks, as private blocks were surveyed for a mean of 29 hours per block for the IBBA whereas public blocks were surveyed for a mean of 62.49 hours per block—more than double the effort for the private blocks. It is quite likely that this difference in survey effort between private and public blocks during the IBBA was due to the difficulty in securing permission from landowners to survey their property—an issue which we also had in some blocks and thus prevented us from being able to sample some blocks within the 11 southernmost Illinois counties. The difficulty in securing landowner permission would mean that the IBBA volunteers had less land, and potentially fewer habitat types, to survey in blocks which were privately owned. In our dataset, we included only blocks in which we were able to secure access to survey 3 sub-blocks per primary block, thus attempting to prevent such bias. It is thus likely that privately owned blocks were under-surveyed during the IBBA, and thus are not as well represented as the publicly owned blocks surveyed during the IBBA. As such, it is likely that comparisons made between publicly owned blocks between the IBBA and 2022 datasets are more accurate than those between privately owned blocks. Nevertheless, an overall increase in species richness was detected even between the publicly owned blocks between the IBBA and 2022 datasets.

For PAM users who hope to streamline the process of excluding false positives from BirdNET outputs in the quickest way possible while still ensuring a high precision and who hope to bypass validating a subset of species detections, setting the BirdNET threshold confidence cut-off to a value of 0.75 may offer a quick and reliable method of filtering out nearly all false positives. Although I found the 0.75 threshold to have a mean precision of 0.99 ± 0.08 SD in my study, it should be noted that my investigation excluded species for which I could not calculate species-specific thresholds for, which was often due to either observing a negative relationship between the probability of a true positive detection and the confidence score of a detection, due to the species-specific threshold never reaching a probability of true detection = 0.95 across any confidence value, or due to the species-specific threshold value being calculated at a value greater than 1.0. As such, without conducting such a pilot study to first ensure that the target species can be reliably detected by the automatic species classifier, it is possible that a precision value lower than that I observed for the 0.75 threshold filter would be attained. Additionally, although I observed the 0.75 threshold as having the lowest mean sensitivity (0.18 ± 0.09 SD), Cole et al. (2022) noted that, when using a confidence-based threshold to filter BirdNET output, the probability of detecting a species that is truly present increases as the length of the recording increases, and thereby making it is possible to combat the low sensitivity of such a conservative confidence score by increasing the amount of recording time during peak avian vocalization hours.

Changes in Species Composition and Richness

It is likely that some portion of the observed increase in warbler and vireo species richness may be due to increased observation hours per block. By using PAM efforts rather than point counts, I was able to increase the average number of hours each block was surveyed, as I achieved an average of 76.71 (SD = 96.34) hours of observation per block whereas the IBBA achieved an average of 47.8 (SD = 11.06) hours of observation per block. Given this increase in the average number of hours of observation, coupled with the fact that a positive relationship exists between the number of hours of observation and number of species detected per block, it is likely that a portion of the observed increase in species richness might be due to the increase in the average number of observation hours per block. Nevertheless, although the average number of warbler/vireo species detected per hour of observation for the IBBA (0.45 ± 0.30 SD) was higher than that of the 2022 dataset (0.18 ± 0.033 SD), this most likely has two explanations. First, less additional species were detected with each additional hour. Additionally, it is possible that human observers may detect bird vocalizations at a greater distance than PAM methods, and thus less hours of observation may be necessary to detect all present species with point counts than with the use of PAM. Furthermore, the fact that a higher number of species were detected per hour of observation for the IBBA data than for the 2022 data further supports that my findings represent a true increase in warbler and vireo species richness over time rather than the observed increase in richness being due to an increase in the average number of observation hours per block.

Furthermore, another possibility which could contribute towards the observed increase in species richness might be the result of some portion of the presence data consisting of false positive species reports. Although this is a possibility, given that the FP-based Thresholds tested to have an average precision = 0.95 (SD = 0.08), I believe it to be unlikely that the observed increase could be solely—or even majorly—explained by false positive species reports. Given that my precision calculation using the FP-based Threshold calculation method was based on species reports rather than individual detections, this would imply that approximately 5% of species detected per block might be accounted for by false positive species errors. Given that the average number of species detected per block was 13.24 (SD = 2.01), this would mean that it is unlikely that more than a single species detected can be attributed to a false positive (21 × 0.05 = 0.66). Furthermore, given that the average precision of the FP-based filtering method greatly exceeds the sensitivity, there is a higher probability of false negative species errors rather than false positive species errors.

In addition to the average 36.38% increase in species richness, a significant shift in warbler and vireo species composition was observed between the IBBA (1986—1991) and the data (2022). The average Jaccard's Similarity Index value of 0.49 was lower than—and thus suggests less similarity—than the average Sørensen's Similarity Index value of 0.63; albeit, both values indicate considerably low similarity in species composition between the same blocks over the 30-year period. The average species turnover rate of 50.82% suggests a significant shift in the warbler and vireo species composition has occurred over the span of the last three decades in Southern Illinois. To compare, a study in which avian species composition and richness between the same sites surveyed 20 years apart in Connecticut documented only a 20% species turnover over the 20-year period (Craig, 2024)—significantly less than the 50.82% species turnover over I documented across the blocks over a 30-year period. The increase in species richness, combined with the fact that the Jaccard's Similarity Index value was lower than the Sørensen's Similarity Index, suggests that the lower Jaccard's value (indicating lower similarity) may be due to the richness increase contributing more towards the dissimilarity in species composition than the "replacement" of species. However, with the added context of the average species turnover rate being so high, this indicates that not only have new warbler and vireo species colonized the blocks, but the warbler and vireo community is very vulnerable to population shifts on a local scale, with Southern Illinois showing high dynamism in avian species composition.

Of the 6 warbler and vireo species which declined in occurrence among the 45 re-sampled blocks, none of these species were found by the BBS to be experiencing statewide declines during the same 30-year time frame, suggesting that although these species may not be experiencing statewide declines, they are still experiencing declines in the Southern Illinois region (Table B.2; Ziolkowski et al., 2022). On the contrary, Illinois BBS data showed a statewide decline in 2 species (Black-and-white Warblers (*M. varia*) and Yellow-throated Vireos (*V. flavifrons*)) while I detected an increase in local occurrence for these species (Table B.2; Ziolkowski et al., 2022). Zooming out to compare my results with continental population trends, 4 of the 6 species for which I detected decreases in occurrence between the IBBA and 2022 were also found to be decreasing on a continental scale by Partners in Flight (2022; Table B.2). Additionally, 2 species (Common Yellowthroats (*G. trichas*) and Prothonotary Warblers (*P. citrea*)) which were found to be decreasing on a continental scale by Partners in Flight were species which I found to either be stable or increasing in occurrence locally (Table B.2; Partners in Flight, 2022). Interestingly, the population trends I observed aligned more closely with the population trends observed on a continental scale by Partners in Flight (2022) than with statewide BBS data for Illinois (Ziolkowski, 2022). Nevertheless, population trends deviating from both the statewide and continental-wide trends were observed in my study (Table B.2), highlighting the importance of conducting long-term, probabilistic population monitoring surveys.

Of the 6 species which declined in occurrence between the IBBA to 2022, 2 of these species (Prairie Warblers (*S. discolor*) and Kentucky Warblers (*G. formosa*)) are on the Yellow Watch List (Partners in Flight, 2020), while 3 are listed as Species of Greatest Conservation Need by the Illinois Wildlife Action Plan, including Kentucky Warblers (*G. formosa*), Prairie Warblers (*S. discolor*), and Blue-winged Warblers (*V. cyanoptera*) (Illinois Wildlife Action Plan, 2022). Interestingly, all species I detected as having declined in number of occupied blocks all preferred either open woodlands or scrub habitats, except for the Kentucky Warbler (*G. formosa*) (Cornell Lab of Ornithology, 2019a-f). Equally interestingly is that, of the 14 species which increased in the number of blocks detected, only 2 of these species prefer open woodlands (Yellow-throated Vireos (*V. flavifrons*) and Golden-winged Warblers (*V. chrysoptera*)), while only 1 preferred scrub habitat (Bell's Vireos (*V. bellii*)) (Cornell Lab of Ornithology, 2019g-u). This pattern may therefore signify that an increase in open woodland and scrubby habitat in Southern Illinois is necessary to prevent further decline for these declining species. Also interesting was that a Red Watch List species—the Golden-winged Warbler (*V. chrysoptera*), increased in the number of blocks detected from the IBBA data to the 2022 data (Partners in Flight, 2020).

Although a portion of the southern Illinois region is considered within the breeding range of the Golden-winged Warbler (*V. chrysoptera*) by the Cornell Lab of Ornithology's range map for the species (Cornell Lab of Ornithology, 2020), the species seems to be somewhat "new" to breeding in Southern Illinois, as it was not documented in the 45 blocks during the IBBA, nor did any confirmed eBird reports exist within the 11 southernmost Illinois counties between mid-May to late June until its first confirmed eBird sighting in Jackson County in 2017 (Sullivan et al., 2009). Since its first confirmed eBird sighting during the breeding season in Jackson County in 2017, it has been sighted during the breeding season every year since then in Jackson County except for 2019 and 2023, thus suggesting that the species is truly present in the region—albeit rare (Sullivan et al., 2009). This expansion into Southern Illinois is quite significant given that the Golden-winged Warbler (*V. chrysoptera*) is a Red Watch List species and is also designated as a Species of Greatest Conservation Need by the Illinois Wildlife Action Program (Partners in Flight, 2020; Illinois Wildlife Action Plan, 2022). Additional species which are of conservation concern for which I observed an increase in the number of blocks detected include Prothonotary Warblers (*P. citrea*) and Ovenbirds (*S. aurocapilla*)—both of which are also listed as Species of Greatest Conservation Need (Illinois Wildlife Action Plan, 2022).

CONCLUSION

Due to the overlap in confidence score values which exist between both true and false positive detections for each species, it is impossible to derive a confidence-based threshold which can successfully exclude all false positive detections without excluding at least some true positive detections—thus resulting in an increase in false negatives. Despite this, by increasing the number of observation hours (by increasing the amount of audio collected during peak vocalization time at a given site), it is possible to increase the probability that at least one detection for the species of interest will have a confidence score high enough to exceed the threshold value being used to filter out false positive detections. Such methods therefore easily lend themselves to investigations inquiring about species richness and composition, but relying solely upon species-specific confidence threshold methods to filter automatic species classifier output may be ill-suited for other potential uses—particularly when research questions require the researcher to understand finer-scale details, such as when and how frequently the species vocalizes, or how many unique individuals are present, for example. Additionally, depending on how many species the researcher is focused on, it may be more fruitful to individually validate all detections for a given species—assuming that the researcher is focused with only one or a few species. Moreover, the purpose of the threshold comparison was not to advise all users of PAM to adopt these methods, but rather, to assess and compare the performance of individual, pre-existing methods of streamlining automatic species classifier output, and to give an example of practical applications for which these methods are an excellent choice. Furthermore, although my research found the FP-based threshold method to be the best choice for my purposes, given my desire to implement whichever threshold tested achieved the highest mean sensitivity while requiring a mean precision ≥ 0.95, I acknowledge that this is not the only adequate threshold type available, and it may be possible to derive a threshold meeting my precision requirements while achieving a greater mean sensitivity than did the FP-based threshold.

Although no "perfect" threshold can exist, either choosing a conservative threshold requiring minimum post-filtering validation as I did, or intentionally selecting a less conservative threshold which may require additional validation post-filtering, are both fine options for researchers interested in streamlining audio validation with the use of confidence-based thresholds. Additionally, unexplored possibilities for streamlining the validation process of automatic species classifier data remain, as the use of confidence-based thresholds alongside the number of detections per unit of time remains unexplored to my knowledge. Potential exists for these methods to work in tandem, such as by decreasing the required species-specific threshold for recordings which have a minimum number of detections for a given species, or perhaps by requiring either a minimum number of detections per recording or a detection exceeding the confidence-based threshold value for a species to be considered present for a given site and sampling date.

I advise against applying my exact species-specific threshold values to other geographic regions or seasons, as there are several factors which could influence the reliability of confidence scores assigned to BirdNET detections for a given species. For example, in this study system, I found that BirdNET's assignment of confidence values to Cerulean Warbler (*S. cerulea*) detections was too unreliable for us to reliably filter false positives using a confidence-based threshold, as the validated Cerulean Warbler (*S. cerulea*) detection with the highest confidence value was found to be a false positive detection rather than a true positive detection. Despite this, I believe that this may be the result of a region-specific issue on account of this study system having an abundance of Northern Parulas (*S. americana*)—a species which frequently was misidentified as a Cerulean Warbler (*S. cerulea*) by BirdNET in my validation dataset, and thus in portions of the Cerulean Warbler (*S. cerulea*) range in which Northern Parulas (*S. americana*) are either absent or less abundant, BirdNET may be more reliable in such circumstances. Additionally, due to the tendency of many bird species with large geographic ranges to have a variety of regional "dialects," it is quite possible that BirdNET may perform better or worse at identifying the same species in different portions of its range (Baker & Cunningham, 1985). In addition, the change in species composition, vocalization behavior, and the impact of juvenile vocalizations (during periods in which a higher proportion of a population is comprised of juveniles), all serve as reasons for why researchers should avoid using the same species-specific threshold values year-round. Given these arguments, I strongly urge researchers who plan to use confidence-based thresholds to conduct a pilot study using audio collected during the time of year and within the geographic location for which they intend to collect the audio data they intend to filter using confidence-based thresholds. In addition to yielding more accurate species-specific threshold values for an individual system and season, this process will also enable the researcher to discern whether their focal species is a good candidate for use with their automatic species classifier of choice.

Although it is likely that some of the observed changes in species richness and composition can be attributed to an increase in observation hours given the positive relationship which exists between the number of observation hours and number of species detected, it is difficult to ascertain exactly how much of this increase can be attributed to the increase in the average number of observation hours per block given the difference in observation methods between the two surveys (point count versus PAM). Regardless, given that an average 36.38% increase in number of species detected per block was observed, I believe it to be unlikely that the IBBA observers missed such a high percentage of species per block on average as to account for the observed increase in species richness. This case study excellently demonstrated some of the challenges, as well as benefits, that can be expected when switching from point count data to PAM data for avian population monitoring. On one hand, PAM enables a small research team to survey a large geographic region in a single season for many more observation hours than would be feasible using traditional point count methods with the same number of people and with the same budget for the project, yet on the other hand, the difference in survey methods can cause uncertainty when attempting to compare historical point count-based data with PAM data. Nevertheless, the promise of PAM is too great to ignore, and in a time when birds are decreasing at such alarming rates (Rosenberg et al., 2019), finding ways to maximize the geographic scope and frequency at which large-scale population monitoring efforts can be conducted is paramount to successful avian conservation. Furthermore, as PAM continues to grow in use, it will eventually become much easier to compare PAM data with other PAM data collected from different time periods, allowing for a more 1:1 comparison to be made for the future of avian population monitoring. Additionally, I hope to see the continuance of the use of PAM for avian population monitoring in Southern Illinois—not only to corroborate the patterns observed in this study, but also to continue to further our understanding of the dynamism of the Southern Illinois avian community.

TABLES AND FIGURES

Tables

Table 2.1: Single factor (Threshold) mixed-effects nested ANOVA to compare effects of threshold type used for filtering data on: a.) the number of true positive species reported per visit, b.) the number of false positive species reported per visit, c.) the number of false negative species reported per visit, d.) sensitivity, e.) precision, f.) specificity, g.) F1 harmonic mean, and h.) MCC (Matthews correlation coefficient). Date (of visit) is nested within site (of visit). Type III ANOVA with Satterthwaite's method was used. Data used was from all 36 site visits conducted between June 21—July 4, 2023, in Jackson County, Illinois, USA.

Single Factor, Mixed-Effects Nested ANOVA

a. True Positive Species

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Threshold	4	1153.7	288.43	155.45	$2.2e^{-16}$

b. False Positive Species

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Threshold	4	1119.1	279.78	156.22	$2.2e^{-16}$

c. False Negative Species

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Threshold	4	1157	289.25	155.92	$2.2e^{-16}$

d. Precision

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Threshold	4	2.79	0.69653	122.15	$2.2e^{-16}$

e. Sensitivity

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Threshold	4	2.69	0.67358	213.67	$2.2e^{-16}$

a. Specificity

	Df	Sum Sq	Mean Sq	F value	Pr(>F)

Threshold	4	0.582	0.1456	145.92	$2.2e^{-16}$

b. F1 Harmonic Mean

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Threshold	4	1.583	0.39585	102.33	$2.2e^{-16}$

c. MCC

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Threshold	4	0.327	0.081802	21.91	$1.158e^{-13}$

Table 2.2: Table containing P-values for all pairwise comparisons resulting from the Tukey's HSD post-hoc test for threshold types, including unfiltered ("unfiltered"/no threshold used), 0.5-filtered, 0.75-filtered, Modeled Threshold, and FP-based Threshold. P-values for the pairwise comparisons contrasted the significant difference between different threshold types on a.) the average number of true positive species reported per visit; b.) the average number of false positive species reported per visit; c.) the average number of false negative species reported per visit; d.) mean precision; e.) mean sensitivity; f.) mean specificity; g.) mean F1 Harmonic Mean; and h.) mean MCC (Matthews Correlation Coefficient).

a. True Positive Species

Threshold Type	Unfiltered	0.5-filtered	0.75-filtered	Modeled Threshold	FP-based Threshold
Unfiltered	__	<0.0001	<0.0001	<0.0001	<0.0001
0.5-filtered	<0.0001	__	<0.0001	0.0001	0.2362
0.75-filtered	<0.0001	<0.0001	__	<0.0001	<0.0001
Modeled Threshold	<0.0001	0.0001	<0.0001	__	0.0767
FP-based Threshold	<0.0001	0.2362	<0.0001	0.0767	__

b. False Positive Species

Threshold Type	Unfiltered	0.5-filtered	0.75-filtered	Modeled Threshold	FP-based Threshold
Unfiltered	__	<0.0001	<0.0001	<0.0001	<0.0001
0.5-filtered	<0.0001	__	0.8075	0.8688	0.9998
0.75-filtered	<0.0001	0.8075	__	0.2439	0.8877
Modeled Threshold	<0.0001	0.8688	0.2439	__	0.7824
FP-based Threshold	<0.0001	0.9998	0.8877	0.7824	__

c. False Negative Species

Threshold Type	Unfiltered	0.5-filtered	0.75-filtered	Modeled Threshold	FP-based Threshold
Unfiltered	__	<0.0001	<0.0001	<0.0001	<0.0001
0.5-filtered	<0.0001	__	<0.0001	0.0001	0.2361
0.75-filtered	<0.0001	<0.0001	__	<0.0001	<0.0001
Modeled Threshold	<0.0001	0.0001	<0.0001	__	0.0767
FP-based Threshold	<0.0001	0.2361	<0.0001	0.0767	__

d. Precision

Threshold Type	Unfiltered	0.5-filtered	0.75-filtered	Modeled Threshold	FP-based Threshold
Unfiltered	__	<0.0001	<0.0001	<0.0001	<0.0001
0.5-filtered	<0.0001	__	0.1682	0.4788	0.9965
0.75-filtered	<0.0001	0.1682	__	0.0016	0.3252
Modeled Threshold	<0.0001	0.4788	0.0016	__	0.2759
FP-based Threshold	<0.0001	0.9965	0.3252	0.2759	__

e. Sensitivity

Threshold Type	Unfiltered	0.5-filtered	0.75-filtered	Modeled Threshold	FP-based Threshold

Unfiltered	__	<0.0001	<0.0001	<0.0001	<0.0001
0.5-filtered	<0.0001	__	<0.0001	<0.0001	0.1022
0.75-filtered	<0.0001	<0.0001	__	<0.0001	<0.0001
Modeled Threshold	<0.0001	<0.0001	<0.0001	__	0.0366
FP-based Threshold	<0.0001	0.1022	<0.0001	0.0366	__

f. Specificity

Threshold Type	Unfiltered	0.5-filtered	0.75-filtered	Modeled Threshold	FP-based Threshold
Unfiltered	__	<0.0001	<0.0001	<0.0001	<0.0001
0.5-filtered	<0.0001	__	0.8425	0.8714	0.9999
0.75-filtered	<0.0001	0.8425	__	0.2801	0.8990
Modeled Threshold	<0.0001	0.8714	0.2801	__	0.8076
FP-based Threshold	<0.0001	0.9999	0.8990	0.8076	__

g. F1 Harmonic Mean

Threshold Type	Unfiltered	0.5-filtered	0.75-filtered	Modeled Threshold	FP-based Threshold
Unfiltered	__	<0.0001	<0.0001	<0.0001	<0.0001
0.5-filtered	<0.0001	__	<0.0001	<0.0001	0.0559
0.75-filtered	<0.0001	<0.0001	__	<0.0001	<0.0001
Modeled Threshold	<0.0001	<0.0001	<0.0001	__	0.0725
FP-based Threshold	<0.0001	0.0559	<0.0001	0.0725	__

h. Matthews Correlation Coefficient

Threshold Type	Unfiltered	0.5-filtered	0.75-filtered	Modeled Threshold	FP-based Threshold
Unfiltered	__	0.9995	<0.0001	0.0081	0.1230

0.5-filtered	0.9995	__	<0.0001	0.0157	0.1931
0.75-filtered	<0.0001	<0.0001	__	<0.0001	<0.0001
Modeled Threshold	0.0081	0.0157	<0.0001	__	0.8589
FP-based Threshold	0.1230	0.1931	<0.0001	0.8589	__

Table 2.3: Table displaying the mean number of true positive (TP), false positive (FP), and false negative (FN) species reported per threshold type, as well as the mean values of precision, sensitivity, specificity, F1 harmonic mean, and MCC (Matthews Correlation Coefficient) for each threshold type used and for the unfiltered data (data pre-filtering).

Threshold Used	TP	FP	FN	Precision	Sensitivity	Specificity	F1	MCC
Unfiltered	11.31	6.58	9.14	0.64	0.55	0.85	0.58	0.42
0.5-filtered	5.58	0.39	14.86	0.94	0.28	0.99	0.42	0.43
0.75-filtered	3.69	0.03	16.94	0.99	0.18	1.00	0.30	0.35
Modeled Threshold	7.08	0.69	13.36	0.91	0.35	0.98	0.50	0.47
FP-based Threshold	6.25	0.33	14.19	0.95	0.31	0.99	0.46	0.46

Table 2.4: The number of blocks each species was detected in the IBBA dataset, the 2022 dataset, and the number of increase or decrease in blocks detected per species from the IBBA to 2022 dataset. Species are listed by current AOS standardized common names and by scientific names.

		Blocks	Blocks	Change in Blocks
Common Name	Scientific Name	Detected	Detected	Detected
		(IBBA)	(2022)	(IBBA to 2022)
American Redstart	*Setophaga ruticilla*	8	38	30
Bell’s Vireo	*Vireo bellii*	3	11	8

Black-and-white Warbler	*Mniotilta varia*	6	19	13
Blue-winged Warbler	*Vermivora cyanoptera*	8	2	-6
Common Yellowthroat	*Geothlypis trichas*	45	45	0
Golden-winged Warbler	*Vermivora chrysoptera*	0	7	7
Hooded Warbler	*Setophaga citrina*	9	22	13
Kentucky Warbler	*Geothlypis formosa*	34	30	-4
Louisiana Waterthrush	*Parkesia motacilla*	26	40	14
Northern Parula	*Setophaga americana*	35	42	7
Ovenbird	*Seiurus aurocapilla*	15	38	23
Pine Warbler	*Setophaga pinus*	9	20	11
Prothonotary Warbler	*Setophaga discolor*	19	35	16
Prairie Warbler	*Protonotaria citrea*	20	11	-9
Red-eyed Vireo	*Vireo olivaceus*	38	44	6
Warbling Vireo	*Vireo gilvus*	28	23	-5
White-eyed Vireo	*Vireo griseus*	41	38	-3
Worm-eating Warbler	*Helmitheros vermivorum*	17	44	27
Yellow Warbler	*Setophaga petechia*	24	2	-22

Yellow-throated Vireo	*Vireo flavifrons*	29	41	12
Yellow-throated Warbler	*Setophaga dominica*	23	44	21

Table 2.5: For each of the 45 IBBA primary blocks sampled, the number of hours of observation (hrs/block) for the IBBA survey and the 2022 survey are listed, as well as warbler and vireo species richness for both the IBBA survey and the 2022 survey, as well as Jaccard's Similarity Index (JSI), Sørensen's Similarity Index (SSI), and the percentage of Species Turnover.

Block ID	Hrs/Block (IBBA)	Hrs/Block (2022)	Richness (IBBA)	Richness (2022)	JSI	SSI	% Species Turnover
278A3	24	64	15	14	0.53	0.69	47.37
281D3	216.5	76	12	14	0.73	0.85	26.67
278D3	16	60	13	14	0.80	0.89	20.00
271D3	8	72	3	18	0.17	0.29	83.33
274D3	26	84	11	13	0.41	0.58	58.82
270A3	103.5	84	17	14	0.63	0.77	36.84
274B3	30	68	10	16	0.44	0.62	55.56
265C3	29	76	16	12	0.75	0.86	25.00
262D3	60.5	88	3	13	0.14	0.25	85.71
270B3	134.9	84	13	16	0.71	0.83	29.41
271B3	39.1	72	14	14	0.56	0.71	44.44

271A3	37	88	15	15	0.76	0.87	23.53
272B3	9.5	120	8	13	0.50	0.67	50.00
272A3	7	76	9	14	0.64	0.78	35.71
273B3	7.7	76	6	13	0.36	0.53	64.29
273A3	19.5	60	5	13	0.20	0.33	80.00
274A3	29.8	60	10	15	0.56	0.72	43.75
270D3	15	72	11	16	0.50	0.67	50.00
266D3	57	72	13	12	0.67	0.80	33.33
265D3	69.1	72	13	12	0.56	0.72	43.75
264C3	116	72	4	9	0.18	0.31	81.82
264D3	14.5	68	9	10	0.36	0.53	64.29
263D3	10.5	72	9	11	0.54	0.70	46.15
262C3	8.5	76	5	16	0.31	0.48	68.75
261C3	16.5	80	1	13	0.08	0.14	92.31
261D3	10	84	2	12	0.17	0.29	83.33
279B3	36	84	10	13	0.53	0.70	46.67
278B3	12	60	12	11	0.53	0.70	46.67
280D3	16	84	7	13	0.43	0.60	57.14

279C3	10	84	11	14	0.39	0.56	61.11
279D3	27.1	72	5	13	0.29	0.44	71.43
278C3	23	68	7	10	0.42	0.59	58.33
283A3	16.5	84	9	13	0.69	0.82	30.77
284B3	13.3	64	6	10	0.33	0.50	66.67
286B3	15.5	68	6	10	0.33	0.50	66.67
286A3	18	72	10	13	0.44	0.61	56.25
271C3	7.5	72	4	15	0.27	0.42	73.33
272C3	26.5	100	9	11	0.43	0.60	57.14
272D3	33	84	12	16	0.75	0.86	25.00
273C3	67.3	84	16	16	0.78	0.88	22.22
273D3	14	72	14	12	0.63	0.77	37.50
274C3	37.5	72	14	13	0.80	0.89	20.00
275C3	21	84	14	14	0.65	0.79	35.29
281A3	623.5	84	18	16	0.89	0.94	11.11
280B3	18	84	6	11	0.31	0.47	69.23

Table 2.6: North American Breeding Bird Survey counts for (data from Illinois routes only) by species for 1991 and 2022. Species listed by both common name and scientific name.

Common Name	Scientific Name	1991	2022
American Redstart	Setophaga ruticilla	4	23
Bell's Vireo	Vireo bellii	2	35
Black-and-white Warbler	Mniotilta varia	1	0
Blue-winged Warbler	Vermivora cyanoptera	0	2
Common Yellowthroat	Geothlypis trichas	566	1115
Golden-winged Warbler	Vermivora chrysoptera	1	0
Hooded Warbler	Setophaga citrina	0	3
Kentucky Warbler	Geothlypis formosa	9	53
Louisiana Waterthrush	Parkesia motacilla	2	5
Northern Parula	Setophaga americana	11	119
Ovenbird	Seiurus aurocapilla	2	2
Pine Warbler	Setophaga pinus	4	4
Prairie Warbler	Setophaga discolor	15	19
Prothonotary Warbler	Protonotaria citrea	4	32
Red-eyed Vireo	Vireo olivaceus	28	138
Warbling Vireo	Vireo gilvus	139	346
White-eyed Vireo	Vireo griseus	45	229

Worm-eating Warbler	Helmitheros vermivorum	1	4
Yellow Warbler	Setophaga petechia	55	70
Yellow-throated Vireo	Vireo flavifrons	24	22
Yellow-throated Warbler	Setophaga dominica	2	19

Figures

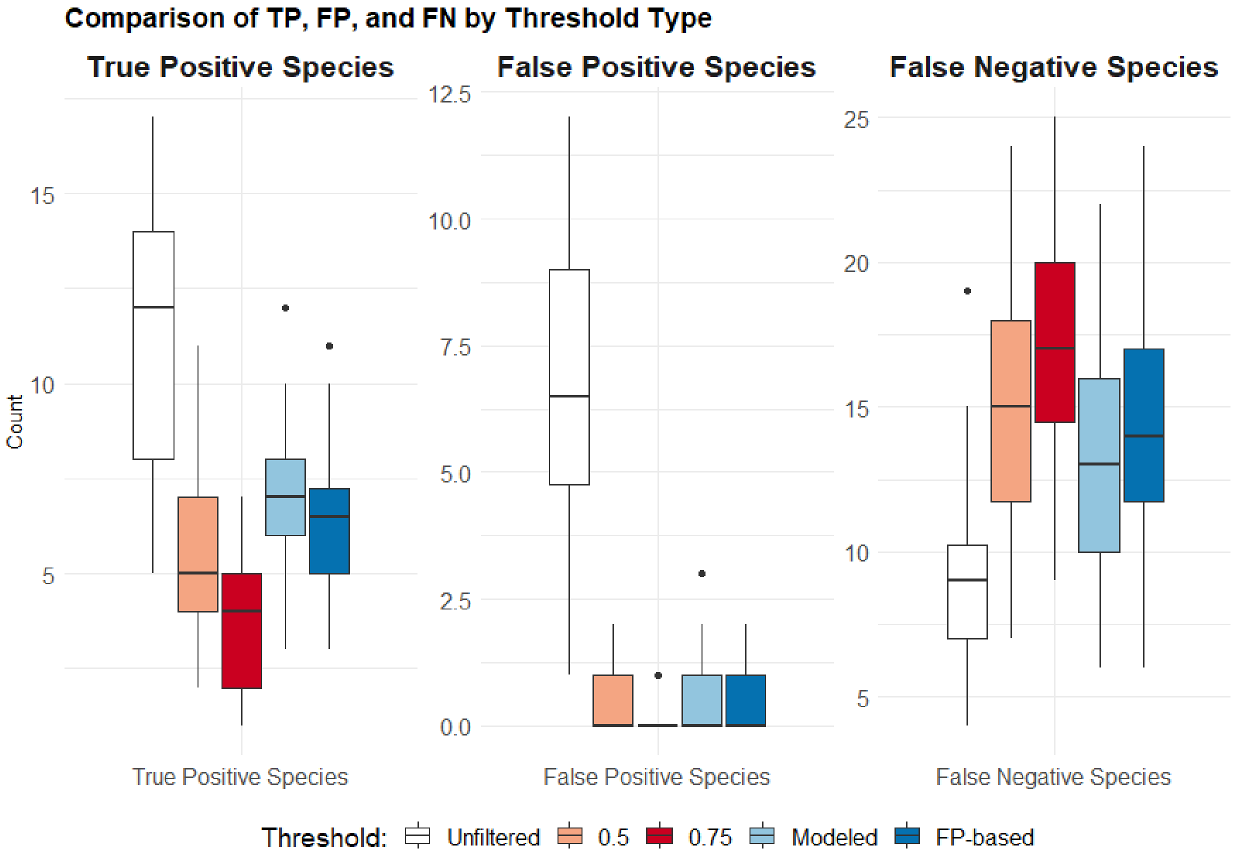

Figure 2.1: Boxplots depicting the number species detected per visit made by BirdNET, including true positive species detected (left), false positive species (middle), and false negative species errors (right). Values depicted represent all threshold types, including, from left to right: Unfiltered (white), 0.5 (orange), 0.75 (red), Modeled (light blue), and FP-based (dark blue). Audio data was validated using point count data collected concurrently for the full duration of each visit. Six sites were visited six times each for a total of 36 site visits used for this data. Data was collected in Jackson County, Illinois, USA, between June 21 to July 4, 2023.

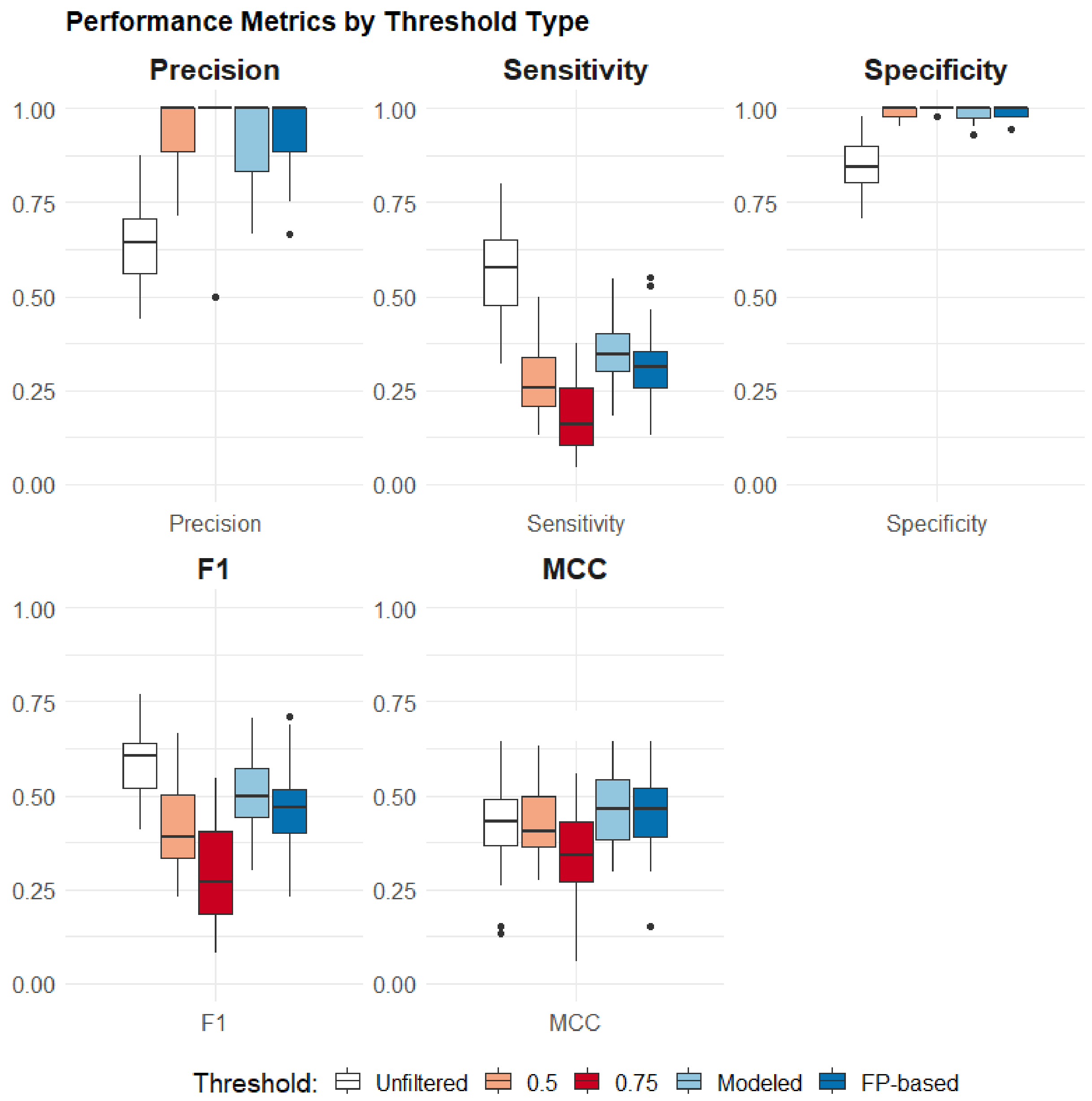

Figure 2.2: Boxplots comparing performance metrics precision, sensitivity, specificity, F1 harmonic mean, and MCC (Matthews Correlation Coefficient). Values depicted for all five threshold types, including unfiltered. Values depicted represent all threshold types, including, from left to right: Unfiltered (white), 0.5 (orange), 0.75 (red), Modeled (light blue), and FP-based (dark blue). Six sites were visited six times each for a total of 36 site visits used for this data. Data was collected in Jackson County, Illinois, USA, between June 21 to July 4, 2023.

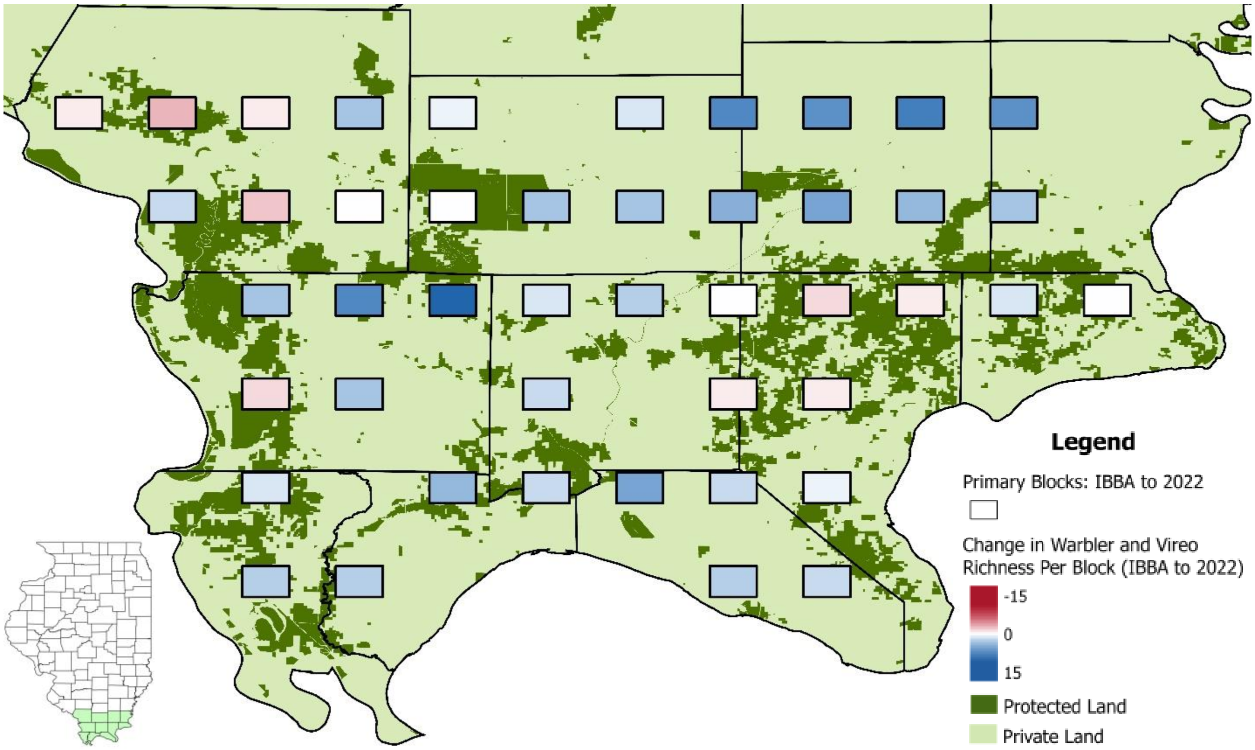
Legend
Primary Blocks: IBBA to 2022
Change in Warbler and Vireo Richness Per Block (IBBA to 2022)
-15
0
15
Protected Land
Private Land

Figure 2.3: Map depicting primary blocks from the Illinois Breeding Bird Atlas which were re-sampled in 2022. Block coloration dependent upon the number of warbler and vireo species lost or gained from the last Breeding Bird Atlas (1986—1991) to 2022. Primary blocks are overlaid over map depicting privately owned, non-protected land (light green) versus publicly owned and protected land (dark green). Map depicts the 11 southernmost Illinois counties.

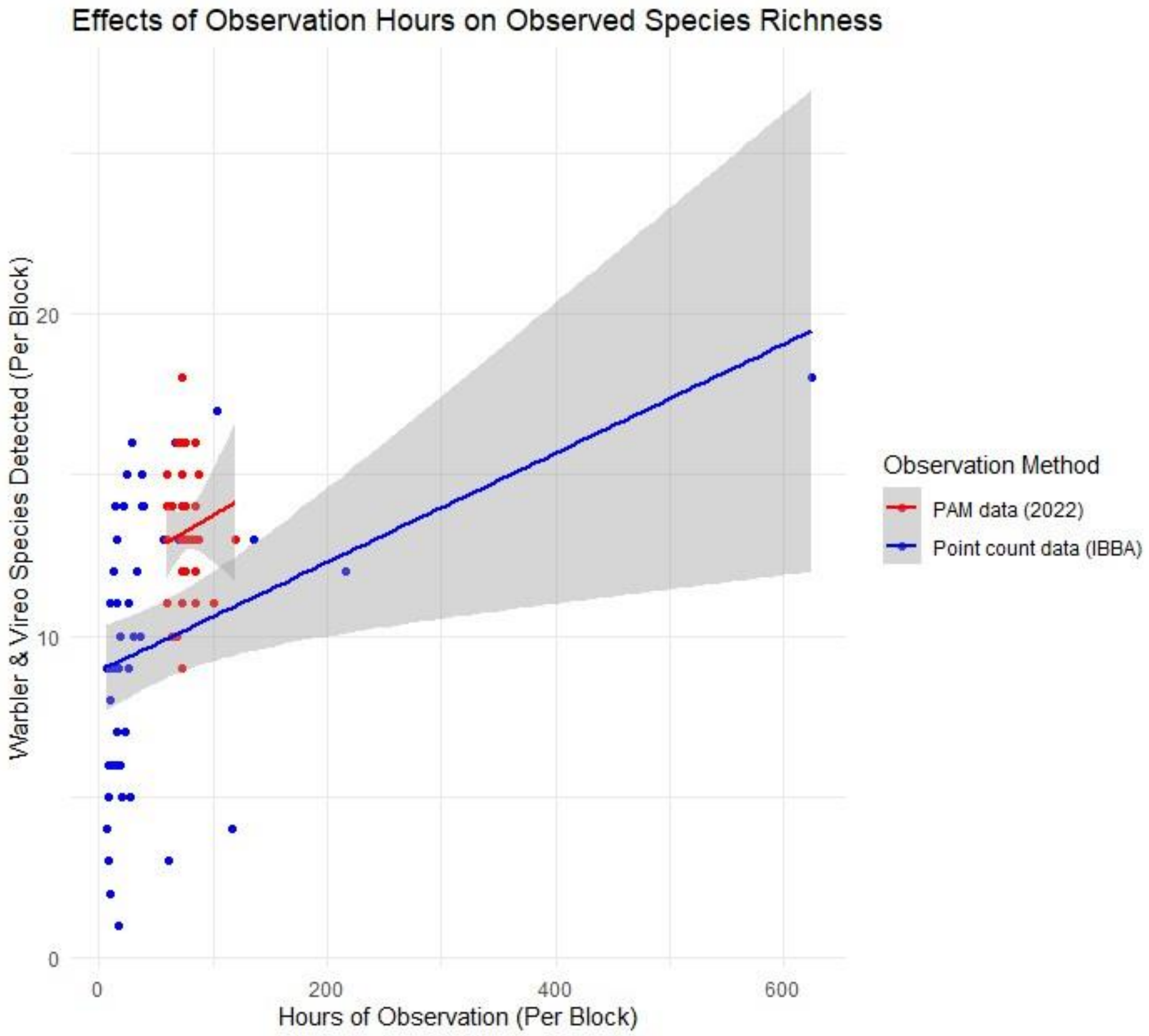

Figure 2.4: Line graph depicting positive linear relationship between the number of observation hours and the observed species richness. Passive acoustic monitoring (PAM) data collected in 2022 shown in red. Point count data collected during the IBBA shown in blue.

LITERATURE CITED

Alquezar, R. D., & Machado, R. B. (2015). Comparisons between autonomous acoustic recordings and avian point counts in open woodland savanna. *The Wilson Journal of Ornithology, 127*(4), 712-723.

Baker, M. C., & Cunningham, M. A. (1985). The biology of bird-song dialects. Behavioral and *Brain Sciences, 8*(1), 85-100.

Bibby, C. J. (2000). Bird census techniques. *Elsevier*.

Bibby, C., Jones, M., & Marsden, S. (2000). Bird surveys. *Birdlife International,* Cambridge, UK, 137.

Bota, G., Manzano-Rubio, R., Catalán, L., Gómez-Catasús, J., & Pérez-Granados, C. (2023). Hearing to the unseen: AudioMoth and BirdNET as a cheap and easy method for monitoring cryptic bird species. *Sensors, 23*(16), 7176.

Cole, J. S., Michel, N. L., Emerson, S. A., & Siegel, R. B. (2022). Automated bird sound classifications of long-duration recordings produce occupancy model outputs similar to manually annotated data. *Ornithological Applications*, *124*(2), duac003.

Cornell Lab of Ornithology. (2019a). All About Birds. Cornell Lab of Ornithology, Ithaca, New York. https://www.allaboutbirds.org/guide/Yellow_Warbler/lifehistory# Accessed 2024.

Cornell Lab of Ornithology. (2019b). All About Birds. Cornell Lab of Ornithology, Ithaca, New York. https://www.allaboutbirds.org/guide/Prairie_Warbler/lifehistory# Accessed 2024.

Cornell Lab of Ornithology. (2019c). All About Birds. Cornell Lab of Ornithology, Ithaca, New York. https://www.allaboutbirds.org/guide/Blue-winged_Warbler/lifehistory# Accessed 2024.

Cornell Lab of Ornithology. (2019d). All About Birds. Cornell Lab of Ornithology, Ithaca, New

York. https://www.allaboutbirds.org/guide/Warbling_Vireo/lifehistory# Accessed 2024.

Cornell Lab of Ornithology. (2019e). All About Birds. Cornell Lab of Ornithology, Ithaca, New York. https://www.allaboutbirds.org/guide/Kentucky_Warbler/lifehistory# Accessed 2024.

Cornell Lab of Ornithology. (2019f). All About Birds. Cornell Lab of Ornithology, Ithaca, New York. https://www.allaboutbirds.org/guide/White-eyed_Vireo/lifehistory# Accessed 2024.

Cornell Lab of Ornithology. (2019g). All About Birds. Cornell Lab of Ornithology, Ithaca, New York. https://www.allaboutbirds.org/guide/American_Redstart/lifehistory# Accessed 2024.

Cornell Lab of Ornithology. (2019h). All About Birds. Cornell Lab of Ornithology, Ithaca, New York. https://www.allaboutbirds.org/guide/Bells_Vireo/lifehistory# Accessed 2024.

Cornell Lab of Ornithology. (2019i). All About Birds. Cornell Lab of Ornithology, Ithaca, New York. https://www.allaboutbirds.org/guide/Black-and-white_Warbler/lifehistory# Accessed 2024.

Cornell Lab of Ornithology. (2019j). All About Birds. Cornell Lab of Ornithology, Ithaca, New York. https://www.allaboutbirds.org/guide/Common_Yellowthroat/lifehistory# Accessed 2024.

Cornell Lab of Ornithology. (2019k). All About Birds. Cornell Lab of Ornithology, Ithaca, New York. https://www.allaboutbirds.org/guide/Golden-winged_Warbler/lifehistory# Accessed 2024.

Cornell Lab of Ornithology. (2019l). All About Birds. Cornell Lab of Ornithology, Ithaca, New York. https://www.allaboutbirds.org/guide/Hooded_Warbler/lifehistory# Accessed 2024.

Cornell Lab of Ornithology. (2019m). All About Birds. Cornell Lab of Ornithology, Ithaca, New

York. https://www.allaboutbirds.org/guide/Louisiana_Waterthrush/lifehistory# Accessed 2024.

Cornell Lab of Ornithology. (2019n). All About Birds. Cornell Lab of Ornithology, Ithaca, New York. https://www.allaboutbirds.org/guide/Northern_Parula/lifehistory# Accessed 2024.

Cornell Lab of Ornithology. (2019o). All About Birds. Cornell Lab of Ornithology, Ithaca, New York. https://www.allaboutbirds.org/guide/Ovenbird/lifehistory# Accessed 2024.

Cornell Lab of Ornithology. (2019p). All About Birds. Cornell Lab of Ornithology, Ithaca, New York. https://www.allaboutbirds.org/guide/Pine_Warbler/lifehistory# Accessed 2024.

Cornell Lab of Ornithology. (2019q). All About Birds. Cornell Lab of Ornithology, Ithaca, New York. https://www.allaboutbirds.org/guide/Prothonotary_Warbler/lifehistory# Accessed 2024.

Cornell Lab of Ornithology. (2019r). All About Birds. Cornell Lab of Ornithology, Ithaca, New York. https://www.allaboutbirds.org/guide/Red-eyed_Vireo/lifehistory# Accessed 2024.

Cornell Lab of Ornithology. (2019s). All About Birds. Cornell Lab of Ornithology, Ithaca, New York. https://www.allaboutbirds.org/guide/Worm-eating_Warbler/lifehistory# Accessed 2024.

Cornell Lab of Ornithology. (2019t). All About Birds. Cornell Lab of Ornithology, Ithaca, New York. https://www.allaboutbirds.org/guide/Yellow-throated_Vireo/lifehistory# Accessed 2024.

Cornell Lab of Ornithology. (2019u). All About Birds. Cornell Lab of Ornithology, Ithaca, New York. https://www.allaboutbirds.org/guide/Yellow-throated_Warbler/lifehistory# Accessed 2024.

Cornell Lab of Ornithology. (2020). Golden-winged Warbler Range Map. Cornell Lab of

Ornithology, Ithaca, New York. https://www.allaboutbirds.org/guide/Golden-winged_Warbler/maps-range. Accessed 2024.

Cornell Lab of Ornithology. (2023b). Macaulay Library. https://www.macaulaylibrary.org/.

Craig, R. (2024). Temporal change in the forest birds of northeastern Connecticut shows partial concordance with predicted effects of climate and habitat change. *Bird Conservation Research*, no. 32.

Crocker, Susan J. (2018). Forests of Illinois, 2017. Resource Update FS-147. Newtown Square, PA: U.S. Department of Agriculture, Forest Service, Northern Research Station. 4 p.

Dong, K., Xu, H., Yongmin, L., Zhang, J. Wang, W., Li, D. (2023). β-diversity Patterns of Bird Communities in Natural Protected Areas in Anhui by Separating the Turnover and Nestedness Components. *Pakistan J. Zool.,* pp., 1—8.

Dunn, E. H., Francis, C. M., Blancher, P. J., Drennan, S. R., Howe, M. A., Lepage, D., Robbins, C. S., Rosenberg, K. V., Sauer, J. R., & Smith, K. G. (2005). Enhancing the scientific value of the Christmas Bird Count. *The Auk*, *122*(1), 338-346.

Fink, D., T. Auer, A. Johnston, M. Strimas-Mackey, S. Ligocki, O. Robinson, W. Hochachka, L. Jaromczyk, C. Crowley, K. Dunham, A. Stillman, I. Davies, A. Rodewald, V. Ruiz-Gutierrez, C. Wood. 2023. eBird Status and Trends, Data Version: 2022; Released: 2023.

Cornell Lab of Ornithology, Ithaca, New York. https://doi.org/10.2173/ebirdst.2022.

Gaston, K. J., Blackburn, T. M., Greenwood, J. J., Gregory, R. D., Quinn, R. M., & Lawton, J. H. (2000). Abundance–occupancy relationships. *Journal of Applied Ecology, 37*, 39-59.

Hill, A. P., Prince, P., Piña Covarrubias, E., Doncaster, C. P., Snaddon, J. L., & Rogers, A. (2018). AudioMoth: Evaluation of a smart open acoustic device for monitoring biodiversity and the environment. *Methods in Ecology and Evolution, 9*(5), 1199-1211.

Hobson, K. A., Rempel, R. S., Greenwood, H., Turnbull, B., & Van Wilgenburg, S. L. (2002). Acoustic surveys of birds using electronic recordings: new potential from an omnidirectional microphone system. *Wildlife Society Bulletin*, 709-720.

Hutto, R. L., & Stutzman, R. J. (2009). Humans versus autonomous recording units: A comparison of point-count results. *Journal of Field Ornithology*, *80*(4), 387-398.

Illinois Department of Natural Resources. (2022). Illinois Wildlife Action Plan 2015 Implementation Guide (Revised October 2022).

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull Soc Vaudoise Sci Nat, 37*, 547-579.

Jones, J. P. (2011). Monitoring species abundance and distribution at the landscape scale. *Journal of Applied Ecology, 48*(1), 9-13.

Jones, J. P., Asner, G. P., Butchart, S. H., & Karanth, K. U. (2013). The 'why', 'what' and 'how' of monitoring for conservation. *Key topics in conservation biology 2*, 327-343.

Kahl, S., Wood, C. M., Eibl, M., & Klinck, H. (2021). BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics, 61*, 101236.

Kleen, V. M., Cordle, L., & Montgomery, R. A. (2004). The Illinois Breeding Bird Atlas. *Illinois Natural History Survey Special Publication No. 26*.

Klingbeil, B. T., & Willig, M. R. (2015). Bird biodiversity assessments in temperate forest: the value of point count versus acoustic monitoring protocols. *PeerJ, 3*, e973.

Lapp, S., Stahlman, N., & Kitzes, J. (2023). A Quantitative Evaluation of the Performance of the Low-Cost AudioMoth Acoustic Recording Unit. *Sensors, 23*(11), 5254.

Laughlin, S. B., Kibbe, D. P., & Eagles, P. F. (1982). Atlasing the distribution of the breeding birds of North America. *American Birds,* Vol. 36, No. 1.

Lauha, P., Somervuo, P., Lehikoinen, P., Geres, L., Richter, T., Seibold, S., & Ovaskainen, O. (2022). Domain-specific neural networks improve automated bird sound recognition already with small amount of local data. *Methods in Ecology and Evolution, 13*(12), 2799-2810.

LeBien, J., Zhong, M., Campos-Cerqueira, M., Velev, J. P., Dodhia, R., Ferres, J. L., & Aide, T. M. (2020). A pipeline for identification of bird and frog species in tropical soundscape recordings using a convolutional neural network. *Ecological Informatics, 59*, 101113.

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Royle, J. A., & Langtimm, C. A. (2002). *Ecology 83*(8), 2248-2255.

MacKenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L., & Hines, J. E. (2017). Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence. *Elsevier*.

Maia R., Chamberlain S., Teucher A., Pardo S. (2023). rebird: R Client for the eBird Database of Bird Observations. R package version 1.3.0 https://docs.ropensci.org/rebird/, https://github.com/ropensci/rebird.

Marsh, D. M., & Trenham, P. C. (2008). Current trends in plant and animal population monitoring. *Conservation Biology*, *22*(3), 647-655.

Montgomery, R. A., Semel, B., & Dundee, I. L. (1987). Illinois breeding bird atlas.

National Audubon Society. (2022). The Christmas Bird Count, 1900-2022.

National Audubon Society. (2024). Answers to your top questions about the Christmas Bird Count. https://www.audubon.org/answers-your-top-questions-about-christmas-bird-count.

Nichols, J. D., & Williams, B. K. (2006). Monitoring for conservation. *Trends in ecology & evolution*, *21*(12), 668-673.

Partners in Flight. (2020). Avian Conservation Assessment Database, version 2020.

Pérez-Granados, C., & Traba, J. (2021). Estimating bird density using passive acoustic monitoring: a review of methods and suggestions for further research. *Ibis, 163*(3), 765-783.

Rosenberg, K. V., Dokter, A. M., Blancher, P. J., Sauer, J. R., Smith, A. C., Smith, P. A., Stanton, J. C., Panjabi, A., Helft, L., Parr, M. & Marra, P. P. (2019). Decline of the North American avifauna. *Science*, *366* (6461), 120-124.

Ruff, Z. J., Lesmeister, D. B., Duchac, L. S., Padmaraju, B. K., & Sullivan, C. M. (2020). Automated identification of avian vocalizations with deep convolutional neural networks. *Remote Sensing in Ecology and Conservation, 6*(1), 79-92.

Sethi, S. S., Bick, A., Chen, M. Y., Crouzeilles, R., Hillier, B. V., Lawson, J., Lee, C., Liu, S., Freitas-Parruco, C., Rosten, C., Somveille, M., Tuanmu, M. & Banks-Leite, C. (2023). Automatic vocalisation detection delivers reliable, multi-faceted, and global avian biodiversity monitoring. *bioRxiv*, 2023-09.

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

Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Biologiske skrifter, 5*, 1-34.

Stowell, D., Wood, M. D., Pamuła, H., Stylianou, Y., & Glotin, H. (2019). Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge. *Methods in Ecology and Evolution, 10*(3), 368-380.

Sugai, L. S. M., Silva, T. S. F., Ribeiro Jr, J. W., & Llusia, D. (2019). Terrestrial passive acoustic monitoring: review and perspectives. *BioScience, 69*(1), 15-25.

Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D., & Kelling, S. (2009). eBird: A citizen-based bird observation network in the biological sciences. *Biological conservation*, *142*(10), 2282-2292.

Tegeler, A. K., Morrison, M. L., & Szewczak, J. M. (2012). Using extended-duration audio recordings to survey avian species. *Wildlife Society Bulletin*, *36*(1), 21-29.

Toenies, M., & Rich, L. N. (2021). Advancing bird survey efforts through novel recorder technology and automated species identification. *California Fish and Wildlife*, *107*, 56-70.

Wildlife Acoustics. (2023). Products. https://www.wildlifeacoustics.com/products.

Wood, C. M., Kahl, S., Chaon, P., Peery, M. Z., & Klinck, H. (2021). Survey coverage, recording duration and community composition affect observed species richness in passive acoustic surveys. *Methods in Ecology and Evolution, 12*(5), 885-896.

Wood, C. M., Barceinas Cruz, A., & Kahl, S. (2023). Pairing a user-friendly machine-learning animal sound detector with passive acoustic surveys for occupancy modeling of an endangered primate. *American Journal of Primatology*, e23507.

Xeno-Canto. (2023). Xeno-Canto: Sharing bird sounds from all over the world. https://www.xeno-canto.org/.

Yip, D., Leston, L., Bayne, E., Sólymos, P., & Grover, A. (2017). Experimentally derived detection distances from audio recordings and human observers enable integrated analysis of point count data. *Avian Conservation and Ecology*, 12(1).

Ziolkowski Jr., D. J., Lutmerding, M., Aponte, V. I., and Hudson, M. A. R. (2022). North American Breeding Bird Survey Dataset 1966 – 2022.

APPENDICES

APPENDIX A

THE EFFECTS OF AUTONOMOUS RECORDING UNIT CHOICE AND BIRDNET-ANALYZER SETTINGS ON BIRDNET PERFORMANCE

Tables

Table A.1: Number of detections per detection source (point count, AudioMoth output, SM4 output, SMMicro output, or SwiftOne output) across 36 site visits between June 21—July 4, 2023, in Jackon County, Illinois, USA.

Common Name	**Scientific Name**	**Point Count**	**AudioMoth**	**SM4**	**SMMicro**	**SwiftOne**
Acadian Flycatcher	*Empidonax virescens*	12	11	10	10	11
American Crow	*Corvus brachyrhynchos*	32	21	14	15	14
American Goldfinch	*Spinus tristis*	15	6	6	6	8
American Kestrel	*Falco sparverius*	1	1	1	1	1
American Redstart	*Setophaga ruticilla*	2	2	1	1	0
American Robin	*Turdus migratorius*	3	0	0	0	0
Baltimore Oriole	*Icterus galbula*	2	2	1	0	0
Barn Swallow	*Hirundo rustica*	5	2	0	2	1
Barred Owl	*Strix varia*	1	0	0	0	0
Belted Kingfisher	*Megaceryle alcyon*	2	1	1	1	1
Blue Groshbeak	*Passerina caerulea*	18	17	17	17	17
Blue Jay	*Cyanocitta cristata*	23	11	6	6	8
Blue-gray Gnatcatcher	*Polioptila caerulea*	11	7	4	3	4
Brown Thrasher	*Toxostoma rufum*	7	6	6	3	5
Brown-headed Cowbird	*Molothrus ater*	12	7	5	4	5
Carolina Chickadee	*Poecile carolinensis*	17	9	8	8	8
Carolina Wren	*Thryothorus ludovicianus*	26	12	8	6	10
Chimney Swift	*Chaetura pelagica*	1	0	0	0	0

Chipping Sparrow	*Spizella passerina*	2	2	1	2	2
Common Grackle	*Quiscalus quiscula*	4	1	2	1	1
Common Yellowthroat	*Geothlypis trichas*	35	17	10	9	17
Dickcissel	*Spiza americana*	12	10	5	6	7
Downy Woodpecker	*Picoides pubescens*	12	8	8	4	7
Eastern Bluebird	*Sialia sialis*	15	10	13	7	12
Eastern Kingbird	*Tyrannus tyrannus*	12	11	10	10	11
Eastern Meadowlark	*Sturnella magna*	7	6	5	5	5
Eastern Phoebe	*Sayornis phoebe*	5	3	1	0	3
Eastern Towhee	*Pipilo erythrophthalmus*	27	21	8	14	9
Eastern Wood-Pewee	*Contopus virens*	17	12	11	11	12
European Starling	*Sturnus vulgaris*	11	4	3	1	6
Field Sparrow	*Spizella pusilla*	16	14	14	15	15
Fish Crow	*Corvus ossifragus*	11	6	7	3	7
Gray Catbird	*Dumetella carolinensis*	6	3	3	3	3
Great Crested Flycatcher	*Myiarchus crinitus*	14	6	4	2	7
Green Heron	*Butorides virescens*	4	3	4	3	4
Hairy Woodpecker	*Leuconotopicus villosus*	2	2	2	0	2
Hooded Warbler	*Setophaga citrina*	3	3	3	3	3
House Finch	*Haemorhous mexicanus*	4	3	1	0	3
House Sparrow	*Passer domesticus*	7	5	2	2	4
Indigo Bunting	*Passerina cyanea*	33	28	25	28	23
Kentucky Warbler	*Geothlypis formosa*	15	15	15	14	15
Killdeer	*Charadrius vociferus*	7	4	3	3	3
Mourning Dove	*Zenaida macroura*	32	16	14	8	15

Northern Bobwhite	*Colinus virginianus*	3	3	3	3	3
Northern Cardinal	*Cardinalis cardinalis*	31	21	10	10	11
Northern Flicker	*Colaptes auratus*	3	0	0	1	1
Northern Mockingbird	*Mimus polyglottos*	6	1	0	0	1
Northern Parula	*Setophaga americana*	21	11	9	7	8
Orchard Oriole	*Icterus spurius*	14	12	11	8	14
Ovenbird	*Seiurus aurocapilla*	4	4	0	1	2
Pileated Woodpecker	*Dryocopus pileatus*	1	0	0	0	1
Pine Warbler	*Setophaga pinus*	3	1	2	1	2
Prairie Warbler	*Setophaga discolor*	7	4	5	7	5
Prothonotary Warbler	*Protonotaria citrea*	1	1	1	1	1
Purple Martin	*Progne subis*	7	7	5	4	6
Red-bellied Woodpecker	*Melanerpes carolinus*	20	12	8	5	11
Red-eyed Vireo	*Vireo olivaceus*	13	5	2	1	3
Red-headed Woodpecker	*Melanerpes erythrocephalus*	4	3	4	2	3
Red-shouldered Hawk	*Buteo lineatus*	10	6	6	6	6
Red-winged Blackbird	*Agelaius phoeniceus*	16	5	2	3	4
Ruby-throated Hummingbird	*Archilochus colubris*	10	4	2	0	3
Scarlet Tanager	*Piranga olivacea*	2	2	2	1	2
Song Sparrow	*Melospiza melodia*	11	6	0	0	0
Summer Tanager	*Piranga rubra*	15	11	13	7	13
Tufted Titmouse	*Baeolophus bicolor*	29	11	5	4	7
Warbling Vireo	*Vireo gilvus*	9	9	7	6	9
White-breasted Nuthatch	*Sitta carolinensis*	13	6	5	3	4

White-eyed Vireo	*Vireo griseus*	21	15	13	10	15
Wild Turkey	*Meleagris gallopavo*	1	0	0	0	0
Wood Thrush	*Hylocichla mustelina*	17	13	9	10	9
Yellow Warbler	*Setophaga petechia*	4	2	0	0	0
Yellow-billed Cuckoo	*Coccyzus americanus*	17	7	8	9	8
Yellow-breasted Chat	*Icteria virens*	18	13	13	14	14
Yellow-throated Vireo	*Vireo flavifrons*	5	5	3	4	4
Yellow-throated Warbler	*Setophaga dominica*	6	6	6	5	5

Table A.2: Average number of true positive species (per visit), false positive species (per visit), false negative species (per visit), precision, sensitivity, specificity, F1 harmonic mean, and MCC (Matthews correlation coefficient) by unit. All values were rounded to the second decimal place. Note that all values displayed were calculated using data from all 36 site visits.

Unit	TP Species	FP Species	FN Species	Precision	Sensitivity	Specificity	F1	MCC
AudioMoth	14.92	11.97	8.56	0.55	0.64	0.88	0.59	0.49
SM4	11.67	8.86	11.81	0.57	0.49	0.91	0.52	0.42
SMMicro	10.36	8.08	13.11	0.57	0.44	0.92	0.49	0.39
SwiftOne	13.03	10.08	10.44	0.57	0.55	0.90	0.55	0.45

Table A.3: The mean number of true positive, false positive, and false negative species reported per visit for each Overlap and Sensitivity setting combination tested. All values were rounded to the second decimal place. Mean values were obtained by aggregating individual site visit values for each metric across all 36 site visits. All audio was collected using an AudioMoth, and was collected in Jackson County, Illinois, USA, between June 21—July 4, 2023.

Overlap Setting	Sensitivity Setting	True Positive Species	False Positive Species	False Negative Species
0	0.5	9.64	3.22	11.50
0	1	11.75	7.22	9.39
0	1.5	19.61	31.06	1.53
0.5	0.5	9.61	3.75	11.53
0.5	1	12.31	8.17	8.83
0.5	1.5	19.83	32.61	1.31
1	0.5	10.08	4.22	11.06
1	1	12.33	8.42	8.81
1	1.5	19.75	33.17	1.39

1.5	0.5	10.53	4.53	10.61
1.5	1	12.89	9.53	8.25
1.5	1.5	20.11	34.47	1.03
2	0.5	10.83	5.53	10.31
2	1	13.67	10.78	7.47
2	1.5	20.17	35.72	0.97
2.5	0.5	11.83	7.08	9.31
2.5	1	14.75	13.50	6.39
2.5	1.5	20.47	38.08	0.67

Table A.4: The mean precision, sensitivity, specificity, F1 (F1 harmonic mean), and MCC (Matthews Correlation Coefficient) for each Overlap and Sensitivity setting combination tested. Setting combination Overlap = 0, Sensitivity = 1 is the default BirdNET settings. All values were rounded to the second decimal place. Mean values were obtained by aggregating individual site visit values for each metric across all 36 site visits. All audio was collected using an AudioMoth, and was collected in Jackson County, Illinois, USA, between June 21—July 4, 2023.

Overlap Setting	Sensitivity Setting	Precision	Sensitivity	Specificity	F1	MCC
0	0.5	0.76	0.46	0.93	0.56	0.46
0	1	0.63	0.56	0.85	0.58	0.42
0	1.5	0.39	0.93	0.36	0.54	0.30
0.5	0.5	0.72	0.46	0.92	0.55	0.44
0.5	1	0.61	0.59	0.83	0.59	0.42
0.5	1.5	0.38	0.94	0.33	0.54	0.29
1	0.5	0.71	0.48	0.91	0.56	0.44
1	1	0.60	0.59	0.83	0.58	0.41
1	1.5	0.38	0.94	0.32	0.53	0.27
1.5	0.5	0.71	0.50	0.91	0.58	0.46
1.5	1	0.59	0.61	0.80	0.59	0.41
1.5	1.5	0.37	0.95	0.29	0.53	0.27
2.0	0.5	0.67	0.51	0.89	0.57	0.44
2.0	1	0.56	0.65	0.78	0.60	0.41
2.0	1.5	0.36	0.95	0.27	0.52	0.25
2.5	0.5	0.63	0.56	0.85	0.59	0.43
2.5	1	0.53	0.70	0.72	0.59	0.39
2.5	1.5	0.35	0.97	0.22	0.51	0.23

APPENDIX B

COMPARING METHODS OF STREAMLINING BIRDNET ANALYZER VALIDATION FOR LONG-TERM AVIAN POPULATION

Tables

Table B.1: List of all threshold values by threshold type for all species included in point counts. Threshold values containing NA could not be calculated from the validation data.

Common Name	Alpha Code	Modeled	FP-based	0.5	0.**75**
Acadian Flycatcher	ACFL	0.48	0.47	0.5	0.75
American Crow	AMCR	0.20	0.27	0.5	0.75
American Goldfinch	AMGO	0.31	0.37	0.5	0.75
American Kestrel	AMKE	0.47	0.40	0.5	0.75
American Redstart	AMRE	0.37	0.37	0.5	0.75
American Robin	AMRO	NA	0.72	0.5	0.75
Baltimore Oriole	BAOR	0.48	0.51	0.5	0.75
Barn Swallow	BARS	0.32	0.38	0.5	0.75
Barred Owl	BADO	NA	0.90	0.5	0.75
Belted Kingfisher	BEKI	NA	0.89	0.5	0.75
Blue Grosbeak	BLGR	0.39	0.45	0.5	0.75
Blue Jay	BLJA	0.33	0.38	0.5	0.75
Blue-gray Gnatcatcher	BGGN	0.10	0.33	0.5	0.75
Brown Thrasher	BRTH	0.93	0.84	0.5	0.75
Brown-headed Cowbird	BHCO	0.25	0.32	0.5	0.75
Carolina Chickadee	CACH	0.24	0.34	0.5	0.75
Carolina Wren	CARW	0.86	0.71	0.5	0.75
Chimney Swift	CHSW	0.44	0.40	0.5	0.75
Chipping Sparrow	CHSP	0.30	0.36	0.5	0.75
Common Grackle	COGR	0.50	0.60	0.5	0.75
Common Yellowthroat	COYE	0.17	0.27	0.5	0.75
Dickcissel	DICK	0.17	0.35	0.5	0.75
Downy Woodpecker	DOWO	0.39	0.40	0.5	0.75
Eastern Bluebird	EABL	0.64	0.69	0.5	0.75
Eastern Kingbird	EAKI	0.13	0.19	0.5	0.75
Eastern Meadowlark	EAME	0.25	0.44	0.5	0.75
Eastern Phoebe	EAPH	0.28	0.37	0.5	0.75
Eastern Towhee	EATO	NA	NA	0.5	0.75
Eastern Wood-Pewee	EAWP	0.10	NA	0.5	0.75
European Starling	EUST	0.16	0.27	0.5	0.75

Field Sparrow	FISP	0.37	0.41	0.5	0.75
Fish Crow	FICR	0.38	0.36	0.5	0.75
Gray Catbird	GRCA	0.45	0.53	0.5	0.75
Great Crested Flycatcher	GCFL	0.10	NA	0.5	0.75
Green Heron	GRHE	NA	0.78	0.5	0.75
Hairy Woodpecker	HAWO	0.60	0.71	0.5	0.75
Hooded Warbler	HOWA	0.59	0.78	0.5	0.75
House Finch	HOFI	0.30	0.43	0.5	0.75
House Sparrow	HOSP	0.21	0.42	0.5	0.75
Indigo Bunting	INBU	0.25	0.31	0.5	0.75
Kentucky Warbler	KEWA	NA	0.96	0.5	0.75
Killdeer	KILL	0.10	0.32	0.5	0.75
Mourning Dove	MODO	0.12	0.19	0.5	0.75
Northern Bobwhite	NOBO	0.65	0.53	0.5	0.75
Northern Cardinal	NOCA	0.50	0.46	0.5	0.75
Northern Flicker	NOFL	0.34	0.39	0.5	0.75
Northern Mockingbird	NOMO	0.51	0.36	0.5	0.75
Northern Parula	NOPA	0.40	0.56	0.5	0.75
Orchard Oriole	OROR	0.28	0.42	0.5	0.75
Ovenbird	OVEN	0.61	0.54	0.5	0.75
Pileated Woodpecker	PIWO	0.92	0.87	0.5	0.75
Pine Warbler	PIWA	0.34	0.50	0.5	0.75
Prairie Warbler	PRAW	0.55	0.79	0.5	0.75
Prothonotary Warbler	PROW	0.58	0.61	0.5	0.75
Purple Martin	PUMA	0.50	0.56	0.5	0.75
Red-bellied Woodpecker	RBWO	0.37	0.70	0.5	0.75
Red-eyed Vireo	REVI	0.22	0.41	0.5	0.75
Red-headed Woodpecker	RHWO	0.64	0.70	0.5	0.75
Red-shouldered Hawk	RSHA	0.27	0.49	0.5	0.75
Red-winged Blackbird	RWBL	0.14	0.22	0.5	0.75
Ruby-throated Hummingbird	RTHU	0.59	0.40	0.5	0.75
Scarlet Tanager	SCTA	0.41	0.37	0.5	0.75
Song Sparrow	SOSP	0.10	NA	0.5	0.75
Summer Tanager	SUTA	0.70	NA	0.5	0.75
Tufted Titmouse	TUTI	0.32	0.43	0.5	0.75
Warbling Vireo	WAVI	0.21	0.27	0.5	0.75
White-breasted Nuthatch	WBNU	0.10	0.49	0.5	0.75
White-eyed Vireo	WEVI	0.21	0.66	0.5	0.75
Wild Turkey	WITU	0.72	0.83	0.5	0.75
Wood Thrush	WOTH	0.30	0.44	0.5	0.75
Yellow Warbler	YEWA	0.57	0.42	0.5	0.75

Yellow-billed Cuckoo	YBCU	0.66	0.93	0.5	0.75
Yellow-breasted Chat	YBCH	0.47	0.44	0.5	0.75
Yellow-throated Vireo	YTVI	0.60	0.77	0.5	0.75
Yellow-throated Warbler	YTWA	0.44	0.47	0.5	0.75

Table B.2: Showing population trends identified by different sources (IBBA—2022, Illinois-only BBS (1991—2022), and Partners in Flight) for all analyzed warbler and vireo species listed by common name and scientific name. Blue cells represent an increase over time, red cells represent a decrease over time, and white cells represent no change or uncertainty. Illinois Breeding Bird Atlas (IBBA) data used was collected between 1986—1991 for the 45 blocks which I resampled in 2022 using ARUs and filtered BirdNET output. The North American Breeding Bird Survey data used was for the state of Illinois only, and from 1991—2022. Partners in Flight (PIF) data show continental-level trends.

Common Name	Scientific Name	IBBA–2022	BBS (IL)	PIF
American Redstart	Setophaga ruticilla	🟦	🟥	🟦
Bell’s Vireo	Vireo bellii	🟦	🟦	🟦
Black-and-white Warbler	Mniotilta varia	🟦	🟥	🟦
Blue-winged Warbler	Vermivora cyanoptera	🟦	🟦	🟥
Common Yellowthroat	Geothlypis trichas	⬜	🟦	🟥
Golden-winged Warbler	Vermivora chrysoptera	🟦	🟥	🟥
Hooded Warbler	Setophaga citrina	🟦	🟦	🟦
Kentucky Warbler	Geothlypis formosa	🟥	🟦	🟥
Louisiana Waterthrush	Parkesia motacilla	🟦	🟦	🟦
Northern Parula	Setophaga americana	🟦	🟦	🟦
Ovenbird	Seiurus aurocapilla	🟦	⬜	🟦
Pine Warbler	Setophaga pinus	🟦	⬜	🟦
Prairie Warbler	Setophaga discolor	🟥	🟦	🟥
Prothonotary Warbler	Protonotaria citrea	🟦	🟦	🟥
Red-eyed Vireo	Vireo olivaceus	🟦	🟦	🟦
Warbling Vireo	Vireo gilvus	🟥	🟦	🟦
White-eyed Vireo	Vireo griseus	🟥	🟦	🟦
Worm-eating Warbler	Helmitheros vermivorum	🟦	🟦	🟦
Yellow Warbler	Setophaga petechia	🟥	🟦	⬜
Yellow-throated Vireo	Vireo flavifrons	🟦	🟥	🟦
Yellow-throated Warbler	Setophaga dominica	🟦	🟦	🟦

Table B.3: Comparing the preferred habitat of each species with its respective gain/loss of primary blocks in occurrence from the Illinois Breeding Bird Atlas to 2022.

Species	**Alpha Code**	**Habitat**	**Nesting**	**Gain/Loss**
American Redstart	AMRE	Forests	Tree	30
Bell's Vireo	BEVI	Scrub	Shrub	8
Black-and-white Warbler	BAWW	Forests	Ground	13

Blue-winged Warbler	BWWA	Open Woodlands	Ground	-6
Common Yellowthroat	COYE	Scrub	Shrub	0
Golden-winged Warbler	GWWA	Open Woodlands	Ground	7
Hooded Warbler	HOWA	Forests	Shrub	13
Kentucky Warbler	KEWA	Forests,	Ground	-4
Louisiana Waterthrush	LOWA	Rivers, and Streams	Ground	14
Northern Parula	NOPA	Forests	Tree	7
Ovenbird	OVEN	Forests	Ground	23
Pine Warbler	PIWA	Forests	Tree	11
Prairie Warbler	PRAW	Scrub	Shrub	16
Prothonotary Warbler	PROW	Forests	Cavity	-9
Red-eyed Vireo	REVI	Forests	Tree	6
Warbling Vireo	WAVI	Open Woodlands	Tree	-5
White-eyed Vireo	WEVI	Scrub	Shrub	-3
Worm-eating Warbler	WEWA	Forests	Ground	27
Yellow Warbler	YEWA	Open Woodlands	Shrub	-22
Yellow-throated Vireo	YTVI	Open Woodlands	Tree	12
Yellow-throated Warbler	YTWA	Forests	Tree	21

Table B.4: Comparing the warbler and vireo species richness between private and public blocks from both the Illinois Breeding Bird Atlas (IBBA) and 2022 survey datasets.

	IBBA		**2022**	
Ownership	**Richness**	**Hours of obs.**	**Richness**	**Hours of obs.**
Private	7	29	12.42	74.95
Public	11.69	62.49	13.85	78.46

Figures

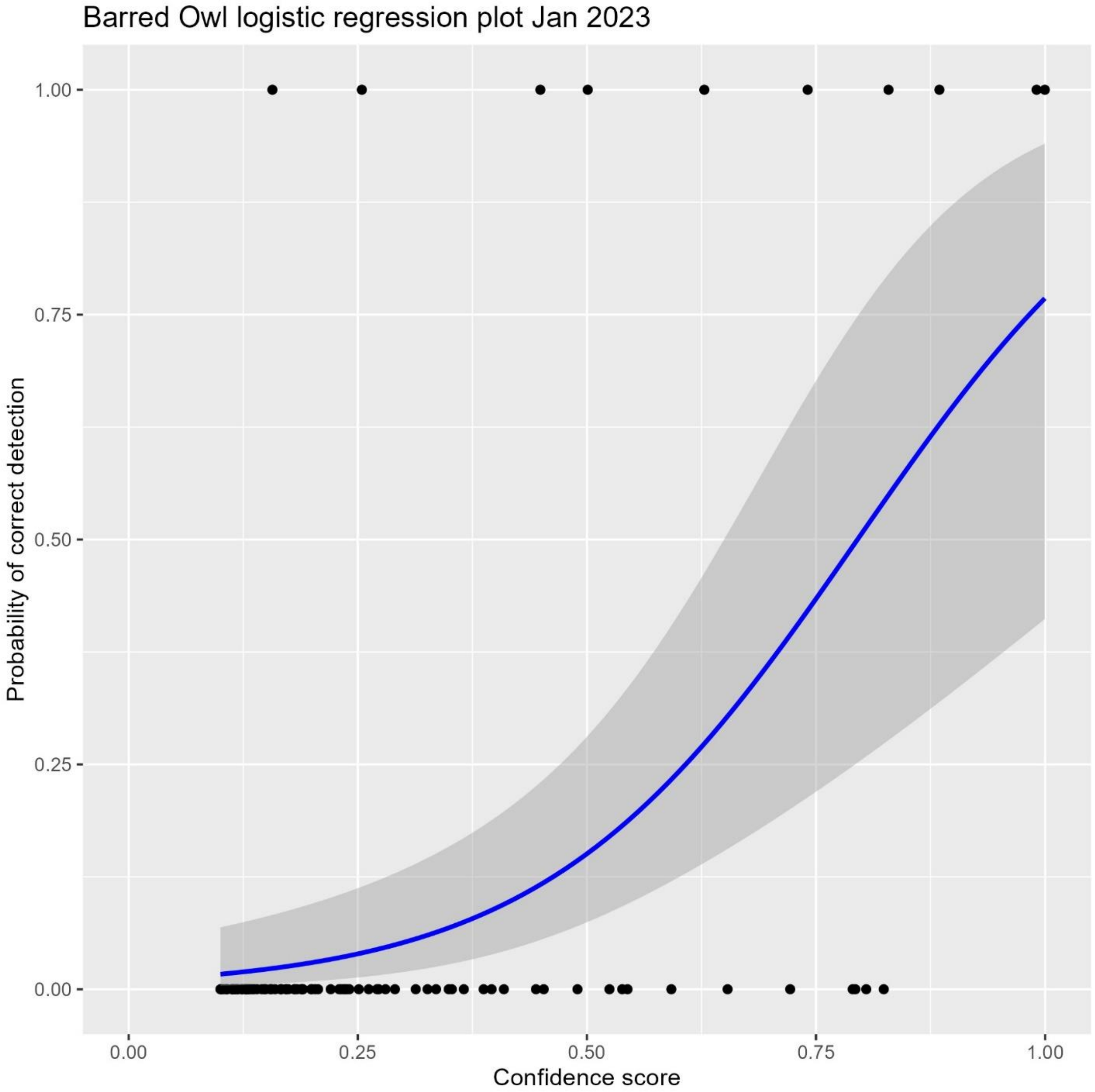

Figure B.1: Logistic regression analysis performed using 100 random detections from BirdNET which were manually validated. The blue line represents the relationship between the probability of a correction detection and the confidence score for a given BirdNET detection for Barred Owls. Since the blue line never reaches a probability of 0.95, a species-specific confidence score could not be derived for this species using the logistic regression analysis.

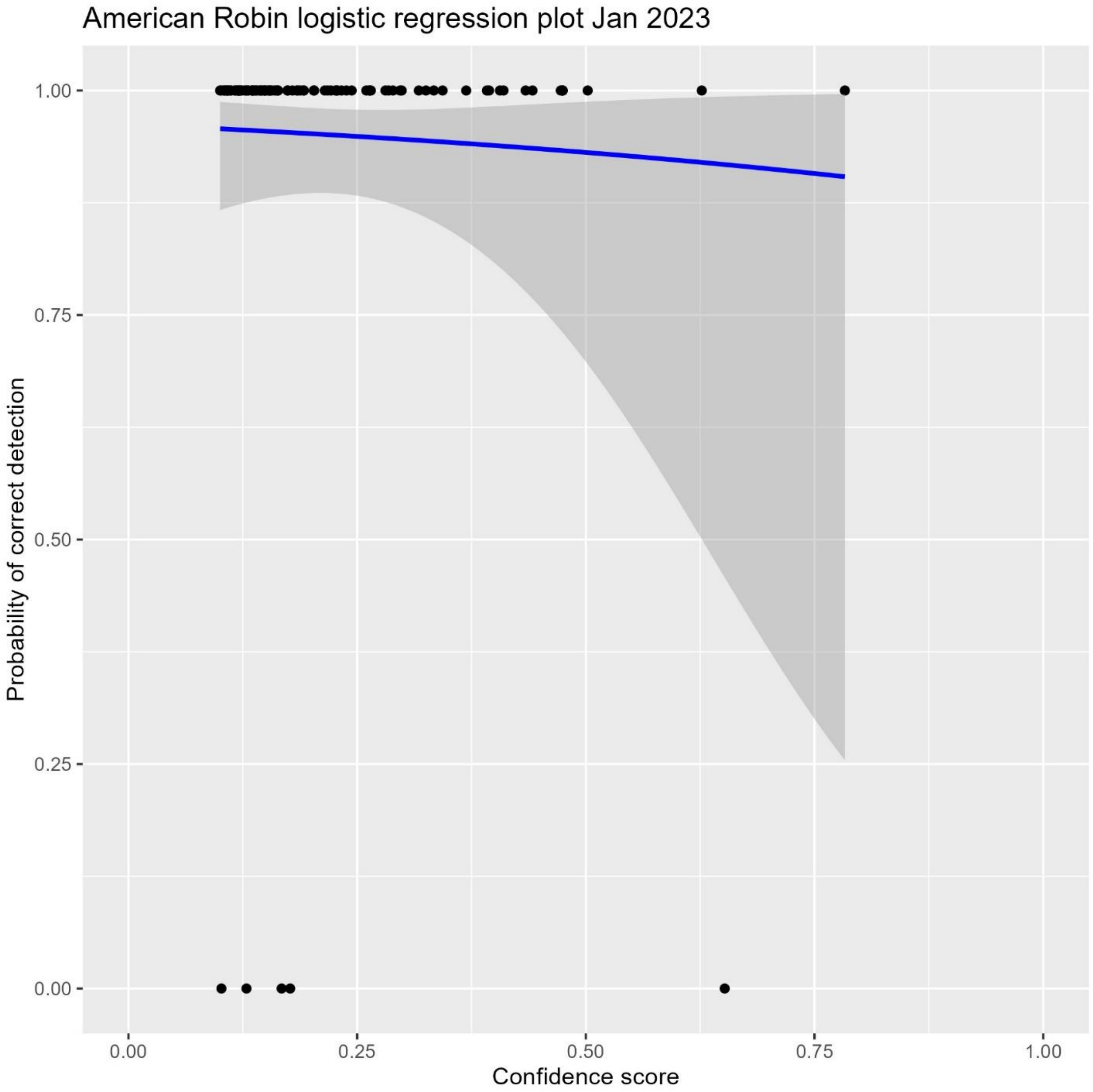

Figure B.2: Logistic regression analysis performed using 100 random detections from BirdNET which were manually validated. The blue line represents the relationship between the probability of a correction detection and the confidence score for a given BirdNET detection for American Robins (*Turdus migratorius*). Since the blue line has a negative slope (implying that the probability of a BirdNET detection being a true positive detection decreases as the confidence score increases), a species-specific confidence score should not be derived for this species using the logistic regression analysis.

VITA

Graduate School
Southern Illinois University Carbondale

Shasta S. W. Corvus

shasta.corvus@siu.edu

Shawnee State University
Bachelor of Science, Biology, May 2019

Thesis Paper Title:

Passive Acoustic Monitoring: Considerations for recording units, BirdNET settings, and filtering methods for long-term avian population monitoring

Major Professor: Brent S. Pease