




Designing a BirdNET classifier for high wind detection in passive acoustic recordings to support wildlife monitoring^{a)}

Danielle T. Fradet,^{1,2,b)}  Megan A. Cimino,³ Easton R. White,¹  and Laura N. Kloepper^{1,2} 

¹Department of Biological Sciences, University of New Hampshire, Durham, New Hampshire 03824, USA

²Center for Acoustics Research and Education, University of New Hampshire, Durham, New Hampshire 03824, USA

³Institute of Marine Science, University of California, Santa Cruz, Santa Cruz, California 95064, USA

ABSTRACT:

Passive acoustic monitoring (PAM) is a powerful tool for ecological research, but recordings can be compromised by background noise such as wind. Addressing wind noise (e.g., clipping and masking) in bioacoustic data remains a challenge, especially as climate change is predicted to increase wind speeds, particularly near the poles. Adélie penguins (*Pygoscelis adeliae*), key indicators of the Antarctic ecosystem, are well-suited for PAM, where large-scale monitoring could assess climate-driven population changes—if wind noise is managed effectively. In this study, the convolutional neural network, BirdNET, inversely identifies unwanted sounds in Adélie penguin colony recordings. Multiple custom models were developed in which the background nontarget noise was Adélie vocalizations, and wind conditions (low, medium, and high) were the target classes. The best-performing model achieved an F-score of 0.43 and accuracy of 0.53. The high wind class within this model had a precision of 0.76 and recall of 0.94. A six-step workflow is presented for creating custom BirdNET models, evaluating their performance and determining an optimal confidence threshold prior to model application on an entire dataset. By automating unwanted sound detection, this approach enables researchers to efficiently identify and remove affected files, streamline data cleaning, and focus on recordings of interest for further analysis. © 2025 Acoustical Society of America.

<https://doi.org/10.1121/10.0036887>

(Received 19 December 2024; revised 14 May 2025; accepted 30 May 2025; published online 20 June 2025)

[Editor: Daniel T. Blumstein]

Pages: 4502–4512

I. INTRODUCTION

The use of passive acoustic monitoring (PAM) in wildlife research is increasing rapidly as a result of its relative ease in collecting large-scale datasets (Shonfield and Bayne, 2017; Sugai *et al.*, 2019). PAM research spans a wide range of topics, including but not limited to abundance (Hart *et al.*, 2021; Kloepper *et al.*, 2016), density (Küsel *et al.*, 2011; Marques *et al.*, 2013), occupancy (Campos-Cerqueira and Aide, 2016; Furnas and Callas, 2015), diversity (Anunciação *et al.*, 2022; Leach *et al.*, 2016), phenology (Bateman *et al.*, 2021; Leach *et al.*, 2016; Oliver *et al.*, 2018), rare and elusive species monitoring (Abrahams and Geary, 2020; Picciulin *et al.*, 2019), human impacts (Brosseau *et al.*, 2024; Nemeth and Brumm, 2010), and behavior (Jahn *et al.*, 2017; Pérez-Granados and Schuchmann, 2021). However, most applications have focused on environments with relatively simple background noise or often fail to consider how background noise impacts animal behavior and detection of target signals. Thus, geophony, or sounds generated through geophysical processes, poses some major challenges when collecting and interpreting acoustic data.

One particularly challenging geophonic sound is wind. Wind moving across a microphone creates turbulence that the equipment records as noise, characterized by high power in lower frequencies with unpredictable peaks in power (Nelke and Vary, 2015). At high wind speeds, recordings may clip, which means that the pressure signal captured exceeds the amplitude range of the device. Clipping distorts the captured signal and causes irreversible data loss. High wind speeds can also mask target signals or alter vocal behavior of animals (Digby *et al.*, 2014), potentially leading to data misinterpretation if windy days are not accounted for in bioacoustic studies.

Current strategies to address wind in PAM studies involve reporting results across wind speed categories (Arneill *et al.*, 2020), including wind speed as a covariate in models (Arneill *et al.*, 2020), restricting recording schedules to less windy parts of the day (Borker *et al.*, 2014), deploying recorders exclusively in wind-sheltered areas (Oppel *et al.*, 2014), or excluding wind-dominated files from analysis (Buxton and Jones, 2012). However, these wind reduction methods at particularly windy sites may not be enough. Zhao *et al.* (2022), who incorporated wind speed in their model, found wind speed and vocal activity rate of Adélie penguin chicks to be significantly correlated, and as a result, they recommended preprocessing data prior to analysis to remove low-quality recordings caused by wind. There remains a need to explore effective data cleaning methods

^{a)}This paper is part of a special issue on Climate Change: How the Sound of the Planet Reflects the Health of the Planet.

^{b)}Email: Danielle.Fradet@ush.edu

for pre-cleaning acoustic datasets (Juodakis and Marsland, 2022) by identifying and removing irreparably distorted recordings in large datasets.

The need for effective methods to identify wind in bioacoustic data is expected to grow as climate change increases the intensity of high winds worldwide (Clarke *et al.*, 2022). Climate change also impacts atmospheric circulation with a general trend of westerly winds strengthening and shifting toward the poles (TS.3.1.2 Spatial Distribution of Changes in Temperature, Circulation and Related Variables, 2007). Antarctica is particularly vulnerable to wind noise, where a combination of atmospheric conditions and topography result in the fastest documented near-surface wind speeds on Earth (Parish, 1988). Climate change exacerbates these winds; increased positive pressure in the Southern Hemisphere annular mode from greenhouse gases and the ozone hole has resulted in a 15%–20% increase in westerly winds (Convey *et al.*, 2009; Marshall, 2003) and an additional increase in near-surface wind speeds in Antarctica (Turner *et al.*, 2005). General circulation models predict continued increase in positive pressure (Turner *et al.*, 2014), further intensifying westerly winds globally over the 21st century (Bracegirdle *et al.*, 2008). Climate change induced increases in these already extreme surface winds will have dramatic effects on the Antarctic environment and soundscape.

One way to monitor climate change effects in vulnerable habitats is by using indicator species to understand wildlife response to environmental shifts (Pearman *et al.*, 2011; Siddig *et al.*, 2016). A key sentinel species in the Antarctic ecosystem is the Adélie penguin (*Pygoscelis adeliae*; Ainley, 2002). Understanding population dynamics and breeding phenology of Adélie penguins is essential because they are one of two penguin species that are true ice obligates, relying on sea-ice for parts of their life cycle (Ancel *et al.*, 2013). Therefore, warming could have devastating consequences for Adélie penguin habitat availability (Ainley *et al.*, 2010; Cimino *et al.*, 2016), making this species particularly vulnerable to climate change. Adélie penguins are already experiencing wide-spread population decline in the West Antarctic Peninsula; for example, colonies have decreased more than 90% near Palmer Station since the mid-1970s (Schofield *et al.*, 2024). Given their vulnerability and conservation significance, it is crucial to develop an easily deployable and scalable method for monitoring population dynamics and breeding phenology of Adélie penguins to provide further mechanistic understanding of climate change impacts.

Penguins comprise 90% of the Southern Ocean's seabird biomass (Black, 2016), making them a dominant component of the soundscape and a potential candidate for PAM. In particular, Adélie penguins breed synchronously in large colonies (e.g., up to hundreds of thousands of individuals; Borowicz *et al.*, 2018) and nest densely (1.4 nest/m²; Beaulieu *et al.*, 2009), allowing their vocalizations to be recorded from fixed positions at colony edges throughout a breeding season. The Adélie penguin breeding strategy, characterized by high nest concentrations and separation of

parental duties, has driven the evolution of a highly vocal species with individualized calls used for resource provisioning and mate choice (Ancel *et al.*, 2013; Brunton *et al.*, 2010; Jouventin, 1982; Speirs and Davis, 1991). Adélie penguins vocalize frequently with no significant daily vocal patterns, although vocal activity has been related to weather conditions (Zhang *et al.*, 2021; Zhao *et al.*, 2022). Human disturbance negatively impacts Adélie penguins (Carney and Sydeman, 1999), suggesting passive monitoring could be a noninvasive form of data acquisition (Nelson and Baird, 2001). Further, this species' circumpolar range makes intensive research at distant breeding sites difficult as a result of the remote location, harsh weather conditions, inherent safety risks, and logistical constraints. These challenges underscore the need for an easily deployable way to collect data. PAM offers a promising solution for monitoring changes in population size (Colombelli-Négrel, 2023; Zhao *et al.*, 2022), distribution, and phenology because acoustic recording units can be deployed across multiple colonies to collect high-resolution data.

Large datasets generated from PAM require an automated pipeline for efficient analysis. Manually annotating large datasets to identify recordings that may need to be excluded from the dataset, such as wind-dominated recordings, presents challenges for automated workflows. Relying on weather station data to identify high-wind events may be undependable for habitats located far from weather stations or where wind conditions vary across the landscape. A potential solution to eliminate files compromised by high wind is to automatically remove all clipped files from a dataset. However, this approach may inadvertently exclude vocalizations of interest for individuals located near a recorder. A method to specifically identify wind from vocalizations is needed. As demonstrated by Terranova *et al.* (2024), convolutional neural networks (CNNs)—the most widely used deep learning networks in bioacoustics (Stowell, 2022)—offer a promising solution for efficiently detecting high-wind recordings in large datasets. BirdNET, a CNN developed by the Cornell Laboratory of Ornithology's Center for Conservation Bioacoustics and the Chemnitz University of Technology, is an open-source deep artificial neural network designed to automatically classify bird sound by species (Kahl *et al.*, 2021). This CNN can now identify over 6000 species of birds and other animals from all 7 continents and has been primarily used by researchers to detect multiple species of interest in datasets (Pérez-Granados, 2023). The latest version of BirdNET graphical user interface (GUI v1.2.0, model v2.4) allows users to build custom models for various target signals—such as mammals, insects, and chainsaws—by adding an embedding layer to the CNN (Symes *et al.*, 2023).

Conducting PAM on Adélie penguins in Antarctica represents an ideal case study in which PAM is justified, but the dataset requires a pre-analysis step to address high-wind conditions. In this study, we present a framework for detecting high wind in a dataset using a custom model built on BirdNET's base algorithm with data from five Adélie

penguin colonies near Palmer Station in the West Antarctic Peninsula. We use BirdNET's custom model to create an inverse application to identify unwanted sounds, where Adélie vocalizations are the nontarget background noise, and various levels of wind are our target classes. We aim to (1) present a method for building a custom wind model, (2) apply the custom model to a large dataset and accurately identify windy data, (3) and explore the impact of minimum confidence threshold on the model's performance.

II. MATERIALS AND METHODS

A. Data collection

We deployed seven Wildlife Acoustics Song Meter Minis (SMMs; Maynard, MA) on recording rigs at five colonies near Palmer Station [Fig. 1(b)]. Each recording rig consisted of a polyvinyl chloride (PVC) structure with a recorder zip-tied to the PVC at a height of 1 m, keeping the SMM out of reach of the Adélie penguins [Fig. 1(c)]. The rigs were positioned at the edge of the colony to minimize disturbance during maintenance, with the microphone facing toward the colony center. Rocks were placed on the base of the rig for added stability. The SMMs recorded for 5 min every hour at 24 kHz sampling rate with 6 dB gain in a 16-bit depth. Data collection took place during the Adélie penguin breeding season from November 27, 2022 to February 27, 2023. Recording began after peak egg date to reduce the risk of nest abandonment during equipment setup (Giese, 1996). In total, we collected 13 932 5-min

recordings. We deployed three rigs at our largest colony because of its size. Four of the SMM rigs recorded data without failure. Elephant seals (*Mirounga angustirostris*) crushed two of the rigs, which field teams replaced with a new recorder within a week of destruction. High wind ripped the recorder off one rig, which we excluded from analysis as a result of lack of data. The 2022–2023 breeding season was one of the windiest on record for Palmer Station.

B. BirdNET custom model workflow

To build a custom model in BirdNET, one must create a training dataset and a test dataset. The training dataset provides example clips of target signals, defined as classes, and nontarget signals, defined as background. BirdNET takes the provided training dataset and randomly splits the data into 80% training data and 20% validation data, using the training data to teach the model and the validation data to automatically fine-tune the hyperparameters of the model. The test dataset is made up of data unseen by the trained model. In this study, we use the test dataset to evaluate BirdNET model performance on new data by manually annotating the test dataset for the classes of interest and then running the BirdNET trained model on the same unseen test dataset. Next, we compare the BirdNET detections to the manual annotations to assess model performance. The BirdNET team provides clear instructions for creating and evaluating a custom model (Symes *et al.*, 2023). Once a model is evaluated, a minimum confidence threshold with a desired

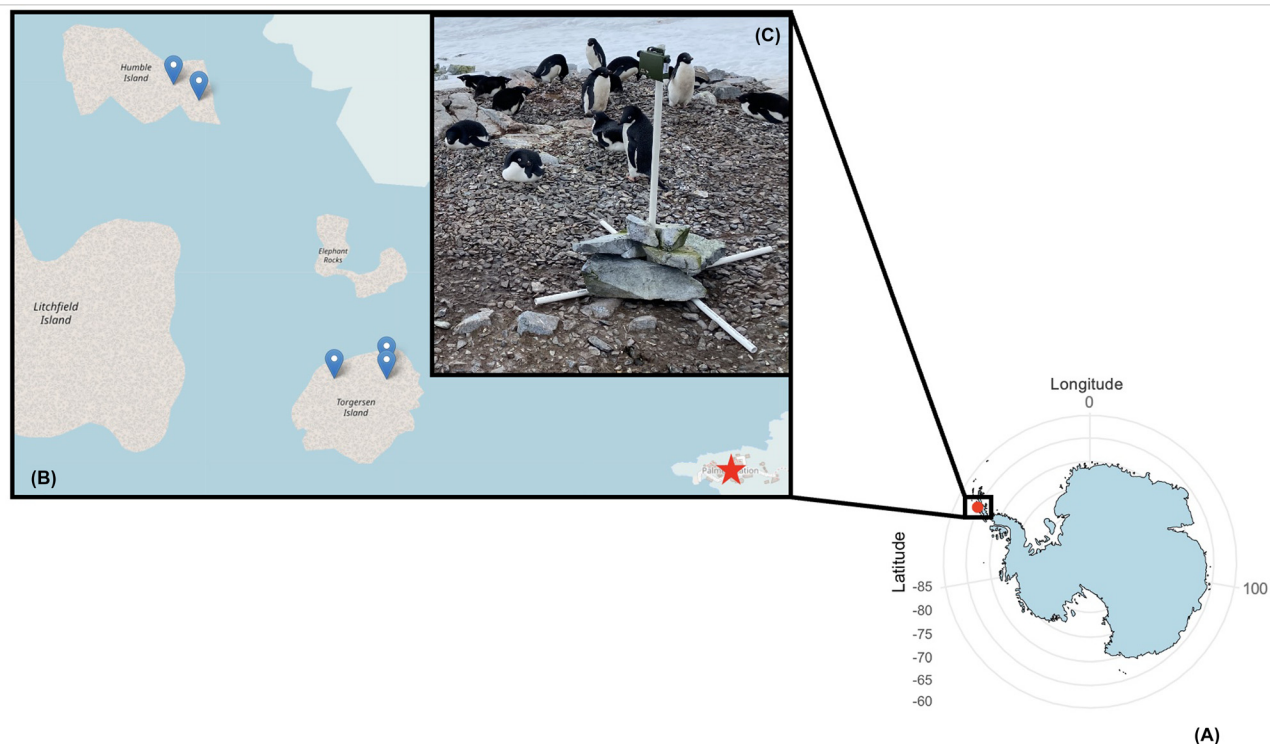


FIG. 1. (A) Map of Antarctica depicts the red point indicating the study area based at Palmer Station in the West Antarctic Peninsula. (B) Inset map shows the five colony locations designated with markers across Humble and Torgersen Islands, with Palmer Station denoted by the star. (C) Typical recording rig is placed near the edge of an Adélie penguin colony. Torgersen and Humble Islands are <1 km apart.

probability level of a correct detection can be determined using the unseen test data before running the model over the full dataset. A visualization of the six-step workflow used in this study is shown in Fig. 2, and a detailed step-by-step procedure, along with our training dataset, can be found in the [supplementary material](#).

1. Step 1: Creating the training dataset

We randomly selected files to extract clips from to build the training dataset of the custom BirdNET models (see the [supplementary material](#)). These files were excluded from the test dataset. A custom model requires training data made up of audio clips that are 3 s or less in duration for each target signal, or class, as well as nontarget signals, or background. Including background training data helps BirdNET distinguish the target signals from nontarget signals in models

with more than one class (Symes *et al.*, 2023). We selected one hundred 3 s clips each of the following classes: low wind, medium wind, and high wind. We also selected one hundred 3 s clips of background, which, in this study, were Adélie penguin adult vocalizations (Symes *et al.*, 2023). Following Terranova *et al.* (2024), we defined low wind as 3 s of continuous wind with no clipping [Fig. 3(e)], medium wind as 3 s of continuous wind where clipping is present but masks less than 50% of the 3 s selection [i.e., <1.5 s; Fig. 3(f)], and high wind as 3 s of continuous wind, where clipping masks $\geq 50\%$ of the 3-s selection [i.e., ≥ 1.5 s; Fig. 3(g)]. All class clips contained only wind-generated noise, excluding penguin vocalizations and other dominant sounds such as seabirds, seals, and rocks. All background clips contained only dominant Adélie penguin adult vocalizations with minimal wind presence [Fig. 3(h)]. When there were no classes of interest present, the model was expected to not make a detection.

We, first, explored simpler data cleaning methods before turning to machine learning. Initially, we assessed the compatibility of wind data collected by Palmer Station weather station located <1 km from our study sites, but site-specific topography caused poor alignment with the wind levels recorded by the SMMs. Next, we tried to remove clipped audio, but manual review showed that penguin vocalizations also caused clipping, hence, removing all clipped segments would have erased key behavioral data. Finally, we attempted to remove wind through frequency filtering. Frequency spectra analysis of the classes and background reveal strong energy overlap with all three wind levels and the fundamental frequency of Adélie penguin vocalizations [Figs. 3(a)–3(d)]. This overlap caused an initial frequency filtering approach to fail. After these methods failed, we applied a CNN to identify and remove data that was compromised by high wind.

2. Step 2: Creating the test dataset

To evaluate the custom model, a test dataset should be developed using unseen data that are separate from the training dataset (Symes *et al.*, 2023). We randomly selected 13 recordings from each of our 6 rigs, totaling seventy-eight 5-min files for our test dataset (Terranova *et al.*, 2024). These recordings represented 0.6% of the total dataset, which is comparable to the 0.5% threshold used by Terranova *et al.* (2024) for their test dataset. We manually annotated the test dataset in Raven Pro (version 1.6.5; Ithaca, NY). To match the 3-s bins that BirdNET uses to process audio data, we generated 3-s fixed-duration selection boxes (see the [supplementary material](#)). Each 5-min file had 100 selections manually annotated for the presence of penguins and wind. Annotators assigned each selection to one of four wind categories: no wind, low wind, medium wind, or high wind, following the same definitions used to select the training clips. They also noted presence of penguin vocalizations and other dominant noise (rocks, seals, and seabirds).

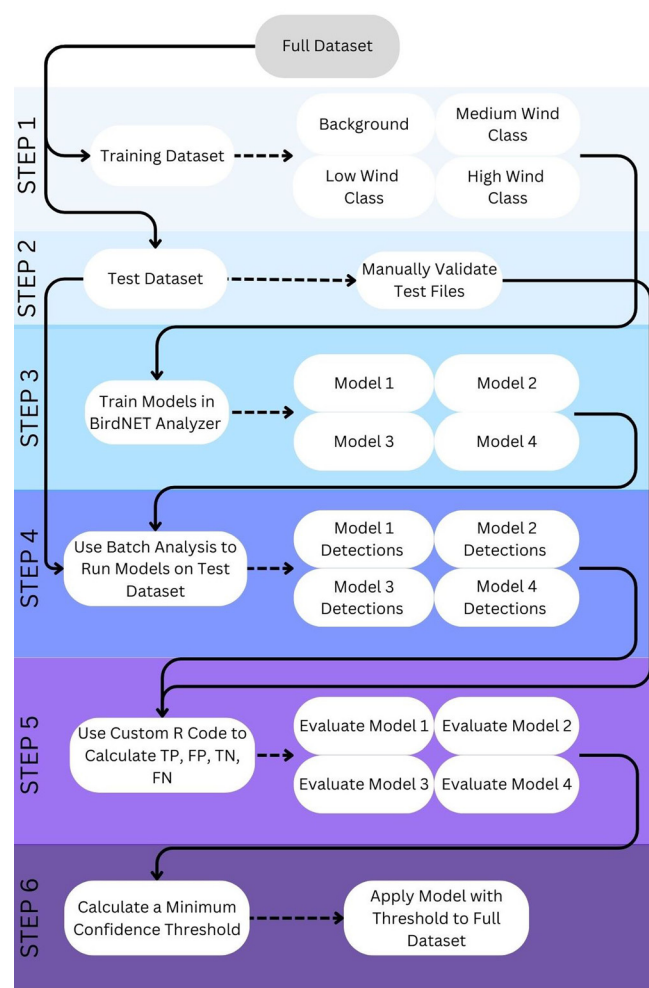


FIG. 2. Workflow to successfully implement our custom wind BirdNET model is depicted using six steps: (1) create a training dataset, (2) generate a test dataset, (3) train custom models, (4) run the custom models on the test dataset, (5) compare the custom model output of the test data to manually annotated output of the test data to evaluate model performance, and (6) calculate a minimum confidence threshold. Dashed lines indicate the output within each step and solid lines show where output from a step incorporates into a secondary step.

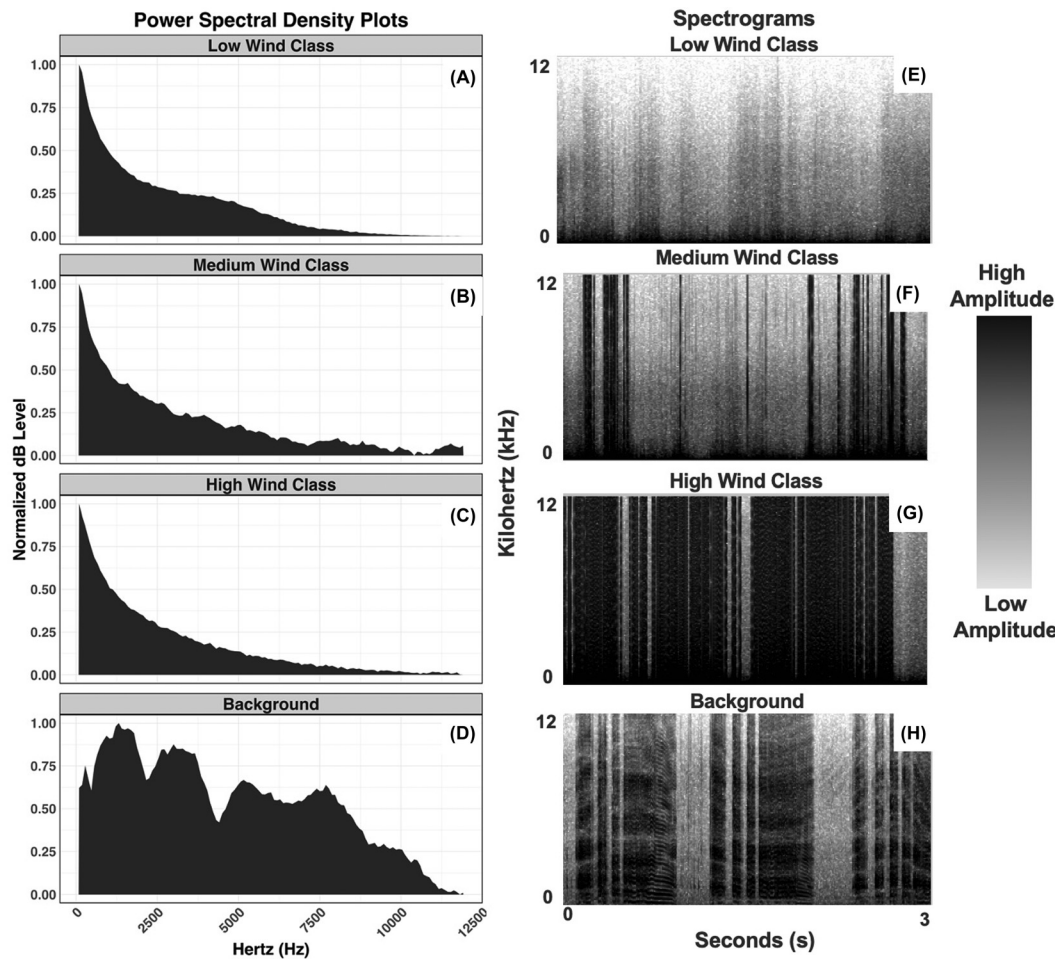


FIG. 3. Example power spectral density plots [(A)–(D)] and spectrograms [(E)–(H)] of 3-s clips from the training dataset. (A), (E) Low wind, in which wind is present but does not clip the recording; (B), (F) medium wind, in which wind is present and clips less than 50% of the recording; (C), (G) high wind, in which wind is present and clips more than 50% of the recording; and (D), (H) background, in which adult Adélie penguin vocalizations are considered “noise” for our wind detector are displayed.

3. Step 3: Training the models in the BirdNET GUI

We followed protocol from Symes *et al.* (2023) and used the BirdNET Analyzer GUI (GUI v1.2.0, model v2.4), to train four models (Table I). According to BirdNET custom model documentation, default parameters are typically sufficient to train a model. However, we included autotune models for comparison because the autotune setting iterates over multiple training runs with different hyperparameters and uses Bayesian optimization to determine ideal hyperparameters for the training data. We also evaluated models with and without the low wind class to assess its impact on model performance as this

TABLE I. Model name, classes, and training parameters for the four custom models. BirdNET training settings used for each model can be found in the supplementary material.

Model	Classes	Training parameters
One	Low, medium, and high winds	Default
Two	Low, medium, and high winds	Autotune
Three	Medium and high winds	Default
Four	Medium and high winds	Autotune

class could be confused with other low-frequency transient sounds (see the [supplementary materials](#) for model training setting). Through the training process, BirdNET splits the training dataset into 80% training data and 20% validation data. Each training event generates a graph showing the area under the precision-recall curve (AUPRC) and the area under the receiver operating characteristic (AUROC) on a scale of 0–1 by assessing how the trained model performs overall (all classes combined) on the validation data over each epoch. BirdNET then compares the true positive (TP) rates and false positives (FPs) rate for the AUROC and precision and recall for AUPRC for each epoch. AUPRCs and AUROCs near one indicate a model that can clearly distinguish between classes, whereas values at 0.5 suggest that a model cannot clearly distinguish between classes better than random chance.

4. Step 4: Generating model detections

To run the trained models on our unseen test dataset, we used the “batch analysis” tab with default settings in the BirdNET Analyzer GUI to apply each custom model to the 78 test audio files (see the [supplementary material](#)). The

batch analysis generates a model output, or detection file, for each test audio file.

5. Step 5: Evaluating the models

To evaluate our models, we compared the BirdNET detections to the manual annotations for each of the 78 test recordings and calculated the number of TPs, FPs, true negatives (TN), and false negatives (FN) using custom *R* code provided in the [supplementary material](#). Then, we assessed the recall, precision, F-score, and accuracy of our custom models (Grandini *et al.*, 2020; Knight *et al.*, 2017; Symes *et al.*, 2023), as well as precision and recall for each class (Pérez-Granados, 2023; Table II). We set the F-score with $\beta = 1$, giving equal weight to precision and recall (Knight *et al.*, 2017).

6. Step 6: Determining a minimum confidence threshold

BirdNET produces a confidence score for each detection that it makes. Importantly, this confidence score should not be interpreted as an indication of the model's certainty about the validity (TP) of the detection (Symes *et al.*, 2023). To have true confidence in the BirdNET detections, one can calculate the minimum confidence threshold required to achieve a 90% probability of a true detection using a binomial logistic regression with protocol from Symes *et al.* (2023). We calculated our minimum confidence threshold by running our best model, model one (Fig. 4, Table III), on the unseen test data through the batch analysis tab with a minimum confidence threshold of 0.0 (see the [supplementary material](#)). We used the resulting detection confidence scores to calculate a minimum confidence threshold for each class using the binomial logistic regression protocol. Once the minimum threshold was determined, this confidence score was applied with model one through batch analysis over the entire dataset (see the [supplementary material](#)), conservatively limiting detections to those most likely to be true (i.e., minimizing FPs).

C. Evaluating the impact of confidence thresholds on model metrics

The impact of confidence thresholds on model metrics is poorly understood (Pérez-Granados, 2023). To address this gap, we calculated model evaluation metrics (Table II) across minimum confidence thresholds values (0.0–0.9 in 0.1 increments) using model one and batch analysis on the unseen test data (see the [supplementary material](#)). These metrics were then plotted across confidence threshold increments for the overall performance of model one (all classes combined). Additionally, precision and recall were plotted for each class across confidence threshold increments.

III. RESULTS

A. Step 3: Model Area Under the Curve training curves

Overall, the AUPRC closely followed but lagged behind the AUROC. For the default models (Fig. 4), both curves converged near 1 well before 50 epochs, indicating that these models could effectively distinguish between class types during training. In contrast, the autotune models did not reach 50 epochs as a result of the validation metrics failing to improve, and their curves did not converge near 1 (Fig. 4). Although all models have AUPRC and AUROC curves above 0.5, indicating performance better than random chance, models 1 and 3 with default training were superior.

B. Step 5: Model evaluation

From our manually annotated 78-file test dataset, which contained a total of 7800 3-s segments, we labeled 3243 segments as low wind, 576 segments as medium wind, and 64 segments as high wind. Table III presents the calculated F-score, accuracy, precision, and recall for each of the four models overall by calculating the total number of TPs, FPs, TNs, and FNs for each model and putting those values into the respective equations (Symes *et al.*, 2023). Precision and

TABLE II. Metric name, formula, description, and interpretation for the four metrics used to evaluate the custom models.

Metric	Equation	Description	Interpretation
Precision	$TP / (TP + FP)$	Precision assesses the model's ability to correctly assign a detected signal of interest to the correct classification (range, 0–1)	A high precision value indicates that the model has correctly assigned each target to its correct category (low number of FPs)
Recall	$TP / (TP + FN)$	Recall assesses how well a model can detect a signal of interest in a dataset (range, 0–1)	A high recall value indicates that the model has detected nearly all target signals in the dataset (low number of FNs)
F-score	$\frac{(\beta^2 + 1) * \text{precision} * \text{recall}}{\beta^2 * \text{precision} + \text{recall}}$	F-score combines the precision and recall into a single value with the ability to weight precision or recall more heavily based off of β being greater than or less than one (Sokolova <i>et al.</i> , 2006; range, 0–1)	An F-score value close to one indicates a well-performing model, where both precision and recall are high
Accuracy	$TP + TN / (TP + FP + TN + FN)$	Accuracy shows the percentage of times that a model correctly identified a detection or non-detection out of the whole test dataset (range, 0–1)	An accuracy value close to one (100%) indicates a model that correctly identifies a detection compared to non-detection and assigns the detection to the correct class

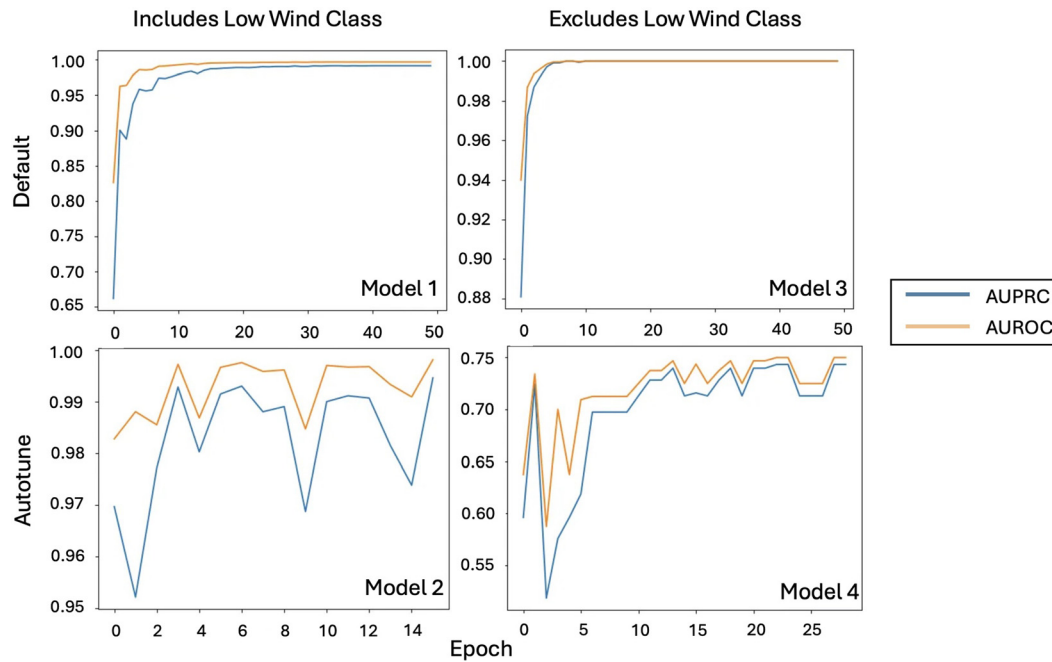


FIG. 4. The AUPRCs and AUROCs for models 1–4.

recall were also calculated for each class in all four models. Based off of the results from Table III, Fig. 4, and the goal of detecting high wind, we selected model one (default training settings with low, medium, and high wind classes)

TABLE III. F-score and accuracy for the models overall (containing all wind classes) and precision, recall, TP, FP, and FN values for the models overall and for each class are reported at the default batch analysis 0.5 minimum confidence threshold. TP, FP, FN, and TN raw values are given. Shaded boxes indicate the highest value for each metric across models (the top two models were highlighted when all four models could be compared). NA, not applicable.

Metric	Classes	Model 1	Model 2	Model 3	Model 4
F-score	Overall	0.43	0.57	0.33	0.30
Accuracy	Overall	0.53	0.49	0.79	0.84
Precision	Overall	0.52	0.42	0.22	0.23
	Low	0.48	0.44	—	—
	Medium	0.92	0.94	0.20	0.23
	High	0.76	0.09	0.71	NA
Recall	Overall	0.37	0.90	0.65	0.43
	Low	0.35	0.91	—	—
	Medium	0.38	0.75	0.62	0.48
	High	0.94	1	0.97	0
TPs	Overall	1393	3159	401	272
	Low	1146	2930	—	—
	Medium	186	165	339	272
	High	61	64	62	0
FPs	Overall	1272	4366	1387	895
	Low	1236	3696	—	—
	Medium	17	11	1362	895
	High	19	659	25	0
FNs	Overall	2399	366	213	359
	Low	2087	307	—	—
	Medium	308	59	211	295
	High	4	0	2	64
TNs	Overall	2753	683	5803	6274

as our best model for detecting wind because of its superior performance during training and relatively high overall performance, especially for the high wind class (recall, 0.94; precision, 0.76).

C. Step 6: Minimum confidence threshold

Model TPs, FPs, and logistic regression determined the following confidence thresholds at which model one predicted a 90% probability of a correct detection for each class: low wind, 4.99 (unachievable as confidence score is bounded by 0 and 1); medium wind, 0.56; and high wind, 0.91 (Fig. 5). The model detected 80 743 3-s segments of medium wind in the full dataset with the 0.56 minimum confidence score. The model detected 11 717 3-s segments of high wind in the full dataset with the 0.91 minimum confidence score. No 3-s segments of low wind were considered as true detections because no confidence score could achieve a 90% probability threshold (Fig. 5).

D. Evaluating the impact of confidence thresholds on model metrics

As the minimum confidence threshold increased, precision increased, F-score and recall decreased, and accuracy plateaued for model one (Fig. 6). Each class had different intersection points where precision and recall were maximized.

IV. DISCUSSION

A. Overall findings

We aimed to determine the validity of applying the BirdNET architecture to identify nontarget wind signals to remove them as part of the data cleaning process for

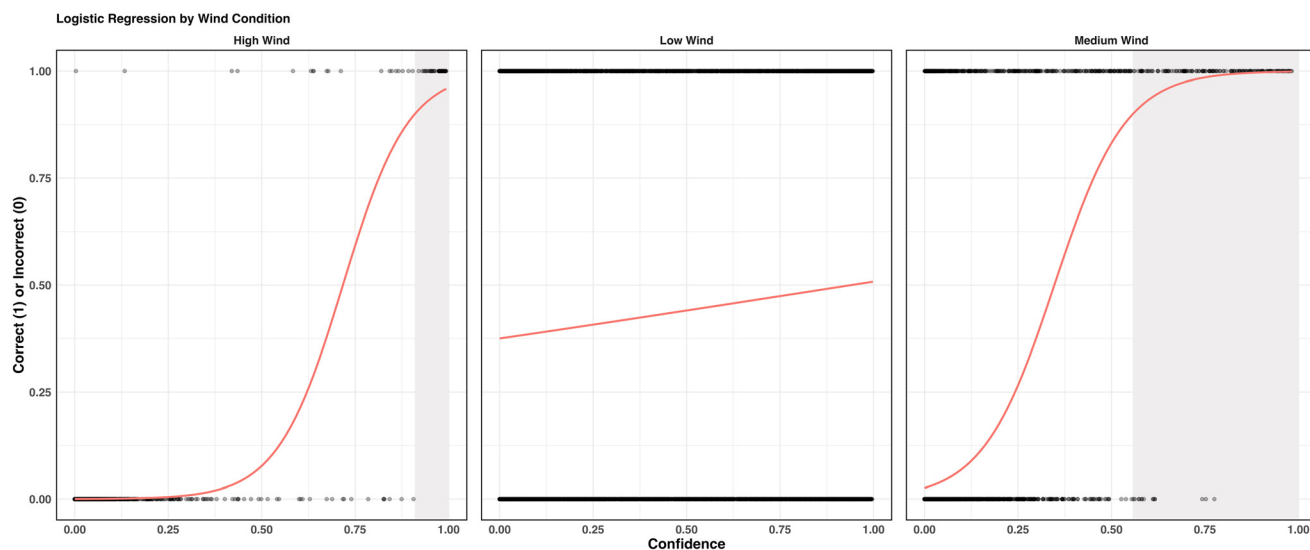


FIG. 5. Binomial logistic regressions fit to model one low, medium, and high wind class detections. Data points represent either a TP (1) or FP (0) on the y axis and the model confidence score assigned to that detection on the x axis. The regression line indicates the probability that a detection is correct at a specific confidence score. The shaded region represents the confidence thresholds at which there is at least a 90% probability of a TP for a specific class.

bioacoustic investigation. Through the creation and evaluation of multiple models, we developed a six-step method for building custom BirdNET models, evaluating their effectiveness and determining a minimum confidence threshold before applying them to an entire dataset. Our results demonstrate that BirdNET can successfully identify high-wind conditions with the best model, model one (default training setting, with low, medium, and high wind classes). This overall model achieved moderate performance scores with an overall F-score of 0.43, accuracy of 0.53, precision of 0.52, and recall of 0.37. However, the high wind class, which was most consequential for our intended application of detecting extreme wind, performed well with a precision of 0.76, recall of 0.94 (Table III), and a 0.91 minimum confidence threshold yielding a 90% probability of a TP. By using BirdNET to automate the detection of unwanted

sound, we can build a pipeline to identify and remove affected files, efficiently cleaning a bioacoustic dataset for analysis to focus on files of interest.

B. Custom model recommendations in BirdNET

The creation and comparison of multiple models revealed best practices for building custom models in BirdNET. Among the three classes, low wind performed the worst, in which precision never exceeded 0.48 (Table III). The original deep architecture of BirdNET Analyzer is built exclusively off of bird calls (focal and non-focal) and soundscapes heavily featuring bird call and songs (Kahl *et al.*, 2021). BirdNET, therefore, will perform best on sounds that have discrete durations that fall below 3 s, are characterized by stereotyped patterns, and fall in the frequency range that

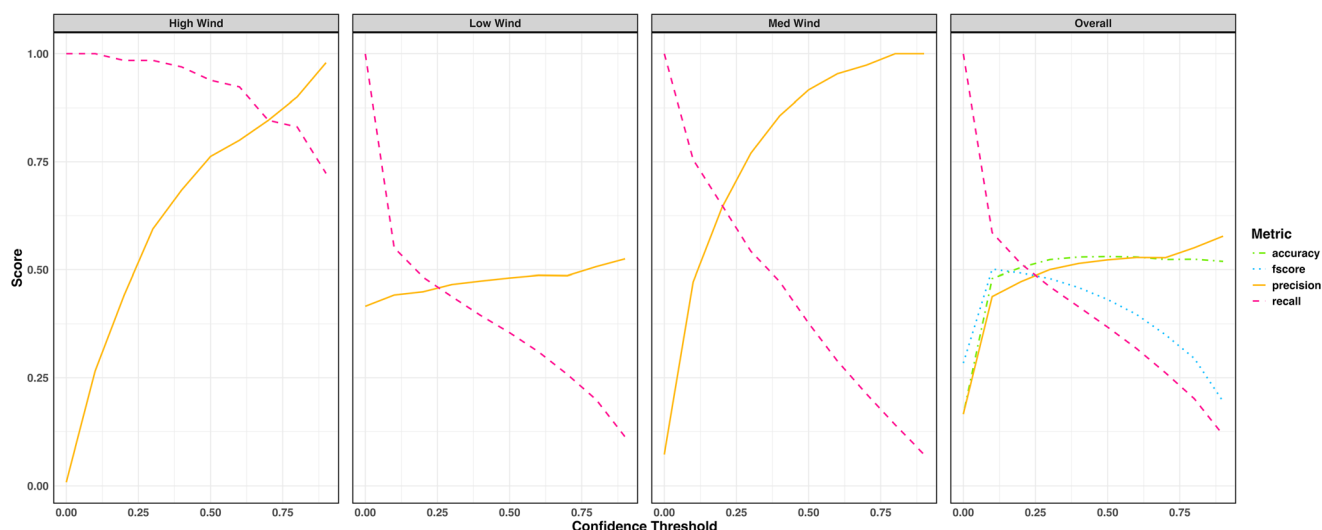


FIG. 6. Model evaluation metrics plotted over minimum confidence threshold for model one for each wind class and the overall model.

is typical of bird songs. Low wind, which lacks a set pattern or duration and is concentrated in low frequencies, can easily be mistaken in a CNN, which analyzes spectrogram images, as other non-discrete low-frequency noises, such as those caused by waves on the shore or machinery, as evidenced by the high ratio of FPs to TPs in the low wind class (Table III). Thus, we recommend using the BirdNET analyzer for detecting sounds that have discrete durations and consistent patterns, such as medium or high-wind conditions with considerable clipping that manifests as energy in frequencies higher than those characteristic of true wind noise [Figs. 3(a)-3(c)].

The BirdNET autotune models—models two and four—did not exhibit smooth AUPRCs or AUROCs. As the autotune setting iterates over multiple training runs with varying hyperparameters in an attempt to optimize performance, the failure to converge near one suggests that an ideal set of hyperparameters could not be identified and the autotune models failed to clearly distinguish between classes during training. The default models, however, could clearly distinguish between class types during training, as both models one and three converged near 1 within 20 epochs. Autotune has been successfully used in other bioacoustic applications of BirdNET focused on species with stereotypical calls (Kimura *et al.*, 2025), suggesting that the lack of convergence with autotune models in our study may be explained by the nature of unpredictable wind noise. We, therefore, recommend training models under the default settings for wind-type noises. This recommendation aligns with Symes *et al.* (2023), who state that default settings will be sufficient for most custom models.

The number of classes used in the model influenced performance. The default model without low wind (model three) did not perform as well in comparison to our default model with low wind (model one; Table III). Model three had a lower F-score and precision and performed poorly on the medium wind class. As the medium wind class encompasses a wide range (i.e., similar to low wind with only one clipped sample or similar to high wind with just under 50% of samples clipped), perhaps the presence of low wind class allows the model to more clearly distinguish the medium wind class. Research has found that the more classes a CNN has, the worse it performs (Luo *et al.*, 2019), but that finding may be negligible in the context of this study because our models only differ by one class.

Last, we recommend using a manually annotated test dataset paired with BirdNET model detection outputs to evaluate model performance rather than the “segments” tab, a built-in model assessment feature provided in the BirdNET Analyzer. Segments allows users to extract detections and assign TPs and FPs, which, with further analysis, can then be used to calculate precision, determine a minimum confidence threshold, and calculate probability of a detection being a TP. However, as the segments tab exclusively uses detections, it does not account for TNs and FNs, which provide essential context to model performance such as the ability to determine recall and accuracy. A manually

annotated test dataset will inherently include TNs and FNs through highlighting when the model fails to detect an annotated detection (FN) and the model and the annotations agree that there is no detection to be made (TN). We also recommend reporting the raw values of TP, FP, FN, and TN as this promotes transparency regarding potential sparsity in the test dataset, imbalances among classes, and their impact on evaluation metrics (Japkowicz, 2006).

C. Confidence threshold impacts on model metrics

As minimum confidence thresholds increased, precision increased and recall decreased. Our results follow an expected trade-off (Pérez-Granados, 2023; Sethi *et al.*, 2021) between FNs (which influence recall) and FPs (which influence precision) as confidence scores increase. In concurrence with Malamut (2022), we found that there are varying degrees to this trend’s intensity among classes in BirdNET (Fig. 6). As precision and recall maximum curves are inconsistent across classes, a single minimum confidence threshold may not be an appropriate decision depending on the research question. Therefore, we strongly recommend determining the 90% probability of a true detection for each class and restricting the detections for each class to those that are greater than or equal to their respective minimum confidence scores. Our minimum confidence thresholds of 0.56 and 0.91 yielded a 90% probability of a true detection for medium and high wind, respectively, and were slightly out of range of the 0.7–0.8 confidence threshold recommended by Sethi *et al.* (2021).

D. Limitations and future work

Limitations to this study include using data from only six recording rigs across one breeding season within a relatively simple soundscape dominated by one species. Including data from other locations to blindly test the model would allow us to test its broad applicability. Although this study helps answer questions related to precision, recall, and confidence thresholds laid out in Pérez-Granados (2023), future research is still needed to explore how to optimize training and batch analysis beyond BirdNET default settings. For example, future work could assess the impact of BirdNET training features such as learning rate, number of epochs, sensitivity, and overlap. We also encourage those who have developed custom models in BirdNET to publish similar results to inform best practices as the use of BirdNET in the bioacoustics community continues to expand.

V. CONCLUSION

The custom wind model’s success lifts a significant roadblock in PAM as a viable tool for monitoring Adélie penguins through the reduction of clipping from wind in a dataset. Although this approach will not perform well on low wind, its influence on the penguin dataset could be easily reduced with a high bandpass filter. Ecological analyses, such as estimating breeding phenology, can now take place on this acoustic dataset to further develop PAM as a

potentially critical conservation tool for monitoring climate change impacts on Adélie penguins. The workflow that we present in this study is, to our knowledge, the first use of BirdNET's custom model feature to detect unwanted sounds as part of a data cleaning step prior to analysis. This workflow will help automate the tedious process of manually identifying files with high background noise that may be irreparably corrupted by clipping or could contain significantly masked vocalizations and altered wildlife behaviors. As climate-induced wind speeds increase, this pipeline furthers the utility of PAM to monitor ecological indicators long-term and provide inferences on climate change impacts.

SUPPLEMENTARY MATERIAL

See [supplementary material](#) for a step-by-step procedure on model training and batch analysis, as well as our training dataset, example R code, an example BirdNET detection file, and an example manually annotated test file.

ACKNOWLEDGMENTS

Funding for this study was provided by support from National Science Foundation EAGER Grant No. OPP 2026045 and Award No. 2226886. Our sincere thanks to Darren Roberts and Megan Roberts for data collection. We thank Sonja Ahlberg for assistance in manually validating the test dataset and Barbara Spiecker for guidance on figure development. We also thank members of the Ecological Acoustics and Behavior Laboratory and the Quantitative Marine Ecology Laboratory at the University of New Hampshire for their encouragement and feedback on the manuscript.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Abrahams, C., and Geary, M. (2020). "Combining bioacoustics and occupancy modelling for improved monitoring of rare breeding bird populations," *Ecol. Indic.* **112**, 106131.

Ainley, D., Russell, J., Jenouvrier, S., Woehler, E., Lyver, P. O., Fraser, W. R., and Kooyman, G. L. (2010). "Antarctic penguin response to habitat change as Earth's troposphere reaches 2°C above preindustrial levels," *Ecol. Monogr.* **80**, 49–66.

Ainley, D. G. (2002). *The Adélie Penguin: Bellwether of Climate Change* (Columbia University Press, New York).

Ancel, A., Beaulieu, M., and Gilbert, C. (2013). "The different breeding strategies of penguins: A review," *C. R. Biol.* **336**, 1–12.

Anunciação, P. R., Sugai, L. S. M., Martello, F., de Carvalho, L. M. T., and Ribeiro, M. C. (2022). "Estimating the diversity of tropical anurans in fragmented landscapes with acoustic monitoring: Lessons from a sampling sufficiency perspective," *Biodivers. Conserv.* **31**, 3055–3074.

Arneill, G. E., Critchley, E. J., Wischniewski, S., Jessopp, M. J., and Quinn, J. L. (2020). "Acoustic activity across a seabird colony reflects patterns of within-colony flight rather than nest density," *Ibis* **162**, 416–428.

Bateman, H. L., Riddle, S. B., and Cubley, E. S. (2021). "Using bioacoustics to examine vocal phenology of neotropical migratory birds on a wild and scenic river in Arizona," *Birds* **2**, 261–274.

Beaulieu, M., Thierry, A.-M., Le Maho, Y., Ropert-Coudert, Y., and Ancel, A. (2009). "Alloparental feeding in Adélie penguins: Why is it uncommon?," *J. Ornithol.* **150**, 637–643.

Black, C. E. (2016). "A comprehensive review of the phenology of *Pygoscelis* penguins," *Polar Biol.* **39**, 405–432.

Borker, A. L., Mckown, M. W., Ackerman, J. T., Eagles-Smith, C. A., Tershy, B. R., and Croll, D. A. (2014). "Vocal activity as a low cost and scalable index of seabird colony size: Automated acoustic seabird monitoring," *Conserv. Biol.* **28**, 1100–1108.

Borowicz, A., McDowall, P., Youngflesh, C., Sayre-McCord, T., Clucas, G., Herman, R., Forrest, S., Rider, M., Schwaller, M., Hart, T., Jenouvrier, S., Polito, M. J., Singh, H., and Lynch, H. J. (2018). "Multi-modal survey of Adélie penguin mega-colonies reveals the Danger Islands as a seabird hotspot," *Sci. Rep.* **8**, 3926.

Bracegirdle, T. J., Connolly, W. M., and Turner, J. (2008). "Antarctic climate change over the twenty first century," *J. Geophys. Res.: Atmos.* **113**, D03103, <https://doi.org/10.1029/2007JD008933>.

Brosseau, J. E., Eddington, V. M., Craig, E. C., White, E. R., and Kloepper, L. N. (2024). "The effect of localized disturbance on the acoustic behavior of the common tern (*Sterna hirundo*)," *JASA Express Lett.* **4**, 091201.

Brunton, D., Rodrigo, A., and Marks, E. (2010). "Ecstatic display calls of the Adélie penguin honestly predict male condition and breeding success," *Behaviour* **147**, 165–184.

Buxton, R. T., and Jones, I. L. (2012). "Measuring nocturnal seabird activity and status using acoustic recording devices: Applications for island restoration," *J. Field Ornithol.* **83**, 47–60.

Campos-Cerqueira, M., and Aide, T. M. (2016). "Improving distribution data of threatened species by combining acoustic monitoring and occupancy modelling," *Methods Ecol. Evol.* **7**, 1340–1348.

Carney, K. M., and Sydeman, W. J. (1999). "A review of human disturbance effects on nesting colonial waterbirds," *Waterbirds: Int. J. Waterbird Biol.* **22**, 68–79.

Cimino, M. A., Lynch, H. J., Saba, V. S., and Oliver, M. J. (2016). "Projected asymmetric response of Adélie penguins to Antarctic climate change," *Sci. Rep.* **6**, 28785.

Clarke, B., Otto, F., Stuart-Smith, R., and Harrington, L. (2022). "Extreme weather impacts of climate change: An attribution perspective," *Environ. Res.: Clim.* **1**, 012001.

Colombelli-Négrel, D. (2023). "Estimating little penguin population sizes using automated acoustic monitoring and citizen science," *Ibis* **165**, 1423–1431.

Convey, P., Bindschadler, R., di Prisco, G., Fahrback, E., Gutt, J., Hodgson, D. A., Mayewski, P. A., Summerhayes, C. P., and Turner, J. (2009). "Antarctic climate change and the environment," *Antarct. Sci.* **21**, 541–563.

Digby, A., Towsey, M., Bell, B. D., and Teal, P. D. (2014). "Temporal and environmental influences on the vocal behaviour of a nocturnal bird," *J. Avian Biol.* **45**, 591–599.

Furnas, B. J., and Callas, R. L. (2015). "Using automated recorders and occupancy models to monitor common forest birds across a large geographic region," *J. Wildl. Manag.* **79**, 325–337.

Giese, M. (1996). "Effects of human activity on Adélie penguin *Pygoscelis adeliae* breeding success," *Biol. Conserv.* **75**, 157–164.

Grandini, M., Bagli, E., and Visani, G. (2020). "Metrics for multi-class classification: An overview," available at <http://arxiv.org/abs/2008.05756> (Last viewed November 5, 2024).

Hart, K. A., Oppel, S., Humphries, G. R. W., Blackburn, A., and Nam, K.-B. (2021). "Estimating streaked shearwater *Calonectris leucomelas* abundance in the Republic of Korea using automated acoustic recorders," *Mar. Ornithol.* **49**, 109–117.

Jahn, O., Ganchev, T. D., Marques, M. I., and Schuchmann, K.-L. (2017). "Automated sound recognition provides insights into the behavioral ecology of a tropical bird," *PLoS One* **12**, e0169041.

Japkowicz, N. (2006). "Why question machine learning evaluation methods? An illustrative review of the shortcomings of current methods," in *21st National Conference on Artificial Intelligence (AAAI, Boston, MA)*, pp. 6–11.

- Jouventin, P. (1982). *Visual and Vocal Signals in Penguins, Their Evolution and Adaptive Characters* (Parey, Berlin, Germany).
- Juodakis, J., and Marsland, S. (2022). "Wind-robust sound event detection and denoising for bioacoustics," *Methods Ecol. Evol.* **13**, 2005–2017.
- Kahl, S., Wood, C. M., Eibl, M., and Klinck, H. (2021). "BirdNET: A deep learning solution for avian diversity monitoring," *Ecol. Inf.* **61**, 101236.
- Kimura, K., Fukuyama, I., and Fukuyama, K. (2025). "Deep learning-based detector of invasive alien frogs, *Polypedates leucomystax* and *Rhinella marina*, on an island at invasion front," *Biol. Invasions* **27**, 95.
- Klopper, L. N., Linnenschmidt, M., Blowers, Z., Branstetter, B., Ralston, J., and Simmons, J. A. (2016). "Estimating colony sizes of emerging bats using acoustic recordings," *R. Soc. Open Sci.* **3**, 160022.
- Knight, E. C., Hannah, K. C., Foley, G. J., Scott, C. D., Brigham, R. M., and Bayne, E. (2017). "Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs," *Avian Conserv. Ecol.* **12**(2), 14.
- Küsel, E. T., Mellinger, D. K., Thomas, L., Marques, T. A., Moretti, D., and Ward, J. (2011). "Cetacean population density estimation from single fixed sensors using passive acoustics," *J. Acoust. Soc. Am.* **129**, 3610–3622.
- Leach, E. C., Burwell, C. J., Ashton, L. A., Jones, D. N., and Kitching, R. L. (2016). "Comparison of point counts and automated acoustic monitoring: Detecting birds in a rainforest biodiversity survey," *Emu - Austral Ornithol.* **116**, 305–309.
- Luo, C., Li, X., Yin, J., He, J., Gao, D., and Zhou, J. (2019). "How does the data set and the number of categories affect CNN-based image classification performance?," *JSW* **14**, 168–181.
- Malamut, E. J. (2022). "Using autonomous recording units and image processing to investigate patterns in avian singing activity and nesting phenology," available at <https://escholarship.org/uc/item/92p9z0gp> (Last viewed November 25, 2024).
- Marques, T. A., Thomas, L., Martin, S. W., Mellinger, D. K., Ward, J. A., Moretti, D. J., Harris, D., and Tyack, P. L. (2013). "Estimating animal population density using passive acoustics," *Biol. Rev.* **88**, 287–309.
- Marshall, G. J. (2003). "Trends in the southern annular mode from observations and reanalyses," *J. Clim.* **16**, 4134–4143.
- Nelke, C. M., and Vary, P. (2015). "Wind noise short term power spectrum estimation using pitch adaptive inverse binary masks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, Australia (April 19–24) (Institute of Electrical and Electronics Engineers, New York), pp. 5068–5072.
- Nelson, J. B., and Baird, P. H. (2001). "Seabird communication and displays," in *Biology of Marine Birds*, edited by E. A. Schreiber and J. Burger (CRC Press, Boca Raton, FL), 52 pp.
- Nemeth, E., and Brumm, H. (2010). "Birds and anthropogenic noise: Are urban songs adaptive?," *Am. Nat.* **176**, 465–475.
- Oliver, R. Y., Ellis, D. P. W., Chmura, H. E., Krause, J. S., Pérez, J. H., Sweet, S. K., Gough, L., Wingfield, J. C., and Boelman, N. T. (2018). "Eavesdropping on the Arctic: Automated bioacoustics reveal dynamics in songbird breeding phenology," *Sci. Adv.* **4**, eaaq1084.
- Oppel, S., Hervias, S., Oliveira, N., Pipa, T., Silva, C., Geraldies, P., Goh, M., Immler, E., and McKown, M. (2014). "Estimating population size of a nocturnal burrow-nesting seabird using acoustic monitoring and habitat mapping," *Nat. Conserv.* **7**, 1–13.
- Parish, T. R. (1988). "Surface winds over the Antarctic continent: A review," *Rev. Geophys.* **26**, 169–180, <https://doi.org/10.1029/RG026i001p00169>.
- Pearman, P. B., Guisan, A., and Zimmermann, N. E. (2011). "Impacts of climate change on Swiss biodiversity: An indicator taxa approach," *Biol. Conserv.* **144**, 866–875.
- Pérez-Granados, C. (2023). "BirdNET: Applications, performance, pitfalls and future opportunities," *Ibis* **165**, 1068–1075.
- Pérez-Granados, C., and Schuchmann, K.-L. (2021). "Passive acoustic monitoring of the diet and annual vocal behavior of the Black and Gold Howler Monkey," *Am. J. Primatol.* **83**, e23241.
- Picciulin, M., Kéver, L., Parmentier, E., and Bolgan, M. (2019). "Listening to the unseen: Passive acoustic monitoring reveals the presence of a cryptic fish species," *Aquat. Conserv. Mar. Freshwater Ecosyst.* **29**, 202–210.
- Schofield, O., Cimino, M., Doney, S., Friedlaender, A., Meredith, M., Moffat, C., Stammerjohn, S., Van Mooy, B., and Steinberg, D. (2024). "Antarctic pelagic ecosystems on a warming planet," *Trends Ecol. Evol.* **39**, 1141–1153.
- Sethi, S. S., Fossøy, F., Cretois, B., and Rosten, C. M. (2021). "Management relevant applications of acoustic monitoring for Norwegian nature—The sound of Norway," Norsk institutt for naturforskning (NINA), Report 2064, Vol. 31, available at <https://brage.nina.no/nina-xmlui/handle/11250/2832294> (Last viewed November 25, 2024).
- Shonfield, J., and Bayne, E. M. (2017). "Autonomous recording units in avian ecological research: Current use and future applications," *Avian Conserv. Ecol.* **12**(1), 14.
- Siddig, A. A. H., Ellison, A. M., Ochs, A., Villar-Leeman, C., and Lau, M. K. (2016). "How do ecologists select and use indicator species to monitor ecological change? Insights from 14 years of publication in Ecological Indicators," *Ecol. Indic.* **60**, 223–230.
- Sokolova, M., Japkowicz, N., and Szpakowicz, S. (2006). "Beyond accuracy, F-Score and ROC: A family of discriminant measures for performance evaluation," in *AI 2006: Advances in Artificial Intelligence, Lecture Notes in Computer Science*, edited by A. Sattar and B. Kang (Springer, Berlin, Germany), Vol. 4304, pp. 1015–1021.
- Speirs, E. A. H., and Davis, L. S. (1991). "Discrimination by Adélie penguins, *Pygoscelis adeliae*, between the loud mutual calls of mates, neighbours and strangers," *Anim. Behav.* **41**, 937–944.
- Stowell, D. (2022). "Computational bioacoustics with deep learning: A review and roadmap," *PeerJ* **10**, e13152.
- Sugai, L. S. M., Silva, T. S. F., Ribeiro, J. W., Jr., and Llusia, D. (2019). "Terrestrial passive acoustic monitoring: Review and perspectives," *BioScience* **69**, 15–25.
- Symes, L. B., Sugai, L. S. M., Gottesman, B. L., Pitzrick, M., and Wood, C. M. (2023). "Acoustic analysis with BirdNET and (almost) no coding: Practical instructions (version 0.9)," available at <https://doi.org/10.5281/zenodo.8357176> (Last viewed December 26, 2024).
- Terranova, F., Betti, L., Ferrario, V., Friard, O., Ludynia, K., Petersen, G. S., Mathevon, N., Reby, D., and Favaro, L. (2024). "Windy events detection in big bioacoustics datasets using a pre-trained convolutional neural network," *Sci. Total Environ.* **949**, 174868.
- TS.3.1.2 Spatial Distribution of Changes in Temperature, Circulation and Related Variables (2007). "AR4 WGI Technical Summary No. IPCC Fourth Assessment Report: Climate Change 2007," Intergovernmental Panel on Climate Change, available at https://archive.ipcc.ch/publications_and_data/ar4/wg1/en/tssts-3-1-2.html (Last viewed October 7, 2024).
- Turner, J., Barrand, N. E., Bracegirdle, T. J., Convey, P., Hodgson, D. A., Jarvis, M., Jenkins, A., Marshall, G., Meredith, M. P., Roscoe, H., Shanklin, J., French, J., Goosse, H., Guglielmin, M., Gutt, J., Jacobs, S., Kennicutt, M. C., II, Masson-Delmotte, V., Mayewski, P., Navarro, F., Robinson, S., Scambos, T., Sparrow, M., Summerhayes, C., Speer, K., and Klepikov, A. (2014). "Antarctic climate change and the environment: An update," *Polar Rec.* **50**, 237–259.
- Turner, J., Colwell, S. R., Marshall, G. J., Lachlan-Cope, T. A., Carleton, A. M., Jones, P. D., Lagun, V., Reid, P. A., and Iagovkina, S. (2005). "Antarctic climate change during the last 50 years," *Int. J. Climatol.* **25**, 279–294.
- Zhang, J., Xia, C., and Zhang, Y. (2021). "The daily vocal pattern of Adélie penguin during brooding period in the polar day," *Chin. J. Ecol.* **40**, 1098–1106.
- Zhao, K., Chen, G., Liu, Y., Møller, A. P., and Zhang, Y. (2022). "Population size assessment of Adélie penguin (*Pygoscelis adeliae*) chicks based on vocal activity rate index," *Global Ecol. Conserv.* **38**, e02263.