Ornithological Applications

AmericanOrnithology.org

Volume 124, 2022, pp. 1–15
https://doi.org/10.1093/ornithapp/duac003

RESEARCH ARTICLE

Automated bird sound classifications of long-duration recordings produce occupancy model outputs similar to manually annotated data

Jerry S. Cole,[1,*] Nicole L. Michel,[2] Shane A. Emerson,[3] and 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
*Corresponding author: jcole@birdpop.org

Submission Date: February 1, 2022; Editorial Acceptance Date: February 1, 2022; Published April 21, 2022

ABSTRACT

Occupancy modeling is used to evaluate avian distributions and habitat associations, yet it typically requires extensive survey effort because a minimum of 3 repeat samples are required for accurate parameter estimation. Autonomous recording units (ARUs) can reduce the need for surveyors on-site, yet their utility was limited by hardware costs and the time required to manually annotate recordings. Software that identifies bird vocalizations may reduce the expert time needed if classification is sufficiently accurate. We assessed the performance of BirdNET—an automated classifier capable of identifying vocalizations from >900 North American and European bird species—by comparing automated to manual annotations of recordings of 13 breeding bird species collected in northwestern California. We compared the parameter estimates of occupancy models evaluating habitat associations supplied with manually annotated data (9-min recording segments) to output from models supplied with BirdNET detections. We used 3 sets of BirdNET output to evaluate the duration of automatic annotation needed to approach manually annotated model parameter estimates: 9-min, 87-min, and 87-min of high-confidence detections. We incorporated 100 3-s manually validated BirdNET detections per species to estimate true and false positive rates within an occupancy model. BirdNET correctly identified 90% and 65% of the bird species a human detected when data were restricted to detections exceeding a low or high confidence score threshold, respectively. Occupancy estimates, including habitat associations, were similar regardless of method. Precision (proportion of true positives to all detections) was >0.70 for 9 of 13 species, and a low of 0.29. However, processing of longer recordings was needed to rival manually annotated data. We conclude that BirdNET is suitable for annotating multispecies recordings for occupancy modeling when extended recording durations are used. Together, ARUs and BirdNET may benefit monitoring and, ultimately, conservation of bird populations by greatly increasing monitoring opportunities.

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

LAY SUMMARY

- Occupancy modeling provides valuable information for understanding bird distributions, but often requires extensive survey effort. Autonomous recording units (ARUs) produce vast amounts of data, yet manually identifying birds on recordings is time-consuming.
- We evaluated the performance of an automated bird sound classifier, BirdNET, by comparing occupancy models that used manually and BirdNET-annotated data for 13 species in northwestern California, USA.
- We manually identified bird species heard during 9-min recordings at 34 sites, and used BirdNET to identify birds during 9–260-min recordings from each site. We also manually verified 100 BirdNET detections for each species.
- BirdNET correctly identified most bird species detected during manual bird identification when data were restricted respectively to nearly all (90% correct) or only high confidence (65% correct) detections. Habitat associations were similar across all models.
- We conclude that BirdNET is a useful tool for automatic annotation of bird vocalization data needed to model bird presence or absence.

Las clasificaciones automatizadas de sonidos de aves de grabaciones de larga duración producen resultados de modelos de ocupación similares a los datos anotados manualmente

RESUMEN

El modelado de la ocupación se utiliza para evaluar las distribuciones de aves y las asociaciones de hábitats, pero normalmente requiere un esfuerzo de estudio extenso porque se necesita un mínimo de tres muestras repetidas para una estimación precisa de los parámetros. Las unidades de grabación autónomas (UGA) pueden reducir la necesidad de censistas en el sitio, pero su utilidad estaba limitada por los costos de hardware y el tiempo requerido para anotar manualmente las grabaciones. Un software que identifique las vocalizaciones de las aves puede reducir el tiempo necesario de los expertos, si la clasificación es lo suficientemente precisa. Evaluamos el desempeño de BirdNET—un clasificador automatizado capaz de identificar vocalizaciones de más de 900 especies de aves de América del Norte y Europa—comparando anotaciones automáticas con anotaciones manuales de grabaciones de 13 especies de aves reproductoras registradas en el noroeste de California. Comparamos las estimaciones de los parámetros de los modelos de ocupación que evalúan las asociaciones de hábitat obtenidas a partir de datos anotados manualmente (segmentos de grabación de nueve minutos) con los resultados de los modelos obtenidos a partir de detecciones de BirdNET. Utilizamos tres sets de resultados de BirdNET para evaluar la duración de la anotación automática necesaria para aproximarse a las estimaciones de los parámetros del modelo de anotación manual: 9 min, 87 min y 87 min de detecciones de alta confianza. Incorporamos 100 detecciones de BirdNET de tres segundos por especie validadas manualmente para estimar las tasas de verdaderos y falsos positivos dentro de un modelo de ocupación. BirdNET identificó correctamente el 90% y el 65% de las especies de aves que un ser humano detectó cuando los datos se restringieron a las detecciones que superaban un umbral de puntuación de confianza bajo o alto, respectivamente. Las estimaciones de ocupación, incluidas las asociaciones de hábitat, fueron similares independientemente del método. La precisión (proporción de verdaderos positivos para todas las detecciones) fue >0.70 para nueve de las 13 especies, y un mínimo de 0.29. Sin embargo, se necesitó el procesamiento de grabaciones más largas para competir con los datos anotados manualmente. Concluimos que BirdNET es adecuado para anotar registros de múltiples especies para el modelado de la ocupación cuando se utilizan registros de duración extendida. Juntos, las UGA y BirdNET pueden beneficiar el monitoreo y, en última instancia, la conservación de las poblaciones de aves al aumentar considerablemente las oportunidades de monitoreo.

Palabras clave: bioacústica, BirdNET, clasificación automatizada de especies, MAP, monitoreo acústico, monitoreo acústico pasivo, red neuronal convolucional, unidad de grabación autónoma

INTRODUCTION

The use of autonomous sound recording units for avian research, also known as passive acoustic monitoring (PAM), has increased rapidly in avian research over the past 20 years (Sugai et al. 2019), with techniques for processing and analyzing large amounts of data generated by acoustic monitoring also improving (Gibb et al. 2019). Early studies using recordings from autonomous recording units (hereafter ARUs) relied on observers 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 trained person-hours required to annotate recordings (Katz et al. 2016a, Kahl 2019).

Vast amounts of audio recording data may be generated during long-term audio deployments, typically requiring researchers to subsample data to shorten the duration of the manual annotation process (Thompson et al. 2017). Automated classifiers process recordings quickly at a low cost and may increase the likelihood of detecting rare species or cryptic individuals because of the relative ease of processing long-duration recordings. Classifiers also allow researchers to avoid difficult decisions about truncating or discarding recording data due to the cost and time needed for manual annotation. However, automated classifiers are not without drawbacks – such as high false-positive and false-negative error rates and, for some applications, a substantial time investment to develop and/or refine classifiers for species of interest (Gibb et al. 2019) or to manually verify at least a subset of automated detections. 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 freely available 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). BirdNET can be run as a Python script executed from a console, used as an application on iPhone or Android, or accessed from an internet browser (https://birdnet.cornell.edu/). BirdNET divides a longer recording into smaller, non-customizable segment lengths of 3 s and analyzes each segment. After a preliminary review of bird vocalization data, the BirdNET team chose a 3-s segment interval because it would likely encompass a complete vocalization for most bird species (Kahl 2019). BirdNET then generates a large list of values (confidence scores) indicating how confident it is that vocalizations of each of the >900 species it can identify are present on the recording. This list of species is then sorted by confidence scores and the three species with the highest scores

are returned as output. Preliminary testing of BirdNET yielded a mean average precision (proportion of true positives to all detections) of 0.79 for single-species recordings across 984 North American and European bird species, indicating that it may be relatively useful for audio annotation of entire bird communities (Kahl et al. 2021). Few independent studies have evaluated BirdNET for its performance relative to manually annotated recordings (Kahl et al. 2021, Wood et al. 2021). Only a handful of studies have evaluated the performance of BirdNET in the western United States (Toenies and Rich 2021, Wood et al. 2021) or with bird sounds from across North America (Arif et al. 2020). Wood et al. (2021) provided a useful demonstration of the duration and spatial breadth of sampling required to adequately sample a western avian community with ARUs and BirdNET. However, none of these studies evaluated how BirdNET output might perform in an occupancy modeling framework relative to data collected using more traditional methods (i.e., human annotation of audio).

Among the tasks necessary for the use of BirdNET in research and monitoring are optimizing classifier performance and using this output to draw inferences about a given species. One way to improve performance is ensuring that the recordings being annotated are sufficiently long (or contain enough vocalizations) to yield an acceptably large number of high-quality detections of the target species. Having a large number of detections allows greater certainty of site occupancy status. BirdNET generates a confidence score from 0 to 1 that indicates how confident the classifier is in each detection. BirdNET users can filter annotation outputs by setting a confidence threshold (a cutoff value that defines detections 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 optimal sampling duration and confidence threshold values for filtering automatically annotated output is critical for replicating human-like annotation performance.

Occupancy modeling has long been used to evaluate bird status, population trends, distributions, and habitat associations (MacKenzie et al. 2006, Kéry et al. 2010, Saunders et al. 2019), but typically requires repeat visits to multiple sites by trained surveyors (MacKenzie et al. 2002). ARUs could provide the extensive data needed to fit robust models, but correct identification of birds from recordings is essential. Modeling false positive detections in an occupancy modeling framework has been well explored (Miller et al. 2011, 2013, Chambert et al. 2015, Banner et al. 2018), and many existing frameworks focus on processing acoustic data. Kéry and Royle (2020) extend the model developed by Chambert et al. (2018) to allow estimation of whether a species' detection is a false positive or true positive based on confidence scores of detections and supplementary information provided via validations. This approach frees the analyst from having to specify a confidence score threshold to filter BirdNET output, with the classification of observation as a true positive or false positive estimated from the data. If the modeling framework described by Kéry and Royle (2020) proves effective for modeling BirdNET output, it could provide a pipeline to gain insights into avian ecology, distribution, and population change through a simple combination of a more robust occupancy model and a state-of-the-art sound classifier.

We evaluated the BirdNET classifier's annotation performance on ARU sound recordings collected within a state park in northwestern California during the bird breeding season. Our goal was to assess whether BirdNET detections modeled in an occupancy framework can provide equivalent model output to human annotation data. More specifically our objectives were to 1) compare occupancy parameter estimates between occupancy models using BirdNET versus manually annotated data, 2) identify the confidence threshold that minimizes false positive detections for focal bird species, and 3) determine the minimum duration of recording required to identify focal species also detected during a 9-min manual annotation. The occupancy models that we compare include the following: 9-min of manual annotation, 9-min of BirdNET annotation, 87-min of BirdNET annotation, and 87-min of BirdNET annotation reduced to output with confidence scores greater than a predefined score.

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. Carnegie is a California State Park comprising >2,000 ha, and is managed for off-highway vehicle (OHV) recreation on designated trails, as well as unrestricted OHV use in small portions of the park. The park is comprised of a mix of upland and riparian plant communities, particularly oak woodland, California annual grassland, and shrublands dominated by coastal sagebrush (*Artemisia californica*) and black sage (*Salvia mellifera*).

Sampling Design

We deployed Song Meter SM4 (Wildlife Acoustics, Massachusetts, USA) ARUs across Carnegie between May 14 and 24, 2018, and sampled each site on a single day from 15 min before local sunrise to 11:00 hr on days without any precipitation. We selected these dates and times because they represent the time of year (breeding season) and the day when birds vocalize most frequently. Access constraints imposed by rough road conditions and proximity to a heavily trafficked road with excessive

TABLE 1. Precision (proportion of true positives relative to all positives) for 13 focal bird species derived from validation of 100 random samples of BirdNET detections of each species, recall (proportion of true positives to all true positives and false negatives) for 9-min recording segments manually annotated and annotated by BirdNET, the maximum confidence score for BirdNET detections that were confirmed to be false positives, and shorthand codes for each bird species (Species code). We also present the threshold above which BirdNET detections were included in the "BN Long Filtered" model.

Common name	Scientific name	Species code	Precision	Recall[a]	Maximum confidence score for false positives	Threshold to retain BirdNET observations
Acorn Woodpecker	*Melanerpes formicivorus*	ACWO	0.84	0.14	0.23	0.28
Ash-throated Flycatcher	*Myiarchus cinerascens*	ATFL	0.63	0.14	0.95	0.95
Bewick's Wren	*Thryomanes bewickii*	BEWR	0.91	0.60	0.49	0.54
California Quail	*Callipepla californica*	CAQU	0.96	0.68	0.08	0.13
California Scrub-Jay	*Aphelocoma californica*	CASJ	0.99	0.20	0.82	0.87
California Towhee	*Melozone crissalis*	CALT	0.99	0.53	0.10	0.15
Common Raven	*Corvus corax*	CORA	0.29	0.11	0.79	0.84
Mourning Dove	*Zenaida macroura*	MODO	0.67	0.09	0.56	0.61
Nuttall's Woodpecker	*Picoides nuttallii*	NUWO	0.95	0.26	0.38	0.43
Oak Titmouse	*Baeolophus inornatus*	OATI	0.91	0.50	0.57	0.62
Spotted Towhee	*Pipilo maculatus*	SPTO	0.74	0.43	0.10	0.15
White-breasted Nuthatch	*Sitta carolinensis*	WBNU	0.69	0.29	0.98	0.95
Wrentit	*Chamaea fasciata*	WREN	0.90	0.40	0.34	0.39

[a]Note that precision and recall are not on the same scale: precision is comparing 3-s recording segments, and recall is comparing whether BirdNET had any detections during a 9-min period at the same sites where a human detected a species during annotation of the same period.

noise interference, along with a random stratified sampling of otherwise useable sites, led us to select 44 sites from among 114 established long-term bird monitoring sites. ARUs were housed within plywood enclosures and mounted on a steel t-post ~1.5 m above ground level (Supplementary Material Figure 1). ARU sampling sites were >150 m from the next site (range = 157–587 m apart) and 75% of sites were >220 m from the next closest site. 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 report results from 34 of the 44 sites because recorded bird vocalizations were difficult to discern due to heavy wind noise at 10 sampling sites, so our recordings may reflect more ideal recording conditions than are typically encountered. We selected 13 focal bird species (Table 1) that were most frequently detected during manual annotation and therefore each presumably provides a relatively broad variety of vocalizations on the recordings (i.e., more detections result in a higher likelihood of different vocalization types) and better facilitate the evaluation of BirdNET's performance on more than a single stereotypical vocalization.

Manual Annotation of Recordings

A single expert observer annotated all bird species heard on a single 9-min recording segment collected by an ARU at each sampling site for a total of 396 min across all sites. A 9-min period was chosen because it could be divided evenly into three 3-min periods for occupancy modeling. We also annotated recordings at a much smaller temporal scale of 3-s to match the length used by BirdNET during annotation. From each ARU, the observer annotated a segment collected during the early morning at a randomly selected start time ranging from 5:55 to 8:30 hr, or 10 to 166 min after local sunrise. The observer used headphones (Model: MDRV6; Sony, Tokyo, Japan) to listen to recordings using Audacity version 2.2.0 (The Audacity Team, https://www.audacityteam.org/download/), and visualized the sound using the spectrogram view. Spectrogram parameters were set to those recommended for songbirds (Lankau et al. 2015). The 9-min segments were initially subdivided into 1-min sections during which the observer identified any vocalizing birds to species (i.e., species presence) because we thought finer-grained annotation might prove useful, though after annotation we recombined the data into 3-min sections. The observer reviewed recordings multiple times and consulted reference sounds (xeno-canto: http://xeno-canto.org, eBird: http://ebird.org/media/catalog?mediaType=a) as needed to confirm species identification. The observer also scored wind noise on recordings on a scale of 0 to 3 (Supplementary Material Table 1) with 0 corresponding to no wind and 3 the heaviest wind. We excluded sites with wind scores of 3.

BirdNET Annotation of Recordings

We used the BirdNET automatic bird sound classifier (Kahl 2019), which is freely available on GitHub (https://github.com/kahst/BirdNET), to annotate the same 9-min recording segments annotated by a human observer. We also used BirdNET to annotate the entire >5 hr recording collected at each sampling point during the day that included the 9-min segment. We ran BirdNET using Python 3.6.7 (Van Rossum and Drake 2009), set it to classify sounds only for the 209 species detected on eBird checklists within a 0.5° latitude by 0.5° longitude grid cell that encompassed Carnegie (Supplementary Material Table 2), and kept all remaining settings at their default values. BirdNET divides recordings into 3-s non-overlapping segments and outputs a text file that provides identities for a maximum of 3 species that BirdNET had the highest confidence – measured in a score ranging from 0 (least confident) to 1 (most confident) – were present on a given segment. We filtered the BirdNET output to only those detections that were ranked 1 in the list of species and had confidence scores >0.01.

Human Validation of BirdNET

We randomly selected 100 BirdNET-generated detections for each of the 13 focal species and validated the 3-s recording segments associated with the purported detections. The observer listened to each 3-s segment and inspected a spectrogram representation of the same recording and then noted whether the focal species was present or absent. The observer was made aware of the focal species that BirdNET identified vocalizing on each set of recordings, and confirmed only whether that species was present on the recording. We used this validation information, along with the confidence scores generated for BirdNET detections, for determining classifier performance.

Classifier performance summaries for focal species. We compared the BirdNET-derived measure of occurrence for each species during a 9-min sampling period to our reference 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 9-min sampling period. For each of the focal species, we calculated a single classification performance metric, precision, which is defined as:

$$\frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

A true positive occurred when both BirdNET and the human verifier detected a given species during a 3-s period, and a false positive occurred when the species was detected by BirdNET but not by the human verifier during the same 3-s period. We assessed BirdNET's precision without filtering the detection data by confidence threshold (i.e., all detections with confidence scores >0.01 were included). We used the validation data (i.e., 100 detections for each species) to calculate precision and determine the maximum BirdNET confidence score for false-positive detections for each species. The maximum confidence score for false positives indicates how stringently to filter BirdNET output to maximize precision. For each species, we added 0.05 to the maximum of all confidence scores for verified false-positive detections and used this as the threshold for filtering BirdNET output by confidence score (Table 1). For species with maximum confidence scores >0.95, we did not add 0.05 because that would result in data too sparse to model. We also calculated the recall metric at a different scale than precision (9-min rather than 3-s) defined as:

$$\frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

where true positives were sites that had at least one manual detection and at least one BirdNET detection of a given species during the manually annotated 9-min period False negatives were sites with a manual detection but no BirdNET detections during the same 9-min period. Again, as with the precision metric, we filtered BirdNET detections at a defined confidence score (Table 1).

Optimal sampling duration for detecting focal species. We also used BirdNET to automatically annotate recordings collected from 5:42 hr (15 min before the latest sunrise) to 10:27 hr from each site on the same day from which we obtained the 9-min manual detections. These long-duration recordings from each site were divided into 26 non-overlapping, 10-min segments. We partitioned our long-duration recordings into 10-min segments rather than 9-min because we wanted to use recording lengths typical of an acoustic deployment (e.g., collecting 10-min recordings each hour). We excluded 20 min of recordings when an in-person visit to each site took place (in case it was disruptive to the bird community). We filtered BirdNET detections of individual bird species using a specified threshold for false positives as described above (Table 1). 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, a site 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. We randomized the unique order of visits per site for BirdNET annotation data 100 times (where each 10-min recording segment is considered a "visit"), using the "sample" function in R, to reduce the likelihood of bias due to time of visit (e.g., birds vocalize more frequently during earlier visits), yielding 100 unique sequences of BirdNET detections.

Traditional classifier metrics are not applicable to BirdNET detection summaries from our extended annotations because we have no knowledge of false

TABLE 2. Description of occupancy models that used a mixture of manual and BirdNET annotation data. The specified values used to filter BirdNET detections are listed in Table 1.

Model name	Number of 3-min annotations included per site	BirdNET annotations with confidence score > 0.01?	Detections with confidence score > specified value?	Included BirdNET annotations?	Confidence score data used for estimation of true positives?
Manual	3	NA	NA	No	No
BN Short	3	Yes	No	Yes	Yes
BN Long	29	Yes	No	Yes	Yes
BN Long Filtered	29	Yes	Yes	Yes	No

positives, false negatives, or even true positives on recording segments that were not annotated by a human, with the exception of our validation data. Instead, we calculated the proportion of sites that BirdNET classified as occupied for each species at each time step that also had manual detections, which we refer to as the "reference site proportion". The reference site proportion provides a baseline to which BirdNET can be compared, assuming species were identified correctly by the human annotator. The reference 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. For each species, we summed all sites with true positives and divided this by the total number of sites with detections during the 9-min manual annotation period (Supplementary Material Figure 2). The reference site proportion can range from 0 representing no sites identified correctly, to 1 representing all sites identified correctly. The reference site proportion is very similar to the traditional recall metric (see Classifier performance section above) such that we define our metric as:

$$\frac{\text{sites with} \geq 1 \text{ BirdNET detection}}{\text{sites with manual detection in 9-min period}}$$

where the numerator could potentially be defined as "true positives" if we knew the status of the site outside the 9-min manual annotation period, and the denominator could be considered the sum of true positives and false negatives (sites where BirdNET has not detected a species but a human has) but we have no insight into species presence outside the 9-min period. The reference site proportion measure does not comprehensively assess false positives or negatives because we cannot be certain of the presence of a species at a given point in time without much more extensive verification (i.e., our manual annotations only give a brief snapshot of where bird species are present). Nonetheless, our metric provides an index of how quickly BirdNET can identify occupied sites relative to a single, short (9-min) manual annotation period.

Modeling occupancy and habitat associations of manual and BirdNET annotations. We modeled focal species occupancy using two data sources with distinct occupancy modeling frameworks. The first data source was derived from manual annotations of a 9-min period at each survey site, with detections of a given species aggregated over three 3-min periods at each site which we refer to as the "manual" data. The remaining data were derived from BirdNET and we only retained the first 3 min of every 10 min of recording per site to match the 3-min manual annotation periods. The subset of BirdNET annotations are as follows: annotations from the same three 3-min periods as the manual data (referred to as "BN Short"), annotations from the same three 3-min periods as the manual data and twenty-six additional 3-min periods at each site (referred to as "BN Long"), and annotations from the same three 3-min periods as the manual data and 26 additional 3-min periods at each site but only including BirdNET detections that exceeded a defined confidence score (Table 1) for each species (referred to as "BN Long Filtered"). Unlike BN Long Filtered, both BN Short and BN Long included all BirdNET detections with confidence scores >0.01. See Table 2 for a brief overview of each model included in our analyses.

The manual data were modeled using an occupancy model of the form detailed in Tingley and Beissinger (2013). Detection (1) or non-detection (0) of a species at site i and period j was represented as y_{ij}. We assumed that manual observations of occupancy were observations of an unobservable occurrence state z_i such that:

$$y_{ij} \sim \text{Bernoulli}(p_{ij} \times z_i)$$

where p_{ij} is the probability of detection at each sampling site and period. We included two covariates on detection, time of day (hours after midnight) and naïve detection of a species (i.e., raw occurrence data) at a site in the previous sampling period, such that:

$$\text{logit}\left(p_{ij}\right) = \alpha_0 + \alpha_1 \times y_{ij-1} + \alpha_2 \times \text{time}_{ij}$$

and where $y_{i0} = 0$ because occurrence is unknown before the first visit. The α_0, α_1, and α_2 coefficients represent the detection intercept, effect of the previous detection, and time of day on detection probability, respectively. We included the α_1 coefficient because of concerns that detections on recordings collected sequentially in time would be

correlated to one another (i.e., 3-min segments collected one after another).

The BN Short, BN Long, and BN Long filtered data were modeled using an occupancy framework that utilizes detection validation information and frequency of detections to inform occupancy estimates (Kéry and Royle 2020: Chp 7). In the case of BirdNET data, we assumed counts of bird vocalizations yBN_{ij} detected at site i during period j were Poisson distributed such that:

$$yBN_{ij} \sim \text{Poisson}(\lambda \times z_i + \omega)$$

where λ is the true positive rate and ω is the false positive rate. Counts of bird vocalizations refer to the total number of 3-s recording segments where BirdNET detected a species of interest within a 3-min period, so the count could range from 0 (no detections) to 60 (every 3-s segment contained a positive detection).

We assume that each BirdNET detection belonged to group $g = 1$ representing a true positive, or group $g = 2$ representing a false positive and had an associated confidence score x_k for each detection of index k. Because the distribution of confidence scores from BirdNET was often strongly skewed (Supplementary Material Figure 3), we chose a Beta distribution to model confidence scores for each group rather than the Normal distribution used by Kéry and Royle (2020). Confidence scores x_k were modeled such that confidence scores for true positives followed the distribution:

$$x_k|g = 1 \sim \text{Beta}(a_1, b_1)$$

and false positives followed the distribution:

$$x_k|g = 2 \sim \text{Beta}(a_2, b_2)$$

Given that we have counts of vocalizations per sampling period yBN_{ij} and an estimate of true occurrence z_i we can generate a prior for the probability that a given BirdNET detection belongs to the true positive group $g = 1$ such that:

$$\Pr(g_{ik} = 1) = \frac{\lambda_i \times z_i}{\lambda_i \times z_i + \omega}$$

and, because the false-positive group makes up the remainder of detections, the probability is simply:

$$\Pr(g_{ik} = 2) = (\Pr(g_{ik} = 1) - 1)$$

An estimate of the true category (true positive or false positive) of a given observation g_k is then made using these prior probabilities and can also be informed by a small validation set where a human has confirmed the presence or absence of a given species on a single recording segment (in this case 3-s). For those observations without validations, g_k are modeled such that:

$$g_k \sim \text{Categorical}(\Pr(g_{1k}), \Pr(g_{2k}))$$

Detection probability for all three BN models was of the form:

$$\text{logit}(p11_{ij}) = \alpha_0 + \alpha_1 \times h_{ij-1}$$

$$p_{ij} = z_i \cdot p11_{ij} + (1 - z_i) \times p10$$

where overall detection probability p_{ij} was comprised of probability that a detection at site i during period j, $p11_{ij}$, was a true positive, and the probability that a detection was a false positive, $p10$. The $p10$ component's prior was drawn from a uniform distribution that ranged from 0 to 1. To keep the model relatively simple we only modeled the effect of naïve detections h_{ij-1} (0 for absence, 1 for presence) from BirdNET data at site i in the previous sampling period $j - 1$, similar to the human-only model.

Modeling habitat associations. We combined vegetation covariates with the detection probabilities for manually and BirdNET-annotated data to estimate occupancy while accounting for habitat associations. We included a cover of three vegetation types within a 100-m radius of each sampling site as a covariate on occupancy probability: sage scrub cover type, grassland cover type, and oak woodland cover type, because these were the predominant vegetation types at the sampling sites. Vegetation data were derived from a park-wide habitat map (AECOM 2012). Both the occupancy model using manually annotated data and the BirdNET data had vegetation cover linked to occupancy probability, ψ_i, such that:

$$\text{logit}(\psi_i) = \beta_0 + \beta_1 \times \text{sage}_i + \beta_2 \times \text{grass}_i + \beta_3 \times \text{oak}_i$$

Where sage_i, grass_i, and oak_i, each respectively represent percent cover of sage scrub, annual grassland, and oak woodland within a 100-m radius of the sampling location. Cover was log-transformed to reduce skewness. Three sampling sites were missing vegetation data so we set these values to the mean of each observed habitat cover value. We retained these sites because our sample size was limited. All covariates were standardized to have a mean of zero and a standard deviation of 1, to maximize convergence and enable a more straightforward comparison between model coefficients.

The "true" occupancy state of a site, z_i, for each of the models (human and BirdNET annotation data) was modeled as a Bernoulli distribution such that:

$$z_i \sim \text{Bernoulli}(\psi_i)$$

We used uninformative priors for all parameters except as detailed in the preceding sections. We assumed model convergence when monitored parameters had Gelman-Rubin statistics <1.1 (Gelman et al. 2004). Each model was run with 3 MCMC chains, had a burn-in of 5,000 iterations,

and a posterior draw of 100,000 iterations thinned to every fourth sample.

Comparing Model Posterior Distributions

We used the "overlap" function from the "overlapping" package in R (Pastore 2018) to determine the proportion that the human model posterior distribution for a subset of coefficient estimates (β_0, β_1, β_2, β_3) overlapped with the posteriors for each BirdNET model (BN Short, BN Long, BN Long Filtered). The output of the "overlap" function returns a value that indicates the amount of overlap of the two distributions and ranges from 0 which indicates no overlap to 1 which is complete overlap. We tested if the collection of overlap values across all 13 focal species was significantly different for a given coefficient (e.g., β_0) between models using a Kruskal–Wallis test. We followed a significant Kruskal–Wallis test result with a pairwise Wilcoxon rank-sum test to determine which if any models were significantly different from one another.

RESULTS

We detected 49 bird species during manual annotation of recordings collected during 9-min sampling periods at each of 34 sampling sites (Supplementary Material Table 2). BirdNET detected 104 and 39 bird species when we retained only species classifications that had the highest confidence score on an individual 3-s segment and 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 44 (90%) and 32 (65%) of species detected by BirdNET matched the manually annotated species list when BirdNET data were filtered at a low and high confidence threshold, respectively (Supplementary Material Table 2). One-hundred species that are known to occur within the general area (per eBird checklists) were not detected by any method (i.e. manual or BirdNET).

The species detected at the greatest percentage of sites during manual annotation were California Scrub-Jay (88.2%), California Quail (82.4%), Bewick's Wren (73.5%), and Ash-throated Flycatcher (64.7%). The species detected by BirdNET at the greatest percentage of sites when filtered at a high threshold, during the same 9-min periods as manual annotation, were California Quail (47.1%), Bewick's Wren (26.5%), Oak Titmouse (20.6%), and Ash-throated Flycatcher, California Scrub-Jay, and Wrentit (each of these species were detected at 17.6% of sites). 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 = 0.78) versus a low confidence threshold (Pearson's r = 0.66). No Bell's Sparrows were detected during manual annotation of the recording data. Upon further inspection we found BirdNET often misclassified loud cricket stridulations on recordings as Bell's Sparrow, leading to the high number of false detections.

Performance Metrics of BirdNET When Identifying Focal Species

We calculated precision and maximum confidence score among false positive BirdNET detections for 13 bird species detected most frequently during manual annotation (Table 1). Precision was >0.70 for 9 of 13 focal species, had a mean of 0.81 (SD = 0.20), and ranged from a low of 0.29 for Common Raven to a high of 0.99 for California Scrub-Jay. Maximum confidence scores for false-positive detections had a mean of 0.49 (SD = 0.49), and ranged from a low of 0.08 for California Quail and a high of 0.98 for White-breasted Nuthatch. We also calculated recall for each species, using a much longer 9-min period (see Methods) and 9 of 13 species had recall values <0.5, meaning that these species were missed at greater than half of all sites with manual detections. Recall values ranged from a low of 0.09 for Mourning Dove to a high of 0.68 for California Quail.

Reducing False Negatives with Increased Sampling Duration

BirdNET correctly identified the presence of each focal species at a mean proportion of 0.8 (SD = 0.1) of sites with manual detections of a species when a mean 246.1 min (SD = 37.5) of recordings were processed (Figure 1). When only a single 10-min period was processed by BirdNET, the 13 focal species were detected at a mean proportion of 0.27 (SD = 0.13) of sites. The proportion of sites with manually annotated detections of a species at which BirdNET correctly classified a species as present, when 250 min of recording were processed, ranged from a low of 0.55 for Mourning Dove to a high of 0.92 for Bewick's Wren, and no species had BirdNET detections at all sites with manual detections. The species for which BirdNET required the least time to reach its peak proportion of correct positive identifications was California Towhee, at 130 min. Across all focal species and 250 min of recording, a mean of 16.4% (SD = 0.09) of sites with BirdNET detections did not have manual detections at the same sites (i.e., BirdNET found a site to be occupied but there were no detections from manual annotation). However, when manual validations of BirdNET detections (also across 250 min) were considered, a mean of 91.2% (SD = 0.1) of sites with BirdNET detections was verified as occupied (Table 3).

Comparison of Occupancy Model Performance Across Sampling Methods

We compared the parameter estimates of a traditional occupancy model using manual annotation data and another more complex occupancy model that used three subsets

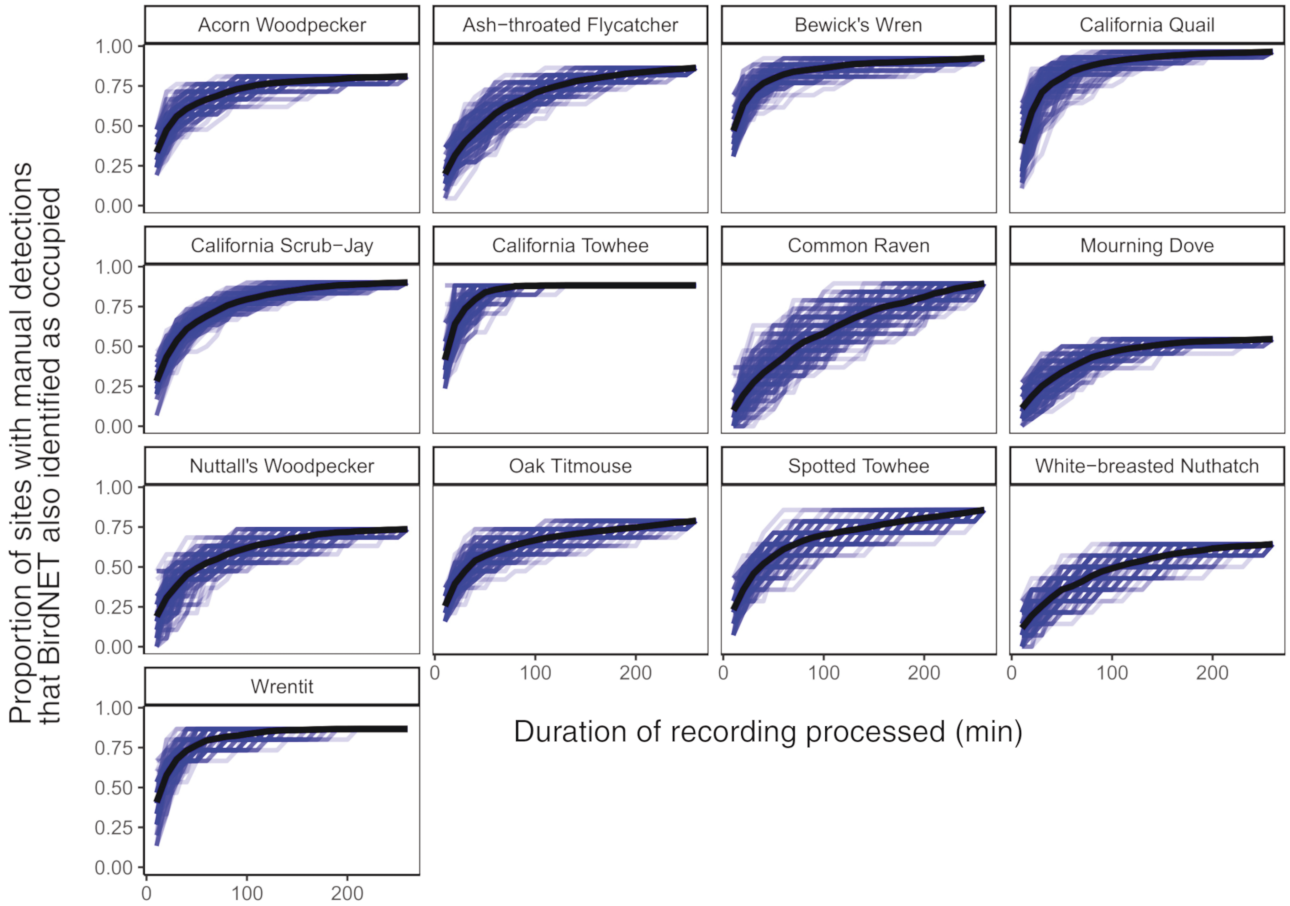

FIGURE 1. Effect of duration of sampling period on the proportion of sites BirdNET was able to correctly identify as occupied by each of 13 focal species. We set each species' threshold for filtering BirdNET data based on the maximum confidence score for false positives in that species (see Table 1). Each blue line represents an individual accumulation curve for a randomly shuffled collection of detections from 26 10-min recording periods (there were 100 randomizations). The thick black line represents the mean proportion across all randomization results.

of BirdNET-derived annotation data, as well as manual validation data, to estimate true and false-positive rates. The occupancy model parameter estimates of the three models using BirdNET-derived data largely mirrored those of the manually derived data (Figure 2). Mean parameter estimates, with the exception of the occupancy intercept (Figure 2), usually had the same sign (positive or negative). The parameter estimates from all models were statistically indistinguishable (Figure 3), with the exception of the BN Long model intercept which had a median overlap with the manually annotated data of 0.43 compared to medians of 0.01 and 0.02 for BN Long Filtered and BN Short. The model using manually-annotated data had the largest intercept across the majority of species, followed by BN Long BirdNET annotation data (i.e., occupancy probability was greatest for these models). All models, with the exception of BN Long, had a smaller estimated proportion of occupied sites than was detected by the human annotator (Figure 4).

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 in turn facilitate spatial and temporal expansion of survey efforts. For instance, when searching recordings for a rare species, BirdNET can quickly annotate a large number of recording hours compared to a human observer. However, automated classifiers may also have drawbacks, such as misclassification of species, or missing species that vocalize infrequently and are often detected visually during surveys (e.g., raptors). We found that these two issues can largely be mitigated for occupancy modeling, at least for our focal species, by using optimal confidence thresholds for the specific application, collecting longer sampling durations, validating a portion of detections, and modeling false-positive rates in the data. Our results demonstrate that the BirdNET annotation output can produce

TABLE 3. Proportion of all sites with BirdNET detections where a species was detected by BirdNET only (i.e., no manual detections at a site during a 9-min period) after analyzing a 260-min recording (Proportion BirdNET only), and the proportion of sites with a BirdNET detection that were verified as occupied (Proportion verified occupied) via validation of 100 randomly selected BirdNET detections.

Common Name	Proportion BirdNET only	Proportion verified occupied
Acorn Woodpecker	0.35	0.96
Ash-throated Flycatcher	0.32	1.00
Bewick's Wren	0.08	1.00
California Quail	0.07	0.90
California Scrub-Jay	0.15	1.00
California Towhee	0.60	1.00
Common Raven	0.53	1.00
Mourning Dove	0.50	0.94
Nuttall's Woodpecker	0.79	0.84
Oak Titmouse	0.33	0.95
Spotted Towhee	0.58	0.95
White-breasted Nuthatch	0.11	1.00
Wrentit	0.15	0.93

occupancy model results comparable to those generated using manual annotation of recordings of 13 relatively common North American bird species.

One drawback of coupling ARUs with an automated classifier is that the estimation of occupancy probability falls far below that of a manual annotation in instances where sampling duration is relatively limited (i.e. BN Short). Whether this lower occupancy estimate is a problem for inference depends on the research question to be answered. If the primary goal is to understand a species' relationship to covariates such as habitat types, then as long as coefficient estimates of those covariates are accurate and precise – and not influenced by habitat structure (e.g., lower detectability within densely vegetated habitats) – the underlying occupancy probability (occupancy model intercept) can be lower than that from a model based on manual annotation with the consequence of imprecise model estimates. Missing the presence of a bird, rather than falsely identifying its presence, might have only limited implications for coefficient estimates if enough points have detections. However, as more true detections are generated by a classifier, coefficient estimates become more accurate

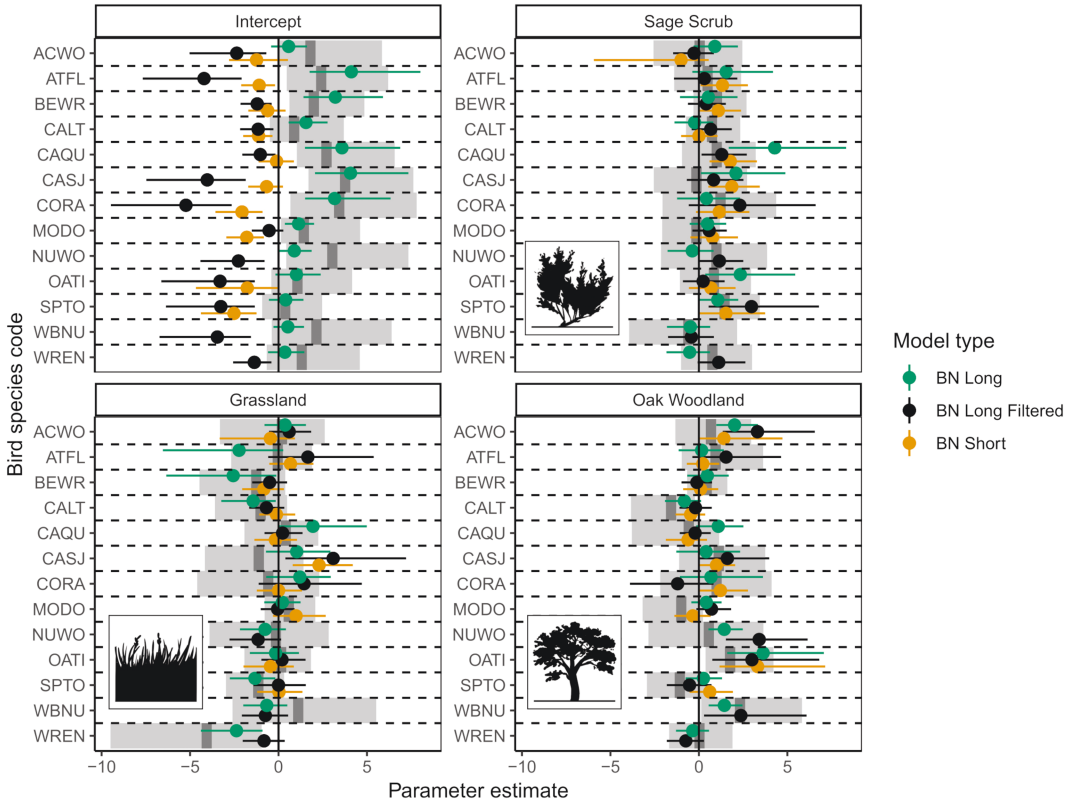

FIGURE 2. Occupancy parameter estimates for habitat associations in occupancy models of 13 focal species using: manual detections from 3 3-min periods (dark gray vertical bar for mean and light gray rectangle for bounds of 95% Bayesian credible intervals) at each site, BirdNET detections from 3 3-min periods (BN Short), BirdNET detections from 29 3-min periods (BN Long), and BirdNET detections filtered to confidence scores greater than a specified threshold from twenty-nine 3-min periods (BN Long Filtered). Mean of parameter estimates are denoted by a point and 95% Bayesian credible intervals by whiskers. Habitat types are listed at the top of each panel. Species codes on the *y*-axis are defined in Table 1. Models that did not converge (NUWO, WBNU, WREN for BN Short) are not presented.

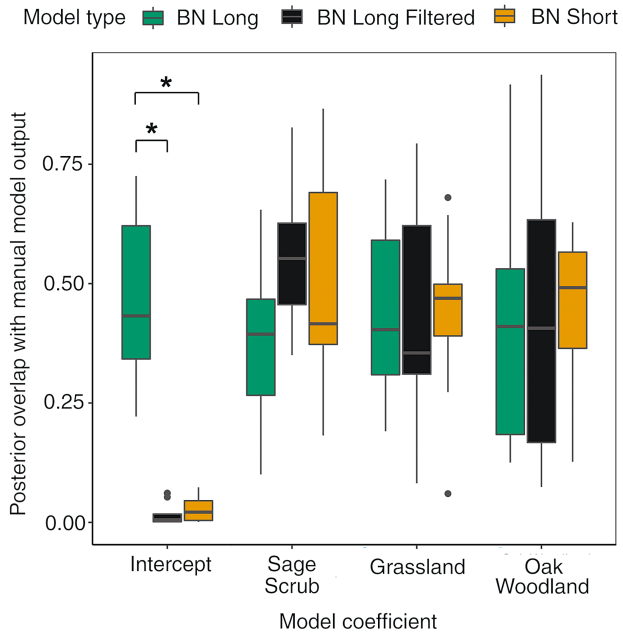

FIGURE 3. Overlap of occupancy parameter estimates for 13 focal species between a model using manually annotated data and 3 models using: BirdNET detections from 3 3-min periods (BN Short), BirdNET detections from 29 3-min periods (BN Long), and BirdNET detections filtered to confidence scores greater than a specified threshold from twenty-nine 3-min periods (BN Long Filtered). Upper and lower bounds of boxplots represent the 1st and 3rd quartiles of overlap estimates and horizontal line in the center of the box represent the median. Dots above and below boxes are considered outliers. Groups with significant differences are denoted with an asterisk.

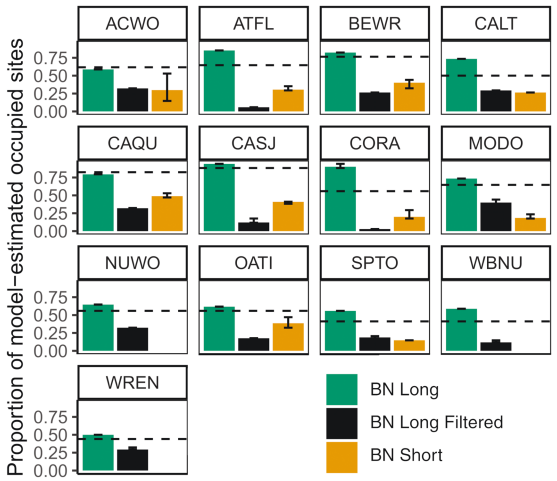

FIGURE 4. Model estimates of the proportion of sites occupied across the study area from occupancy model output using three sets of BirdNET annotation data (colors indicated in legend). Black error bars represent 95% Bayesian credible intervals (BCIs). BCIs are very small for some models (i.e. BN Long Filtered) and may not be apparent on plots. Horizontal dotted line indicates the proportion of sites where each species was manually detected during a 9-min period. Bird species codes that label each panel are listed in Table 1. Models that did not converge (NUWO, WBNU, WREN for BN Short) are not presented.

and precise. Another potential pitfall is placing ARUs close enough together to detect the same bird simultaneously at two points, which may inflate estimates of habitat use. If the study objective is to identify all points with a given species with relative certainty, then much longer duration recordings may be required. Ultimately, simulation is perhaps the best method for determining the minimum sampling required to meet a given objective (see Wood et al. 2021).

BirdNET Classifier Performance

BirdNET performed relatively well at identifying 9 of 13 focal species and had relatively high precision (i.e. precision > 0.7). BirdNET was worst at identifying Ash-throated Flycatcher (*Myiarchus cinerascens*), Common Raven (*Corvus corax*), Mourning Dove (*Zenaida macroura*), and White-breasted Nuthatch (*Sitta carolinensis*), with precisions of 0.63, 0.29, 0.67, and 0.69 respectively. Indeed, precision values were generally high relative to values previously reported (Kahl 2019; Supplementary Material Table 3). For example, Nuttall's Woodpecker (*Picoides nuttallii*) had a high precision value of 0.95 compared with 0.36 reported by Kahl. A notable exception was Common Raven, with a precision value of 0.29 compared with 0.66 reported by Kahl. The majority of Common Raven false positives were misidentifications of heavy traffic noise from a nearby paved road. In the case of Nuttall's Woodpecker, habitat in our study area habitat was more suitable for this species than most other woodpeckers, so even low confidence detections labeled as Nuttall's Woodpecker (such as drumming) were likely to be correct. We caution against generalizing our results to other study areas because of potential regional differences in vocalizations and recording environment. Some characteristics that may lead to high precision values among particular species may be frequent and/or loud vocalizations, occurrence at high densities, or frequent movement within home ranges such that an ARU has a higher likelihood of obtaining a high-quality recording at least once during a long-duration recording. Our metric of precision was similar to those used in other classifier studies which defined true positives as classifier detections that are temporally very near (i.e., within 1 s) 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). While our metric of precision does not ensure instantaneous alignment of manual and classifier annotations, it should be fine-scaled enough to accurately measure the false-positive rate of the classifier.

We hypothesize that the relatively long duration required to detect focal species was driven by birds vocalizing too far from the ARU and a low signal-to-noise ratio, resulting in low BirdNET confidence scores. A human's

ability to discern recorded bird vocalizations is well known to decline with the distance between the bird and recording device (Yip et al. 2017b) and a similar relationship appears to hold true for automated classifiers (Knight and Bayne 2019). In our study, the ability of BirdNET to identify a species presumably declines faster with distance than a human's identification ability. Time to detection would have likely been even greater if we analyzed recordings with strong wind noise, given that wind noise can adversely affect recall by obscuring the audio signal of bird vocalizations (Stowell et al. 2019). If multiple bird species are vocalizing during a single 3-s period then the loudest species will likely be given the highest confidence score, and be ranked 1, with softer calling or more distant species ranked >1. If a louder and more vocal species consistently vocalize and obscure another softer calling species, the likelihood of detection for the latter will be lower, but the species may eventually be detected with sufficiently extended sampling duration. For situations where manually 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. Alternatively, an estimate of anthropogenic noise – one component of the Normalized Difference Soundscape Index defined by Kasten et al. (2012) – could be generated by the "ndsi" function in the R package "soundecology" (Villanueva-Rivera and Pijanowski 2018). This index could be used to exclude recordings above a certain anthropogenic noise threshold, or the noise estimate could be used as a covariate on false and true positive rate, or as a covariate on detection probability (Knight et al. 2021).

Because we were able to model true and false-positive rates, we did not need to filter our data at a confidence threshold, which would likely have delayed the average time to the first detection. When true and false positive rates are modeled, all detections can be included in the model and the model estimates which detections are true positives, rather than removing detections falling below a certain confidence score and then modeling these filtered data. The drawback of modeling all BirdNET detections, versus filtering and then modeling, is that models require much more processing time because of the large number of detections that must be estimated as a true or false positive. Also, although BirdNET validations are not required for occupancy modeling, they do help the model determine the false positive rate with greater precision. In our analysis, using a confidence threshold to exclude likely false positives yielded substantially lower estimates of occupied sites than manual annotation, but the occupancy model ran much more quickly. Barre et al. (2019) proposed filtering data using at least two significantly different thresholds, then running models using those data, and looking for consistency (or lack thereof) in model results – rather than searching for an optimal threshold. Occupancy models that account for false-positive detections may provide accurate results with as little as 1% of detections verified by a human observer (Chambert et al. 2018). In instances where there are very few detections, 50 validated detections for a given species are required for accurate estimation of the false-positive rate (Chambert et al. 2018). In our study, we chose to be conservative and validate 100 detections.

The validation process might become relatively time-consuming when BirdNET is used in species-rich study areas. In our study, approximately 30 min per species was required to validate 100 detections; validating all 104 species detected by BirdNET would thus have taken about 52 person-hours. Rare species (i.e., fewer than 100 detections) for which all BirdNET detections are validated, can be modeled using a more typical occupancy model that excludes estimation of true and false positive rates because all true positive detections are verified. Overall, when using the eBird species filtering process we implemented, or if supplying BirdNET with a list of species known to occur in a region, the required validation should be relatively tractable. The limited effort invested in the validation allows effective modeling of false-positive error, without the extensive effort required to validate days or months of recordings.

Our occupancy model comparisons suggest that analyzing BirdNET data using occupancy models that account for false positives can produce results that are similar to those produced by manual annotation. The occupancy model that generated results most similar to the manual annotation was based on long-duration BirdNET data in which the occupancy intercept was closest to the human model output, meaning sites occupied by a species were correctly identified and habitat characteristics of those sites influenced the model output. Nevertheless, any of the BN models might be improved further by adding covariates on the true- and false-positive rates. For example, external factors (e.g., wind, habitat type, temperature, time of day, diversity of vocalizations) likely influence the ability of BirdNET to correctly identify a species. Wind noise could be evaluated for a given duration of recording by another prebuilt classifier, and a metric of vocal diversity such as the Acoustic Complexity Index (Pieretti et al. 2011) could be tested for its relationship to false-positive rate. Presumably, more acoustically complex communities present more opportunities for BirdNET to incorrectly identify a vocalization.

Another issue related to the analysis of BirdNET data is how to deal with species missed due to distance from the microphone. Researchers interested in understanding the effective sampling area of an ARU for a given bird species

(Turgeon et al. 2017, Yip et al. 2017a) could investigate the relationship between distance to ARU and BirdNET confidence score, ARU model, and habitat type, as well as the frequency and volume of the vocalization. Determining the effective sampling area of an ARU may be particularly useful in deciding the radius within which vegetation characteristics may be relevant, or restricting counts of birds to a given habitat type using the loudness of vocalizations (Hedley et al. 2021). Estimating an effective sampling area can be accomplished with playback experiments to attract birds to an ARU array (Knight and Bayne 2019), or by playing recorded vocalizations at a series of distances from ARUs (Yip et al. 2017a) and modeling the relationship between classifier confidence score and distance. Others have successfully generated estimates of abundance (though not density) by coupling human surveys with ARU annotations through a form of N-mixture model (Doser et al. 2021) and via a statistical offset (Van Wilgenburg et al. 2017). The latter method could also be applied to BirdNET data if detections were partially validated or the data stringently filtered. Density has successfully been estimated with ARUs when the cue rate of a species is known, weather conditions are good, and extensive data on the relationship between distance and power of vocalization has been modeled (Sebastián-González et al. 2018). Though promising, this method may be quite time-consuming for a full bird community.

Conclusion

Our results demonstrate that in a study area with an estimated bird richness of 49 species (based on manual annotation), BirdNET output data for the top 25% most abundant species can produce occupancy modeling results similar to the manual annotation of bird recordings, provided that a sample of observations are validated by a human and sampling duration is substantially increased relative to conventional point counts using human observers. We provide a demonstration of the process required to use BirdNET – from analyzing raw recordings to generating occupancy model output. Other methods of automated annotation such as single species classifiers (whether using a convolutional neural network or not) are compatible with the BN Long model so long as an estimate of classifier confidence (e.g., correlation coefficient for template matching) is generated by the method. We encourage the continued evaluation of the BirdNET classifier in other geographic regions, with a variety of ARU models, habitat types, and bird communities. We are optimistic about the capacity of BirdNET to annotate bird acoustic recordings in diverse contexts, including poorly surveyed regions (e.g., Van Wilgenburg et al. 2020) and provide valuable standardization across studies. The dual developments of affordable ARUs and automated classifiers such as BirdNET provide the opportunity to greatly expand bird research and monitoring efforts and increase the accuracy and precision of occupancy, abundance, and population trend estimates, improving our understanding of population status and dynamics, and our ability to conserve vulnerable species.

SUPPLEMENTARY MATERIAL

Supplementary material is available on *Ornithological Applications* online.

ACKNOWLEDGEMENTS

We extend our appreciation to Tara Kerss for logistical and study design assistance, the staff of Carnegie SVRA for facilitating access to sampling locations, and Morgan Tingley for providing helpful suggestions that improved this manuscript. We thank the three anonymous reviewers for their valuable input that also improved the quality of our manuscript. Lauren Helton provided the illustrations contained in the figures and graphical abstract. This is Contribution Number 725 of The Institute for Bird Populations.

Funding statement: This research was funded by the California Department of Parks and Recreation, Off-Highway Motor Vehicle Recreation Division under agreement numbers C15V0023 and C19V0014. The funder did not influence the content of the submitted or published manuscript. The funder did not require approval of the final manuscript to be published.

Ethics statement: We followed the recommendations in Guidelines to the Use of Wild Birds in Research.

Author contributions: J.S.C., N.L.M., S.A.E., and R.B.S. formulated the questions; J.S.C. collected data and performed analyses; N.L.M., S.A.E., and R.B.S. supervised the research; N.L.M. and S.A.E. provided funding; J.S.C., N.L.M., and R.B.S. wrote the paper, which was reviewed by all authors.

Data depository: Analyses reported in this article can be reproduced using the R code and data provided by Cole et al. (2022). For access to the full recording files that were annotated for this manuscript please contact the authors. The files are too large to make publicly available.

LITERATURE CITED

AECOM (2012). Vegetation Classification and Mapping Report. California Departments of Parks and Recreation, Sacramento, CA, USA.

Arif, M., R. Hedley, and E. Bayne (2020). Testing the Accuracy of a BirdNET, Automatic Bird Song Classifier. https://doi.org/10.7939/r3-6khb-kz18

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.

Barré, K., I. Le Viol, R. Julliard, J. Pauwels, S. E. Newson, J. Julien, F. Claireau, C. Kerbiriou, and Y. Bas (2019). Accounting for automated identification errors in acoustic surveys. Methods in Ecology and Evolution 10:1171–1188.

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 3and optimization of recordings post-processing. Methods in Ecology and Evolution 9:560–570.

Cole, J. S., N. L. Michel, S. A. Emerson, and R. B. Siegel (2022). Data from: Automated bird sound classifications of long-duration recordings produce occupancy model outputs similar to manually annotated data. Ornithological Applications 124:duac003. https://doi.org/10.5061/dryad.x95x69pkr

Doser, J. W., A. O. Finley, A. S. Weed, and E. F. Zipkin (2021). Integrating automated acoustic vocalization data and point count surveys for estimation of bird abundance. Methods in Ecology and Evolution 12:1040–1049.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2004). Bayesian Data Analysis, 2nd edition. CRC/Chapman & Hall, Boca Raton, FL, USA.

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.

Hedley, R. W., S. J. Wilson, D. A. Yip, K. Li, and E. M. Bayne (2021). Distance truncation via sound level for bioacoustic surveys in patchy habitat. Bioacoustics 30:303–323.

Kahl, S. (2019). Identifying Birds by Sound: Large-scale Acoustic Event Recognition for Avian Activity Monitoring. Universitätsverlag Chemnitz, Bonn, Germany.

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

Kasten, E. P., S. H. Gage, J. Fox, and W. Joo (2012). The remote environmental assessment laboratory's acoustic library: An archive for studying soundscape ecology. Ecological Informatics 12:50–67.

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.

Kéry, M., and J. A. Royle (2020). Applied Hierarchical Modeling in Ecology: Analysis of Distribution, Abundance, and Species Richness in R and BUGS. Volume 2: Dynamic and Advanced Models. Academic Press, New York, NY, USA.

Kéry, M., J. A. Royle, H. Schmid, M. Schaub, B. Volet, G. Haefliger, and N. Zbinden (2010). Site-occupancy distribution modeling to correct population-trend estimates derived from opportunistic observations. Conservation Biology 24:1388–1397.

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., R. M. Brigham, and E. M. Bayne (2021). Specialist or generalist? It depends. Context-dependent habitat relationships provide insight into forest disturbance effects for a boreal bird species. Forest Ecology and Management 502:119720.

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.

MacKenzie, D. I., J. D. Nichols, J. A. Royle, K. H. Pollock, L. L. Bailey, and J. E. Hines (2006). Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence. Academic Press, San Diego, CA, USA.

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.

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:e65808.

Pastore, M. (2018). Overlapping: A R package for estimating overlapping in empirical distributions. Journal of Open Source Software 3:1023.

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

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

Saunders, S. P., K. A. Hall, N. Hill, and N. L. Michel (2019). Multiscale effects of wetland availability and matrix composition on wetland breeding birds in Minnesota, USA. The Condor: Ornithological Applications 121:duz024.

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 13:7.

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.

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

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.

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

Tingley, M. W., and S. R. Beissinger (2013). Cryptic loss of montane avian richness and high community turnover over 100 years. Ecology 94:598–609.

Turgeon, P. J., S. L. Van Wilgenburg, and K. L. Drake (2017). Microphone variability and degradation: implications for monitoring programs employing autonomous recording units. Avian Conservation and Ecology 12:9.

Villanueva-Rivera, L. J., and B. C. Pijanowski (2018). soundecology: Soundscape Ecology – R package. http://ljvillanueva.github.io/soundecology/

van Rossum, G., and F. L. Drake (2009). Python 3 Reference Manual. CreateSpace, Scotts Valley, CA, USA.

Van Wilgenburg, S. L., P. Sólymos, K. J. Kardynal, and M. D. Frey (2017). Paired sampling standardizes point count data from humans and acoustic recorders. Avian Conservation and Ecology 12:13.

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, 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.

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

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

Yip, D. A., L. Leston, E. M. Bayne, P. Sólymos, and A. Grover (2017b). Experimentally derived detection distances from audio recordings and human observers enable integrated analysis of point count data. Avian Conservation and Ecology 12:11.