2020 Bird Survey Season Report:

California State Vehicular Recreation Area Avian Monitoring for Habitat Conditions and Disturbance Effects

January 8, 2021

Jerry S. Cole and Rodney B. Siegel

The Institute for Bird Populations
P.O. Box 518
Petaluma, CA 94953

Table of Contents

Introduction

Research using automated recording units (ARUs) has increased dramatically in the last two decades (Darras et al. 2019), likely due to decreased cost of units and advances in methods for rapidly analyzing acoustic data (Gibb et al. 2019). ARUs can be deployed by a technician that is untrained in bird identification, and can collect recordings for extended periods of time. We believe the flexibility provided by having an untrained observer collecting monitoring data may be especially beneficial for State Vehicular Recreation Areas (SVRAs), many of which are distant from larger cities and may not be easily accessible by skilled bird surveyors. Recordings can be collected on site and transferred to a skilled observer who performs a "virtual survey" by listening and annotating the audio remotely.

Ultimately it might be most beneficial to have an automated processing system that can use recordings collected at SVRAs to generate species summaries with little human intervention, and then those data can be used by park personnel to make inferences about trends in bird populations. Such automated systems do exist, though until very recently were limited to a small suite of bird species (e.g., Shonfield et al. 2018). However, a prototype system known as BirdNET was recently developed by the Cornell Lab of Ornithology in collaboration with the Chemnitz University of Technology, that can identify nearly 1000 North American and European bird species (Kahl 2020).

At the beginning of 2020, the California State Parks Off-highway Motor Vehicle Recreation (OHMVR) Division entered in to agreement C19V0014 with The Institute for Bird Populations (IBP) in partnership with the National Audubon Society (NAS), where IBP and NAS were tasked with providing guidance on monitoring bird populations within State Vehicular Recreation Areas (SVRAs). More specifically, IBP was tasked with providing recommendations for improving current bird survey designs and for expanded testing of ARUs as a supplement to in-person surveys. Here we report our progress to date on this work.

In the first year of this study we collaborated with SVRA Environmental Scientists to collect acoustic recordings at SVRAs across the state with a low-cost ARU (i.e., Audiomoth) and a high-end ARU (i.e., Wildlife Acoustics Song Meter 4) with the goal of evaluating how well the BirdNET software was able to identify bird vocalizations captured by the two ARUs. In addition, human annotators identified all vocalizing bird species on the same recordings. We plan to use the human annotations as baseline to which we will compare the output of BirdNET. Using data collected during this process we plan to determine whether a low-cost ARU unit might be a suitable alternative for bird acoustic data collection and establish a set of best practices when using ARUs for bird monitoring at SVRAs.

Methods

Study Area

We provided ARU equipment and training protocols to all SVRAs that were interested in participating in the ARU project. The SVRAs that participated were: Clay Pit, Hollister Hills, Hungry Valley, Onyx Ranch, and Prairie City. These parks spanned a variety of habitat types including oak woodland, annual grassland, chaparral, oak woodland, and desert scrub. We provided park personnel with instruction in recording audio using ARUs at the Habitat Monitoring System (HMS) locations that have been seasonally surveyed in-person using a point count protocol.

Survey Protocol

We designed and built a tripod system (Figure 1) and provided SVRA staff with tripods that they could transport to each sampling location within their park. We also produced and provided a sampling protocol detailing the process required for collecting (Appendix 1) and archiving (Appendix 2) recordings using the equipment we provided.

Figure 1. ARU tripod deployed at a Habitat Monitoring System sampling point at Onyx Ranch SVRA in eastern Kern County. Photo courtesy of Shane Emerson.

Park staff were instructed to audibly announce the beginning and end of the survey to the ARU unit and stand a sufficient distance away from the units so ARUs would not pick up anthropogenic noise (i.e., papers rustling, pen scraping). The units were left to record at each HMS location for the duration of the park's standard in-person survey (e.g., 5 min, 7 min, 10 min, etc.) and then transported to the next

location within the SVRA. Staff were given the choice of conducting their own in-person bird survey alongside the ARUs. In addition to the standard survey protocol, Hungry Valley deployed ARU units recording at the center point of all survey transects for the duration of a morning.

Analysis

Recordings collected from both the Audiomoth and SM4 units at each park are currently being annotated by a team of skilled annotators. Recordings will also be processed by the BirdNET software to determine if ARU model has a significant effect on the effectiveness of BirdNET to detect and identify bird species correctly. Observed species richness for the two ARU units which will be derived from human annotation and BirdNET annotation, and compared. We also search for patterns in the bird species that are detected at a lower rate by the Audiomoth versus the SM4. We will use the long-duration recording data from Hungry Valley to determine the rate of species accumulation for the Audiomoth and the SM4 to identify the optimal sampling duration required to characterize the bird community at a survey location.

Results

With the generous assistance of the Environmental Scientists at SVRAs across California, we were able to collect recordings from the following parks:

- Clay Pit
- Hollister Hills
- Hungry Valley
- Prairie City – 2 rounds of sampling
- Onyx Ranch

We have now processed and archived recordings from these parks, including insuring that recordings are trimmed to the defined survey period of interest, and aligned so recordings from the Audiomoth and SM4 are synced in time. We have completed human annotation of recordings for the first round of surveys at Prairie City and Onyx Ranch, with efforts in progress for Hungry Valley and Hollister Hills. We estimate recording annotations and summary information should be completed before the start of spring bird surveys in 2021. Given the current outbreak of the Covid-19 pandemic in California we do not expect to be able to provide trained bird surveyors to SVRAs in 2021, but instead will again rely on conducting 'virtual point counts' using our ARU protocol. We will continue to provide technical assistance and ARU equipment to park Environmental Scientists in the coming year to support their bird surveying efforts.

During 2020, we were also able to continue analysis the ARU data we collected at Carnegie SVRA in 2018 to determine best practices for summarizing the automated detections generated by the BirdNET software (Cole et al. in review). The paper is currently under peer review by the ornithology journal *Condor: Ornithological Applications* and is attached in Appendix 3. The lay summary we provided in our paper is reproduced below:

- Manually identifying bird species on recordings is time consuming so we evaluated how well an automatic bird sound identification software called BirdNET.

- We manually identified bird species heard during a 10-minute recording at 34 sites and compared that to the species detected by BirdNET during recording periods spanning from 10 to 310 minutes at each site.

- BirdNET correctly identified 93% of the bird species detected during manual bird identification during a 10 min span, but only after BirdNET processed 260 minutes of recording from the same site.

- We demonstrate that BirdNET may be a suitable alternative to manual annotation of sound recordings and we provide recommendations that will help scale up bird surveys in an economical and reproducible way.

The main conclusion of our research was that when ARUs are set to record at a site for a long period of time (i.e. > 4 hr) BirdNET can begin to approach human-like performance at identifying the full complement of aurally detectable species. Based on our research, it is likely most efficient to establish long-term sampling stations at HMS survey points, if SVRAs are interested in continuing to collect bird monitoring data via ARU.

Establishing a long-term sampling station would entail driving in a t-post or other semi-permanent post into the ground and attaching a weatherproof enclosure for an ARU. Then SVRA staff could deploy a large number of ARUs (perhaps as many as 20 or 25) the day before the intended survey date and set the units to record the following day, and then collect units that afternoon. Then the ARUs could be moved to the next set of established stations. Staff could repeat this process until all stations are visited. Additional in-person surveys at HMS plots would be beneficial for detecting species that rarely vocalize or soar above plots (e.g., raptors).

Discussion

The first field season of bird sampling proved to be more difficult than expected because the coronavirus pandemic that precluded in-person visits by our staff to SVRAs. Fortunately, we still were able to collect data thanks in large part to the cooperation of Environmental Scientists at SVRAs. The number of parks with ARU deployments was more limited than we initially expected – 5 were surveyed instead of the intended 8 – but the data we collected will provide a wealth of information to evaluate the effectiveness of ARU models for long term, annual avian sampling within the SVRAs. The data previously collected in 2018 at Carnegie allowed us to further develop the analytical techniques required for process large volumes of audio data.

We look forward to collecting more recordings at all 5 participating SVRAs during spring 2021 and analyzing the audio data collected during spring 2020. We would also like to test the effectiveness of using Audiomoths to track vehicle usage within a SVRA. This vehicle traffic monitoring component of the contract had to be postponed in spring 2020 because the in-person collaboration that was required to establish sampling stations would have been unsafe. However, we are currently developing an alternative process for deploying the units for vehicle traffic monitoring while respecting Covid-related social distancing guidelines.

In addition to the bird survey support we have provided during the first year of the monitoring contract, we have also provided additional bird monitoring advice and analytical support to three of the SVRAs. We consulted with Stephanie Little of Oceano Dunes SVRA about the analysis of shorebird data

that has been collected by the park in recent years, and will be completing a report detailing our findings in early in 2021. We provided advice to Jessica Vannatta of Hungry Valley SVRA regarding expansion of bird survey locations within the park, and revision of the in-person bird survey protocol. Additionally, Prairie City SVRA consulted us regarding the design of a nest searching protocol for determining if any bird species were still nesting within a management unit slated for a prescribed burn. We have enjoyed the collaborative relationship between ourselves and the Environmental Scientists of the OHMVR Division and believe both parties have benefitted from the consulting portion of our current agreement.

Acknowledgments

We thank Shane Emerson for helping to coordinate our remote bird monitoring across all the SVRAs and collecting audio data from Onyx Ranch and Clay Pit SVRAs. A special thanks is due to each of the Environmental Scientists at participating parks including: Nicolas Somilleda at Hollister Hills, Jessi Vannatta and AJ Heredia at Hungry Valley, Tricia Farmer at Onyx Ranch, and McKenzie Boring and Peter Jones at Prairie City. We thank Tara Kerss at Carnegie SVRA for supporting our 2018 pilot ARU monitoring project.

Literature Cited

Cole, J. S., N. L. Michel, S. A. Emerson, and R. B. Siegel (in review). Longer duration recordings and optimized data filtering enable effective automated sound classification for multispecies bird surveys.

Darras, K., P. Batáry, B. J. Furnas, I. Grass, Y. A. Mulyani, and T. Tscharntke (2019). Autonomous sound recording outperforms human observation for sampling birds: a systematic map and user guide. Ecological Applications 29:e01954.

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.

Kahl, S. (2020). Stefan Kahl Identifying Birds by Sound: Large-scale Acoustic Event Recognition for Avian Activity Monitoring. Dissertation. Chemnitz University of Technology, Chemnitz, Germany.

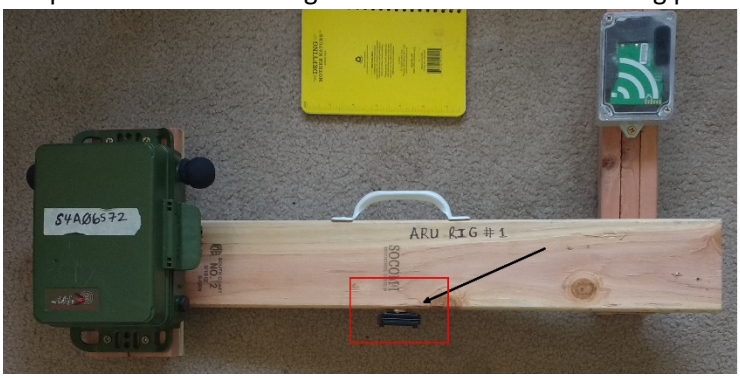
Shonfield, J., S. Heemskerk, and E. M. Bayne (2018). Utility of Automated Species Recognition for Acoustic Monitoring of Owls. Journal of Raptor Research 52:42–55.

Appendix 1: ARU tripod survey protocol

SVRA Park Staff ARU Deployment Protocol ver. 2

This protocol is intended to provide a set of instructions for untrained park staff that are conducting an in-person survey of a location by using a portable ARU tripod that holds two automated recording units (ARUs) that are running in tandem. **NOTE: Units are programmed to record from 6 AM to 12 PM.** If your survey is outside this time window contact jcole@birdpop.org for alternative scheduling files. Steps for surveying a bird sampling point are laid out in order below.

1. Check to make sure that the tripod mounting plate and screw are still solidly attached to the ARU mounting rig **before departing** for surveys in your vehicle. If screw is loose, use a flathead screwdriver to tighten. Or if using a unit with a plastic mounting plate **ONLY** tighten by hand. See the photos below showing the location of the mounting plate and screw.



2. Open the SM4 unit and press the “Schedule Stop” button.
3. From the Main Menu arrow down to Settings and press Enter
4. Select Audio and press enter
5. Arrow down to Sample Rate and make sure it is **48000Hz**
    a. **IF NOT** arrow to the right and arrow up or down until the display reads 48000Hz
    b. Press Enter
6. Turn the SM4 unit power switch from **INT to EXT**. The unit is being powered off for transport to reduce the likelihood of corrupting files saved to the SD card.
7. Fill the top portion of the **Park Staff ARU Deployment Sheet,** making sure to note the **ARU Rig #**, the **SM4 ID** (SongMeter 4 – the large green unit) which should have the ID taped to the front of the unit, and the **AudioMoth ID** (the smaller unit contained within the clear box), which should have the ID written on the unit and should be visible through the box cover. The Park Name field should be filled out with the name of your park (e.g., Carnegie), the Survey Duration field filled out with the length of an individual survey (e.g., 10 for 10 min survey), and the Observer field filled with the first and last name of the person conducting the survey.

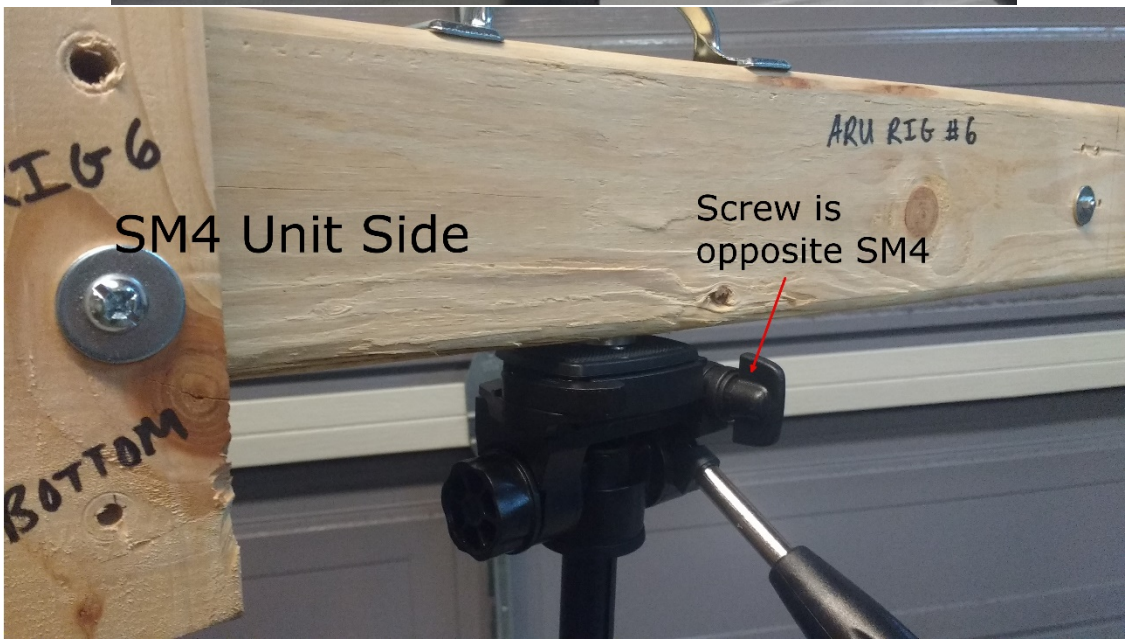
8. Make sure the AudioMoth is positioned correctly within the enclosure. Very little pink tape from the AudioMoth should be visible. See photo below for correct positioning. **MAKE SURE** the microphone symbol (seen below with arrow pointing to it) is inside of the large drilled hole in the case. This is important for getting a good recording.

9. Once you have arrived near your sampling point and have finished driving, begin programing your SM4 unit.
10. Open the SM4 case and flip the power switch from EXT to INT.
11. From the Main Menu select “Quick Start” and press “Enter”
12. Select “Record Always” from the next menu and press “Enter”
13. Press “Schedule Start” and wait for the green led to start blinking. The unit should now say “Currently Recording”.
14. Close the SM4 case and head to the survey point.
15. Check for blinking lights on the SM4 and AudioMoth units. The SM4 should have a blinking **green light** when it is recording and the AudioMoth a blinking **red light** when recording. Both units should be flashing because the units will have already begun recording according to their scheduling.
    a. **ONLY IN THE EVENT THAT** the AudioMoth unit is not blinking red or instead blinking green, remove the unit from the enclosure and further inspect the unit to see if there are actually no lights blinking.
        i. If so, switch the unit to USB/OFF and inspect the battery compartment – ensuring that all batteries are secure.
        ii. Then switch the unit to DEFAULT, and note the time that you started the recording on your datasheet. You should now only see a blinking red led light indicating the unit is recording.
        iii. The AudioMoth should now be ready to go for all following points.
16. Carry the ARU rig by the handle, though be careful because it is unbalanced. Try not to hit vegetation with the units when walking to the survey point.

17. When at the survey point set up the tripod and adjust the legs to level the tripod. If there is a slope, the legs should be fully extended on the downslope side to make the unit level. **Try to keep the unit as far away from ground vegetation as possible.** Extend the center post until you are about 4" above the tripod legs. Before extending the post, loosen the securing screw highlighted below, then when extended to the correct height retighten.

18. Level the tripod mounting head and tighten all adjustment screws **BEFORE** attaching the ARU head unit.

19. Attach the ARU rig to the tripod by opening the mounting plate securing clip on the tripod (see first photo below) and then insert the mounting plate from the ARU head unit into the tripod head. **Making sure** to position the heavy SM4 unit on side opposite the upper screw (see second photo).



20. Once you are sure the units are recording correctly (flashing green light for SM4 and flashing red light for Audiomoth) then continue with the survey protocol.

21. Fill out a blank row on the **Park Staff ARU Deployment Sheet**, noting the HMS point number under "Survey Point", the date of the survey MM/DD/YYYY, skip the "Time" field until the survey begins, circle Y or N to indicate if an observer will be silently conducting a traditional point count survey at the same time, circle Y or N to indicate if there was some ARU malfunction during the survey and note the details in comments.
22. Walk approximately 4 meters away from the ARU rig
23. Now note the start time of the survey in the "Time" column, use 24 hr time format (e.g., 13:00 for 1 PM). Begin your in-person survey **by audibly announcing** the following information:
    a. (Insert surveyor name) at survey point (insert point number) at (insert park name) on (insert date), start survey now (beep watch to start survey time)
    b. Example: Jerry Cole at survey point 23 at Carnegie SVRA on May fourteenth twenty twenty, start survey now (beep)
24. Start your watch or timing device to track the duration of the survey. Make the survey duration the same as the park's protocol for a traditional bird survey.
25. If conducting your own tandem survey, start your survey concurrent with the ARUs, taking care to be silent during the survey period.
26. Once the survey period has elapsed wait a few seconds before reading the following script:
    a. Survey point (insert point number) at (insert park name) end survey now
    b. Example: Survey point 23 at Carnegie SVRA end survey now
27. Open the SM4 enclosure and press the "Schedule Stop" button. The green led should stop flashing.
28. Flip the SM4 power switch from INT to EXT to power off the unit.
29. Disconnect the ARU Rig from the tripod by flipping the mounting clip. Return to the vehicle and proceed to surveying the next point starting again at **step 9**.
30. Once the time period for surveys has passed (e.g., stopping surveys after 10 AM) return to the office and follow the **Park Staff ARU File Storage Protocol** to download files from the units, and then follow the **Park Staff ARU Scheduling Protocol** to prepare the units for deployment the next day.

Scan or photograph all data sheets at the end of each day to ensure against accidental data loss.

Appendix 2: ARU tripod data archiving protocol

SVRA Park Staff ARU File Storage Protocol

This is the protocol for retrieving data from both the Song Meter 4 (SM4) and AudioMoth (AM) units after a day of bird surveys. You **should always download data** from the SD cards to a computer at the end of each day and preferably transfer these files to IBP electronically or download recordings to the supplied USB flash drive(s). Contact Jerry Cole jcole@birdpop.org for more information.

At the end of an ARU deployment day you will fill out the **Park Staff ARU Archiving Sheet**. You will fill separate sheets for each ARU Tripod system deployed. For instance, if a park has ARU Rig #1 and ARU Rig #2 both in use at the park, then you will have 2 separate forms for those units. The form can be used across multiple days, just fill in the Survey Date (MM/DD/YYYY) and File Name information, **making sure** the ARU Rig number is still correct.

For the Wildlife Acoustics SM4 unit

- Make sure to press the "Schedule Stop" button prior to switching the unit off or you risk corrupting the recordings.
- To power off the unit flip the power switch from INT to EXT and then remove the SD card. See the following photo for the location of the switch.

- Remove the SD card from the SM4 and insert in to your computer's SD card reader, or an external SD card reader
- In File Explorer you should see 2 files and one folder visible on the card

- Open the “Data” folder on the card
- If you haven’t already, create a Folder in your Documents on your work computer to hold the raw recording files named ARU Recordings.
- Copy the audio file(s) from the card to your ARU Recordings folder and create subfolders named after the rigs present within the parks, Create a new sub folder for recordings from that day named YYYY_MM_DD_Recordings, so for instance April 14, 2020 will be saved within 2020_04_14_Recordings.
- Briefly play the copied audio file to make sure you have a good transfer.
- Record the name of this copied file in the File Name column of the **Park Staff ARU Archiving Sheet**, and check the File Saved field to indicate task completion, and when the file has been transferred to IBP electronically or via USB drive check the File Sent field.
- After files have been copied to your computer’s folder then copy these files to the USB flash drive supplied. You can use the same folder naming system. An “ARU Recordings” folder and “Scanned Datasheets” folder. You can just simply copy the folder holding that days recordings on the laptop to the ARU Recordings folder on the flash drive. Also please scan in the data sheets for that day and place them within the Scanned Datasheets folder on the flash drive.
- Once data has been successfully transferred to the work computer and the flash drive, then format the SD card and reload the schedule by following the **Park Staff ARU Scheduling Protocol**.

For the AudioMoth unit

- **MAKE SURE** to flip the switch on the AudioMoth from “CUSTOM” to “USB/OFF”, wait for the red light to flash briefly and then go dark. The switches are fragile so please be gentle.
- Once lights are no longer flashing on the unit then you may remove the microSD card
- Insert the microSD into the SD card adapter and load the card on your computer

microSD card inserted

- Copy the audio file from the SD card to your already created ARU Recordings folder and the subfolder for the sampling date. **Important! If you are sampling with more than 1 ARU rig the**

AudioMoth files will be named the same across multiple units! In this instance, when you have already copied a set files from one AudioMoth for a given day the second set of files you copy will conflict with the files existing in the folder. Rename the newly copied file with a "_RigNumber" suffix **instead of overwriting the existing file.** So for instance if the newly copied file was named "20200414_053000" and came from ARU Rig 4 then you would rename file you're copying in "20200414_053000_4".

- Briefly play the audio from the copied file to ensure that you have a good transfer.
- Record the name of this copied file in the File Name column of the **Park Staff ARU Archiving Sheet**, check the file saved field, and when the file has been transferred to IBP check the File Sent field.
- Once data has been transferred to the computer folder and the USB flash drive you can format the card and reload the schedule using the **Park Staff ARU Scheduling Protocol**.

Transfer of files to IBP

- When all surveys have been completed during a sampling session (e.g., 3 days of surveys Apr 1, 2, and 3) then mail the USB flash drive to:
    - Jerry Cole
      1624 Joan Drive
      Petaluma, CA 94954

  Please retain the recordings, datasheets, and datasheet scans until IBP has confirmed that it has received the files. Thanks!

- Contact Jerry Cole at IBP, jcole@birdpop.org to coordinate transfer of files.

Apppendix 3: ARU manuscript in review at The Condor: Ornithological Applications

RESEARCH ARTICLE

Longer duration recordings and optimized data filtering enable effective automated sound classification for multispecies bird surveys

Jerry S. Cole[1], Nicole L. Michel[2], Shane A. Emerson[3], Rodney B. Siegel[1]

1. The Institute for Bird Populations, Petaluma, California, USA
2. National Audubon Society, New York, New York, USA
3. California State Parks, Off-highway Motor Vehicle Recreation Division, Sacramento, California, USA

ABSTRACT

The use of autonomous recording units (ARUs) for wildlife surveys has increased in the last decade as hardware costs have declined. Until very recently, few software options existed for automatically identifying birds from recordings, making annotation time-consuming. We tested BirdNET - an automated classifier capable of identifying vocalizations from >900 bird species across North America and Europe. We assessed the performance of BirdNET by comparing automated to manual annotations of recordings of breeding birds in northwestern California. We tested two procedures for filtering automated bird detections and evaluated the sampling duration needed for BirdNET to identify a similar number of species detected during manual annotation. BirdNET generated a high proportion of true positives and moderate number of false negatives, meaning some sites with manual bird detections had no detections by BirdNET. At the maximum F1 score (a metric which balances true positives and false negatives) for all species, precision (proportion of true positives to all detections) was 0.68 and recall (proportion of true positives to all true positives and false negatives) 0.48 for the more effective BirdNET filtering method. Most focal species required >100 min of recording to achieve the highest proportion of species detections that matched manual annotations (median of 90%). Nearly all bird species (93%) were correctly detected by BirdNET after >260 min of recording at each site. BirdNET was capable of correctly classifying 13 focal species, and was effective at identifying all focal bird species manually detected over a 10-min recording period. We conclude that BirdNET is suitable for annotating multispecies recordings provided that the sampling duration is sufficient – which may require much longer sampling times than are needed for studies using manual annotation. Together, ARUs and BirdNET may benefit monitoring and, ultimately, conservation by greatly increasing monitoring opportunities at minimal cost.

Keywords: acoustic monitoring, automated species classification, autonomous recording unit, bioacoustics, BirdNET, California, convolutional neural network

LAY SUMMARY

- Wildlife sound recorders produce vast amounts of data which can increase our understanding of bird communities.

- Manually identifying bird species on recordings is time consuming so we evaluated how well an automatic bird sound identification software called BirdNET performed the task.
- We manually identified bird species heard during a 10-minute recording at 34 sites and compared that to the species detected by BirdNET during recording periods spanning from 10 to 310 minutes at each site.
- BirdNET correctly identified 93% of the bird species detected during manual bird identification, but only after BirdNET processed 260 minutes of recording from the same site.
- We demonstrate that BirdNET may be a suitable alternative to manual annotation of sound recordings and provide recommendations that will help scale up bird surveys in an economical and reproducible way.

INTRODUCTION

The use of autonomous sound recording units for avian research has increased rapidly in the past 20 years (Darras et al. 2019), with techniques for processing and analyzing the large amounts of data generated by acoustic sampling also improving (Gibb et al. 2019). Early studies using recordings from autonomous recording units (hereafter ARUs) relied solely on humans skilled in identifying bird songs and calls to 'manually' annotate the recordings (Haselmayer and Quinn 2000, Celis-Murillo et al. 2009). However, advances in automated processing technologies have substantially reduced the person-hours required to annotate recordings (Katz et al. 2016a, Kahl 2019).

Automated bird sound classifiers are critical for enabling widespread, systematic, and inexpensive avian surveys carried out via ARU. Because vast amounts of audio recording data may be generated during long-term audio deployments, researchers typically subsample data to shorten the duration of the manual annotation process (Thompson et al. 2017). Automated classifiers have the potential to process recordings quickly at low cost, may increase the likelihood of detecting additional species or individuals because of the relative ease of processing long-duration recordings, and allow researchers to avoid difficult decisions about truncating or discarding data. However, automated classifiers are not without drawbacks – such as high false-positive and false-negative errors and, for some applications, a substantial time investment to develop and/or refine classifiers for species of interest (Gibb et al. 2019). The recent development of a classifier for >900 bird species across both North America and Europe (Kahl 2019) is therefore an exciting next step in automated bird sound processing. A joint effort between the Cornell Lab of Ornithology and Chemnitz University of Technology resulted in BirdNET, a comprehensive classifier that uses a convolutional neural network algorithm to rapidly identify a large suite of bird species in small segments of longer audio recordings (Kahl, 2019). To our knowledge this classifier has not been independently evaluated for its performance relative to a human annotator.

Classifier performance may be optimized in multiple ways. One way is ensuring that the recordings being annotated are sufficiently long (or contain enough vocalizations) to yield acceptably large number of high-quality detections of a given species. BirdNET, like many other classifiers, generates a confidence score from 0 to 1 that indicates how confident the classifier is for each annotation, with scores near 0 indicating little confidence and scores close to 1 near certainty. Filtering data by confidence score is thus another way to optimize classification performance. After generating a BirdNET annotation output users can filter these data by setting a confidence

threshold (a cutoff value that defines annotations with confidence scores greater than the threshold as valid) to fit their needs, either low to minimize the risk of missing detections (at a cost of increasing the frequency of false positives) or high to reduce the risk of false positives (at a cost of increasing the frequency of false negatives). Identifying the optimal sampling duration and confidence threshold values for filtering automatically annotated output is critical for replicating human-like annotation performance or using statistical methods to account explicitly for any shortcomings. Few studies have addressed finding optimal thresholds for particular classifiers, however see Knight et al. (2017) and Knight and Bayne (2019), nor included explicit justification for threshold choices, despite calls for more standardized reporting of such measures (Katz et al. 2016b, Knight et al. 2017). Reporting the threshold values used in a given study when annotating recordings with BirdNET will at the very least make results more reproducible, and at best give other researchers default threshold values they can use for their own applications.

We evaluated the BirdNET classifier's annotation performance on ARU sound recordings collected within a state park in northwestern California during the breeding season. Our objectives were to 1) identify the optimal confidence threshold that balanced false negative and false positive annotations for focal bird species and all bird species present, and 2) determine the minimum duration of recording required to identify the same number of species detected during a 10-min manual annotation. Based on our results, we present a generalized methodology for optimizing the performance of BirdNET.

METHODS

Study Area

We deployed ARUs to record bird vocalizations within Carnegie State Vehicular Recreation Area (37.6263°N, 121.5536°W), hereafter Carnegie, in northwestern California (Figure 1). Carnegie is a California State Park managed for off-highway vehicle (OHV) recreation on designated trails, as well as unrestricted riding in small portions of the park. Carnegie comprises > 2,000 ha of which approximately 1,400 ha are closed to vehicle recreation. The park is comprised of a mix of upland and riparian plant communities including oak woodland, California annual grassland, and shrublands dominated by coastal sagebrush (*Artemisia californica*) and black sage (*Salvia mellifera*).

Sampling Design

We deployed Song Meter 4 (Wildlife Acoustics, Massachusetts, USA) Autonomous Recording Units (ARUs) at 44 sampling sites across Carnegie between May 14 and May 24, 2018 and sampled each site on a single day from 15 min before local sunrise to 1100. We selected these dates and times because they represent the time of year (breeding season) and day when birds vocalize most frequently. ARUs were housed within a plywood enclosure and mounted on a steel t-post ~1.5 m above ground level (Appendix Figure 8**)**. The sound sampling rate was set to 44.1 kHz, left and right gain set to 16 dB, and preamplifier gain set to 26 dB. We used random stratified sampling to select 72 sites from among 114 established long-term bird monitoring sites, stratifying by OHV use type (36 sites within OHV use and 36 non-OHV use). We excluded sites within riparian habitat from the potential sampling sites because of high vehicle traffic noise from a nearby road. We were only able to sample 44 of the 72 selected sites because of rough road conditions that precluded access. We report results from 34 of the 44 sites (Figure 1) because recorded bird vocalizations were difficult to discern due to heavy wind noise at 10 sampling sites. We suggest that our limited sample of recordings is a suitable test of the practicality of using an automated classifier to make

inferences from data on the scale at which a typical land manager may work (i.e., a single park of moderate size).

Manual Annotation of Recordings

A single observer annotated all bird species heard on a single 10-min recording segment collected by an ARU at each sampling site. The observer annotated a segment collected during the early morning at randomly selected start times ranging from 0555 to 0830 hours, or 10 to 166 min after local sunrise. The observer used Sony MDRV6 headphones (Sony, Tokyo, Japan) to listen to recordings using Audacity 2.2.0 (The Audacity Team, https://www.audacityteam.org/download/), and visualized the sound using the spectrogram view. The 10-min segments were subdivided into 1-min sections during which the observer identified any vocalizing birds to species (i.e., species presence). The observer reviewed recordings multiple times and consulted reference sounds databases Xeno-canto (http://xeno-canto.org) and eBird (http://ebird.org/media/catalog?mediaType=a) as needed to confirm species identification. The observer also scored wind noise levels on recordings on a scale of 0 to 3 following Lankau et al. (2015) (see Appendix Table 4 for details) with 0 corresponding to no wind and 3 the heaviest wind.

BirdNET Annotation of Recordings

We used the BirdNET automatic bird sound classifier (Kahl 2019) to annotate the same 10-min recording segments reviewed by a human observer. We also used BirdNET to annotate the entire >5 hr recording collected at each sampling point. We ran BirdNET, which is freely available on GitHub (https://github.com/kahst/BirdNET), using Python 3.6.7, set it to classify sounds only to those species detected on eBird checklists within a 0.5° latitude by 0.5° longitude grid cell that encompassed Carnegie SVRA, and left all other settings at their default values. BirdNET by default partitions a recording into 3-sec non-overlapping segments and outputs a text file that provides identities for a maximum of 3 species that BirdNET had the highest confidence were present on a given segment. The BirdNET output also provides a confidence score that ranges from 0.1 to 1 for each species it identifies on a segment. BirdNET defaults to discarding any species with a confidence score < 0.1, however this is adjustable.

Classifier performance summaries for focal species. We assessed the performance of BirdNET at identifying 13 bird species that were most frequently detected during manual annotation because classification metrics for less frequently identified species were not very informative. Comparisons of bird abundance between annotations generated by BirdNET and a human are complicated because BirdNET rarely returns >1 detection of a single species on a 3 sec segment, thus effectively yielding a measure of occurrence (0 or 1) rather than abundance. We compared the BirdNET-derived measure of occurrence for each species during the 10-min sampling period to what we defined as the true measure of species presence (i.e., detection during manual annotation). A focal species was thus considered to be present at a sampling location if it was detected at least once by the human observer during the 10-min sampling period (Figure 2). We aggregated the BirdNET confidence scores for the same 10-min recording at each sampling site that was manually annotated to produce an estimate of the probability that at least one detection of a given species was a true positive following the procedure outlined in Balantic and Donovan (2019) such that:

$$tp\_prob_i = 1 - \prod_{d=1}^{total\ det} (1 - BNconf_{di})$$

Where tp_prob_i is the probability that at least one BirdNET detection during the 10-min recording period at sample site *i* is a true positive, $BNconf_{d,i}$ is the confidence score generated by BirdNET for detection *d* of a given focal species at site *i* (note that if there are no detections of a given species then tp_prob_i is set to 0), and *total det* is the total number of detections of a species at site *i*. For example, if BirdNET detected California Towhee (*Melozone crissalis*) at site 5 twice and the BirdNET confidence scores were 0.5 and 0.25 for those two detections, we would obtain tp_prob_5= 1 – (1 – 0.5) x (1 – 0.25) = 0.625. We refer to *tp_prob* hereafter as "aggregate detection confidence".

We calculated three classification performance metrics – recall, precision, and F1 (defined in Table 1) – for each of the focal species. Performance metrics for machine learning models are typically calculated at the most granular level of classification – in this case, 3-sec segments of a recording processed by BirdNET. However, because we were interested in how well BirdNET performed at determining site occupancy for species during a much longer survey period (10 min), we instead compared BirdNET-derived site detections to manually-derived site detections which we assumed for the purposes of this study to be free of errors. We assessed BirdNET performance metrics across a range of thresholds (hereafter confidence threshold) of aggregate detection confidence values spanning 0.1 to 0.9 because we wanted to determine the threshold that optimized both recall and precision. A site had a valid detection if the BirdNET aggregate detection confidence for a species at a site was greater than the user-supplied threshold. For example, a site with *tp_prob* = 0.4 was considered to have a valid detection if the threshold was 0.2, but had no detection if the threshold was set at 0.5. True positives were the total number of sites where a human and BirdNET both reported a detection, and false positives are where BirdNET had a detection but the human did not. Precision, recall, and F1 all range from 0 to 1, with 1 representing the best performance. We determined the optimal recall and precision for each focal species by finding the confidence threshold value between 0.1 and 0.9 where the F1 score was at its maximum.

Optimal sampling duration to reach human-like performance at detecting focal species. We also used BirdNET to automatically annotate recordings collected from 0542 to 1052 hours using the settings detailed previously from the same sites with 10-min manual annotations. We selected the recording period for annotation to ensure the same duration of sampling time across all ARUs, because units were set to record starting 15 min before sunrise, which varied slightly across sampling days (i.e., 0542 hours was the latest recording start time of all days). The recordings from each site were divided into 31 non-overlapping, 10-min segments. We filtered BirdNET detections of individual bird species at the optimal threshold for a given species, previously determined using the 10-min segments with manual annotations. We wanted to answer two questions with our longer BirdNET annotation procedure: 1) what is the minimum sampling period needed to correctly identify all sites with manual detections of each species? and 2) what is the maximum proportion of sites where BirdNET site classifications match manual annotations over the full 310-min period we evaluated? We calculated the aggregate detection confidence for each 10-min segment following the procedure outlined in the classifier performance section above (i.e., segment spanning 0 to 10 min had an independent aggregate detection confidence from segment spanning 10 to 20 min). A site was considered to have a valid detection if a species was detected at least once during any time period up to and including the one under consideration. For example, site 1 would be considered occupied after 30 min of sampling if there were any BirdNET detections during either the first, second, or third 10-min recording segments at that site.

Traditional classifier metrics are not applicable to BirdNET detection summaries from our extended annotations because we have no knowledge of false positives, false negatives, or even true positives for 10 min segments that were not annotated by a human. Instead, we calculated the proportion of sites that BirdNET correctly classified as have a detection at each time step, which we refer to as the "correct site proportion". The term correct site proportion is not meant to imply that species detected during manual annotation are all the species actually present at the site, but instead a measure of how well BirdNET can correctly identify the same species that were detected during manual annotation (which we assume is acceptably close to truth). The correct site proportion was defined such that if a human and BirdNET each detected a species at a given site then that was counted as a true positive. We summed all sites with true positives and divided this by the total number of sites with detections during the 10-min manual annotation period. The correct site proportion can range from 0 representing no sites identified correctly, to 1 representing all sites identified correctly. The correct site proportion measure does not assess false positives or negatives because we cannot be certain of presence of a species at a given point in time without human verification. Nonetheless our metric provides an index of how quickly BirdNET can approach human-like performance with regard to reducing false negatives at the site level.

Classifier performance summaries for all species. We estimated classifier performance for all bird species with a procedure similar to the focal species classifier performance calculation. However, many species other than our 13 focal species were detected at few sites by either the human annotator, BirdNET, or both, diminishing the utility of performance metrics. We instead calculated composite classifier performance metrics that took into consideration all species detected during surveys. The precision formula was the same as for focal species, except true positives were defined as species that were detected during manual annotation and detected by BirdNET at a given site, summed over all sites (i.e., a tally of how many species identifications were correct per site - for example if Wrentit (*Chamaea fasciata*) was detected at sites 2 and 5 by BirdNET and at the same sites by the human annotator then that would count as two true positives). To be clear, our measure of true positives does not compare simple richness, but instead site level species detections. False positives were defined as the species detected by BirdNET and not detected during manual annotation at a given site, summed over all sites. False negatives were defined as the species detected during manual annotation and not detected by BirdNET at a given site, summed over all sites.

We used two alternative methods for filtering BirdNET data via confidence thresholds: 1) we applied a uniform confidence threshold across all species which we refer to as the "uniform threshold" method and 2) we used a uniform threshold for all species except for the focal species, which we set to the optimal thresholds that we calculated previously, and refer to this as the "focal threshold" method. We calculated the maximum F1 value and the corresponding optimal confidence threshold value (the value which corresponds to the largest F1) for each method.

Optimal sampling duration to reach human-like performance – all species. We calculated the minimum sampling period required to maximize the proportion of manually annotated individual species detected by BirdNET using a procedure similar to the one used for focal species. Instead of calculating the proportion of sites correctly identified as occupied by BirdNET, we calculated the proportion of species identities correctly identified by BirdNET at each site. We plotted the accumulation of correct species identities when we filtered BirdNET output at 3 different confidence thresholds (0.1, 0.9, and thresholds that optimized F1) and 2 threshold methods (uniform and focal) for a total of 6 comparisons.

RESULTS

We detected 50 bird species during manual annotation of recordings collected during 10-min sampling periods at each of 34 sampling sites (Table 2). BirdNET detected 129 and 52 bird species when detections were filtered respectively at a low confidence threshold (confidence score > 0.1) and a high confidence threshold (confidence score > 0.9) during the same period. A total of 46 and 37 species detected by BirdNET matched the manually annotated species list when BirdNET data were filtered at a low and high confidence threshold, respectively (Figure 3 and Appendix Table 5). The species detected at the greatest percentage of sites during manual annotation were California Scrub-Jay (88%), California Quail (82%), Bewick's Wren (77%), and Ash-throated Flycatcher (65%). The species detected by BirdNET at the greatest percentage of sites when filtered at a high threshold were California Quail (47%), Bewick's Wren (44%), Bell's Sparrow (35%), and California Scrub-Jay (26%). The percentage of sites with manual and BirdNET species-level detections was more strongly correlated when filtered at a high confidence threshold (Pearson's r = 0.66) vs a low confidence threshold (Pearson's r = 0.56)(Figure 3). No Bell's Sparrows were detected during human annotation of the recording data. Upon further inspection we found BirdNET often misclassified cricket chirping on recordings as Bell's Sparrow, leading to the high number of false Bell's Sparrow detections.

Performance Metrics of BirdNET when Identifying Focal Species based on 10-min Samples

We calculated precision, recall, and F1 for 13 bird species detected most frequently during manual annotation. The BirdNET classifier had a precision of >0.75 at the lowest confidence threshold (0.1) for 10 of the 13 species, indicating that BirdNET reliably identified the focal species during 10-min recorded intervals (Figure 4**)**. However, recall – even at the lowest confidence threshold – was <0.50 for 6 of the 13 species. The optimal F1 score for most species was obtained at the lowest evaluated confidence threshold (Table 2). Maximum F1 scores ranged from 0.43 (Mourning Dove) to 0.85 (Bewick's Wren).

Reducing False Negatives with increased Sampling Duration for Focal Species

The BirdNET classifier correctly identified the presence of each focal species at nearly all sites (ranging from 84 to 100%) with manual detections of a species when sampling periods were extended sufficiently beyond 10 min (Figure 5). The proportion of sites with manually annotated detections of a species at which BirdNET correctly classified that species as present rarely reached peak value in less than 60 min (3 species) and the majority (8 species) required at least 100 min. The species for which BirdNET required the most time to reach its peak proportion of correct positive identifications was Mourning Dove, at 250 minutes. For all species, the proportion of sites with correct positive identifications plateaued long before the maximum proportion was obtained, suggesting that substantially shorter sampling durations may be nearly as effective at improving detection (Figure 5).

BirdNET Classification Performance for all Species

The use of optimal confidence thresholds for the 13 focal species (Table 2) improved the recall, precision, and F1 metrics for the entire bird community relative to the use of a single confidence threshold for all species (Figure 6, Table 3). The recall statistic for the focal threshold method declined more slowly with increasing confidence threshold relative to the uniform threshold method. The percentage of manually annotated species in the 10-min annotated recordings detected by BirdNET was 41% and 46% for the focal and uniform threshold methods, respectively.

Reduction in Species missed by BirdNET with increased Sampling Duration

The BirdNET classifier correctly identified a maximum of 93% of the species identified during 10 min of manual annotation after 260 minutes of sampling when using the focal or uniform threshold method (Figure 7). The percentage of manually annotated species detected by BirdNET was relatively similar between the focal and uniform threshold methods when we used a low confidence threshold of 0.1. When all BirdNET detections of non-focal species were filtered using a high confidence threshold (0.9), the focal threshold method resulted in a substantially larger proportion of correctly detected species (85%) vs the uniform threshold method (76%).

DISCUSSION

The use of an effective bird sound classification system to rapidly process ARU recordings could dramatically reduce the cost and time needed to conduct bird surveys, which could facilitate spatial and temporal expansion of survey efforts. However, automated classifiers may also have drawbacks, such as misclassification of species and missing species that vocalize infrequently. We found that these two issues can largely be mitigated by using optimal confidence thresholds for the specific application and longer sampling durations. With these methods employed, our results demonstrate that the BirdNET classifier can achieve performance comparable to manual annotations when annotating recordings collected within a moderately diverse western North American bird community.

BirdNET Classifier Performance for Focal Species

The BirdNET classifier performed relatively well at identifying 13 focal species. The classifier had relatively high precision across the full range of confidence threshold values that we tested. The focal species precision was high relative to the precision values previously reported (appendix D in Kahl 2019; Appendix Table 6). Some characteristics that may lead to the observed high precision among these species may be frequent vocalization, occurrence at high densities, or frequent movement within their home ranges so an ARU has a higher likelihood of obtaining a high-quality recording. We defined BirdNET detections during a 10 min period as survey-level events to match how in-person point occupancy surveys are carried out (i.e., a site is considered occupied if a species vocalizes at least once during a survey period). Therefore, if BirdNET missed identifying multiple vocalization events (i.e., 3 sec recordings with a bird vocalizing) during a survey, but managed to correctly identify at least one event then the species would still be considered detected and the false negatives effectively hidden in our measure of recall. Our metric of precision was much more liberal than those used in similar studies which defined true positives as classifier detections that are temporally very near (i.e., within 1 sec) manual annotations (Katz et al. 2016b) or were recording clips of vocalizations detected by a classifier and verified as correct by a human (Sebastián-González et al. 2018). In contrast, we considered an aggregated measure of presence during a 10-min period, rather than evaluating BirdNET output during the much shorter 3-sec periods for which it produced classifications.

Many of the focal species (6 of 13) had relatively low recall regardless of the confidence threshold. We hypothesize that 10-min recorded segments with false negatives likely had birds vocalizing farther too far from the ARU and a low signal-to-noise ratio, resulting in low BirdNET confidence scores. The relationship between a human's ability to discern bird vocalizations on a recording is well known to decline with distance (Yip et al. 2017) and a similar relationship appears to hold true for automated classifiers (Knight and Bayne 2019). Recall would have likely been even lower if we analyzed recordings with strong wind noise, given wind noise can adversely affect recall by

obscuring the audio signal of bird vocalizations (Stowell et al. 2019). For situations where selectively filtering recordings to only those with low to moderate wind noise is not feasible, evaluation of the performance of BirdNET after the application of denoising techniques (i.e., methods which remove background noise to allow better identification of the bird audio signal), such as wavelet decomposition (Priyadarshani et al. 2016) could be beneficial. Recall may be a lesser priority relative to precision when annotating ARU recordings automatically if the data will be analyzed using traditional occupancy modeling, which assumes no false positives and is specifically designed to account for false negatives (Mackenzie et al. 2002). However, models that explicitly account for false-positives and negatives in a more comprehensive occupancy framework (Miller et al. 2011, Miller et al. 2013, Chambert et al. 2015) also exist, including some that were explicitly developed for the analysis of automatically annotated data (Banner et al. 2018; Chambert et al. 2018). Occupancy models that account for false-positives have been demonstrated to provide accurate results with as little as 1% of automatic annotations verified by a human observer (Chambert et al. 2018).

We evaluated the influence of increased sampling time on the proportion of sites BirdNET correctly identified as occupied (i.e., a modified measure of recall) because we observed low recall values for the focal species. The majority of focal species (8 of 13) required >150 min of sampling time – 15 times longer than a human – to correctly identify the presence of a given species at a site at least once. However, sampling duration can be increased at little or no additional cost or human intervention and subsequently processed by an automated classifier. We hypothesize that the number of sites with true positives for a given species increased with greater sampling time for two main reasons: individual birds moved within their territory and vocalized closer to the ARU at least once; and more time yielded greater likelihood that a species would produce at least one vocalization that didn't overlap with vocalizations of other species. Quantifying the level of sampling effort required to reach human-like annotation performance by BirdNET, or any automated classifier, is important for reducing bias when integrating automated annotation output with manual annotated data. Our results indicate that a longer sampling period than is typically used for in-person surveys is necessary for surveying a bird community using ARUs and automated annotation via BirdNET.

Recall is directly related to the confidence threshold used for filtering a classifier output, with lower thresholds resulting in higher recall but also lower precision. Therefore, recall could potentially be improved by lowering the confidence threshold rather than extending the sampling period, but a correction would be needed to account for increased false positives and misclassification of species (Banner et al. 2018, Wright et al. 2020). Classifier confidence score generally declines with distance due to sound attenuation (Knight and Bayne 2019) and consequently the effective sampling area for a species is directly related to the confidence threshold used (i.e., a higher confidence threshold results in fewer valid detections of distant species and yields a smaller sampling area). Researchers interested in understanding the effective sampling area of an ARU for a given bird species might consider determining the relationship between distance to ARU and BirdNET's confidence score, ARU model, and habitat type. This can be accomplished with playback experiments to attract a bird to an ARU array (Knight and Bayne 2019), or by playing vocalizations of a species at a series of distances from an ARU (Yip et al. 2017) and modeling the relationship between classifier confidence score and distance.

A general-purpose bird sound classifier is perhaps most useful for ARU recording processing if it detects an equivalent number of bird species as a human observer. In pursuing this, we found that

determining the optimal confidence threshold for focal species was relatively straightforward, but setting thresholds for rarely detected species was difficult because of the limited sample size available. A method that appeared to improve the BirdNET classifier's performance was using confidence thresholds derived in earlier analytical steps and then determining an optimal uniform threshold for all remaining bird species. Manual annotation of recordings to identify an optimal confidence threshold for every study species in each new study area will likely be infeasible in most instances. Fortunately, confidence threshold relationships identified in our study area will likely hold true for other regions if the same ARU model is used under similar field conditions. Alternatively, optimal confidence threshold values may be estimated from a smaller, representative sample of the data (e.g., audio surveys collected across the range of habitat types and time periods), if a dataset is too large to manually annotate completely. We encourage researchers to follow previously established best practice guidelines (Knight et al. 2017), which recommend reporting the classifier performance metrics for individual bird species across a range confidence threshold values so that other researchers can determine optimal thresholds for their species of interest. Ideally, each bird species present within a study area would have a fine-tuned confidence threshold that matches the goals of the researcher (i.e., maximizing recall or precision, or finding an optimal balance between the two). Our method of using specific confidence thresholds for more frequently detected species provided an improvement in both recall and precision over a method that only used a uniform threshold across all species.

Optimal confidence thresholds provided by the developer of BirdNET (Kahl 2019) could be used for filtering annotations for a given species of interest. Precision, recall, and F0.5 (an F-score that more highly prioritizes precision than F1) calculated by Kahl (2019) were evaluated against weakly labeled recordings (i.e., a species of interest vocalized at some point during the recording) and likely contained unlabeled background species (Kahl 2019). Users who wish to use the default confidence threshold values provided by Kahl (2019) will be locked into prioritizing precision because they reported confidence thresholds at the maximum F0.5 value rather than F1. Thresholds identified by Kahl (2019) might not be the most optimal for the specific ARU model used in a given research program due to differences in microphone characteristics. BirdNET's default confidence thresholds differ from those we report because they are used to filter BirdNET confidence scores for detections at the 3-sec scale (i.e., the most granular level). For comparison, we report optimal threshold values using F0.5 alongside those reported by Kahl (2019) for our focal species (Appendix Table 6). We calculated our confidence thresholds to optimize aggregate confidence scores based on confidence scores from multiple 3-sec detections over a 10-min period. Therefore, if BirdNET default confidence thresholds are used to filter BirdNET detections, a single detection by BirdNET that exceeds the confidence threshold within a given survey period (in our case 10 min) would classify a detection at a site as valid.

Conclusion

Our results demonstrate that in a study area with moderate bird diversity the BirdNET classifier can survey a bird community nearly as effectively as a manual annotation of bird recordings, provided that an optimal confidence threshold is used and sampling duration is substantially increased relative to conventional point counts using human observers. We provide an independent assessment of the BirdNET classifier and give explicit guidance about the amount of time required for sufficient sampling of a bird community using BirdNET. Researchers using Wildlife Acoustics Song Meter 4 units for surveying western bird communities can use the precision-recall curves provided for the 13 focal species to set confidence thresholds to optimal levels for their survey

objectives. We encourage the continued evaluation of the BirdNET classifier in other geographic regions, with a variety of ARU models and field conditions (e.g., high wind and traffic noise). Development of a searchable database which compiles BirdNET performance metrics across multiple studies could prove useful for establishing standard confidence thresholds for all species identifiable by BirdNET. We are optimistic about the capacity of the BirdNET classifier for annotating bird acoustic recordings across North America and providing valuable standardization across studies and we look forward to more applications of this technology to ecological research. The twin developments of affordable ARUs and automated classifiers such as BirdNET provide the opportunity to greatly expand bird monitoring efforts and increase the accuracy and precision of abundance and trend estimates, improving our understanding of population status and dynamics. Moreover, ARUs are enabling scientists to scale-up monitoring in poorly surveyed regions (e.g., Van Wilgenburg et al. 2020), improving our ability to conserve vulnerable birds in these regions.

ACKNOWLEDGEMENTS

See the title page for this information.

LITERATURE CITED

Balantic, C., and T. Donovan (2019). Dynamic wildlife occupancy models using automated acoustic monitoring data. Ecological Applications 29:e01854.

Banner, K. M., K. M. Irvine, T. J. Rodhouse, W. J. Wright, R. M. Rodriguez, and A. R. Litt (2018). Improving geographically extensive acoustic survey designs for modeling species occurrence with imperfect detection and misidentification. Ecology and Evolution 8:6144–6156.

Celis-Murillo, A., J. L. Deppe, and M. F. Allen (2009). Using soundscape recordings to estimate bird species abundance, richness, and composition. Journal of Field Ornithology 80:64–78.

Chambert, T., D. A. W. Miller, and J. D. Nichols (2015). Modeling false positive detections in species occurrence data under different study designs. Ecology 96:332–339.

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

Darras, K., P. Batáry, B. J. Furnas, I. Grass, Y. A. Mulyani, and T. Tscharntke (2019). Autonomous sound recording outperforms human observation for sampling birds: a systematic map and user guide. Ecological Applications 29:e01954.

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.

Haselmayer, J., and J. S. Quinn (2000). A comparison of point counts and sound recording as bird survey methods in amazonian southeast Peru. The Condor 102:887–893.

Kahl, S. (2019). Stefan Kahl Identifying Birds by Sound: Large-scale Acoustic Event Recognition for Avian Activity Monitoring. Dissertation. Chemnitz University of Technology, Chemnitz, Germany.

Katz, J., S. D. Hafner, and T. Donovan (2016a). Tools for automated acoustic monitoring within the R package monitoR. Bioacoustics 25:197–210.

Katz, J., S. D. Hafner, and T. Donovan (2016b). Assessment of Error Rates in Acoustic Monitoring with the R package monitoR. Bioacoustics 25:177–196.

Knight, E. C., and E. M. Bayne (2019). Classification threshold and training data affect the quality and utility of focal species data processed with automated audio-recognition software. Bioacoustics 28:539–554.

Knight, E. C., K. C. Hannah, G. J. Foley, C. D. Scott, R. M. Brigham, and E. Bayne (2017). Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs. Avian Conservation and Ecology 12:art14.

Lankau, H.E., A. MacPhail, M. Knaggs, and E. Bayne. (2015). Acoustic recording analysis protocol. Bioacoustic Unit, University of Alberta and Alberta Biodiversity Monitoring Institute, Edmonton, Alberta, Canada.

Mackenzie, D. I., J. D. Nichols, G. B. Lachman, S. Droege, J. A. Royle, and C. A. Langtimm (2002). Estimating Site Occupancy Rates When Detection Probabilities Are Less Than One. Ecology 83:2248–2255.

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.

Miller, D. A. W., J. D. Nichols, J. A. Gude, L. N. Rich, K. M. Podruzny, J. E. Hines, and M. S. Mitchell (2013). Determining Occurrence Dynamics when False Positives Occur: Estimating the Range Dynamics of Wolves from Public Survey Data. PLoS ONE 8.

Priyadarshani, N., S. Marsland, I. Castro, and A. Punchihewa (2016). Birdsong denoising using wavelets. PLoS ONE 11:e0146790.

Sebastián-González, E., R. J. Camp, A. M. Tanimoto, P. M. de Oliveira, B. B. Lima, T. A. Marques, and P. J. Hart (2018). Density estimation of sound-producing terrestrial animals using single automatic acoustic recorders and distance sampling. Avian Conservation and Ecology 13.

Stowell, D., M. D. Wood, H. Pamuła, Y. Stylianou, and H. Glotin (2019). Automatic acoustic detection of birds through deep learning: The first Bird Audio Detection challenge. Methods in Ecology and Evolution 10:368–380.

Thompson, S. J., C. M. Handel, and L. B. McNew (2017). Autonomous acoustic recorders reveal complex patterns in avian detection probability. The Journal of Wildlife Management 81:1228–1241.

Van Wilgenburg, S. L., L. C. Mahon, G. Campbell, L. McLeod, M. Campbell, D. Evans, W. Easton, C. M. Francis, S. Haché, C. S. Machtans, C. Mader, et al. (2020). A cost efficient spatially balanced hierarchical sampling design for monitoring boreal birds incorporating access costs and habitat stratification. PLoS ONE 15:1–28.

Wright, W. J., K. M. Irvine, E. S. Almberg, and A. R. Litt (2020). Modelling misclassification in multi-species acoustic data when estimating occupancy and relative activity. Methods in Ecology and Evolution 11:71–81.

Yip, D. A., E. M. Bayne, P. Sólymos, J. Campbell, and D. Proppe (2017). Sound attenuation in forest and roadside environments: Implications for avian point-count surveys. The Condor: Ornithological Applications 119:73–84.

APPENDIX

Table 4. Criteria for scoring wind noise heard on ARU recordings. Reproduced from Lankau et al. (2015).

Code	Description
0	No wind
1	Rustling leaves/trees creaking (background noise), affects ability to detect distant/faint species
2	Begins to muffle microphones (frequency and decibel rates begin to spike), occasionally affects the ability to detect nearby species
3	Always muffles microphones, frequency and decibel graphs spike constantly (sometimes cuts out due to noise threshold)

Table 5. Number of sites where bird species were detected during either human annotation or automated annotation using the BirdNET classifier of 10-min recordings collected at Carnegie. Number of sites with BirdNET detections are summarized using a low confidence threshold of 0.1 and a high confidence threshold of 0.90. Species with no detections using a given method are denoted with "-".

			Sites with BirdNET detections	
Common name	Scientific name	Sites with human detections	Confidence threshold = 0.1	Confidence threshold = 0.90
Acorn Woodpecker	*Melanerpes formicivorus*	21	10	3
American Crow	*Corvus brachyrhynchos*	1	-	-
American Goldfinch	*Spinus tristis*	-	9	-
American Kestrel	*Falco sparverius*	6	7	4
American Pipit	*Anthus rubescens*	-	11	1
American Wigeon	*Mareca americana*	-	4	-
Anna's Hummingbird	*Calypte anna*	8	2	1
Ash-throated Flycatcher	*Myiarchus cinerascens*	22	16	7
Bald Eagle	*Haliaeetus leucocephalus*	-	2	-
Band-tailed Pigeon	*Patagioenas fasciata*	-	2	-
Barn Owl	*Tyto alba*	-	2	-
Bell's Sparrow	*Artemisiospiza belli*	-	25	12
Belted Kingfisher	*Megaceryle alcyon*	-	3	-
Bewick's Wren	*Thryomanes bewickii*	26	21	15
Black-headed Grosbeak	*Pheucticus melanocephalus*	13	10	3
Black-necked Stilt	*Himantopus mexicanus*	-	6	-
Black-throated Gray Warbler	*Setophaga nigrescens*	-	2	-
Black Phoebe	*Sayornis nigricans*	1	3	1

Blue-gray Gnatcatcher	*Polioptila caerulea*	2	3	1
Brewer's Blackbird	*Euphagus cyanocephalus*	-	3	1
Brown-headed Cowbird	*Molothrus ater*	1	1	-
Brown Creeper	*Certhia americana*	-	1	-
Bufflehead	*Bucephala albeola*	-	2	-
Bullock's Oriole	*Icterus bullockii*	1	7	-
Bushtit	*Psaltriparus minimus*	7	9	5
California Quail	*Callipepla californica*	28	20	16
California Scrub-Jay	*Aphelocoma californica*	30	14	9
California Thrasher	*Toxostoma redivivum*	6	13	4
California Towhee	*Melozone crissalis*	17	10	9
Canada Goose	*Branta canadensis*	-	1	-
Canyon Wren	*Catherpes mexicanus*	3	-	-
Cedar Waxwing	*Bombycilla cedrorum*	-	6	-
Chestnut-backed Chickadee	*Poecile rufescens*	-	3	-
Chipping Sparrow	*Spizella passerina*	-	1	-
Cliff Swallow	*Petrochelidon pyrrhonota*	-	3	1
Common Merganser	*Mergus merganser*	-	3	-
Common Raven	*Corvus corax*	19	10	2
Common Yellowthroat	*Geothlypis trichas*	-	2	-
Cooper's Hawk	*Accipiter cooperii*	-	1	-
Dark-eyed Junco	*Junco hyemalis*	6	7	1
Downy Woodpecker	*Picoides pubescens*	-	4	1
Dunlin	*Calidris alpina*	-	4	-
Eurasian Wigeon	*Mareca penelope*	-	2	-
European Starling	*Sturnus vulgaris*	6	8	3
Fox Sparrow	*Passerella iliaca*	-	1	-
Golden-crowned Kinglet	*Regulus satrapa*	-	1	-
Golden-crowned	*Zonotrichia atricapilla*	-	5	-

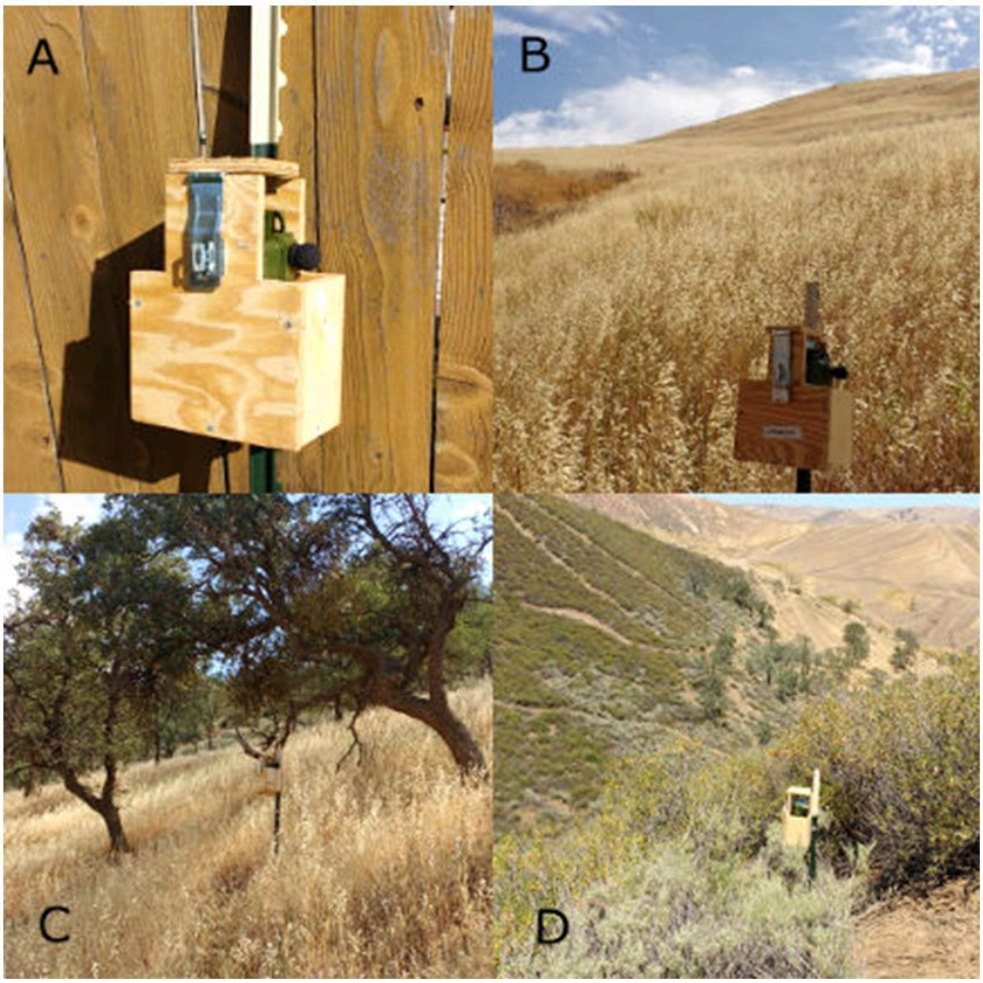
Sparrow				
Great Blue Heron	*Ardea herodias*	-	2	-
Great Egret	*Ardea alba*	-	3	-
Great Horned Owl	*Bubo virginianus*	-	3	-
Greater Roadrunner	*Geococcyx californianus*	-	8	-
Green-winged Teal	*Anas crecca*	-	2	-
Hairy Woodpecker	*Picoides villosus*	-	3	1
Hammond's Flycatcher	*Empidonax hammondii*	-	2	-
Hermit Thrush	*Catharus guttatus*	-	2	-
Hooded Merganser	*Lophodytes cucullatus*	-	4	1
Hooded Oriole	*Icterus cucullatus*	-	7	1
Horned Lark	*Eremophila alpestris*	-	15	2
House Finch	*Haemorhous mexicanus*	10	9	4
House Sparrow	*Passer domesticus*	-	1	-
House Wren	*Troglodytes aedon*	3	1	1
Hutton's Vireo	*Vireo huttoni*	-	1	-
Killdeer	*Charadrius vociferus*	3	3	2
Lark Sparrow	*Chondestes grammacus*	2	24	4
Lawrence's Goldfinch	*Spinus lawrencei*	3	15	5
Lazuli Bunting	*Passerina amoena*	-	11	1
Least Sandpiper	*Calidris minutilla*	-	1	-
Lesser Goldfinch	*Spinus psaltria*	3	4	1
Lesser Yellowlegs	*Tringa flavipes*	-	7	-
Lewis's Woodpecker	*Melanerpes lewis*	-	4	1
Lincoln's Sparrow	*Melospiza lincolnii*	-	4	-
Loggerhead Shrike	*Lanius ludovicianus*	1	8	2
Long-billed Curlew	*Numenius americanus*	-	2	1
Mallard	*Anas platyrhynchos*	-	1	-
Marsh Wren	*Cistothorus palustris*	-	2	-

Merlin	*Falco columbarius*	-	1	-
Mourning Dove	*Zenaida macroura*	22	7	2
Northern Flicker	*Colaptes auratus*	9	2	-
Northern Mockingbird	*Mimus polyglottos*	-	1	-
Northern Rough-winged Swallow	*Stelgidopteryx serripennis*	-	2	-
Nuttall's Woodpecker	*Picoides nuttallii*	19	12	5
Oak Titmouse	*Baeolophus inornatus*	19	13	8
Orange-crowned Warbler	*Oreothlypis celata*	-	4	-
Pacific-slope Flycatcher	*Empidonax difficilis*	6	3	2
Phainopepla	*Phainopepla nitens*	3	8	2
Pine Siskin	*Spinus pinus*	-	3	-
Purple Finch	*Haemorhous purpureus*	1	4	-
Red-breasted Sapsucker	*Sphyrapicus ruber*	-	4	-
Red-winged Blackbird	*Agelaius phoeniceus*	-	1	-
Ring-necked Duck	*Aythya collaris*	-	9	1
Rock Wren	*Salpinctes obsoletus*	-	4	1
Rufous-crowned Sparrow	*Aimophila ruficeps*	3	12	7
Rufous Hummingbird	*Selasphorus rufus*	-	2	-
Savannah Sparrow	*Passerculus sandwichensis*	2	10	-
Song Sparrow	*Melospiza melodia*	-	7	-
Spotted Towhee	*Pipilo maculatus*	14	7	6
Steller's Jay	*Cyanocitta stelleri*	-	1	-
Swainson's Thrush	*Catharus ustulatus*	2	-	-
Swamp Sparrow	*Melospiza georgiana*	-	4	-
Townsend's Solitaire	*Myadestes townsendi*	-	5	-
Townsend's Warbler	*Setophaga townsendi*	1	3	-
Tree Swallow	*Tachycineta bicolor*	-	9	-

Varied Thrush	*Ixoreus naevius*	-	1	-
Vaux's Swift	*Chaetura vauxi*	-	2	1
Violet-green Swallow	*Tachycineta thalassina*	-	4	-
Virginia Rail	*Rallus limicola*	-	1	-
Warbling Vireo	*Vireo gilvus*	4	1	-
Western Bluebird	*Sialia mexicana*	8	8	2
Western Grebe	*Aechmophorus occidentalis*	-	3	-
Western Kingbird	*Tyrannus verticalis*	4	-	-
Western Meadowlark	*Sturnella neglecta*	11	14	4
Western Screech-Owl	*Megascops kennicottii*	-	1	-
Western Tanager	*Piranga ludoviciana*	5	4	2
Western Wood-Pewee	*Contopus sordidulus*	6	3	1
White-breasted Nuthatch	*Sitta carolinensis*	14	9	4
White-crowned Sparrow	*Zonotrichia leucophrys*	-	11	-
White-throated Swift	*Aeronautes saxatalis*	-	1	-
Willow Flycatcher	*Empidonax traillii*	-	2	-
Wilson's Snipe	*Gallinago delicata*	-	1	-
Wilson's Warbler	*Cardellina pusilla*	9	7	4
Wrentit	*Chamaea fasciata*	15	10	6
Yellow-billed Magpie	*Pica nuttalli*	1	5	-
Yellow Warbler	*Setophaga petechia*	4	1	-

Table 6. Summary classifier performance values for 13 focal bird species detected at Carnegie derived from the BirdNET classifier for the threshold at which the F0.5 metric was maximized. We also provide the same metrics provided by Kahl (2019) in appendix D based on their model validation data. Recall was not reported in the supplementary data provided in Kahl (2019).

Common name	Confidence threshold Carnegie	Max F0.5 Carnegie	Precision at max F0.5 Carnegie	Recall at max F0.5 Carnegie	Confidence threshold BirdNET	Max F0.5 BirdNET	Average precision at max F0.5 BirdNET
Acorn Woodpecker	0.10	0.82	1.00	0.48	0.15	0.85	0.85
Ash-throated Flycatcher	0.10	0.93	1.00	0.73	0.16	0.78	0.75
Bewick's Wren	0.20	0.92	1.00	0.69	0.11	0.72	0.75
California Quail	0.10	0.88	1.00	0.59	0.18	0.68	0.65
California Scrub-Jay	0.10	0.93	1.00	0.71	0.13	0.74	0.75
California Towhee	0.10	0.81	1.00	0.47	0.07	0.77	0.75
Common Raven	0.10	0.68	0.80	0.42	0.12	0.67	0.66
Mourning Dove	0.15	0.65	1.00	0.27	0.08	0.53	0.52
Nuttall's Woodpecker	0.10	0.67	0.75	0.47	0.07	0.33	0.36
Oak Titmouse	0.80	0.82	1.00	0.47	0.27	0.80	0.76
Spotted Towhee	0.10	0.71	0.86	0.43	0.17	0.59	0.61
White-breasted Nuthatch	0.25	0.83	1.00	0.50	0.19	0.78	0.74
Wrentit	0.15	0.88	1.00	0.60	0.10	0.78	0.79



Appendix Figure 8

Figure 8. Autonomous recording unit (ARU) secured within a plywood box and attached to a steel t-post (A). ARUs deployed within California annual grassland (B), oak woodland (C), and coastal sagebrush – black sage (D) habitat within the study area

TABLES

Table 1. Definitions of classifier performance metrics.

Term	Definition	Formula
Precision	Proportion of identifications made by the BirdNET classifier that were correct	$\frac{true\ positives}{true\ positives + false\ positives}$
Recall	Proportion of all identifications (identified by manual annotation) identified correctly by the BirdNET classifier	$\frac{true\ positives}{true\ positives + false\ negatives}$
Fβ-score	Measure of classifier accuracy that considers both recall and precision. The β term adjusts the weighting to prioritize either precision (β smaller than 1), recall (β larger than 1), or neither (β equal to 1).	$(1 + \beta^2) \cdot \frac{recall \cdot precision}{(\beta^2 \cdot recall) + precision}$
F1	Measure of accuracy that equally weights recall and precision	$2 \cdot \frac{recall \cdot precision}{recall + precision}$
F0.5	Measure of accuracy that weights precision more highly than recall	$(1.25) \cdot \frac{recall \cdot precision}{(0.25 \cdot recall) + precision}$

Table 2. Performance metric values for 13 focal bird species derived from the BirdNET classifier for the threshold at which the F1 metric was maximized. The terms precision, recall, and F1 are defined in the methods section of the text.

Common name	Scientific name	Optimal confidence threshold	Max F1	Precision at max F1	Recall at max F1
Acorn Woodpecker	*Melanerpes formicivorus*	0.10	0.65	1.00	0.48
Ash-throated Flycatcher	*Myiarchus cinerascens*	0.10	0.84	1.00	0.73
Bewick's Wren	*Thryomanes bewickii*	0.10	0.85	0.95	0.77
California Quail	*Callipepla californica*	0.10	0.83	1.00	0.71
California Scrub-Jay	*Aphelocoma californica*	0.10	0.64	1.00	0.47
California Towhee	*Melozone crissalis*	0.10	0.74	1.00	0.59
Common Raven	*Corvus corax*	0.10	0.55	0.80	0.42
Mourning Dove	*Zenaida macroura*	0.15	0.43	1.00	0.27
Nuttall's Woodpecker	*Picoides nuttallii*	0.10	0.58	0.75	0.47
Oak Titmouse	*Baeolophus inornatus*	0.10	0.69	0.85	0.58
Spotted Towhee	*Pipilo maculatus*	0.10	0.57	0.86	0.43
White-breasted Nuthatch	*Sitta carolinensis*	0.25	0.67	1.00	0.50
Wrentit	*Chamaea fasciata*	0.15	0.75	1.00	0.60

Table 3. Classifier performance metrics for BirdNET annotations of all bird species when using the focal and uniform confidence threshold methods. The uniform threshold method varies the confidence threshold across all species simultaneously (i.e., every species has threshold set to 0.1, 0.15, 0.20, etc.). The focal threshold method varies the confidence threshold across all non-focal species simultaneously, but focal species (i.e., the 13 species defined in text) are set to their optimal confidence thresholds (reported in Table 2).

Confidence threshold method	Optimal confidence threshold	Max F1	Precision at max F1	Recall at max F1
Uniform	0.5	0.49	0.58	0.42
Focal	0.65	0.56	0.68	0.48

FIGURES

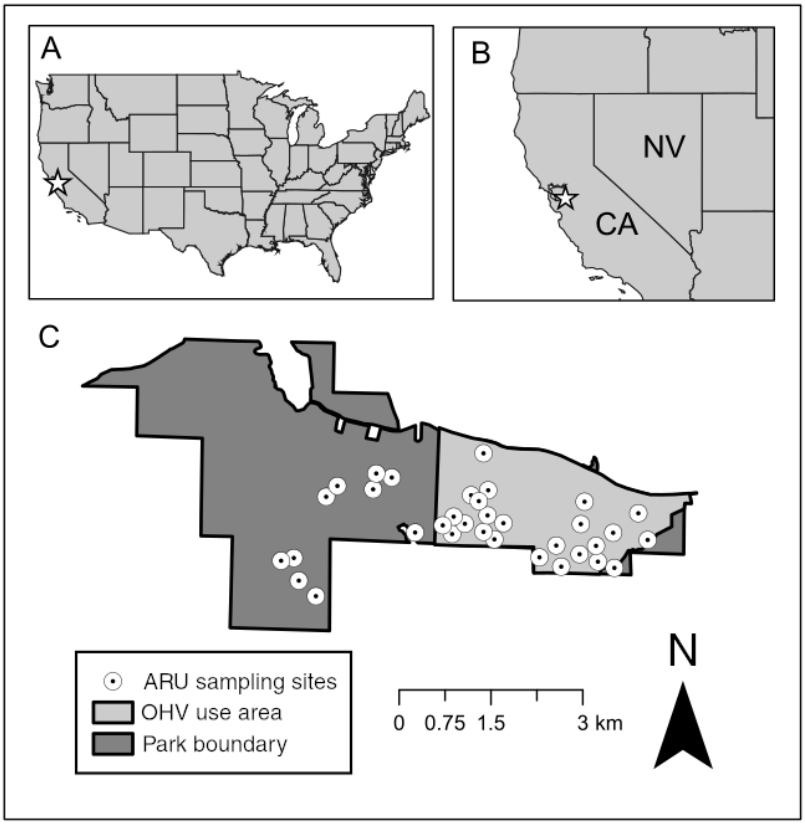


Figure 1

Figure 1. Location of study area (white star) within the continental United States (**A**) and within the southwestern United States (**B**). Distribution of 34 sampling sites within Carnegie State Vehicular Recreation Area in California, USA (**C**). State codes are the following: CA = California, NV = Nevada.

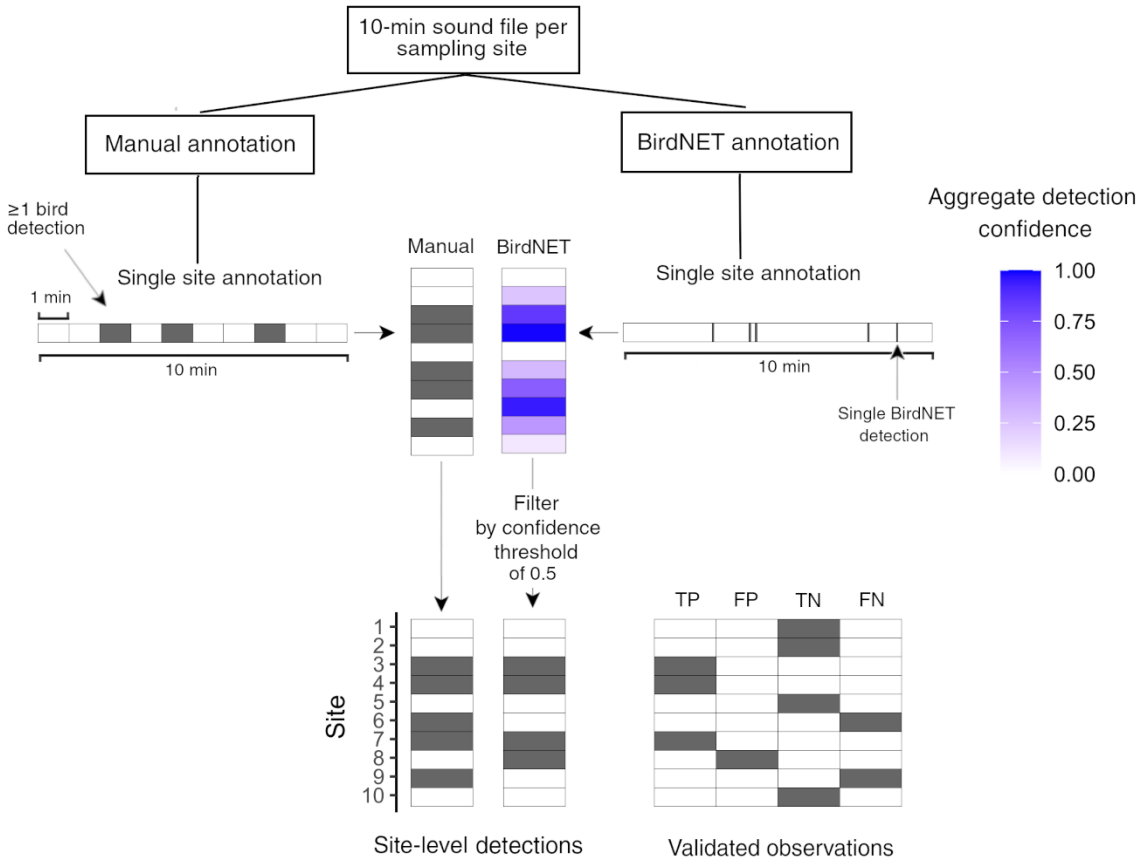


Figure 2

Figure 2. Workflow diagram for processing 10-min audio segments collected by ARUs at each sampling site. Each site had a single 10-min segment annotated manually by a human observer and automatically by BirdNET. A site is considered occupied during human annotation if a bird species is detected at least once (dark cells in the vertical column). The output of BirdNET annotation is summarized by an aggregate detection confidence (see Methods) for each site that ranges from 0 (no birds detected) to 1 (BirdNET was certain the species was present). The BirdNET aggregate detection confidence values were filtered by a user-specified threshold value (0.5 in the example) that converted any sites with values >0.5 to a value of 1 (indicating detection) or a value of 0 (indicating non-detection). Sites with detections by BirdNET were compared to sites with manual detections (considered truth) to determine true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) at each site.

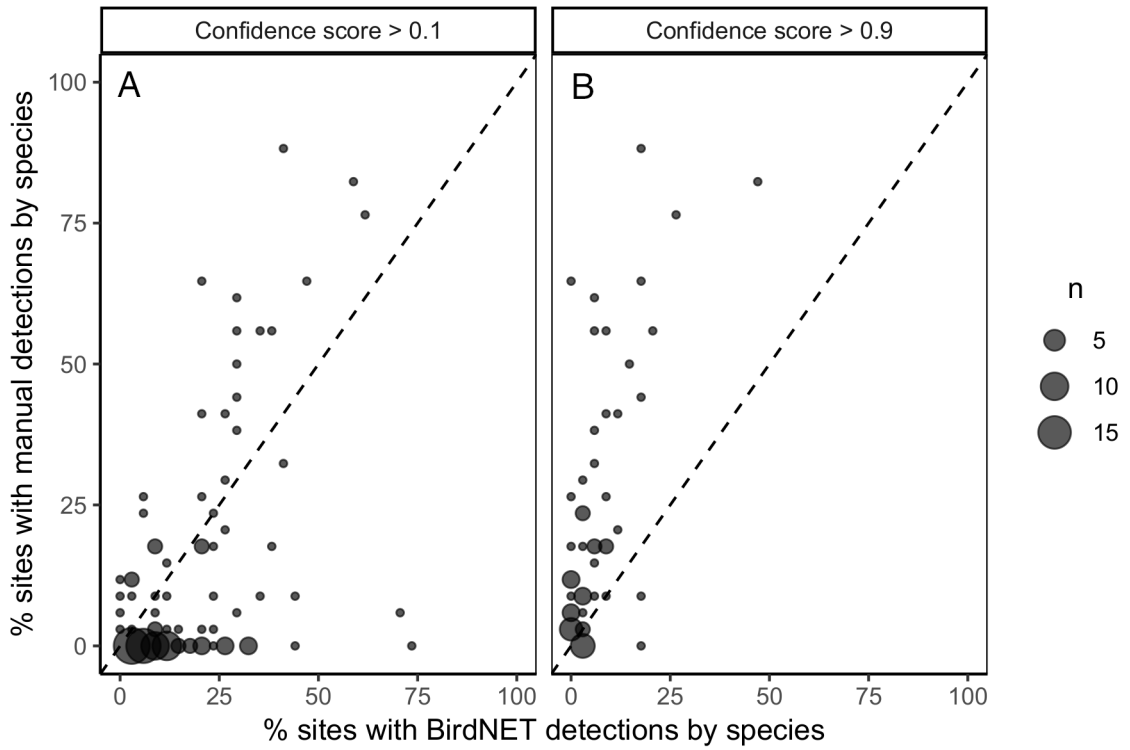


Figure 3

Figure 3. Comparison of the percentage of sampling sites with a minimum of one detection for a given species during either manual annotation (y-axis) or BirdNET annotation (x-axis) of recordings during a 10-min period at 34 sampling locations and BirdNET annotations filtered at a low confidence score (**A**) and high confidence score (**B**). Point diameter represents the number of bird species (n) with the same manual and BirdNET site percentage and the diagonal dashed line represents a 1:1 relationship. Points are not labeled with species identities for ease of presentation, but specific counts for site detections per species are provided in Appendix Table 5.

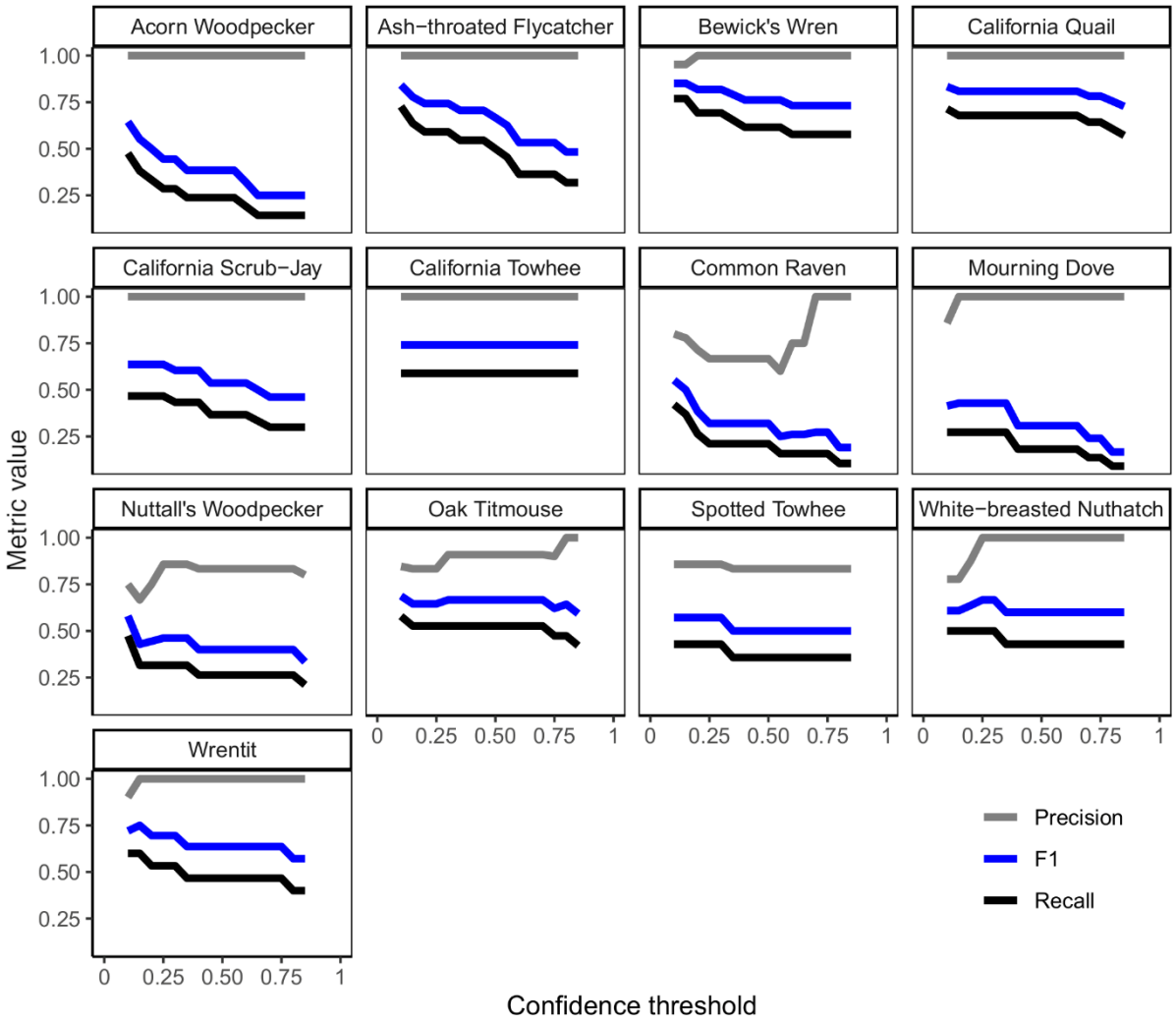


Figure 4

Figure 4. Precision, F1, and recall metrics for the BirdNET classifier for 13 bird species in response to varying confidence threshold values used to filter BirdNET detections. See Methods section for details about metric calculation.

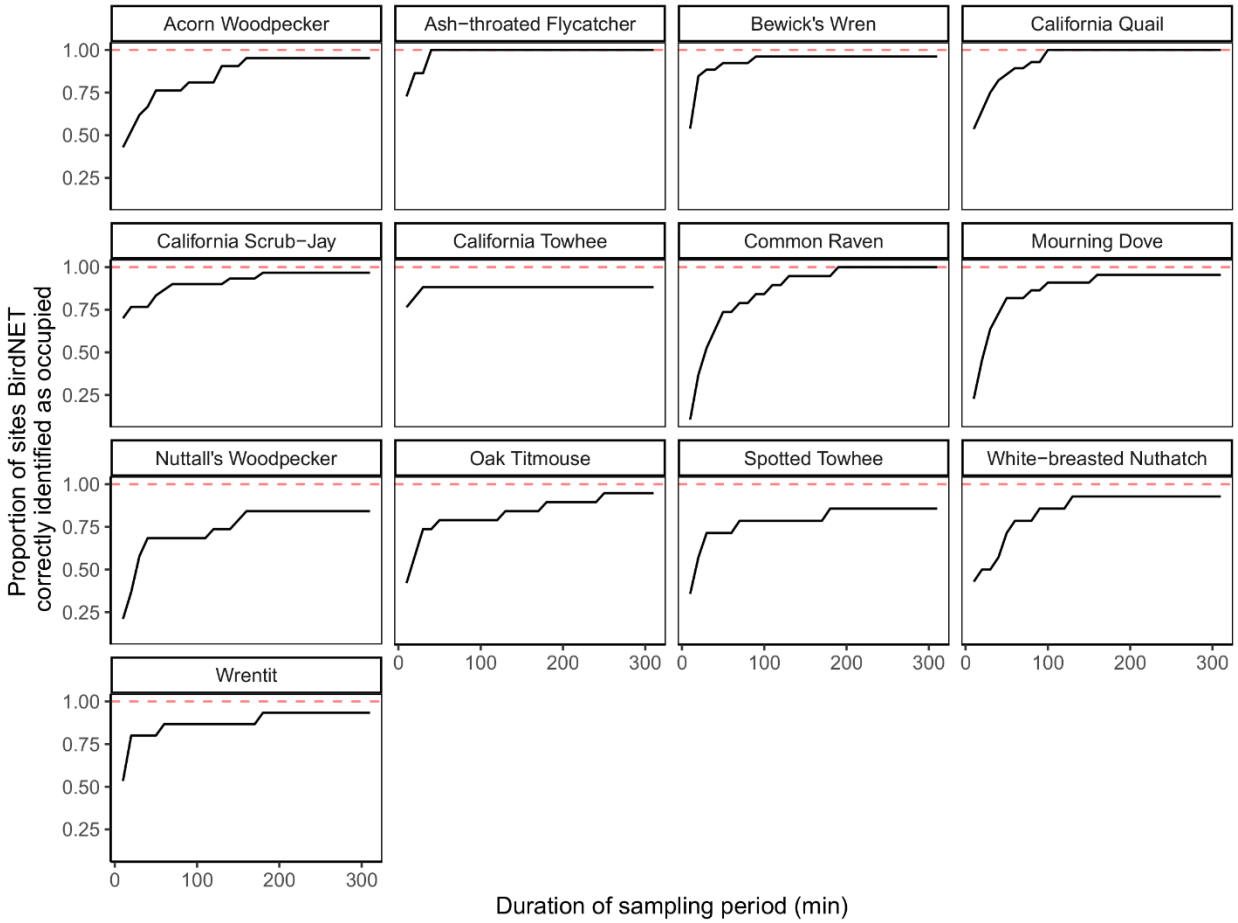


Figure 5

Figure 5. Effect of duration of sampling period on the proportion of sites BirdNET was able to correctly identify as occupied by each of 13 focal species. We set each species' threshold for filtering BirdNET data to the value that maximized F1 (see Table 2). The dotted horizontal line in each panel denotes the level where all sites classified as occupied during human annotation are also classified as occupied by BirdNET.

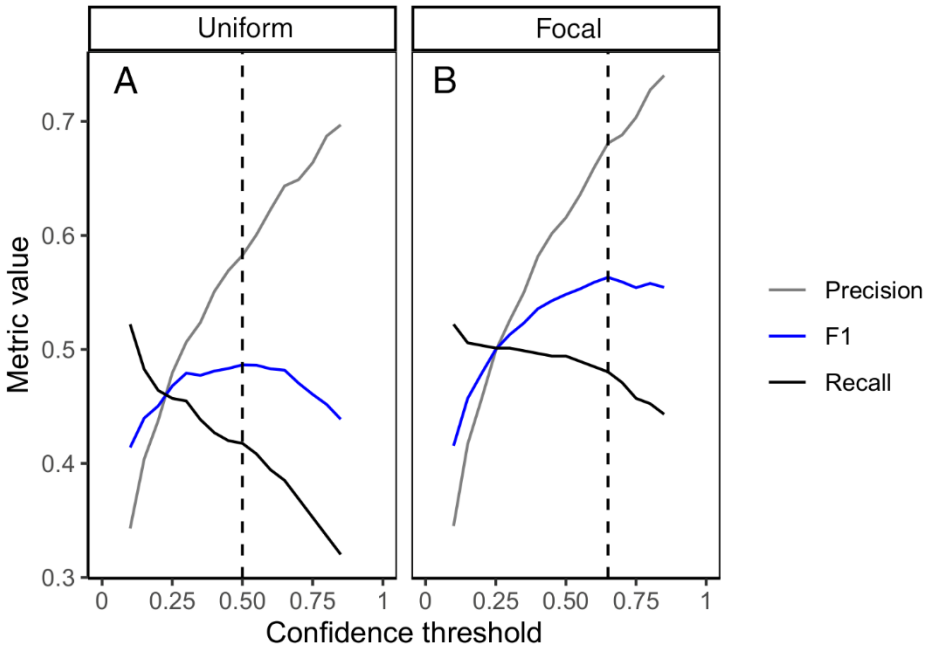


Figure 6

Figure 6. Precision, F1, and recall metrics for the BirdNET classifier across all bird species in response to varying confidence threshold values used to filter BirdNET detections. See methods section for details about statistic calculation. The uniform panel (**A**) displays the metrics with confidence thresholds for all species set to the same level (denoted on the x-axis). The focal panel (**B**) displays the metrics with confidence thresholds for the 13 most frequently detected focal species set to their optimal values (see Table 2) and confidence thresholds for all remaining species set to the same level (denoted on the x-axis). Vertical dotted line denotes the threshold value at the maximum F1 value.

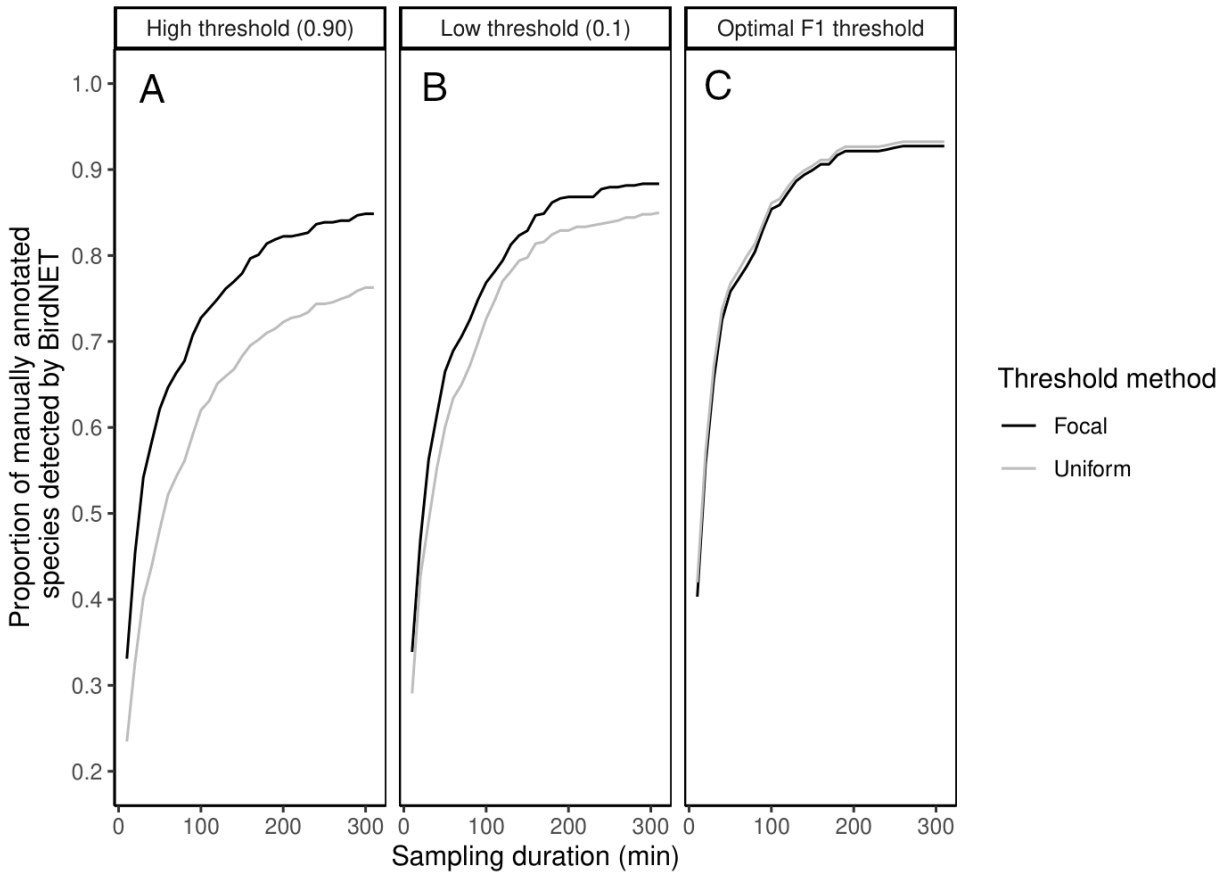


Figure 7

Figure 7. Effect of duration of sampling period on the proportion of species BirdNET was able to correctly identify as present across all sites, based on species detected during the 10-min human annotation. Proportion of species detected using the focal and uniform confidence threshold methods are denoted by dark and light gray lines, respectively. High (**A**) and low (**B**) confidence threshold panels display the results when BirdNET results for species were filtered at 0.90 and 0.1, respectively. The rightmost panel (**C**) displays the results with BirdNET output filtered at the optimal confidence thresholds identified in Figure 6 (Focal = 0.65, Uniform = 0.5) and with focal species thresholds set to their optimal levels, but only for the focal threshold method.