Received: 18 May 2023 | Accepted: 13 November 2023

DOI: 10.1111/2041-210X.14272

Methods in Ecology and Evolution

RESEARCH ARTICLE

Nighthawk: Acoustic monitoring of nocturnal bird migration in the Americas

Benjamin M. Van Doren[1,2] | Andrew Farnsworth[1,3] | Kate Stone[4] | Dylan M. Osterhaus[5] | Jacob Drucker[6,7] | Grant Van Horn[1,8]

[1]Cornell Lab of Ornithology, Cornell University, Ithaca, New York, USA; [2]Department of Natural Resources and Environmental Sciences, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA; [3]Actions@EBMF, New York, New York, USA; [4]MPG Ranch, Missoula, Montana, USA; [5]Department of Biology, New Mexico State University, Las Cruces, New Mexico, USA; [6]Committee on Evolutionary Biology, University of Chicago, Chicago, Illinois, USA; [7]Negaunee Integrative Research Center, Field Museum of Natural History, Chicago, Illinois, USA and [8]Manning College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, Massachusetts, USA

Correspondence
Benjamin M. Van Doren
Email: vandoren@illinois.edu

Funding information
Cornell Lab of Ornithology; Cornell Presidential Postdoctoral Fellowship; University of Chicago Steiner Fund; Actions@EBMF; Field Museum of Natural History H. B. Conover Fund

Handling Editor: Camille Desjonquères

Abstract

1. Animal migration is one of nature's most spectacular phenomena, but migratory animals and their journeys are imperilled across the globe. Migratory birds are among the most well-studied animals on Earth, yet relatively little is known about in-flight behaviour during nocturnal migration. Because many migrating bird species vocalize during flight, passive acoustic monitoring shows great promise for facilitating widespread monitoring of bird migration.
2. Here, we present Nighthawk, a deep learning model designed to detect and identify the vocalizations of nocturnally migrating birds. We trained Nighthawk on the in-flight vocalizations of migratory birds using a diverse dataset of recordings from across the Americas.
3. Our results demonstrate that Nighthawk performs well as a nocturnal flight call detector and classifier for dozens of avian taxa, both at the species level and for broader taxonomic groups (e.g. orders and families). It achieves an average precision score above 0.80 for 50 species and a mean average precision of 0.96 across 4 orders. The model accurately quantified nightly nocturnal migration intensity (80% variation explained) and species phenology (78% variation explained) and performed well on data from across North America. Incorporating modest amounts of additional annotated audio (50–120 h) into model training yielded high performance on target datasets from both North and South America (average precision on order Passeriformes >0.99).
4. By monitoring the vocalizations of actively migrating birds, Nighthawk provides a detailed window onto nocturnal bird migration that is not presently attainable by other means (e.g. radar or citizen science). Scientists, managers and practitioners could use acoustic monitoring with Nighthawk for a number of applications, including: monitoring migration passage at wind farms; studying airspace usage

during migratory flights; monitoring the changing migrations of species susceptible to climate change; and revealing previously unknown migration routes and behaviours. Overall, this work will empower diverse stakeholders to efficiently monitor migrating birds across the Western Hemisphere and collect data in aid of science and conservation.

KEYWORDS
acoustic monitoring, artificial intelligence, bioacoustics, bird migration, machine learning, machine listening, movement ecology

1 | INTRODUCTION

Seasonal migration is fundamental to the life histories of countless organisms. Migrating animals have captured human imagination for millennia, and movement is a key mechanism by which organisms can adjust to rapid environmental change (Van Doren et al., 2021). Movement is a fundamental mediator of organisms' interactions with their environment and with each other. As anthropogenic pressures increasingly influence these interactions, knowledge of movement is increasingly necessary for guiding conservation action (Davy et al., 2017; Fraser et al., 2018). This is particularly true for animals that use the atmosphere—a key habitat where volant organisms interact, forage and even rest (Diehl, 2013; Liechti et al., 2013). For species constantly on the move, their life histories and conservation threats are inextricably tied to ever-changing spatiotemporal distributions. Despite the need for movement data to inform science and conservation, migratory animals are often challenging to monitor. They may be rare, secretive, sensitive to disturbance, too small to carry a tracking device or too expensive to monitor at sufficient numbers or resolution (Kays et al., 2015).

Migratory birds, for instance, are among the world's most well-studied moving organisms, but scientists still lack important information about the movements of most species. There is great urgency to develop better methods for monitoring migratory birds, as populations of these global travellers have declined precipitously over the last half century (Rosenberg et al., 2019). Available tools for monitoring migratory birds include Doppler radar networks (Bauer et al., 2019; Dokter et al., 2011; Gauthreaux et al., 2003) and widespread citizen science projects such as eBird (Sullivan et al., 2014). These tools are invaluable: radar can measure the fluxes of moving birds even at continental extents (Nussbaumer et al., 2021; Van Doren & Horton, 2018), and data from citizen scientists can support detailed spatial models used by scientists and practitioners globally (Fink et al., 2020; Reynolds et al., 2017). However, these tools also carry major shortcomings. Doppler weather radars cannot discern individuals or species identities, and they are generally stationary, expensive and easily limited by mountainous topography. Scientific use of radar data is also frequently impeded by bureaucratic, political or technical issues (Shamoun-Baranes et al., 2022), and radar is still a developing tool outside of North America and the Western Palearctic. Citizen science databases such as eBird (Sullivan et al., 2014) can provide information at finer taxonomic scales, but survey coverage can be highly variable away from human population centres. In addition, most bird species migrate at night, so crowd-sourced human observations have limited ability to actively monitor nocturnal avian movements. Fortunately, radar, citizen science and other data sources are highly complementary, and a growing number of studies integrate these and other multimodal data to infer unseen behaviour (Bota et al., 2020; Shipley et al., 2018; Van Doren, Lostanlen, et al., 2023). However, no current approaches provide a scalable and affordable path towards monitoring the in-flight behaviour of billions of nocturnally migrating birds at species and individual resolution.

Acoustic methods are frequently used by bat researchers to monitor species presence and activity and increasingly by ornithologists to detect breeding species (Sugai et al., 2019), but their potential to study animals on the move is often overlooked. Many bird species actively vocalize during nocturnal flights, and ground-based microphones can capture the ‘flight calls’ of migrating individuals (Farnsworth, 2005). Because flight calls often encode species identity, acoustic monitoring provides an accessible path towards portable, inexpensive and widely deployable systems for species- and individual-level monitoring (Evans & Rosenberg, 2000). Recent work has demonstrated the potential of acoustic monitoring to document the nightly passage rates of migratory birds (Van Doren, Lostanlen, et al., 2023), and acoustic information could enable the study of intra- and inter-species interactions during migration (Gayk & Mennill, 2023). The primary impediment to the widespread adoption of acoustic monitoring of bird migration is the large time investment needed to transform hours of recorded audio into counts of identified vocalizations. In recent years, however, the outlook has greatly improved due to advances in machine learning technology for sound detection and classification (Kahl et al., 2021; Lostanlen et al., 2018; Van Doren, Lostanlen, et al., 2023). The current state-of-the-art is the BirdVox model, which was trained to classify 14 species of frequently heard migrant species (Lostanlen et al., 2018). However, important challenges remain that prevent the broad application of acoustic systems for avian migration monitoring. Existing systems are rarely trained on large numbers of nocturnal flight calls and are frequently based on datasets that are restricted in taxonomic or geographic scope, and therefore these systems may not generalize well to other areas.

Merlin is a smartphone application that uses machine learning to identify bird vocalizations (https://merlin.allaboutbirds.org/). The sound identification module of Merlin is powered by a convolutional neural network that applies computer vision principles to audio spectrograms. Merlin Sound ID is a high-performing system with great promise for a range of acoustic monitoring applications, but the current system is not available or downloadable for research use, and it is not ideal for migration monitoring applications for several reasons. Merlin outputs species detections on 3-s chunks of audio, but flight calls are much shorter than this duration (often lasting only 50–250 ms; Evans & O'Brien, 2002; Lanzone et al., 2009). Temporal resolution is therefore limited to 3 s; the use of smaller chunk sizes could increase temporal resolution. Furthermore, many flight calls cannot be easily identified by species and Merlin does not capture this taxonomic uncertainty (but see Cramer et al., 2020). However, because neural networks are highly configurable, adjustments to model architecture and training data could address these challenges.

Here, we present *Nighthawk*, a deep learning model based on Merlin Sound ID that is designed to detect and identify the vocalizations of nocturnally migrating birds. We trained Nighthawk on in-flight vocalizations from a diverse collection of recordings collected across the Americas. In this research paper, we examine the performance of the Nighthawk model across spatial and taxonomic scales. We evaluate model performance across a taxonomic hierarchy of 82 species, 18 families and 4 orders of birds that vocalize during nocturnal migration. We apply Nighthawk to >6000 h of continuous audio recordings collected across an entire migration season and investigate how well the model generalizes across spatial scales. Finally, we discuss best practices for using Nighthawk to monitor bird migration in the Americas.

2 | METHODS

2.1 | Overview

Our analysis comprised three parts (Figure 1). In Part 1, we assembled a dataset of annotated recordings and trained Nighthawk to predict the presence of flight calls from different avian taxa in an audio segment. We refer to this dataset as the 'Core dataset'. In Part 2, we trained additional versions of Nighthawk and applied them to data from locations not included in training (hereafter, 'target datasets'). We tested target recordings from regions well represented in the Core dataset ('in-domain' data) and regions poorly represented ('out-of-domain' data). In Part 3, we applied Nighthawk to continuous recordings collected across an entire migration season—the intended use case of Nighthawk.

2.2 | Acoustic data

We obtained annotated recordings from 25 data sources, comprising published and unpublished datasets from across the Americas. Geographically, most data came from eastern North America, with some from western North America and a small sample from northern South America. A summary of data sources is in Table S1. These data primarily comprised passive nocturnal audio recordings (~72%); the remainder were general sound datasets (~18%), recordings of captive birds (~2%) or not easily placed in one category (~8%). These datasets were generated using varied approaches, and we anticipated variation in their error rates. We checked representative samples of each dataset to verify error rates were less than 5%, and we corrected errors where encountered.

2.3 | Data annotation

To assign recorded vocalizations to specific avian taxa, we visualized audio as spectrograms and annotated the onset and offset of each vocalization event. When a bird made repeated vocalizations, we considered sounds occurring more than 0.5 s apart as separate events. We used multiple software platforms that support temporal annotation, including Audacity, Raven Pro and a custom web interface (https://merlinvision.macaulaylibrary.org).

Identification uncertainty is a fundamental part of acoustic monitoring; many flight calls can be classified by species, but others cannot (e.g. due to recording quality or overlap with similar-sounding species; Evans & O'Brien, 2002; Landsborough et al., 2019). We addressed this challenge by explicitly modelling taxonomic uncertainty. Annotators labelled flight calls at the finest taxonomic resolution possible. We used species, families and orders as defined in the Clements Checklist, the avian taxonomy used by the eBird database (Sullivan et al., 2014). We recorded species-level classifications using the abbreviated codes defined in this taxonomy (e.g. 'amered' is the code for American Redstart *Setophaga ruticilla*). For species that have broadly similar flight calls, we created an additional 'group' classification level. We defined a group as two or more similar-sounding species in the same family. Each species can belong to at least one group. For example, Swamp Sparrow (*Melospiza georgiana*) and Lincoln's Sparrow (*M. lincolnii*) give highly similar flight calls, and we treat them together here as the 'SWLI' group. Thus, we used a four-level taxonomic hierarchy: species, group, family and order. Table S2 lists all defined groups and their member species. Table S3 gives the proportion of annotations provided at family and species levels for each dataset.

2.4 | Taxa included

We focused on the following taxa: (1) nocturnally migrating landbirds in the orders Passeriformes and Cuculiformes; (2) nocturnally migrating shorebirds in the order Charadriiformes; and (3) nocturnally migrating waterbirds in the order Pelecaniformes. We included taxa known to migrate at night and vocalize during flight (Evans & O'Brien, 2002). For example, we did not include the Passeriformes families of Fringillidae and Hirundinidae because

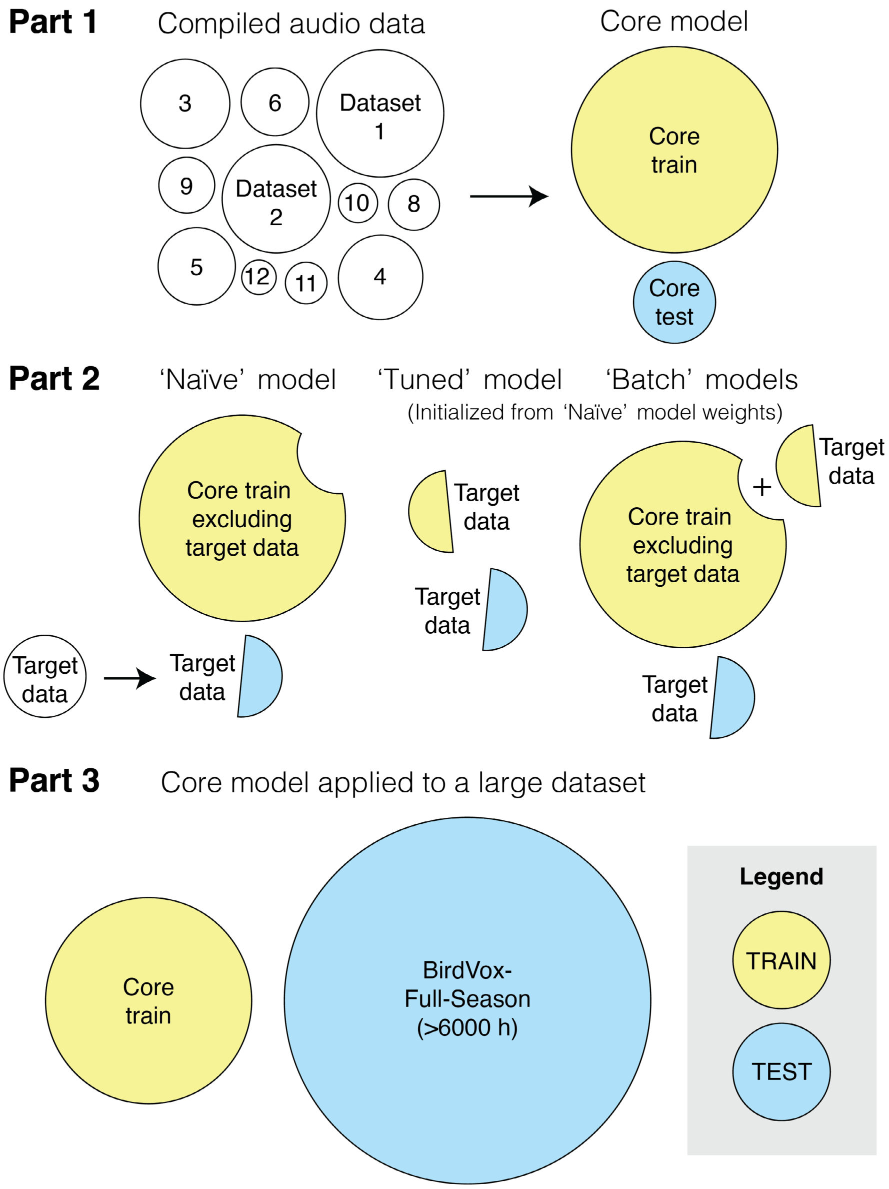

FIGURE 1 Analysis overview. In Part 1, we compiled annotated acoustic datasets and split acoustic data into nonoverlapping 'Core' test and train sets, from which we trained the Core model. In Part 2, we applied Nighthawk to three target datasets (from Pennsylvania, New Mexico and Colombia) using three different model construction strategies (Naïve, Tuned and Batch). For Tuned and Batch model types, we initialized model weights to those of the Naïve model, which was trained on data excluding those from the target dataset. In Part 3, we applied the Core model to >6000 h of continuous audio from a migration monitoring array in central New York State, USA.

these taxa primarily migrate diurnally. We included taxa for which we were able to compile at least 50 training examples and 20 testing examples, totalling 82 species, 17 groups, 18 families and 4 orders.

2.5 | Part 1: Assembling the Core dataset

After compiling annotated recordings, we focused on building a balanced and accurate test dataset. We used three steps: first, we constructed a Core test split that minimized spatiotemporal overlap with the Core train split; second, we subsampled test data to obtain a taxonomically balanced and representative sample; and third, we manually reviewed all annotations for accuracy.

To ensure separation between Core test and train splits, we randomly assigned audio to train or test splits based on recording location wherever possible, such that each recording site was included in *either* train or test sets but not both, thus preventing the model from using idiosyncratic information about recording location or microphone setup (i.e. data leakage). We were unable to subdivide 33% of the data solely by location, either because data came from a single recording location (e.g. captive recordings from a single research station) or because location was not consistently reported. For these remaining data, we randomly assigned each audio file to either the train or test split based on the recording session, such that no data from the same session was included in both train and test splits.

Next, we prepared test data for manual review by subsampling the data. For each taxonomic class, we randomly selected an annotated call from each of our source datasets in sequence until we either reached 500 calls or ran out of annotations for that class. We then randomly sampled an additional 10,000 negative examples (e.g. ambient noise or vocalizations that were not flight calls) following the same procedure. Our dataset includes many audio files containing more than one vocalization, and we annotated all flight call vocalizations of non-target taxa, so in practice, our test set includes more than 500 examples from frequently recorded taxa. See Table S4.

Finally, two authors (BMVD and AF) reviewed all Core test set annotations to correct any errors. If the two reviewers disagreed about the identity of a vocalization, we used a more general taxonomic categorization for that vocalization. For example, if there was agreement that a call was made by a thrush (family Turdidae) but not on its species identity, we used 'Turdidae' as the classification. After review, the Core train and test sets consisted of 428,009 and 47,305 annotations, respectively (Table S4).

It was most important for us to closely review the test set because the conclusions of this study rely on an accurate assessment of model performance (Northcutt et al., 2021; Van Horn et al., 2015). Although the compiled training dataset was too large to manually review every annotation, we also checked representative samples of training data to verify error rates were less than 5%, and we corrected errors where encountered.

For Core model training and evaluation, we extracted all audio segments containing flight calls as well as an approximately equal number of audio segments containing no flight calls. We refer to this approach as a 'Balanced Cropped Vocalization' analysis.

2.6 | Part 1: Building the Core Nighthawk model

To build the Core *Nighthawk* model, we adapted the existing *Merlin Sound ID* deep learning framework. Merlin is trained to identify bird vocalizations but is not optimized for the short-duration calls prevalent in recordings of nocturnal bird migration. Merlin uses a deep neural network to predict the bird species present in short audio segments. Merlin takes as input a 3-s audio clip and outputs a vector that represents learned features relevant to the classification task at hand. These features are then used as inputs to a logistic regression model to predict the presence of each species. The neural network backbone of Merlin is a MobileNet (Howard et al., 2019), a compact network designed to perform well on mobile devices.

We modified the Merlin Sound ID architecture in three important ways (Figure 2). First, we decreased the input duration from 3 to 1 s because flight calls have shorter vocalizations (generally 50–250 ms) than those typically evaluated by Merlin (Evans & O'Brien, 2002; Farnsworth, 2005). Second, because we were not constrained to smartphone deployments, we replaced the MobileNet backbone with a ResNet-34 backbone (He et al., 2016). ResNet-34 is a larger architecture that can achieve higher performance than MobileNet but requires more computational resources. Third, to facilitate classification at our four taxonomic levels (species, group, family and order), we replaced Merlin's single, species-level output layer with four output layers, each of which received a shared feature vector as input (Figure 2). With this multi-head output, Nighthawk can simultaneously make predictions of the species, group, family and order classes of every input.

Altogether, Nighthawk takes as input a 1-s audio sample at a sample rate of 22,050 Hz. It then generates a mel spectrogram with a frequency range of 100–11,025 Hz using a hop length of 128 samples, an FFT length of 512 samples and 128 mel bands. This is rendered as a single-channel 171 × 128 pixel grayscale image, which is rescaled to decibel scale (max of 80 dB), normalized and rescaled to a maximum value of 255. This serves as the input to the ResNet-34 backbone, which ends in a global max pooling operation and outputs a 512-dimension feature vector. The feature vector is separately passed directly to each of four fully connected layers with linear activations, outputting logit values for each taxonomic class. We regularize the network with L2 regularization throughout and a dropout layer after global max pooling with a dropout rate of 0.2. During training, we apply several augmentation techniques, including mixup within each batch, randomly masking 5% of frequency values, randomly shifting the time axis by up to 90%, randomly shifting frequency values by up to 5% and randomly applying one of the following filter operations with a probability of 0.3: high pass, low pass, band pass or band stop. These augmentation parameters were determined from initial experiments. Model training and evaluation were implemented in Python using the Tensorflow framework.

We trained Nighthawk on the Core training set of 428,009 annotations. During training, any examples with an unknown classification at one or more taxonomic levels were masked so that they did not incur loss at those taxonomic levels. In this way, the model could incorporate information from partially labelled examples (e.g. when order and family are known but not species). We did not implement separate noise classes in the model. Instead, all non-flight call sounds were pooled and treated as negative data (i.e. loss incurred across

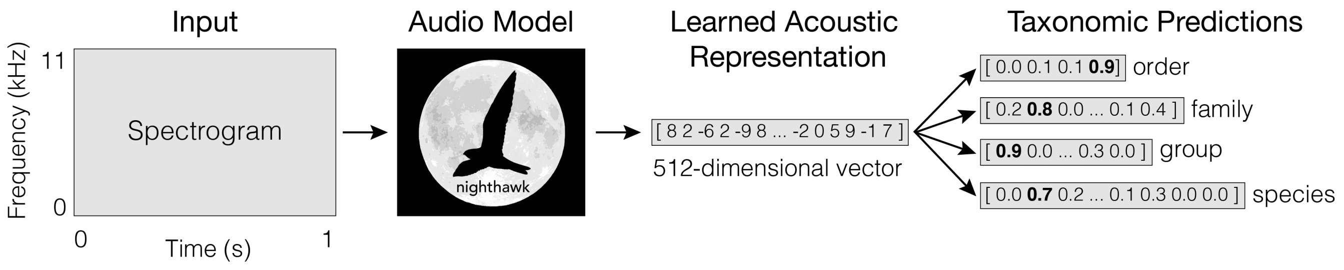
Input
Frequency (kHz)
11
0
Spectrogram
0
Time (s)
1
Audio Model
nighthawk
Learned Acoustic Representation
[8 2 -6 2 -9 8 ... -2 0 5 9 -1 7]
512-dimensional vector
Taxonomic Predictions
[0.0 0.1 0.1 **0.9**] order
[0.2 **0.8** 0.0 ... 0.1 0.4] family
[**0.9** 0.0 ... 0.3 0.0] group
[0.0 **0.7** 0.2 ... 0.1 0.3 0.0 0.0] species

FIGURE 2 Nighthawk takes as input a 1-s spectrogram and outputs a high-dimensional feature vector that is a learned representation of the input. The final layers of the model apply logistic regression to this feature vector to make predictions across multiple taxonomic levels.

all classes); these sounds included environmental and mechanical sounds, non-bird animals and bird vocalizations that were not flight calls. Hyperparameter sweeps revealed consistent performance after training for 100 epochs (batch size = 64) using stochastic gradient descent optimization with cosine learning rate decay, starting at a learning rate of 0.1.

2.7 | Part 1: Evaluating the Core model

We evaluated Core model performance on the Core test dataset of Balanced Cropped Vocalizations with *average precision* (AP) scores. Average precision is a metric favoured by the machine learning community because AP captures overall performance without setting an arbitrary score threshold. AP summarizes the precision-recall curve for a single class and takes a value between 0 and 1, where 1 represents a system with perfect precision and perfect recall for that class. Precision refers to the proportion of classifications that are correct, and recall refers to the proportion of true calls that are correctly classified. The *mean average precision* (mAP) is the mean of AP scores across multiple classes and summarizes overall model performance. Here, we evaluate models by calculating mAP scores separately for species, group, family and order levels.

2.8 | Part 2.1: Constructing target datasets

After evaluating the Core model (Part 1), we investigated Nighthawk's performance on data from locations not included in its training dataset ('target datasets'). We constructed and evaluated models with three target datasets, respectively, from (1) Pennsylvania, USA; (2) New Mexico, USA; and (3) Colombia. In each case, we split target datasets into train and test halves by recording sessions.

2.9 | Part 2.2: Evaluating target models

We evaluated all target models on the holdout (test) portion of the target data. We conducted two types of evaluations. In the first, we constructed Balanced Cropped Vocalization datasets for model training, in which we extracted all target audio segments containing flight calls and roughly 2× as many audio segments without flight calls. We evaluated these data using average precision (AP), as for the Core evaluation.

In field applications, we expect to deal with very unbalanced datasets, in which there are many more sections of audio without flight calls than those with flight calls. Therefore, in a second evaluation, we ran Nighthawk on the full-length, continuous recordings comprising the target dataset. All target audio files were fully annotated by an author or collaborator. For this continuous listening application, we evaluated 1-s windows incremented by 0.2 s and retained predictions that exceeded a score threshold as 'detections'. We set score thresholds using the precision-recall curve from the corresponding Balanced Cropped Vocalizations evaluation, selecting a threshold that achieved high precision (0.99) on the target data.

We postprocessed detections from continuous listening in three steps: first, we enforced taxonomic consistency by only retaining detections that are consistent across species, group (if applicable), family and order at the given score threshold. Second, when using 1-s windows incremented by 0.2 s, we expect each call to trigger multiple model 'hits' since each audio sample is included in multiple overlapping windows. Thus, in the continuous listening application, we dropped single, uncorroborated detections, as these were usually false positives. Third, we simplified output by merging overlapping detections up to a maximum duration of 5 s per detection.

To quantify performance for continuous listening, we calculated precision and recall across all files using existing annotations. In addition, we manually reviewed predictions in case some detected calls had been missed by the original annotator. Here, we focused on evaluations of the Passeriformes class because it broadly captures the model's ability to separate flight calls from background noise.

2.10 | Part 2.3: Well-represented target dataset

The first target dataset came from central Pennsylvania, USA (PA) (dataset 18 in Table S1, hereafter 'PA dataset'). These recordings were made in the northeast US (a well-represented region in our training dataset), but >200 km from our other large datasets (e.g. datasets 5, 6, 7, 8, 9 and 10). The PA dataset consists of 119 h of annotated audio, from which we extracted 2069 annotated calls and randomly sampled 5064 audio clips without flight calls. We used 3532 of these annotations for additional model training and set aside 3600 for evaluation, as described below.

We then constructed target models using three different strategies (Part 2 in Figure 1). First, we trained a new model on non-PA Core training data using the same hyperparameters as the Core model. We refer to this as the 'Naïve' model. It is similar to the Core model, but its training data includes no PA data. We then experimented with two approaches to fine-tuning a target model to perform well on PA data. In approach 1, we initialized model weights to those of the Naïve model and trained further on the PA data not used for evaluation (3532 annotations) (hereafter the 'Tuned' model). In approach 2, we trained two Nighthawk models, in which we again initialized model weights to those of the Naïve model, then took the data used to train the Naïve model and augmented it with PA data (3532 annotations). To emphasize performance on PA data, we employed a custom batch construction strategy where we filled half of each batch with PA data. For the first model (hereafter the 'Batch–More' model), we filled half of each batch from the PA data used in the Tuned model; in the second (hereafter the 'Batch–Less' model), we filled half of each batch from a smaller pool of 1000 examples of PA data.

For Tuned, Batch–More and Batch–Less models, we started training by initializing model weights to those of the Naïve model. For the Tuned model, we performed 30 epochs of training using stochastic gradient descent optimization with cosine learning rate decay starting at a learning rate of 0.001. For Batch–More and Batch–Less, we used the same procedure with one adjustment: we trained for 15 epochs because each epoch consisted of approximately twice as many examples due to the batch construction strategy. Based on our results, we also introduced a version of the Batch–More and Batch–Less models trained for only 1 epoch instead of 15 because our results suggested that 1 epoch might be sufficient.

2.11 | Part 2.4: Underrepresented target dataset

We next used a target dataset from a region poorly represented in the Core data: White Sands Missile Range, New Mexico, USA (dataset 17 in Table S1, hereafter 'NM dataset'). Recording locations at White Sands Missile Range are >1000 km from the recording locations of our other large audio datasets. The NM dataset consists of 45 h of audio annotated primarily at the order level, from which we extracted 2174 annotated calls and randomly generated 5061 clips of audio without flight calls. We used 3412 of these annotations for additional fine-tuning and set aside 3823 for evaluation.

To construct the NM dataset, we followed the same procedure as for PA. The NM dataset is annotated primarily at the order level (virtually all Passeriformes), so for this analysis, we only evaluated performance for the Passeriformes class.

2.12 | Part 2.5: Out-of-domain target dataset

Nearly all our Core training data come from North America, but many of the migratory species we record in North America have migration routes that traverse Central and South America. Therefore, Nighthawk could be useful in these out-of-domain areas, although it would likely encounter soundscapes that are very different from those in the Core dataset (e.g. different ambient noises and bird species).

We used a target dataset from six recording stations in Colombia, located in or near the cities of Bogotá, Barrancabermeja and San José del Guaviare (hereafter 'CO dataset'). Some data from these locations are included in datasets 12 and 19 (Table S1), so we excluded these datasets from training and trained a new model on the remaining data ('Naïve' model). The CO dataset consists of 51 h of annotated audio, annotated primarily at the order level, from which we extracted 5650 annotated calls and randomly generated 10,132 clips of audio without flight calls. We used 9990 of these annotations for additional fine-tuning and set aside 5791 for evaluation.

To construct the CO dataset, we followed the same procedure as for PA and NM.

2.13 | Part 3: Evaluating continuous listening for migration monitoring

Lastly, we applied Nighthawk to a large dataset of continuous recordings from an entire autumn migration season; this scenario represents our envisioned use case of Nighthawk for continuous nocturnal listening. This dataset comprises >6000 h of audio from a regional microphone array, recorded in autumn 2015 in Tompkins County, New York, USA (the BirdVox-full-season dataset; Farnsworth et al., 2022). Previous work used this dataset to relate acoustic measures of bird migration to independent radar and citizen science measures (Van Doren, Lostanlen, et al., 2023). Their model, called BirdVoxDetect (Lostanlen & Salamon, 2022; Salamon et al., 2016; Van Doren, Lostanlen, et al., 2023), was trained to detect and classify the flight calls of 14 bird species. As a comparison, we ran the Core Nighthawk model on the BirdVox-full-season dataset. Following Van Doren, Lostanlen, et al. (2023), we trimmed our analysis to only include nocturnal periods (between civil twilight dusk and dawn). We calibrated model scores by fitting class-specific logistic regression models using the Core test dataset, effectively treating the Core test dataset as a validation split for this analysis. Calibrating model scores makes the resulting scores more comparable across classes, as long as sufficient validation data are used for calibration and they are representative of the test data. If scores are not calibrated, the same score may have very different performance properties in different classes. After calibration, we extracted outputs for the >6000 h of audio and kept all detections that exceeded a calibrated score of 0.8.

We then replicated the analysis from Van Doren, Lostanlen, et al. (2023) using the original analysis code. Briefly, we (a) related acoustic measures of nightly migration intensity to concurrent measures of migration intensity from nearby Doppler radar; (b) compared how well data generated with Nighthawk and BirdVoxDetect explained radar measures of migration; and (c) generated species-specific migration timing estimates from acoustic data and compared them to migration timing estimates derived from citizen science observations. See Van Doren, Lostanlen, et al. (2023) for details. For analysis (c), we followed that study by including all species with average daily eBird reporting frequencies of at least 1% of checklists and average nightly call rates of at least 0.25 calls per hour.

3 | RESULTS

3.1 | Core model performance (Part 1)

We trained Nighthawk on the Core training dataset and evaluated performance on the Core test dataset using mean average precision (mAP) (Table 1). Nighthawk achieved performance of AP > 0.80 on 50 (61%) species classes, 17 (100%) group classes, 13 (72%) family classes and 4 (100%) order classes (Table 1). Full performance metrics for all taxa are in Table S4, and confusion matrices for each taxonomic level are in Figures S1–S4.

There was a strong association between training sample size and model performance (Figure 3, Figures S5–S7). Species classes with at least 1000 training examples performed very well (mAP = 0.92; *n* = 26). For example, see results for White-throated Sparrow in Figure 4 and for all other classes in Figure S10 through Figure S129. In contrast, species classes with fewer than 200 training examples often performed poorly (mAP = 0.42; *n* = 17).

3.2 | Performance on well-represented target dataset (Part 2.3)

We evaluated Nighthawk on target audio from Pennsylvania, USA (dataset 18 in Table S1; Figure 5). All Nighthawk models performed well, but the Tuned and Batch–More models performed best on PA data. However, there was a difference in how these two top models performed on the Core test data: Batch–More maintained good performance (−0.008 species mAP compared to Naïve), while Tuned lost performance (−0.063 species mAP compared to Naïve). This indicates that fine-tuning only on target data came at the cost of lower performance on the Core dataset. See Table S5 and Figure 5.

For the continuous listening evaluation, tuning with Batch–More decreased the false positive rate from 33.4% (*N* = 315) to 0.7% (*N* = 6) while *increasing* the percentage of successfully detected annotator-marked calls from 59% to 78%. A manual review of detections revealed that some calls Nighthawk found had been missed by the annotator. See Table S5.

TABLE 1 Summary of model performance across taxonomic levels on the Core test dataset. Mean average precision (mAP) scores are given for all evaluated classes and for the subset of classes with at least 1000 training examples.

Level	No. classes	mAP	mAP (*N* > 1000)	No. classes (AP > 0.8)
Order	4	0.96	0.96	4
Family	18	0.86	0.95	13
Group	17	0.94	0.95	17
Species	82	0.75	0.92	50

3.3 | Performance on underrepresented target dataset (Part 2.4)

We next evaluated Nighthawk on target audio from the White Sands Missile Range in New Mexico (NM), USA (dataset 17 in Table S1; Figure 6). Tuned and Batch–More models performed best on NM data and were substantially better than the Naïve model. However, we again saw a difference in how these two top models performed on our original Core test dataset: Batch–More maintained good performance (−0.003 species mAP compared to

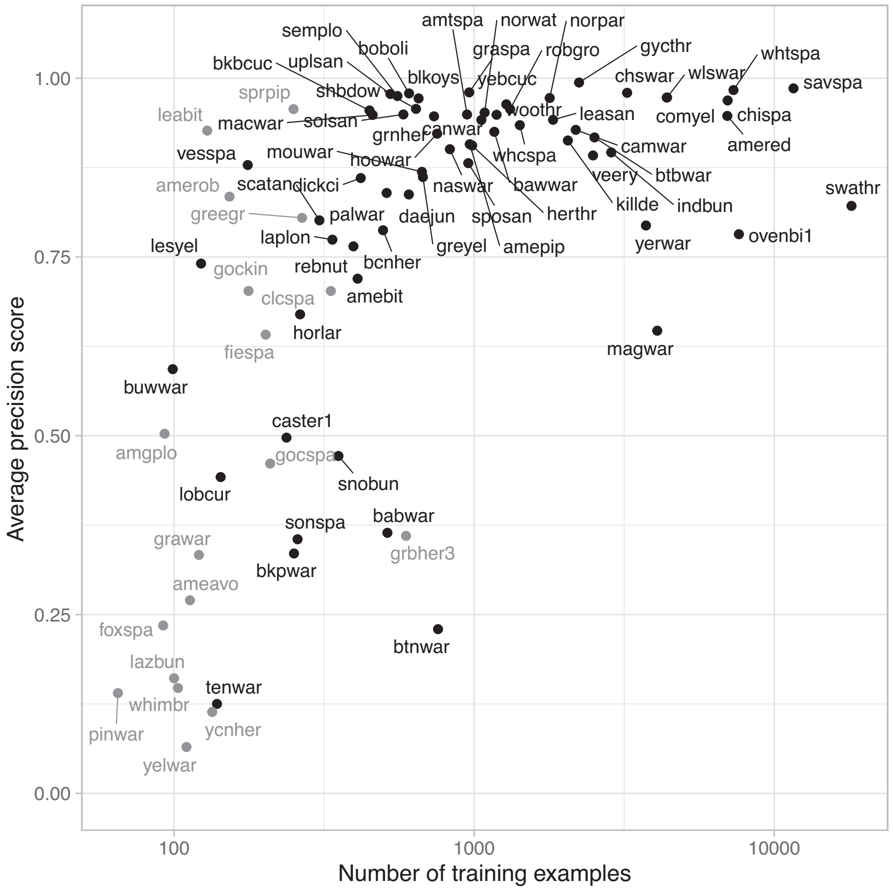

FIGURE 3 Model performance for species classes on Core test data. Performance is measured by average precision (AP) and plotted against the number of training examples. Species with more training examples generally performed better. Classes plotted in grey have less than 20 testing examples, so their reported performance is subject to increased uncertainty.

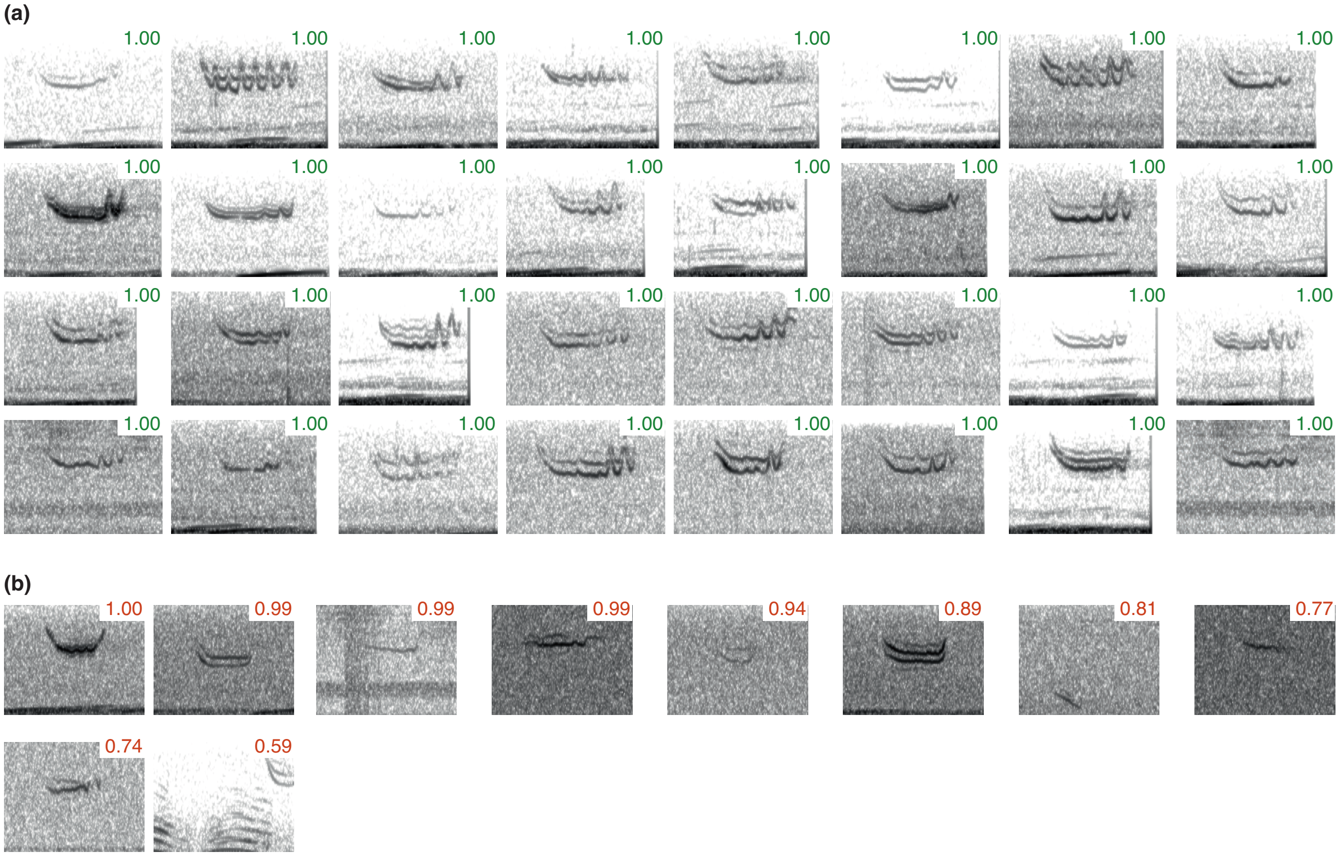

FIGURE 4 Classification examples for White-throated Sparrow drawn from the test dataset, shown as cropped spectrograms. The numbers in the upper-right corners are calibrated probability values returned by the model. (a) Test examples scoring highest for this species. Any incorrect classifications are outlined in red. (b) Incorrectly classified examples scoring highest for this species (i.e. the most confusing cases for the model). Spectrogram parameters: *x*-axis: 0–0.3 s; *y*-axis: 3–11 kHz; window type: Hanning; window length: 200 samples; hop size: 20 samples; dynamic range floor: −60 dB).

Naïve), while Tuned lost performance (−0.06 species mAP compared to Naïve), indicating that fine-tuning only on target data came at a cost of lower performance on the Core dataset. See Table S6 and Figure 6.

For the continuous listening evaluation, tuning with Batch–More decreased the false positive rate from 18.8% (*N* = 228) to 3.5% (*N* = 33) while keeping constant the percentage of successfully detected annotator-marked calls (69% and 69%). A manual review of detections again revealed that some calls Nighthawk found had been missed by the annotator. See Table S6.

3.4 | Performance on out-of-domain target dataset (Part 2.5)

Our final target evaluation was based on target audio from Colombia. Although the Naïve model performed poorly, fine-tuning with CO data yielded a high-performing model for both Tuned and Batch–More strategies (Figure 7). Again, there was a difference in how these two top models performed on our Core test dataset: Batch–More lost some performance (−0.037 species mAP compared to Naïve), while Tuned substantially lost performance (−0.108 species mAP compared to Naïve). See Table S7 and Figure 7.

For the continuous listening evaluation, tuning with Batch–More decreased the false positive rate from 24.7% (*N* = 21) to 9.3% (*N* = 25) while *increasing* the percentage of successfully detected annotator-marked calls from 14% to 51%. See Table S7.

3.5 | Continuous listening in New York State, USA (Part 3)

We used Core Nighthawk output to model nightly nocturnal migration intensity in central New York State. Acoustic detections alone explained 69% of the variation in radar-based migration intensity (Figure 8 A; methods in Van Doren, Lostanlen, et al. (2023)). Our ability to explain migration intensity using acoustics improved after incorporating wind variables and time of season, which yielded a model explaining 80% of the variation in nightly migration intensities (Figure 8b). Models using Nighthawk detections outperformed otherwise identical models using detections generated by BirdVoxDetect (from Van Doren, Lostanlen, et al. (2023); Figure 8c,d). Evaluation on an annotated

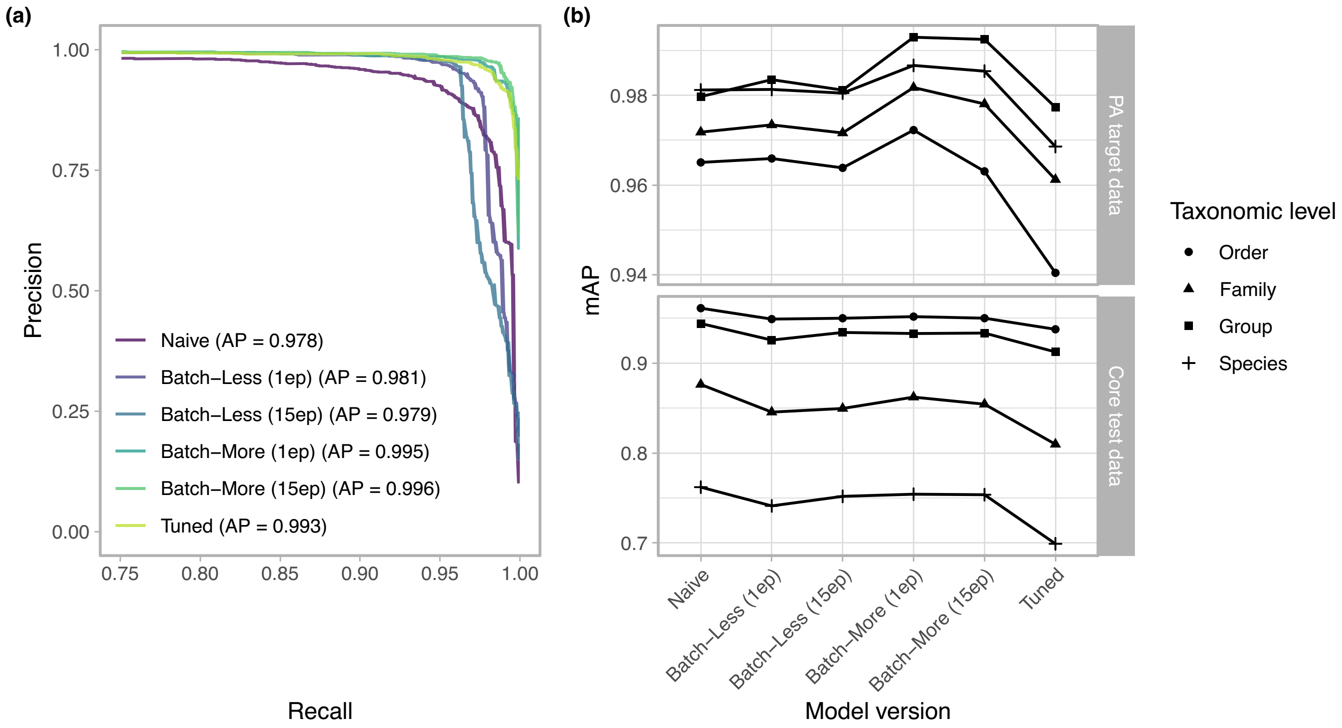

FIGURE 5 Performance on the central Pennsylvania holdout dataset. (a) Precision-recall curves for order Passeriformes. Recall values are displayed only from 0.75 to 1.00 to emphasize the subtle differences in performance between models. (b) Performance across taxonomic levels. Shown are mean average precision (mAP) scores for taxa with at least 20 examples in the holdout test set.

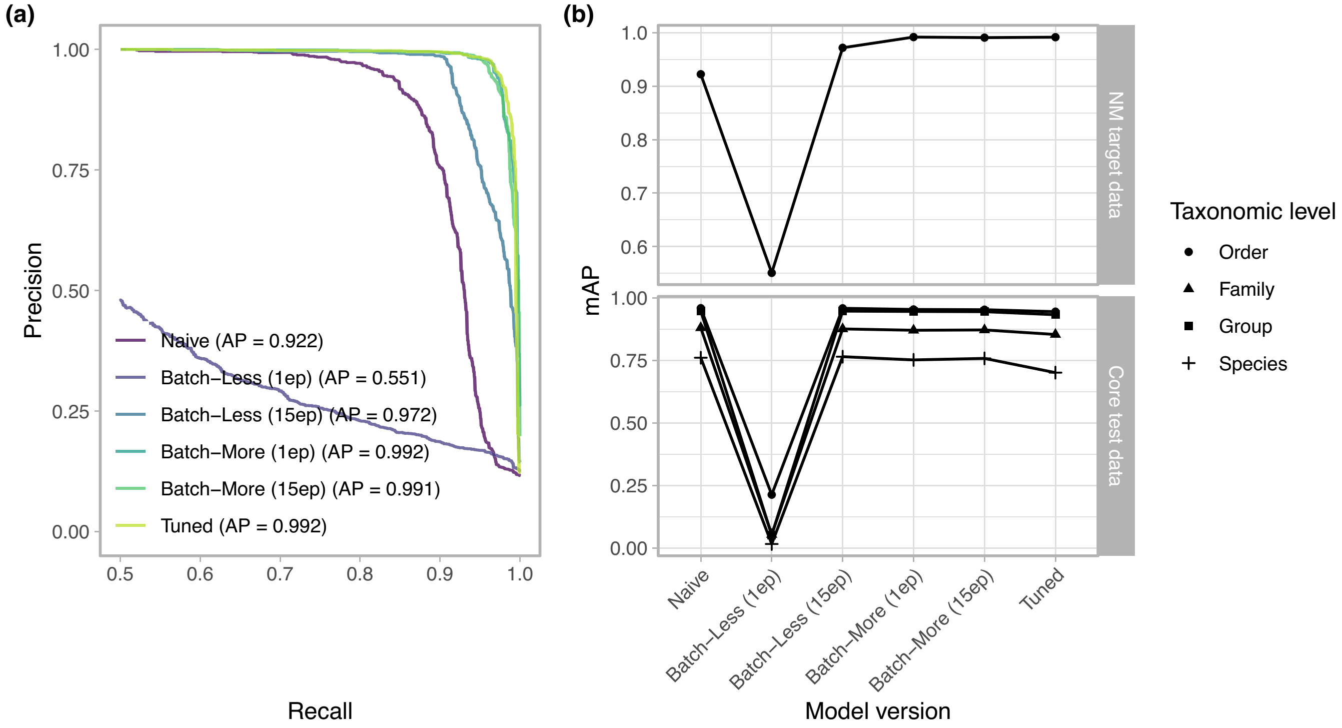

FIGURE 6 Precision-recall curves for order Passeriformes on a dataset from the White Sands Missile Range, NM, USA. Note that recall values are displayed only from 0.5 to 1.00 to emphasize the subtle differences in performance between models.

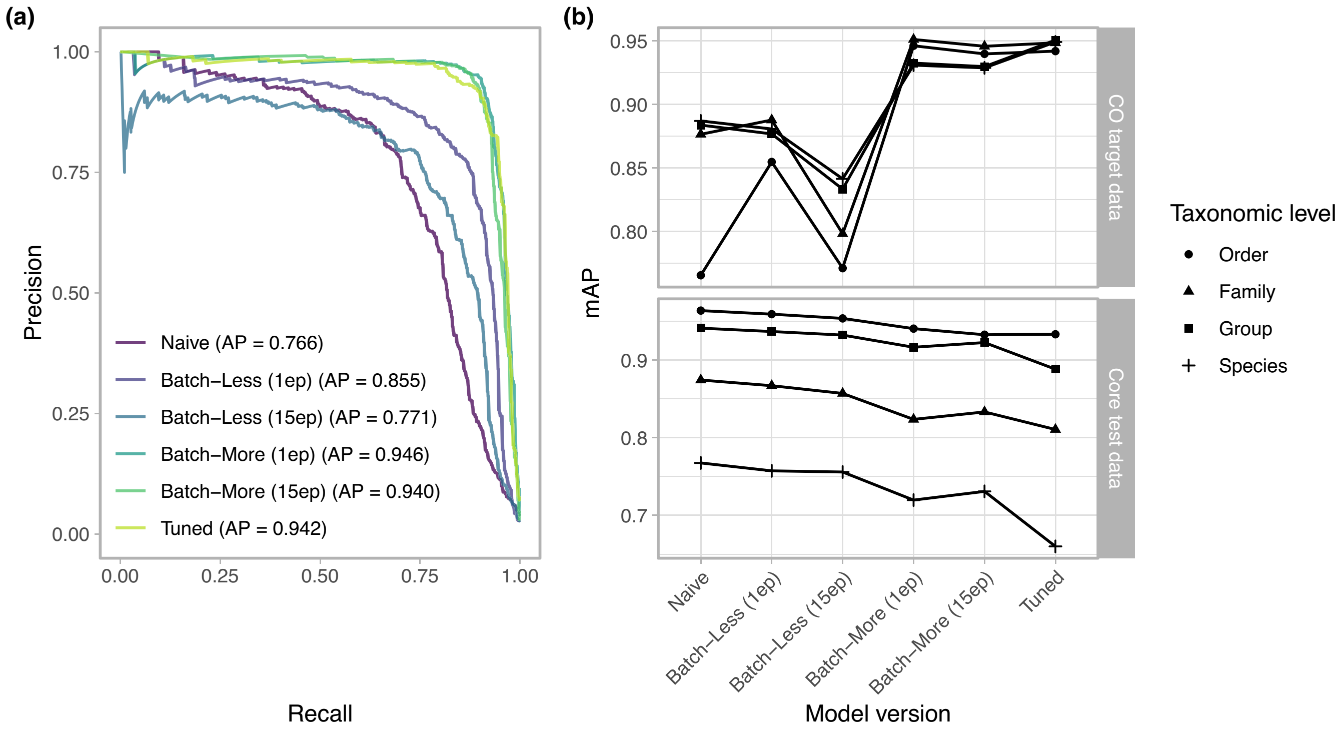

FIGURE 7 Performance on the Colombia holdout dataset. (a) Precision-recall curves for order Passeriformes. (b) Performance across taxonomic levels. Shown are mean average precision (mAP) scores for taxa with at least 20 examples in the holdout test set.

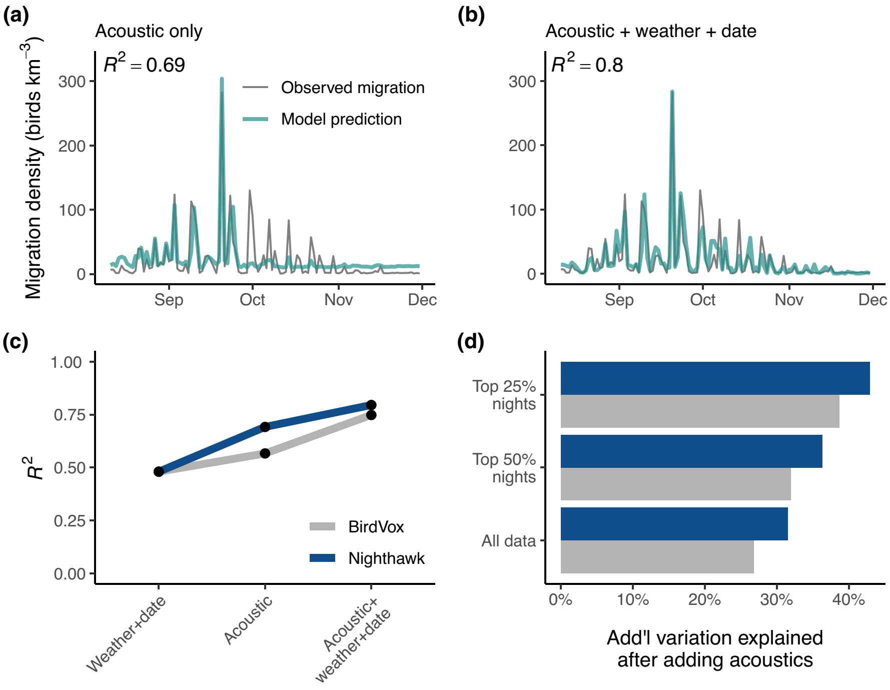

FIGURE 8 Using Nighthawk to monitor the intensity of nocturnal bird migration over Tompkins County, New York, USA, in autumn 2015. (a) Explaining migration intensity with acoustics. The grey line shows migration densities observed by radar; the teal lines show cross-validated predictions made with a generalized linear model comprising only flight call counts. (b) Explaining migration intensity with acoustics, weather and date information. Same as (a), with weather and date added to the generalized linear model. (c) R-squared values for three generalized linear models of migration intensity, comparing acoustic data generated with BirdVoxDetect and Nighthawk. (d) Observed increase in model fit (*R*-squared) after adding acoustic data to a model including weather and date variables for acoustic data generated with BirdVoxDetect (grey) and Nighthawk (blue).

subset of the dataset (BirdVox-296h: Farnsworth et al. (2021)) for stations in the test dataset yielded a precision of 0.51 and a recall of 0.89.

Acoustic detections generated with Nighthawk captured fine temporal variation in migration behaviour among species (Figure 9 and Figure S8). Nightly time series of acoustic detections showed that many species used airspace over central New York during only a handful of nights in the fall season. See Figure S8 for plots of all included species.

Species-specific migration timing estimates derived from Nighthawk detections closely matched timing estimates from the independent eBird database (Figure 9b). This close correspondence was quantified by the estimated slope of the relationship between these two measures, which did not significantly differ from 1 (slope = 1.04; 95% CI [0.84–1.24]). Acoustic timing estimates explained 78% of the variation in eBird-derived timing estimates. The species included in the timing analysis had a mean AP score on the Core test set of 0.90 (SD 0.13).

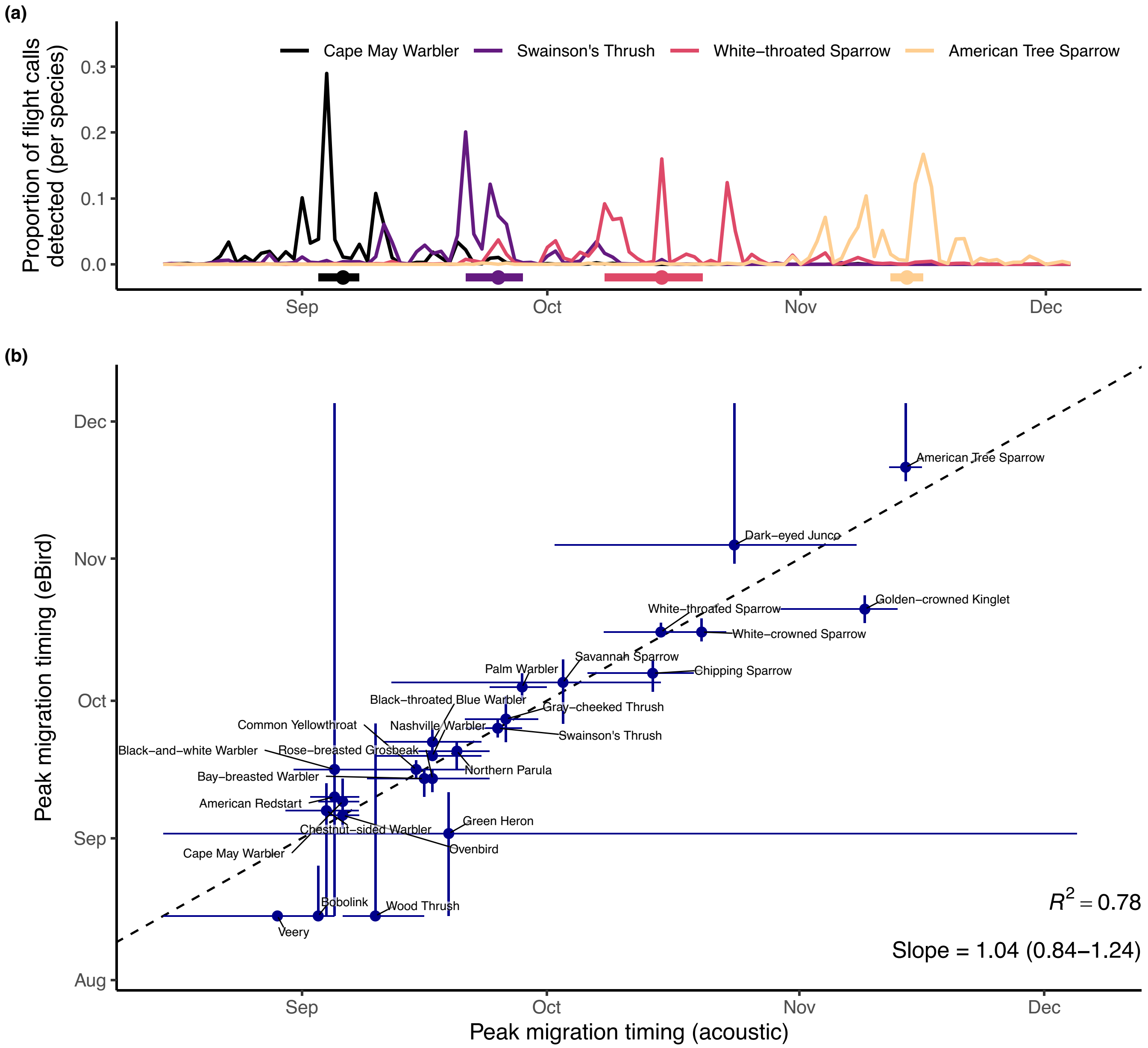

FIGURE 9 Acoustic monitoring of migration timing at the species level, estimated with acoustic and eBird data, from Tompkins County, New York, USA, in fall 2015. (a) Nightly flight call detections made by Nighthawk for four representative migrant species. The *y*-axis shows the proportion of the season's calls detected for that species on a given night. For example, 29% of all Cape May Warbler calls recorded during this season were detected on 4 September. The coloured lines and points beneath the scatterplot show estimated the peak migration timing and 95% confidence interval. (b) Comparison of migration timing at the species level, estimated with acoustic and eBird data. Each point represents one species, with confidence bars showing 95% credible intervals for acoustic and eBird estimates. The dashed line is the identity line; when the acoustic confidence interval (*x*-axis) intersects the identity line, the 95% CI in acoustic timing overlaps the eBird estimate.

4 | DISCUSSION

Our results demonstrate that Nighthawk performs very well as a nocturnal flight call detector and classifier for dozens of avian taxa. The taxonomic breadth, geographic coverage and quantity of training data exceed previously published tools (e.g. compare BirdVoxDetect, trained on 14 species and focused on one migration season in one US county; Lostanlen and Salamon (2022), Lostanlen et al. (2019) and Van Doren, Lostanlen, et al. (2023)). Taxa with at least 1000 training examples consistently performed well, though performance was variable for classes with less training data. Total Nighthawk detections accurately captured nightly nocturnal migration intensity and species phenology in our migration monitoring evaluation. The target dataset analyses show that the Core Nighthawk model performed well on data from North American locations not included in model training, and that incorporating modest amounts of additional annotated audio (50–120 h) can yield high performance on custom datasets within North America and beyond.

Our analyses also show that gathering meaningful data about bird migration with acoustics does not require a perfect system. In our migration monitoring evaluation (Part 3), the combination of acoustics, weather and date explained 80% of the nightly variation in migration intensity, and acoustics alone explained 69%. These associations were drawn from a Nighthawk run that achieved a precision of 'only' 0.51 (recall: 0.89) on a subset of these data. The results outperform those obtained with BirdVoxDetect and analysed using identical methods. This result is encouraging because it suggests that even moderately skilled flight call classifiers may yield high-quality data if the objective is to measure overall migration passage rates.

We note that Nighthawk differs from BirdVoxDetect in several important ways, some of which make direct comparison of these models challenging. BirdVoxDetect uses a combined detector and classifier approach, while Nighthawk uses only a classifier. BirdVoxDetect can therefore achieve higher temporal precision in its detections. The training datasets also differ between these two tools, with Nighthawk incorporating substantially more training data. At present, Nighthawk is trained on substantially more data from a greater variety of locations, and Nighthawk includes many more taxa. It is likely that augmenting BirdVoxDetect with additional training data would improve its performance.

Our migration monitoring evaluation using a regional array of acoustic stations (Part 3) also highlights that the bulk of many migratory species may use airspace over a given region during only a small number of nights. For example, 29% of all Cape May Warbler detections occurred on 4 September, and 38% of all White-throated Sparrow detections occurred on three dates (8, 15 and 23 October) (Figure 9a). Previous studies using radar observations have shown that the bulk of avian migrants pass a given location on only a handful of nights in a given season (Horton et al., 2021). Our results suggest that this pattern may also hold true at the species level. This finding is particularly relevant for species of concern, as it suggests that most migratory species utilize a given aerial habitat during only a small number of nights each season—nights that may represent particularly important targets of conservation action (e.g. reducing light pollution or pausing wind turbines).

Species-specific migration timing estimates derived from acoustics (Part 3) closely matched those derived from independent citizen science observations, validating the accuracy of acoustic measures. The slope of the relationship between acoustic and eBird-derived timing measures was not distinguishable from 1, but there were a handful of species that showed larger deviations. Of particular interest are Veery (*Catharus fuscescens*) and Wood Thrush (*Hylocichla mustelina*), species that are difficult to detect visually during migration periods. eBird observations peaked at the very start of August and declined smoothly thereafter, likely reflecting dwindling detection of breeding birds and negligible detection of migrants. However, with acoustics, we were able to identify the subsequent passage periods of these secretive species. Another advantage of acoustic monitoring over citizen science data is the standardization of sampling effort. Autonomous recording units collect standardized samples while citizen scientists submit varying numbers of checklists from varying locations from day to day. Even in Tompkins County, NY, one of the most heavily birded counties in the world, the number of checklists submitted to eBird in autumn 2015 varied greatly from day to day and declined at the beginning and end of the season (Figure S9). For long-term analyses over many years (e.g. Fink et al. (2020)), such variation can be overcome, but daily variation in observer effort poses a much greater challenge to daily analyses of migrating birds. For applications that require standardized estimates of nocturnal migration at fine taxonomic levels, acoustic monitoring may have great utility.

The Core Nighthawk model achieved AP > 0.8 on 50 species, highlighting its potential for monitoring entire communities of birds on the move. However, there were some species where the model distinctly underperformed, even with robust training sample sizes (Figure 3). These species included Magnolia Warbler (*Setophaga magnolia*), Black-throated Green Warbler (*Setophaga virens*), Bay-breasted Warbler (*Setophaga castanea*), Blackpoll Warbler (*Setophaga striata*) and Song Sparrow (*Melospiza melodia*). These species share an important characteristic: their calls are very similar to those of other species and are readily confused by the model. Fortunately, our modelling approach explicitly accounted for species-level identification uncertainty by joining similar-sounding species in 'groups'. For example, the 'ZEEP' group includes Magnolia Warbler, Bay-breasted Warbler and Blackpoll Warbler, and it is one of the highest-scoring classes in the Core evaluation (AP = 0.97) (Figure S5, Table S4). Similarly, Black-throated Green Warblers belong to the 'DBUP' group (AP = 0.91), and Song Sparrows belong to the 'SFHS' group (AP = 0.97). All groups achieved AP scores >0.80 on Core test data, further highlighting the usefulness of the 'group' construction: even if Nighthawk is not able to accurately classify some vocalizations at the species level, it can still confidently classify these signals to a small group of closely related species, maximizing information gain.

We tested Nighthawk on recordings from three target locations outside the Core training dataset: Pennsylvania, New

Mexico and Colombia. The performance of 'Naïve' models on target data varied; Naïve models performed best on in-domain data from central Pennsylvania and worst on out-of-domain data from Colombia. However, in all cases, we successfully fine-tuned a high-performing model using a modest amount of annotated audio data from target locations. Performance on target data after fine-tuning was comparable to (or better than) performance on locations included in the Core training set. We observed dramatic performance improvements when using a custom batch construction strategy in which half of each batch came from the target dataset. Remarkably, Batch–More models appeared to need only one epoch of fine-tuning to achieve excellent performance on target data. Although we could achieve similar, if not better, performance by initializing the model to 'Naïve' weights and doing further training *only* on target data ('Tuned' models), the Batch–More models retained better performance on the Core test set and sometimes on the target dataset. This behaviour could be advantageous in cases where a researcher tunes Nighthawk for a particular target dataset but may not have a fully representative sample of annotated audio. In these cases, custom batch construction could boost overall performance by maintaining skill on species not included in the target training data.

For all three target datasets, Nighthawk detected a meaningful number of calls that had been overlooked during the original annotation process, thus exceeding the abilities of skilled humans. Annotating hours of continuous audio is a tedious process, and humans may miss, for example, faint vocalizations that Nighthawk is able to detect. These results highlight both the difficulty of constructing a high-quality evaluation split and the advantages of systems that integrate both humans and machines in iterative data processing pipelines (Branson et al., 2014). Although it may be tempting to just 'set the computer loose' on a dataset, human input and quality control are still essential to producing a high-performing system. In the case of evaluating performance on target datasets, having a human review the final Nighthawk output drastically changed our interpretation of performance. For example, on PA target data, our estimate of continuous listening precision on order Passeriformes went from 0.73 before human review to 0.99 after review. Thus, including a human in the loop changed our view of the Nighthawk system from one that performed poorly (precision = 0.73) to one that performed impressively well (precision = 0.99). Although we did not explicitly assess the different types of errors made by Nighthawk, we observed that insects and non-flight call bird vocalizations were frequent sources of false positive detections. In the future, incorporating additional training data and analysing these confusions may lead to model improvements. One challenge with applying our model in a continuous listening context is that the model encounters a substantially higher proportion of negative data than we included in the training dataset. For example, some full-night recordings may contain hours of audio without any flight calls. We found that our post-processing methods of enforcing taxonomic consistency and requiring corroborating, overlapping detections helped reduce the number of false positives encountered in continuous listening.

4.1 | Using Nighthawk

By monitoring the vocalizations of actively migrating birds, Nighthawk provides a detailed window onto nocturnal bird migration that is not presently attainable by other means (e.g. radar or citizen science). Important advantages of this method include its ability to collect data at simultaneously fine spatial, temporal and taxonomic resolutions. Scientists, managers and practitioners could use acoustic monitoring for a number of applications, including: monitoring migration passage at wind farms at the species level; studying how bird species use different parts of the landscape during migratory flights; monitoring the changing arrival, departure and passage times of species susceptible to climate change; and revealing previously unknown migration routes and behaviours. These applications could be most important in areas lacking adequate radar or citizen science coverage. See the Data Availability section for information on downloading Nighthawk.

Our results suggest that Nighthawk's Core model will function well for monitoring nocturnal bird migration in much of North America, especially in the eastern half of the continent. The current public release of Nighthawk includes the Core model and makes it easy for users to run this model on their own data. Support for full model retraining and fine-tuning is left as future work. In our analyses of target data, annotators reviewed approximately 50–120 h of audio. To achieve the greatest benefit from fine-tuning and in order to function properly as validation data, researchers should incorporate target training data that is as representative as possible of the broader target dataset across locations, dates, seasons and species. In our target datasets, we split annotated data into two approximately equal-sized sets for training and validation. Validation data allow researchers to select score thresholds that fit their needs on a larger dataset. Detection results are sensitive to the choice of threshold; some applications may call for higher precision (e.g. fully autonomous monitoring) and others for higher recall (e.g. locating rare species with some manual review). We intend to continue building our library of annotated recordings and expanding the capabilities of Nighthawk to more species and locations.

AUTHOR CONTRIBUTIONS

Benjamin M. Van Doren conceived of the study, acquired funding and data, performed analyses and wrote the first manuscript draft. Grant Van Horn created the original Merlin Sound ID model and oversaw this study. Andrew Farnsworth shaped the study and contributed audio data and annotations. Kate Stone, Dylan M. Osterhaus and Jacob Drucker also provided audio data and annotations. All authors contributed to the final written draft of the manuscript.

ACKNOWLEDGEMENTS

Special thanks to David Benvent, William R. Evans, Mike Farmer, Jenn Foote, Eric R. Gulson-Castillo, Joe Gyekis, Jacob Job, Blaine Landsborough, Michael Lanzone, Debbie Leick, Nicole Liao, Jay McGowan, Sara R. Morris, Dan Mennill, Drew Weber, Ryan Zucker and countless contributors to the Macaulay Library (macaulayli brary.org) for providing audio recordings and annotations. Thanks to Steve Kelling, Juan Pablo Bello, Aurora Cramer, Vincent Lostanlen and Justin Salamon for their work on the BirdVox effort and for helpful feedback on this project. Thanks to Benjamin Hoffman and Harold Mills for their invaluable technical contributions, expertise and assistance. Connor Wood and one anonymous reviewer provided helpful feedback on the manuscript. This work was funded by the Cornell Lab of Ornithology, Actions at EBMF, the University of Chicago Steiner Fund, the Field Museum of Natural History H. B. Conover Fund, and the Cornell Presidential Postdoctoral Fellowship.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.14272.

DATA AVAILABILITY STATEMENT

The Nighthawk model, trained on the Core dataset presented here, is archived on Zenodo (Van Doren, Mills, et al., 2023) and available for download on GitHub (https://github.com/bmvandoren/Nighthawk/), along with Python code demonstrating its proper use. We welcome questions, feedback and collaboration inquiries by email to the corresponding author. Users can also use Nighthawk by installing the program Vesper (https://github.com/HaroldMills/Vesper) and using a plugin (https://github.com/HaroldMills/vesper-nighthawk). Vesper is designed for the management and processing of audio recordings for nocturnal bird migration monitoring and is maintained by Harold Mills (https://github.com/HaroldMills).

ORCID

Benjamin M. Van Doren https://orcid.org/0000-0002-7355-6005
Andrew Farnsworth https://orcid.org/0000-0002-9854-4449
Dylan M. Osterhaus https://orcid.org/0000-0002-9044-1090
Jacob Drucker https://orcid.org/0000-0002-2193-4879
Grant Van Horn https://orcid.org/0000-0003-2953-9651

REFERENCES

Bauer, S., Shamoun-Baranes, J., Nilsson, C., Farnsworth, A., Kelly, J., Reynolds, D. R., Dokter, A. M., Krauel, J., Petterson, L. B., Horton, K. G., & Chapman, J. W. (2019). The grand challenges of migration ecology that radar aeroecology can help answer. *Ecography*, *42*, 861–875. https://doi.org/10.1111/ecog.04083

Bota, G., Traba, J., Sardà-Palomera, F., Giralt, D., & Pérez-Granados, C. (2020). Acoustic monitoring of diurnally migrating European bee-eaters agrees with data derived from citizen science. *Ardea*, *108*, 139–149. https://doi.org/10.5253/arde.v108i2.a3

Branson, S., Van Horn, G., Wah, C., Perona, P., & Belongie, S. (2014). The ignorant led by the blind: A hybrid human–machine vision system for fine-grained categorization. *International Journal of Computer Vision*, *108*, 329.

Cramer, J., Lostanlen, V., Farnsworth, A., Salamon, J., & Bello, J. P. (2020). *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 901–905). IEEE. https://doi.org/10.1109/ICASSP40776.2020.9052908

Davy, C. M., Ford, A. T., & Fraser, K. C. (2017). Aeroconservation for the fragmented skies. *Conservation Letters*, *10*, 773–780. https://doi.org/10.1111/conl.12347

Diehl, R. H. (2013). The airspace is habitat. *Trends in Ecology & Evolution*, *28*, 377–379. https://doi.org/10.1016/j.tree.2013.02.015

Dokter, A. M., Liechti, F., Stark, H., Delobbe, L., Tabary, P., & Holleman, I. (2011). Bird migration flight altitudes studied by a network of operational weather radars. *Journal of the Royal Society Interface*, *8*, 30–43. https://doi.org/10.1098/rsif.2010.0116

Evans, W. R., & O'Brien, M. (2002). *Flight calls of migratory birds: Eastern North American landbirds [CD-ROM]*. Old Bird Inc.

Evans, W. R., & Rosenberg, K. V. (2000). *Acoustic monitoring of night-migrating birds: A progress report* (R. Bonney, D. N. Pashley, R. J. Cooper, & L. Niles (Eds.), 15 p.). U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station.

Farnsworth, A. (2005). Flight calls and their value for future ornithological studies and conservation research. *The Auk*, *122*, 733–746. https://doi.org/10.1093/auk/122.3.733

Farnsworth, A., Kelling, S., Lostanlen, V., Salamon, J., Cramer, A., & Bello, J. P. (2021). BirdVox-296h: A large-scale dataset for detection and classification of flight calls. https://doi.org/10.5281/zenodo.4603643

Farnsworth, A., Van Doren, B. M., Kelling, S., Lostanlen, V., Salamon, J., Cramer, A., & Bello, J. P. (2022). BirdVox-full-season: 6672 hours of audio from migratory birds. https://doi.org/10.5281/zenodo.5791744

Fink, D., Auer, T., Johnston, A., Ruiz-Gutierrez, V., Hochachka, W. M., & Kelling, S. (2020). Modeling avian full annual cycle distribution and population trends with citizen science data. *Ecological Applications*, *30*, e02056. https://doi.org/10.1002/eap.2056

Fraser, K. C., Davies, K. T. A., Davy, C. M., Ford, A. T., Flockhart, D. T. T., & Martins, E. G. (2018). Tracking the conservation promise of movement ecology. *Frontiers in Ecology and Evolution*, *6*. https://doi.or/10.3389/fevo.2018.00150

Gauthreaux, S. A., Belser, C. G., & van Blaricom, D. (2003). *Using a network of WSR-88D weather surveillance radars to define patterns of bird migration at large spatial scales* (P. Berthold, E. Gwinner, & E. Sonnenschein (Eds.) (pp. 335–346). Springer.

Gayk, Z. G., & Mennill, D. J. (2023). Acoustic similarity of flight calls corresponds with the composition and structure of mixed-species flocks of migrating birds: Evidence from a three-dimensional microphone array. *Philosophical Transactions of the Royal Society, B: Biological Sciences*, *378*, 20220114. https://doi.org/10.1098/rstb.2022.0114

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. (pp. 770778).

Horton, K. G., Van Doren, B. M., Albers, H. J., Farnsworth, A., & Sheldon, D. (2021). Near-term ecological forecasting for dynamic aeroconservation of migratory birds. *Conservation Biology*, *35*, 1777–1786. https://doi.org/10.1111/cobi.13740

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., & others. (2019). Searching for MobileNetV3. (pp. 13141324).

Kahl, S., Wood, C. M., Eibl, M., & Klinck, H. (2021). BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, *61*, 101236. https://doi.org/10.1016/j.ecoinf.2021.101236

Kays, R., Crofoot, M. C., Jetz, W., & Wikelski, M. (2015). Terrestrial animal tracking as an eye on life and planet. *Science*, *348*, aaa2478. https://doi.org/10.1126/science.aaa2478

Landsborough, B. J., Foote, J. R., & Mennill, D. J. (2019). Decoding the 'zeep' complex: Quantitative analysis of interspecific variation in the nocturnal flight calls of nine wood warbler species (Parulidae spp.). *Bioacoustics*, *28*, 555–574. https://doi.org/10.1080/09524622.2018.1509373

Lanzone, M., Deleon, E., Grove, L., & Farnsworth, A. (2009). Revealing undocumented or poorly known flight calls of warblers (Parulidae) using a novel method of recording birds in captivity. *The Auk*, *126*, 511–519. https://doi.org/10.1525/auk.2009.08187

Liechti, F., Witvliet, W., Weber, R., & Bächler, E. (2013). First evidence of a 200-day non-stop flight in a bird. *Nature Communications*, *4*, 2554. https://doi.org/10.1038/ncomms3554

Lostanlen, V., & Salamon, J. (2022). BirdVox/birdvoxdetect: 1.0. *Zenodo* https://doi.org/10.5281/zenodo.7414934

Lostanlen, V., Salamon, J., Farnsworth, A., Kelling, S., & Bello, J. P. (2018). *2018 IEEE International Conference on acoustics, speech and signal processing (ICASSP)* (pp. 266–270). IEEE. https://doi.org/10.1109/ICASSP.2018.8461410

Lostanlen, V., Salamon, J., Farnsworth, A., Kelling, S., & Bello, J. P. (2019). Robust sound event detection in bioacoustic sensor networks. *PLoS One*, *14*, e0214168. https://doi.org/10.1371/journal.pone.0214168

Northcutt, C. G., Athalye, A., & Mueller, J. (2021). Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv* preprint:arXiv:2103.14749.

Nussbaumer, R., Bauer, S., Benoit, L., Mariethoz, G., Liechti, F., & Schmid, B. (2021). Quantifying year-round nocturnal bird migration with a fluid dynamics model. bioRxiv:2020.10.13.321844 https://doi.org/10.1101/2020.10.13.321844

Reynolds, M. D., Sullivan, B. L., Hallstein, E., Matsumoto, S., Kelling, S., Merrifield, M., Fink, D., Johnston, A., Hochachka, W. M., Bruns, N. E., Reiter, M. E., Veloz, S., Hickey, C., Elliott, N., Martin, L., Fitzpatrick, J. W., Spraycar, P., Golet, G. H., McColl, C., … Morrison, S. A. (2017). Dynamic conservation for migratory species. *Science Advances*, *3*, e1700707. https://doi.org/10.1126/sciadv.1700707

Rosenberg, K. V., Dokter, A. M., Blancher, P. J., Sauer, J. R., Smith, A. C., Smith, P. A., Stanton, J. C., Panjabi, A., Helft, L., Parr, M., & Marra, P. P. (2019). Decline of the North American avifauna. *Science*, *366*, 120–124. https://doi.org/10.1126/science.aaw1313

Salamon, J., Bello, J. P., Farnsworth, A., Robbins, M., Keen, S., Klinck, H., & Kelling, S. (2016). Towards the automatic classification of avian flight calls for bioacoustic monitoring. *PLoS One*, *11*, e0166866. https://doi.org/10.1371/journal.pone.0166866

Shamoun-Baranes, J., Bauer, S., Chapman, J. W., Desmet, P., Dokter, A. M., Farnsworth, A., van Gasteren, H., Haest, B., Koistinen, J., Kranstauber, B., Liechti, F., Mason, T. H. E., Nilsson, C., Nussbaumer, R., Schmid, B., Weisshaupt, N., & Leijnse, H. (2022). Meteorological data policies needed to support biodiversity monitoring with weather radar. *Bulletin of the American Meteorological Society*, *103*, E1234–E1242. https://doi.org/10.1175/BAMS-D-21-0196.1

Shipley, J. R., Kelly, J. F., & Frick, W. F. (2018). Toward integrating citizen science and radar data for migrant bird conservation. *Remote Sensing in Ecology and Conservation*, *4*, 127–136. https://doi.org/10.1002/rse2.62

Sugai, L. S. M., Silva, T. S. F., Ribeiro, J. W. J., & Llusia, D. (2019). Terrestrial passive acoustic monitoring: Review and perspectives. *Bioscience*, *69*, 15–25. https://doi.org/10.1093/biosci/biy147

Sullivan, B. L., Aycrigg, J. L., Barry, J. H., Bonney, R. E., Bruns, N., Cooper, C. B., Damoulas, T., Dhondt, A. A., Dietterich, T., Farnsworth, A., Fink, D., Fitzpatrick, J. W., Fredericks, T., Gerbracht, J., Gomes, C., Hochachka, W. M., Iliff, M. J., Lagoze, C., La Sorte, F. A., … Kelling, S. (2014). The eBird enterprise: An integrated approach to development and application of citizen science. *Biological Conservation*, *169*, 31–40. https://doi.org/10.1016/j.biocon.2013.11.003

Van Doren, B. M., Conway, G. J., Phillips, R. J., Evans, G. C., Roberts, G. C. M., Liedvogel, M., & Sheldon, B. C. (2021). Human activity shapes the wintering ecology of a migratory bird. *Global Change Biology*, *27*, 2715–2727. https://doi.org/10.1111/gcb.15597

Van Doren, B. M., & Horton, K. G. (2018). A continental system for forecasting bird migration. *Science*, *361*, 1115–1118. https://doi.org/10.1126/science.aat7526

Van Doren, B. M., Lostanlen, V., Cramer, A., Salamon, J., Dokter, A., Kelling, S., Bello, J. P., & Farnsworth, A. (2023). Automated acoustic monitoring captures timing and intensity of bird migration. *Journal of Applied Ecology*, *60*, 433444. https://doi.org/10.1111/1365-2664.14342

Van Doren, B. M., Mills, H., & Stafford, S. (2023). bmvandoren/Nighthawk: v0.3.0. *Zenodo* https://doi.org/10.5281/zenodo.10215021

Van Horn, G., Branson, S., Farrell, R., Haber, S., Barry, J., Ipeirotis, P., Perona, P., & Belongie, S. (2015). Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. (pp. 595604).

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Supporting Information S1. Supplementary Tables and Figures.

How to cite this article: Van Doren, B. M., Farnsworth, A., Stone, K., Osterhaus, D. M., Drucker, J., & Van Horn, G. (2024). *Nighthawk*: Acoustic monitoring of nocturnal bird migration in the Americas. *Methods in Ecology and Evolution*, *15*, 329–344. https://doi.org/10.1111/2041-210X.14272