Ornithological Applications, 2025, 00, 1–8
https://doi.org/10.1093/ornithapp/duaf049
Advance access publication 05 August 2025
Special Feature

Humans outperform Merlin Sound ID in field-based point-count surveys

Brooke D. Goodman,[1,2,*] Kyle A. Lima,[1] Seth Benz,[1] Nicholas A. Fisichelli,[1] and Justin Kitzes,[2]

[1]Schoodic Institute at Acadia National Park, Winter Harbor, Maine, USA
[2]University of Pittsburgh, Biological Sciences Department, Pittsburgh, Pennsylvania, USA

*Corresponding author: brookedgoodman23@gmail.com

ABSTRACT

Point counts are integral to bird population monitoring, and multiple methods have been developed to improve their reliability. These methods include multiple-observer protocols, double-sampling, and surveys using autonomous recording units. However, these improved protocols are time-consuming, expensive, and often require hiring surveyors with inconsistent abilities. Merlin Sound ID, an automated bird-sound classifier that can run in the field on mobile devices, could be a part of the solution to the problems associated with point counts by functioning as a no-cost observer with consistent performance. Here, we compare the accuracy of Merlin and human observers in 144 paired field-based point-counts. Our goals were to evaluate how the species identified, number of individual detections, and precision for humans and Merlin compare. We also examined the consistency of Merlin across different devices. Human observers reported 382 individual detections over the course of the study, 72% more than the 222 detections reported by Merlin. Precision was similar with 92% and 86% for humans and Merlin, respectively. Merlin detected 4 species missed by humans but also detected 12 false positive species. Merlin disagreed with itself on 57% of point counts when running on 2 different devices at the same time. We propose Merlin not as a replacement for human point counters, but as an aid to increase detection probability of point-count surveys as secondary observers, with humans acting as reviewers and arbitrating disagreements between devices when needed. With further research into this methodology, technology like Merlin could help address some of the challenges associated with point-count-based research.

Keywords: acoustic monitoring, artificial intelligence (AI), bioacoustics, machine learning, Merlin Sound ID, point counts, survey techniques

How to Cite

Goodman, B. D., K. A. Lima, S. Benz, N. A. Fisichelli, and J. Kitzes (2025). Humans outperform Merlin Sound ID in field-based point-count surveys. *Ornithological Applications* 127:duaf049.

LAY SUMMARY

- Point counts are a bedrock of bird population monitoring, and multiple advancements in survey technique have been made to improve their reliability. However, these improved protocols are time-consuming, expensive, and often require hiring surveyors, whose abilities can be inconsistent. Merlin Sound ID, a free phone-based bird-sound classifier, may resolve some of these issues.
- We compare the accuracy of Merlin and humans in 144 paired field-based point-counts. Our goals were to examine differences in species identified and individual detections and understand Merlin's consistency on different devices.
- Humans detected 72% more individual birds than Merlin. Merlin detected 4 species missed by humans and 12 false positive species. Merlin disagreed with itself on 57% of point counts when running on 2 different devices simultaneously.
- We propose Merlin not as a replacement for human point counters, but to increase detection probability of surveys as a second observer, especially for rare species, with humans acting as reviewers and arbitrating disagreements between devices when needed.

Los humanos superan a Merlin Sound ID en relevamientos de conteos por puntos realizados a campo

RESUMEN

Los conteos por puntos son fundamentales para el monitoreo de poblaciones de aves, y se han desarrollado múltiples métodos para mejorar su confiabilidad. Estos métodos incluyen protocolos con múltiples observadores, doble muestreo y relevamientos con unidades de grabación autónomas. Sin embargo, estos protocolos mejorados requieren mucho tiempo, son costosos y a menudo requieren contratar observadores con habilidades inconsistentes. Merlin Sound ID, un clasificador automático de sonidos de aves que puede ejecutarse en campo en dispositivos móviles, podría ser parte de la solución a los problemas asociados con los conteos por puntos, al funcionar como un observador sin costo y con un rendimiento consistente. Aquí comparamos la precisión de Merlin y de observadores humanos en 144 conteos por puntos pareados realizados a campo. Nuestros objetivos fueron evaluar cómo se comparan las especies identificadas, el número de detecciones individuales y la precisión entre humanos y Merlin. También examinamos la consistencia de Merlin en distintos dispositivos. Los observadores humanos reportaron 382

Submission Date: January 31, 2025; Editorial Acceptance Date: July 25, 2025
© The Author(s) 2025. Published by Oxford University Press for the American Ornithological Society.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/condor/advance-article/doi/10.1093/ornithapp/duaf049/8222742 by guest on 05 January 2026

detecciones individuales durante el estudio, un 72% más que las 222 detecciones reportadas por Merlin. La precisión fue similar, con un 92% para los humanos y un 86% para Merlin, respectivamente. Merlin detectó cuatro especies que los humanos pasaron por alto, pero también detectó 12 especies con falsos positivos. Merlin no coincidió consigo mismo en el 57% de los conteos por puntos cuando se ejecutó simultáneamente en dos dispositivos diferentes. Proponemos que Merlin no reemplace a los observadores humanos en conteos por puntos, sino que funcione como una ayuda para aumentar la probabilidad de detección en los relevamientos, actuando como observador secundario, con los humanos como revisores y árbitros cuando sea necesario ante desacuerdos entre los dispositivos. Con más investigaciones sobre esta metodología, tecnologías como Merlin podrían ayudar a abordar algunos de los desafíos asociados con la investigación basada en los conteos por puntos.

Palabras clave: aprendizaje automático, bioacústica, conteos por puntos, inteligencia artificial (IA), Merlin Sound ID, monitoreo acústico, técnicas de relevamiento

Downloaded from https://academic.oup.com/condor/advance-article/doi/10.1093/ornithapp/duaf049/8222742 by guest on 05 January 2026

INTRODUCTION

Point counts are used by researchers to understand temporal and spatial changes in bird populations and are a bedrock of bird population monitoring. A point count involves an observer silently recording birds they detect aurally and visually for a predetermined amount of time (Ralph et al. 1995). Since their inception, point counts have illuminated species distributions, habitat use, and population status (Ralph et al. 1995, Hudson et al. 2017, Sauer et al. 2017, Rosenstock et al. 2002). Multiple advancements in survey technique have been made to improve upon unadjusted point-count surveys (Thompson 2002). For instance, multiple-observer methods allow for the calculation of observer-specific detection probabilities, improving measures of abundance (Nichols et al. 2000), distance sampling provides direct estimates of bird density that are not confounded by detectability (Rosenstock et al. 2002), and double sampling provides unbiased density and trend estimates (Bart and Earnst 2002). These protocols have all greatly increased the power of point-count data.

Recently, passive acoustic monitoring (PAM) has also emerged as a successful method for conducting population surveys (Blumstein et al. 2011, Gibb et al. 2019). PAM often involves placing an autonomous recorder at a point to record the soundscape, then retrieving it weeks to months later. Most birds detected during an in-person survey are detected aurally, making PAM a valid way to monitor bird populations (Darras et al. 2018). Recordings can then be reviewed through audio-based ex situ point counts or using predeveloped species detection software. The extensive spatiotemporal scale at which PAM surveys can be conducted has unlocked new avenues for studying birds and their behavior (Sugai et al. 2019). Novel findings on species trends, habitat use by hard-to-detect species, and occupancy over large geographic scales have been illuminated using PAM (Wood et al. 2019, Larkin 2023, Chronister et al. 2024). PAM also comes with well-documented benefits to survey technique, like equal or higher species presence/absence accuracy than traditional human point counts (Shonfield and Bayne 2017, Darras et al. 2018, Hoefer et al. 2023). Advancements in both human- and machine-based survey techniques have led to dramatic improvements in ecological knowledge.

Despite these advancements, issues remain with both human- and machine-based surveys. Some of these issues are material; point counts rely on skilled observers with a strong knowledge of bird identification in a specific region (Kepler and Scott, 1981), who are often difficult to find or train in a cost-effective manner (Ralph et al. 1995, Francis et al. 2009, Hutto and Stutzman 2009). Additionally, multiple-observer, repeat, and double-survey methods (Nichols et al. 2000, Bart and Earnst 2002, Thompson 2002) are strained by limited budgets, access to skilled observers, and time constraints (Hutto and Stutzman 2009). Other issues are more abstract; observers have vastly different identification skills and knowledge baselines, which can lead to inaccurate measures of occupancy and abundance (Kepler and Scott 1981, Simons et al. 2009, Farmer et al. 2012, Schmidt et al. 2023). More advanced point-count protocols are also only appropriate when certain assumptions are met, which is often hard to evaluate. Additionally, the implementation of PAM is hampered by unique issues. A single audio recorder and associated supplies often cost more than 100 USD. PAM methods require specialized knowledge and use of automated classification methods (e.g., BirdNET, OpenSoundscape; Kahl et al. 2021, Lapp et al. 2023), or a substantial time commitment for manual review of audio files. These learning curves and costs discourage researchers from using bioacoustics (Shonfield and Bayne 2017). Despite advances in human-based and PAM survey techniques, significant barriers remain in implementation.

Merlin Sound ID (hereafter "Merlin", released in 2021) is an automated bird-sound classifier that runs on mobile devices and could be a solution to the problems associated with data from point-count surveys and PAM. Merlin records audio locally on a device (e.g., mobile phone or tablet), while making species-level identifications in real time (Cornell Lab of Ornithology 2024). Merlin detects species using image features it recognizes from training data in spectrograms, allowing a user to simply press record on their device and view bird identifications from the soundscape around them. The Merlin app is free and can be operated on devices likely already owned by surveyors, making it a no-cost addition to survey methods. The mobility and wide availability of devices like smartphones, the simplicity of Merlin's interface, its real-time identifications, and the possible consistency of an automated detector compared to a human make Merlin a potential tool to improve the reliability of point-count data.

Here, we compare the accuracy of Merlin and intermediate-level human observers in field-based point-count surveys. We conducted repeated-visit point counts at 10 points in a northern coastal spruce-fir forest in Maine, USA, using a paired human observer-Merlin design. Our specific objectives were to (1) evaluate the effectiveness and ease of use of Merlin in field-based point counts; (2) determine how the species identified, individual detections, and precision for the two survey methods compare; and (3) understand Merlin's consistency on different devices. We then offer recommendations for further research into the integration of Merlin and point counts.



FIGURE 1. Schoodic Forest, a 687.97-ha coastal spruce-fir forest, is managed by Maine Coast Heritage Trust and located on the Schoodic Peninsula in Winter Harbor, Maine, USA.

METHODS

Study Area

This study took place at the Schoodic Forest in Winter Harbor, Maine, USA (Figure 1). This 687.97-ha forest parcel is owned and managed by Maine Coast Heritage Trust and co-managed with Schoodic Institute. The Schoodic Forest property is a coastal spruce-fir forest with abundant *Acer rubrum* (Red Maple) that experienced partial harvest last in the late 1990s. The land is used by the local community for hunting, hiking, and other recreation. The survey was conducted along Curtis Road, a small gravel road that is now predominantly maintained as an ATV trail and access road and is the only road through the property.

Survey Methods

Field surveys

The survey route consisted of 10 points along the unpaved Curtis Road, starting ~500 m off from the paved road (Figure 2). Due to turns in the road, not all points were separated by 500 m, but no 2 points were closer than 300 m. Surveys were conducted beginning at point one and continuing in increasing order along Curtis Road to point 10 over the course of one day, except one instance, where we began the survey at point 10 and continued in decreasing order to point 1. We conducted these surveys from July to October in 2023, and visited all 10 points during each weekly survey, unless conditions deteriorated, to where detection probability would be impaired as indicated in the National Parks Service Breeding Landbird Monitoring Volunteer Resource Manual (Faccio et al. 2024). This happened on one survey day, when we only conducted counts at points 1–4 before ending the survey. Whenever possible, surveys were scheduled in accordance with ideal weather conditions. Each point count consisted of a 1-min settlement period, during which no data were collected, when observers stood silently to allow their presence to minimize disturbance on bird behavior. At the end of this 1 min, we began a 3-min point count, where a human observer auditorily and visually identified birds alongside another observer operating Merlin on a handheld device (see Merlin Sound ID Software and Recorder Hardware for device make and models). Observers stood within 1–2 m of each other. The Merlin operator aimed the microphone of their handheld device away from themselves and remained still for 3 min to minimize sound interference. The human observer did not consult with the Merlin operator during the count regarding birds identified by the software. Occasionally there were 2 Merlin operators on separate devices during the same count. There were 2 human observers over the course of the study, both deemed intermediately skilled as they were considered level-two birders according to the standards outlined in the Birder Certification Online program (Howe and Giese 2021). All point counts were conducted between 1100 and 1800 UTC.

Merlin sound id software and recorder hardware

Six device types were used over the course of the study: Apple iPhone SE, Samsung Galaxy Tab A, Apple iPhone 12 Pro, Apple iPhone 14, Apple iPhone 13 Mini, and Apple iPhone 13. Three versions of the Merlin software (versions 2.1.8, 3.0, and 3.0.1) were used during the 4-month study period on Apple iPhones, and 1 (version 2.1.4) on the Samsung Galaxy. To our knowledge, no changes to the classifier used in the Sound ID feature were made during the duration of the study. Over the course of the study (144 point counts), we generated 214 3-min audio files, accounting for point counts where we had 2 devices running Merlin at the same time. Points 1–4 had 22 audio files each, whereas points 5–10 had 21 each, resulting in 642 min of audio in total.

Analytical methods

We used R version 4.4.1 (R Core Team, 2025) for all analyses and plotting. To examine how the human observers and Merlin compared in point counts, a local expert who did not participate in the fieldwork reviewed the recording from each individual point count and labeled each identification from human observers and Merlin as correct, incorrect, or unverifiable. The

Downloaded from https://academic.oup.com/condor/advance-article/doi/10.1093/ornithapp/duaf049/8222742 by guest on 05 January 2026

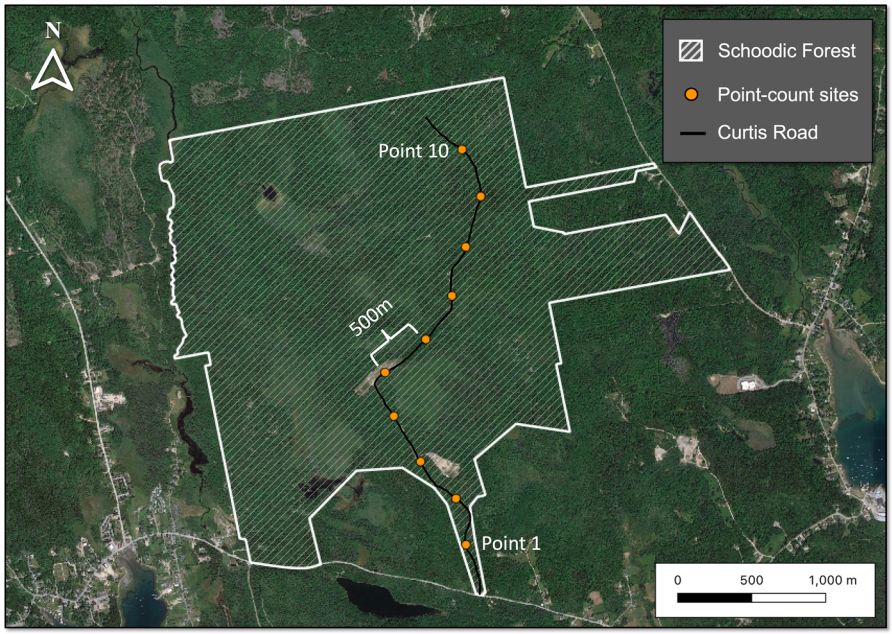

FIGURE 2. Survey route on Curtis Road through Schoodic Forest. The 10 points are distributed ~ every 500 m along Curtis Road, a gravel access road, in the Schoodic Forest. The white polygon with diagonal line fill represents the Schoodic Forest boundary. Orange markers represent the points, numbered from one to 10, starting with the southernmost marker.

reviewer listened to the audio from the point counts and also examined spectrograms on a phone within the Merlin app. Each audio file was reviewed only once, but the reviewer would repeat portions of the recording in the moment if needed for an identification. The reviewer was deemed an expert, as they were considered a level-three birder according to the standards outlined in the Birder Certification Online program (Howe and Giese 2021). Vocalizations were considered unverifiable if the human observer detected a species during a point count, but neither its vocalization, nor a similar-sounding species, could be identified in the spectrogram. A common example of this was *Corvus brachyrhynchos* (American Crow), which often vocalized from distances far enough that its call did not appear on the spectrogram in the Merlin app, but the human observer was able to hear it in the field. Unverifiable and visual observations by human observers were considered correct for the purpose of our analysis.

An observation from the human observer was considered incorrect if it was a misidentification of one species for another; for example, the human observer detected *Vireo olivaceus* (Red-eyed Vireo) at a point where the reviewer only detected *Vireo solitarius* (Blue-headed Vireo) in the recording. An identification by Merlin was considered incorrect if the species it detected in the recording was not detected by the reviewer in the same recording. This provided a way to measure the precision of the 2 methods. We defined precision as the percentage of individual detections that had the correct species identification. Precision was calculated as:

$$\text{Precision} = \frac{\text{true positive detections}}{\text{true positives} + \text{false positives}} \text{x} 100$$

An individual detection was defined as the identification of the presence of a species during a point count, ignoring abundance. For example, if the human observer reported 2 *V. olivaceus*, 5 *C. brachyrhynchos*, and 1 *Cardinalis cardinalis* (Northern Cardinal) during a point count, it would represent 3 individual detections. The total number of individual detections for a method, such as human point counts, was calculated as the sum of individual detections across all 144 point counts. We did not compare abundance, since Merlin cannot generate any comparative measure of abundance. We explored total species count and number of individual detections from human observers and Merlin over the entire study with no observations removed and also when false positives were removed as determined by the expert reviewer. On survey days when 2 devices ran Merlin at the same time at each point count, one device was randomly chosen to be the Merlin representative for calculation of individual detections and comparison to the human observer for the survey day.

We also examined the 70 point counts when 2 devices were running Merlin at the same time (140 separate counts). We compared the species lists produced by the 2 devices and counted the number of times that the devices produced the same species list (agreements) and different species lists (disagreements). For disagreements, we calculated the mean number of species the lists differed by. We then removed false positives as determined by the expert reviewer, allowing us to examine errors of omission. When false positives were removed, the lists would only differ if one device picked up a species that truly was there, and another device failed to detect that same bird. We again counted the number of agreements and disagreements in species lists between devices and calculated the mean number of species the lists differed by.

RESULTS

Differences in Detection Between Methods

Human observers had 382 individual species detections throughout the course of the study, 351 of which were determined to be correct by the expert reviewer, resulting in a precision of 92% (Table 1). Eight of those individual detections

Downloaded from https://academic.oup.com/condor/advance-article/doi/10.1093/ornithapp/duaf049/8222742 by guest on 05 January 2026

TABLE 1. Species count and individual detections for each observation type (Merlin, Human, and Reviewer) across scenarios of raw data (false positives included) and false positives removed.

	Observation type	Raw data	False positives removed
Species	*Merlin*	47	35
	Human	40	39
	Reviewer	45	
Individual detections	*Merlin*	222	192
	Human	382	351
	Reviewer	385	

were visual only. Merlin had 222 individual detections over the course of the study, 192 of which were determined to be correct by the expert reviewer, a precision of 86%. Merlin reported 12 false positive species: *Pluvialis squatarola* (Black-bellied Plover), *Setophaga striata* (Blackpoll Warbler), *Cardellina canadensis* (Canada Warbler), *Setophaga tigrina* (Cape-may Warbler), *Dumetella carolinensis* (Gray Catbird), *Passerina cyanea* (Indigo Bunting), *C. cardinalis*, *Buteo jamaicensis* (Red-tailed Hawk), *Agelaius phoeniceus* (Red-winged Blackbird), *Melospiza georgiana* (Swamp Sparrow), *Hylocichla mustelina* (Wood Thrush), and *Setophaga petechia* (Yellow Warbler). Human observers reported one false positive of *C. cardinalis*.

Variability Among Species

There was high variability among species identifications for human observers and Merlin. *Corvus brachyrhynchos*, *Cyanocitta cristata* (Blue Jay), *Sitta canadensis* (Red-breasted Nuthatch), and *V. olivaceus* were the species with the largest difference in frequency of detections, with human observers having more detections than Merlin in all 4 cases (Figure 3A). The most extreme case was *C. brachyrhynchos*; human observers detected this species 36 times over the course of the study, while Merlin detected it only 7 times. When examining species with over 5 detections, *Setophaga coronata* (Yellow-rumped Warbler) and *Bombycilla cedrorum* (Cedar Waxwing) were the only species where Merlin had a higher frequency of correct identifications than human observers, but by only 1 observation in both cases (Figure 3A). When examining species with fewer than five detections, Merlin performed better for 8 species and human observers performed better for 16 (Figure 3B). There were 4 species that only Merlin identified correctly over the course of the study: *Spizella passerine* (Chipping Sparrow), *Sayornis phoebe* (Eastern Phoebe), *Setophaga palmarum* (Palm Warbler), and *Melospiza melodia* (Song Sparrow). There were 8 species that only humans identified correctly over the course of the study: *Strix varia* (Barred Owl), *Branta canadensis* (Canada Goose), *Astur cooperii* (Cooper's Hawk), *Larus smithsonianus* (American Herring Gull), *Setophaga americana* (Northern Parula), *Setophaga pinus* (Pine Warbler), Buteo platypterus (Broad-winged Hawk), and *Dumetella carolinensis* (Gray Catbird). *Strix varia, A. cooperii, and S. pinus* were visual only detections.

Differences in Detection Between Devices

When examining uncorrected Merlin output from the 70 point counts (140 separate counts), where 2 devices ran Merlin at the same time, the devices agreed 30 times and disagreed 40 times. From these 40 disagreements, the lists differed by 1 species 29 times, 2 species 6 times, and 3 species 5 times, for an average of 1.4 species. The corrected lists with false positives removed agreed 44 times and disagreed 26 times. Out of 26 disagreements, the lists differed by 1 species 20 times and 2 species six times, for an average of 1.2 species.

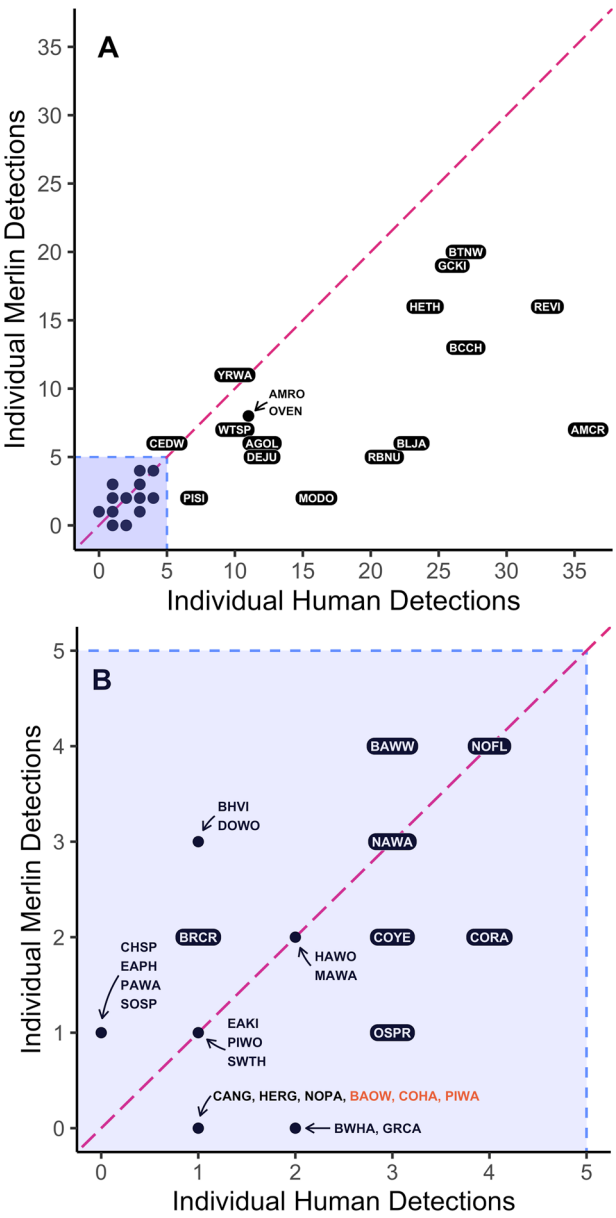

FIGURE 3. A comparison of the number of individual detections per species by human observers and Merlin with false positive identifications removed. The pink dashed line is a line of parity to compare performance. Species are represented by their four-letter alpha codes, a key to which is included in the Supplemental Material. (**A**) Species with number of individual detections > 5. The blue dashed lines represent the border of Figure 3B. (**B**) Species with number of individual detections < 5. *S. varia* (BAOW), *A. cooperii* (COHA), *and S. pinus* (PIWA) were the three species only ever detected visually, represented by orange text.

DISCUSSION

Human observers outperformed Merlin in precision and detected more individual birds and overall species over the

Downloaded from https://academic.oup.com/condor/advance-article/doi/10.1093/ornithapp/duaf049/8222742 by guest on 05 January 2026

course of the study. Human observers had 382 individual detections over the course of the study, compared to Merlin's 222 (Table 1). This means that when a bird vocalized, the human observer was 72% more likely to detect it than Merlin. The likelihood that a detected bird was identified correctly was similar between the 2 methods: 86% for Merlin and 92% for human observers. Merlin reported 12 false positive species, significantly more than the one from the human observers. For the 10 species where Merlin outperformed humans, there was usually a difference of only 1 detection, and never more than 2. When human observers outperformed Merlin it was by, in the most extreme case of *C. brachyrhynchos*, 29 detections (Figure 3A). Merlin frequently disagreed with itself when running on 2 different devices at the same time; out of 70 point counts where 2 people operated Merlin, the 2 devices generated the same species list at 30 points and disagreed by an average of 1.4 species at 40 points. When Merlin false positives were removed, the 2 devices disagreed at only 26 points. These 26 points were errors of omission for species that truly were present in the field, indicating that different devices have differing detection probabilities, possibly due to microphone quality or positioning during point counts.

Our results illuminate the differing, but complementary, abilities of human observers and Merlin. There were 8 species only detected by humans, 3 of which were visual detections. Human observers better detected species, such as *C. brachyrhynchos*, that vocalized from farther distances, generally during flyovers. Human observers also better identified species with low-frequency vocalizations, such as *Zenaida macroura* (Mourning Dove), a species noted to have especially low frequency song (0–1 kHz; Pieplow 2019). While Merlin is capable of recording vocalizations at this frequency, most anthropogenic sound pollution exists at this level, which can mask bird vocalizations on spectrograms (Goodwin and Podos 2013). The Merlin model is trained to identify both song and calls for many species, an advantage over human observers, who are often less confident identifying calls and focus on identifying singing birds during point counts. Merlin performed better with vocalizations from species that the human observer mentioned personally struggling to identify, such as *S. coronata* and *B. cedrorum*, and successfully detected four species that appeared only once over the course of the study and were not detected by the human observers. If Merlin had not been recording during those point counts, those species would have been completely missed.

Our results are relevant to our study system, but may not be fully generalizable to other environments. Overlap in time and pitch between the sound of a bird and other sounds in the environment may mask spectrogram features, leading to false negatives or positives. This exemplifies an issue with Merlin in loud environments; a point count in a city, for example, may present a particular challenge for the software. There was no significant variation in background noise at our field site with which to examine the impacts of noise on Merlin's performance. Another consideration is that different phones have different microphones, likely altering Merlin's detection probabilities. Survey protocols using multiple devices must evaluate inter-device variability by examining microphone response to standardized broadband frequency-playback. Future research should explore the use of external smartphone and tablet microphones to both standardize audio input between device types, and improve audio quality, with the goal of improving classifier performance. Even surveys using one device should examine changes in microphone sensitivity year-to-year. Users must also be wary of comparing results gathered from different versions of the classifier, but this information is not currently made clearly accessible. While performance varies in different habitats and on different devices, the neural network remains the same, a consistency not found with the variation in skill between humans. Our study focused on humans with intermediate song identification skills, but the identifications of a local expert may not be markedly improved by the integration of an automated classification tool, such as Merlin. Though, even if not used in statistical analysis, recordings collected in the app can function as valuable historical records of the soundscape. Nevertheless, future work should examine the potential for Merlin to improve the detection probability of point count surveys in different environments and with surveyors of varying skill levels.

Despite Merlin's flaws, we suggest future studies explore Merlin as a second observer in double-observer point counts, which may be its most appropriate use. In double-observer point counts, a primary observer records birds heard and seen at a point, while a secondary observer records the birds detected by the primary observer, while also recording the species missed by the primary observer (Nichols et al. 2000). The role of primary and secondary observer typically alternates over the course of the study, but it may be most appropriate to use Merlin exclusively as a secondary observer. This allows the primary observer to arbitrate any disagreements between their identifications and Merlin's and remove obvious false positives from Merlin's output. This may be necessary as Merlin reported 12 false positive species throughout our study, and occupancy models are not robust to false positives as they are to false negatives (Miller et al. 2011).

While obvious false positives may be easy to catch and remove, false positive detections of expected local species may prove more insidious. To avoid these false positives introducing bias, users should rely on a review of the audio from a count to determine if a Merlin identification was correct, and reject unverifiable identifications. Users of Merlin for research must also consider that unlike a human observer who can estimate the number of birds detected during a point count and thus measure abundance, Merlin lacks any comparative way to convey the number of birds it has detected. The app displays which species it has detected in a recording period, but cannot count how many different individuals within that species have vocalized. This is a significant drawback to Merlin's use, and is one of the reasons we do not recommend using Merlin alone to conduct surveys. Using Merlin for research requires novel methodology for implementation compared to other established strategies, but the potential benefits to biodiversity monitoring make this research worth pursuing.

Point counts are and will likely continue to be a go-to method for surveying birds, but the drawbacks of the data produced by them is well documented. A possible method to improve the quality of data collected via point-counts is automated sound identification methods such as Merlin, but questions have emerged about its validity in scientific research. Merlin being free, relatively easy to operate, and popular among novice and expert birders alike, makes it the logical choice for potential use in field-based surveys. We do not propose Merlin be used as a primary observer in point-count surveys, as its frequent

Downloaded from https://academic.oup.com/condor/advance-article/doi/10.1093/ornithapp/duaf049/8222742 by guest on 05 January 2026

false positives and missed detections make it unreliable alone and unreviewed. However, it could function as an aid to increase detection probability of rare species as a secondary observer. With further research into this methodology, Merlin could address some of the challenges associated with point-count based research.

Supplementary material

Supplementary material is available at *Ornithological Applications* online.

Acknowledgments

We would like to thank the seasonal staff and volunteers that assisted in the field data collection: Ella Halbert, Abby Bar, Evelyn Dyer, Jack Kovarik, Jamie Attanasio, Kate Moore, Catherine Devine, and Zoë Kennedy.

Funding statement:

Maine Coast Heritage Trust and the Cathy and Jim Gero Acadia Early-Career Fellowship provided funding.

Ethics statement

This study required no ethical permits.

Conflict of interest statement

The authors declare no conflicts of interest.

Author contributions

B.D.G designed and performed the experiment, wrote the paper, developed methods, and analyzed the data. K.L. conceived the idea, supervised research, substantially edited the paper, and analyzed data. S.B. conceived the idea and supervised research. N.A.F. contributed substantial resources to the manuscript. J.K. substantially edited the paper.

Data availability

Analyses reported in this article can be reproduced using the data provided by Goodman et al. (2025).

LITERATURE CITED

Bart, J., and S. Earnst (2002). Double sampling to estimate density and population trends in birds. *The Auk* 119:36–45.

Blumstein, D. T., D. J. Mennill, P. Clemins, L. Girod, K. Yao, G. Patricelli, J. L. Deppe, A. H. Krakauer, C. Clark, K. A. Cortopassi et al. (2011). Acoustic monitoring in terrestrial environments using microphone arrays: Applications, technological considerations and prospectus. *Journal of Applied Ecology* 48:758–767.

Chronister, L. M., J. T. Larkin, T. A. Rhinehart, D. King, J. L. Larkin, and J. Kitzes (2024). Evaluating the predictors of habitat use and successful reproduction in a model bird species using a large-scale automated acoustic array. *Ecography* 2025:e06940.

Cornell Lab of Ornithology. (2024). Identify bird songs and calls with Sound ID. Cornell Lab Merlin Bird ID. https://merlin.allaboutbirds.org/sound-id/

Darras, K., P. Batáry, B. Furnas, A. Celis-Murillo, S. L. Van Wilgenburg, Y. A. Mulyani, and T. Tscharntke (2018). Comparing the sampling performance of sound recorders versus point counts in bird surveys: A meta-analysis. *Journal of Applied Ecology* 55:2575–2586.

Faccio S., E. L. Sharron, B. R. Mitchell, A. Kozlowski, and M. Johnson (2024). NPS Region 1—Breeding Landbird Monitoring Volunteer Resource Manual. National Park Service. Northeast Temperate Network, Woodstock, VT, USA. https://irma.nps.gov/DataStore/Reference/Profile/2303728

Farmer, R. G., M. L. Leonard, and A. G. Horn (2012). Observer effects and avian-call-count survey quality: Rare-species biases and overconfidence. *The Auk* 129:76–86.

Francis, C. M., P. J. Blancher, and R. D. Phoenix (2009). Bird monitoring programs in Ontario: What have we got and what do we need?*The Forestry Chronicle* 85:202–217.

Gibb, R., E. Browning, P. Glover-Kapfer, and K. E. Jones (2019). Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. *Methods in Ecology and Evolution* 10:169–185.

Goodman, B. D., K. A. Lima, S. Benz, N. A. Fisichelli, and J. Kitzes (2025). Data from: Humans outperform Merlin Sound ID in field-based point-count surveys. *Ornithological Applications* 127:duaf049. https://doi.org/10.5061/dryad.3ffbg79xb [Dataset].

Goodwin, S. E., and J. Podos (2013). Shift of song frequencies in response to masking tones. *Animal Behaviour* 85:435–440.

Hoefer, S., D. T. McKnight, S. Allen-Ankins, E. J. Nordberg, and L. Schwarzkopf (2023). Passive acoustic monitoring in terrestrial vertebrates: A review. *Bioacoustics* 32:506–531.

Howe, R., and E. Giese (2021). *Certification Levels*. Birder Certificate Online. https://birdercertification.org

Hudson, M.-A. R., C. M. Francis, K. J. Campbell, C. M. Downes, A. C. Smith, and K. L. Pardieck (2017). The role of the North American Breeding Bird Survey in conservation. *The Condor: Ornithological Applications* 119:526–545.

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

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

Kepler, C. B., and J. M. Scott (1981). Reducing bird count variability by training observers. In *Estimating Numbers of Terrestrial Birds* (C. J. Ralph, and J. M. Scott, Editors). Studies in Avian Biology6:366–371.

Lapp, S., J. L. Larkin, H. A. Parker, J. T. Larkin, D. R. Shaffer, C. Tett, D. J. McNeil, C. J. Fiss, and J. Kitzes (2023). Automated recognition of ruffed grouse drumming in field recordings. *Wildlife Society Bulletin* 47:e1395.

Lapp, S., T. A. Rhinehart, L. Freeland-Haynes, J. Khilnani, A. Syunkova, and J. Kitzes (2023). OpenSoundscape: An open-source bioacoustics analysis package for Python. *Methods in Ecology and Evolution* 14:2321–2328.

Larkin, J. T. (2023). A multi-regional assessment of Eastern Whip-Poor-Will (Antrostomus vociferus) occupancy in managed and unmanaged forests using autonomous recording units. MSc thesis, University of Massachusetts, Amherst, MA, USA. https://doi.org/10.7275/35866786.0

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

Nichols, J. D., J. E. Hines, J. R. Sauer, F. W. Fallon, J. E. Fallon, and P. J. Heglund (2000). A double-observer approach for estimating detection probability and abundance from point counts. *The Auk* 117:393–408.

Pieplow, N. (2019). *Peterson Field Guide to Bird Sounds of Western North America*. Houghton Mifflin Harcourt, Boston, MA, USA.

R Core Team. (2025). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/

Downloaded from https://academic.oup.com/condor/advance-article/doi/10.1093/ornithapp/duaf049/8222742 by guest on 05 January 2026

Ralph, C. J., J. R. Sauer, and S. Droege (1995). *Monitoring bird populations by point counts*. General Technical Report PSW-GTR-149, U.S. Department of Agriculture, Forest Service, Pacific Southwest Research Station, Albany, CA, USA.

Rosenstock, S. S., D. R. Anderson, K. M. Giesen, T. Leukering, and M. F. Carter (2002). Landbird Counting Techniques: Current practices and an alternative. *The Auk* 119:46–53.

Sauer, J. R., K. L. Pardieck, D. J. Ziolkowski, A. C. Smith, M.-A. R. Hudson, V. Rodriguez, H. Berlanga, D. K. Niven, and W. A. Link (2017). The first 50 years of the North American Breeding Bird Survey. *The Condor: Ornithological Applications* 119:576–593.

Schmidt, B. R., S. S. Cruickshank, C. Bühler, and A. Bergamini (2023). Observers are a key source of detection heterogeneity and biased occupancy estimates in species monitoring. *Biological Conservation* 283:110102.

Shonfield, J., and E. Bayne (2017). Autonomous recording units in avian ecological research: Current use and future applications. *Avian Conservation and Ecology* 12:14.

Simons, T., K. Pollock, J. Wettroth, M. Alldredge, K. Pacifici, and J. Brewster (2009). Sources of measurement error, misclassification error, and bias in auditory avian point count data. *Modeling Demographic Processes in Marked Populations* 3:237–254.

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

Thompson, W. L. (2002). Towards reliable bird surveys: Accounting for individuals present but not detected. *The Auk* 119:18–25.

Wood, C. M., V. D. Popescu, H. Klinck, J. J. Keane, R. J. Gutiérrez, S. C. Sawyer, and M. Z. Peery (2019). Detecting small changes in populations at landscape scales: A bioacoustic site-occupancy framework. *Ecological Indicators* 98:492–507.

Downloaded from https://academic.oup.com/condor/advance-article/doi/10.1093/ornithapp/duaf049/8222742 by guest on 05 January 2026