Deep learning for passive acoustic monitoring: how to study changing phenology in remote areas

Sylvain Christin[1], Éric Hervet[2], Paul A. Smith[3], Ray Alisauskas[4], Dominique Berteaux[5, 6], Glen Brown[7], Kyle Elliott[8], Jannik Hansen[9], Sandra Lai[5, 6], Jean-Francois Lamarre[10], Richard B. Lanctot[11], Christopher Latty[12], Audrey Le Pogam[5], Douglas MacNearney[3], Vijay P. Patil[13], Jennie Rausch[14], Daniel R. Ruthrauff[13], Sarah T. Saalfeld[11], Niels M. Schmidt[9], Andrew Tam[15], Francois Vézina[5], Øystein Varpe[16], Paul Woodard[14], Glenn Yannic[17], and Nicolas Lecomte*[1]

[1]Canada Research Chair in Polar and Boreal Ecology and Centre d'Études Nordiques, Department of Biology, University of Moncton, Moncton, NB, Canada

[2]Department of Computer Science, University of Moncton, Moncton, NB, Canada

[3]National Wildlife Research Centre, Environment and Climate Change Canada, Ottawa, ON K1A 0H3, Canada

[4]Environment and Climate Change Canada, Saskatoon, Saskatchewan, Canada

[5]Centre for Northern Studies, Quebec Centre for Biodiversity Science, Université du Québec à Rimouski, Rimouski, QC, Canada

[6]Canada Research Chair on Northern Biodiversity, Université du Québec à Rimouski, Rimouski, QC, Canada

[7]Ontario Ministry of Natural Resources & Forestry, Wildlife Research and Monitoring Section, Peterborough, ON, Canada

[8]Department of Natural Resource Sciences, McGill University, Montreal, QC, Canada

[9]Aarhus University, Department of Ecoscience, Roskilde, Denmark

[10]Polar Knowledge Canada, Cambridge Bay, Nunavut, Canada

[11]U.S. Fish and Wildlife Service, Migratory Bird Management, Anchorage, AK, USA

[12]U.S. Fish and Wildlife Service, Arctic National Wildlife Refuge, Fairbanks, Alaska, USA

[13]US Geological Survey, Alaska Science Center, 4210 University Drive, Anchorage AK, 99508, USA

[14]Canadian Wildlife Service, Environment and Climate Change Canada, Yellowknife, NT, Canada

[15]Department of National Defence, 8 Wing Canadian Forces Base Trenton, Astra, ON, Canada

[16]Department of Biological Sciences, University of Bergen, Bergen, Norway and Norwegian Institute for Nature Research, Bergen, Norway.

[17]Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LECA, 38000 Grenoble, France

Running title: Studying audio phenology with Deep learning

Keywords: Acoustics, deep learning, phenology, automated sound detection, avian soundscape

Type of article: Method

Word count:

- Abstract: 148
- Main text: 4751

Number of references: 51

Number of:

- Figures: 3

- Tables: 2
- Text boxes: 0

Corresponding author: Nicolas lecomte.

- E-mail: nicolas.lecomte@umoncton.ca.
- Address: Département de Biologie, Pavillon Rémi Rossignol, Université de Moncton, 18 Antonine Maillet, E1A 3E9, Moncton, NB, Canada
- Tel: (1) 506-858-4291

Author Contributions

This work is part of SC's PhD thesis supervised by NL and EH. NL and SC designed the study. SC developed the algorithms with inputs from NL and EH. SC and NL led the data sampling coordination. Model testing was led by SC, EH, and NL. NL and PAS provided the audio recording platform. SC led the writing of the first manuscript's version with input from NL. All authors contributed data, commented, and revised the manuscript.

Data Availability

All the code in this paper is available at https://github.com/vin985/BioSoundNet and will be made available on the Open Science Framework at https://dx.doi.org/10.17605/OSF.IO/4SFCR website upon acceptation of the manuscript.

Abstract

Understanding how species adjust to seasonality is fundamental in ecology, especially with rapidly increasing global air temperatures. Bioacoustic monitoring offers promise for tracking shifts in seasonal timing of vocal species, as recent automated sound recorders enable large-scale and long-term data collection. Yet, analyzing vast datasets necessitates automation and innovative detection methods. Here, we introduce BioSoundNet, a deep learning model designed for bird vocalization detection. Trained on field data and open-access databases, BioSoundNet achieved AUC scores of 0.88-0.93 and average precisions of 0.87-0.97 across five datasets spanning various ecosystems, and effectively captured the temporal patterns of avian acoustic activity at different time scales. Our findings underline the importance of evaluating models in ecological contexts and address the potential consequences of missing detections. Operating efficiently on standard computers, BioSoundNet is a robust tool for automated bird vocalization detection, providing a valuable resource for ecological phenology studies and acoustic dataset analysis.

1 Introduction

Understanding how species adjust their behaviour to seasonality is fundamental in many aspects of ecology and evolution (e.g. Varpe 2017). Deciphering the responses of species to abiotic variation is even more pressing with rapidly increasing global air temperatures caused by climate change (Pecl et al. 2017; Parmesan and Yohe 2003). At the forefront is phenology, i.e. the timing of life events such as flowering or migration (Menzel et al. 2020; Charmantier and Gienapp 2014). Yet, tracking how species adjust phenology is challenging, given the multitude of species involved, each characterized by complex trophic interactions and the need for repeated sampling across years to distinguish directional trends from inter-annual variability (Thackeray et al. 2016; Kwon et al. 2019; Visser and Both 2005; Schmidt et al. 2023).

As inadequate phenological responses can have detrimental effects on individual performance and even on populations, accurately quantifying how species adjust their timing is essential (Stenseth and Mysterud 2002; Samplonius et al. 2021). For instance, migratory birds often time their egg laying so that the peak availability of food coincides with the highest energy demands of their young (Visser, Holleman, and Gienapp 2006; Visser and Gienapp 2019; Saalfeld et al. 2021). Any mismatch between prey availability and these critical energy-dependent life stages can lead to adverse effects such as reduced offspring growth or reproductive success, ultimately resulting in population declines or disruption in food webs (Ross et al. 2017; Senner, Stager, and Sandercock 2017; Visser and Gienapp 2019; Nakazawa and Doi 2012; Samplonius et al. 2021).

In species that use sounds to communicate, one way to track phenological behaviour is with bioacoustic monitoring, which is the study of the sounds made by living organisms (Ross et al. 2023; Aide et al. 2013). Acoustic activity can be linked to important life stages for many taxa. For instance in birds, vocal activity of breeding individuals is closely linked to their nesting stages (Slagsvold 1977). By analyzing variations in vocal activity within and across seasons, it becomes possible to infer the nesting stage of birds through their level of vocal activity. In particular, identifying maxima of vocal activity could help determine the egg-laying period (Slagsvold 1977). This approach is facilitated by recent advances in bioacoustic technology that enable the deployment of small, robust, and cost-effective recorders in harsh environments (Christin and Lecomte 2023). This allows large-scale data collection over extended periods, even in remote areas (Gibb et al. 2019; Christin et al. 2023).

Yet, automated bioacoustic technology creates a new problem, as it generates vast amounts of data, which can span years of sound recordings. Manual analysis of sound recordings becomes unfeasible, necessitating the use of automated tools. Acoustic assessments of phenology have been done at the soundscape level, i.e. the whole acoustic landscape that encompasses all vocal species, by using acoustic indices, such as the Acoustic Complexity Index or the Bioacoustic Index (Farina, Pieretti, and Piccioli 2011; Buxton et al. 2016; Bateman, Riddle, and Cubley 2021; Krause and Farina 2016). These acoustic indices provide useful and fast analyzes of large amounts of sound data, but are hampered by other ambient noises such as weather or sounds related to anthropogenic activities (Buxton et al. 2018; Fairbrass et al. 2017). Another approach for analyzing large sound datasets involves the use of machine learning tools such as deep learning, which has demonstrated high accuracy in classification

and detection tasks (Kahl et al. 2021; LeCun, Bengio, and Hinton 2015; Christin, Hervet, and Lecomte 2019). Deep learning has been successfully applied to phenological studies of flowers and migration in birds, for instance (Mann et al. 2022; Van Doren et al. 2023). So far, however, most deep learning models have only focused on a few species and can be resource intensive. Moreover, deep learning tools are rarely presented in ecological contexts and mostly report classification accuracy metrics such as the F1-score, or average precision, without validating their ecological application (Kahl et al. 2021; Lostanlen et al. 2022; Ruff et al. 2020).

Here we introduce BioSoundNet, a reliable and fast performing deep learning avian vocalization detector, created for avian soundscape studies. By using carefully curated datasets tailored for this purpose, we demonstrate the effectiveness of our method for phenological studies. Furthermore, we stress the significance of considering ecological relevance in addition to model performance. Finally, we demonstrate that our approach is also particularly useful for segmentation, i.e. for isolating acoustic events, especially in regions with fewer singing species. This capability allows BioSoundNet to function as a versatile pre-processing tool for classification or the annotation of audio samples.

2 Material and Methods

2.1 Databases

To enhance the BioSoundNet model's applicability, we used a combination of publicly available sound databases and sound data gathered from our study sites (Table 1). We exclusively selected strongly annotated databases. Here, strongly annotated means that every vocaliza-

tion in an audio file is identified to the species level, and precise start and end times are specified. These databases served multiple purposes, including model training and performance assesment, as outlined in Table 1. In cases where a database was partitioned into training and evaluation sets, the evaluation data was never exposed to the models during the training process (Christin, Hervet, and Lecomte 2021). The next section describes a list of the databases used in this study:

2.1.1 Manually created databases

- **ArcticBirdSounds+**: This database expands upon the initial ArcticBirdSound database (Christin et al. 2023; Christin and Lecomte 2023) by incorporating additional annotated files derived from field data collected in 2018 and 2019 in the Arctic. The complete details of annotations and sites added can be found in Supporting Information Section 1.1. This aggregated database is coined ArcticBirdSound+ in Table 1.

- **ArcticChecked**: This database was created via a multi-step process. First, a previous iteration of the BioSoundNet model was applied to our field data. We then extracted audio samples corresponding to what the model predicted as bird vocalizations. Finally, we manually annotated them. This was done both to assess the performance of previous model iterations and to identify false positive cases. False positives were then treated as hard negatives (Delplanque et al. 2022; Shrivastava, Gupta, and Girshick 2016) to improve model performance. Sampling involved random selection from model detections, with a 0.5-second buffer added before and after each detection to streamline annotations. We ensured that selected samples provided a minimum of 10 minutes of

audio recordings for each deployment site. Detailed information on the content and origin of the dataset can be found in Supporting Information 1.2.

- **NABS_Gen**: This database was generated using excerpts sourced from the North American Bird Species (NABS) Database (Zhao 2018). The NABS database contains short, isolated vocalizations from 11 North-American birds species, extracted from field recordings available on the Xeno-Canto database (https://xeno-canto.org). As these extracts are quite short, we compiled them into 30-second audio recordings using the following rules:

    1. Create an empty 30-second audio file filled with random gaussian noise.
    2. Select a random excerpt from the NABS database and insert it seamlessly into our audio file while simultaneously creating an annotation file to document the vocalization's position and class.
    3. Repeat step 2 with a new random sample, ensuring no overlap between audio excerpts. This step is repeated until the file is deemed full, which occurs when the total duration of bird vocalizations is greater than a predetermined threshold. This threshold is different for every file and is randomly selected between 0% and 30% of the total duration of the file. This threshold has been selected to introduce intervals of silence between vocalizations.
    4. Once the file is full, repeat step 1 until all samples from the NABS dataset are exhausted.
    5. Repeat steps 1-3 for a second round to augment the sample count.

In total, following these procedures, we obtained 769 files, for a total of 6.3 hours of audio and 17,038 annotations from the NABS databse.

- **ArcticSummer**: This is a new strongly annotated database containing field recordings collected in Igloolik, NU, Canada during the summer of 2018. We chose thirty second audio samples at random every hour between June 11th 2018 and July 23rd 2018 and annotated them. This database was created for evaluation purposes only, providing a comprehensive overview into the fluctuations of vocal activity throughout an entire summer.

2.1.2 Publicly available databases

- **CityNet**: This is the database used to create the CityNet deep learning audio detector (Fairbrass et al. 2019). We used the same training and evaluation datasets as the original paper.

- **NIPS4b**: The NIPS4Bplus database was developed and described by (Morfi et al. 2019). In our study, 20% of NIPS4Bplus was first set aside for evaluation. The remaining data were split into training and validation datasets with a 80:20 ratio.

- **ENAB**: This is the database published by Chronister et al. (2021). It was used only for evaluation.

2.2 Model workflow

The BioSoundNet model has its roots in the CityNet model (Fairbrass et al. 2019), a model that provided good detection results in the London, UK urban soundscape. BioSoundNet

Table 1: Description of the datasets used during the training and evaluation steps of the BioSoundNet model creation. Number of classes, duration (h) and number of annotations (N) are presented for each dataset in the relevant category.

Dataset	Location	N classes	Training		Evaluation	
			Duration (h)	N	Duration (h)	N
ArcticBirdSounds+	8 sites across the Arctic	44	20.9	13,303	-	-
NABS_Gen	Northern United States	12	6.3	17,038	-	-
Citynet	London, UK	7	18.3	11442	0.67	789
NIPS4B	France and Spain	88	0.6	4,575	0.2	988
ArcticChecked	16 sites across the Arctic	30	6.3	2718	1.5	877
ArcticSummer	Igloolik, NU, Canada	37	-	-	8.6	5,260
ENAB	Northeastern United States	48	-	-	6.42	16,052

was created using the Mouffet framework (Christin and Lecomte 2022) and leveraged its flexibility for training and evaluating models. Here we present the steps used to train and evaluate our model. This workflow is summarized in Figure 1.

2.2.1 Data pre-processing

We loaded audio files using the librosa v0.9.2 (McFee et al. 2022) python library and transformed them into mel-scaled spectrograms with 32 bands. We resized all spectrograms using the Pillow python image library v9.4.0 (https://pypi.org/project/Pillow/) to ensure every spectrogram had a resolution of 100 pixels per second on the time axis. Complete pre-processing details can be found in Supporting Information, Section 2.

For each audio file, all annotations were filtered to keep only sounds of a biotic nature. These mostly consisted of bird vocalizations, but occasionally included sounds from insects, mammals, or even the sound of wingbeats. Because bird vocalizations represented 98.8% of the total dataset, we considered the datasets to only contain bird vocalizations in this paper.

Annotations were then converted into a one dimensional array of presence/absence of biotic sound with the same temporal resolution as the spectrograms (see below).

2.2.2 Model architecture

Our objective was to create a model capable of rapidly and efficiently processing large volumes of sound data, serving as a valuable preprocessing tool. Consequently, we designed a streamlined model that prioritized precision, accuracy and speed. Our model has only 530,000 parameters and 16 layers (figure SuppInfo.2), which is far lower than the millions of parameters found in recent bird sound classifiers (e.g. BirdNet: 157 layers and 27 millions parameters (Kahl et al. 2021), Bird@Edge: 12 millions parameters (Höchst et al. 2022)).

2.3 Model training

The data used for training purposes was partitioned into training and validation subsets, adhering to the nomenclature outlined in Christin, Hervet, and Lecomte (2021), with an 80:20 ratio allocated to each subset, respectively.

Following a similar approach outlined by Fairbrass et al. (2019), we extracted one-second (1s) windows from the spectrograms as inputs for our model. Our spectrograms featured 32 Mel bands along the frequency axis and were resized to have a time axis of 100 pixels per second. Consequently, each extracted window was represented as a 32*100 pixel image. These windows were extracted with a 95% overlap.

For each 1-second window, a presence/absence score was automatically assigned to indicate the presence (1) or absence (0) of vocalisations based on annotations. A score of 1 was attributed if a vocalization was detected at any point within the window. We refer to the

1-second windows containing bird vocalizations as 'present samples', while those without any vocalizations are referred to as 'absent samples'.

Due to the initial class imbalance in our training datasets, which were skewed towards absent samples, we created a balanced representation of both present and absent samples in each training iteration by using all present samples and an equivalent number of absent samples randomly selected from the pool of available absent samples.

Data augmentation is a well-established method for increasing the amount of training data and mitigating overfitting (Christin, Hervet, and Lecomte 2019). To this end, we first implemented simple data augmentation techniques, specifically time and frequency masking, which have demonstrated effectiveness in audio recognition tasks (Park et al. 2019). These techniques involve masking certain portions of the spectrogram in either the temporal or frequency domain. Based on our internal testing, we decided to apply these techniques 75% of the time during augmentation. Additionally, we applied the same normalization techniques described in Citynet to get a 4-layer 32x100 image (Fairbrass et al. 2019).

We trained the model for 50 iterations, with a learning rate of 0.01. We employed early stopping, a common technique used to address overfitting, as its halts the training when no further progress is observed. The model was designed to return a probability score ranging from 0 to 1, indicating the likelihood of the presence of a bird vocalization for each spectrogram.

Training involved two steps. We first combined the training datasets of the ArcticBirdSounds+, NABS_Gen, CityNet and NIPS4B into a single training dataset (Training 1). Results of this model were evaluated to create the ArcticChecked database. This dataset

was then added to the Training 1 dataset to create a second training dataset (Training 2). The final model was then trained using this model and by making sure that all false postive results from the ArcticChecked dataset were seen on each iteration of the model.

2.4 Model evaluation

Evaluation of the model was performed using data that were not part of the training. We assessed the performance of our model using different methods to best identify its strengths and weaknesses. Prior to evaluation, we smoothed the temporal pattern of acoustic activity predicted by our model using a rolling average with a window of three to remove some of the noise in prediction scores. We used two evaluation methods as explained below.

2.4.1 Evaluation methods

- **Direct evaluation**: This method categorizes a prediction as present if the probability assigned by the model exceeds a specific threshold.
- **Event detection**: This method focuses on detecting and isolating sound events to evaluate the model's ability to accurately segment vocalizations. To achieve this, we defined a starting threshold above which an event begins and an end threshold below which an event concludes. Additionally, a duration threshold was set to discard events considered too brief. A vocalization event is considered a true positive if any part of it overlaps with an annotation. This approach allows our detector to effectively isolate vocalization events.

2.4.2 Detection evaluation

For all evaluation methods, we calculated the F1-Score and average precision (AP) to assess the performance of our model. For the direct evaluation method, we also calculated the area under the receiver operator characteristic curve (AUC).

Using the ArcticSummer dataset, we evaluated our model's performance in detecting arctic birds species. To gauge the model generalizability and robustness, we performed assessments on the test subsets of CityNet, NIPS4Bplus, and ENAB databases.

2.4.3 Phenology evaluation

We evaluated the ability of our model to detect temporal trends on two datasets: first, we used the complete ArcticSummer dataset, consisting of recordings spanning 43 days, to assess the model's ability to capture vocal activity patterns over an entire breeding season. Additionally, we selected a 3-hour recording from the ENAB dataset to investigate if the model could identify activity patterns at a smaller temporal scale.

In our analysis, we initially computed the total duration of vocalizations within specific time windows. For the Arctic summer dataset, we utilized a day-long time window, while for the ENAB dataset, we employed 5-minute segments. Subsequently, we applied the rolling trends method from the pandas package (Reback et al. 2022) to calculate the moving average across seven consecutive windows. This approach enabled us to observe the overall trend in total vocal activity throughout the dataset. We calculated trends for both the ground-truthed data and the predictions generated by our model. We then measured the similarity between these two curves by computing the Euclidean distance between them. As the shape

of the curves holds greater significance for phenology assessment than the absolute values, we normalized them to alleviate the potential effects of over- or under-estimation. To do this, we removed the mean from each values and divided by the standard deviation of all estimates. This allowed us to capture the fundamental shape and pattern of the curves, allowing for more robust and insightful comparisons.

2.4.4 Segmentation evaluation

One of our primary objectives was to optimize our model for fast and efficient audio segmentation, ensuring the rapid extraction of bird vocalization data. To achieve this, we employed an event detection methodology that involved setting probability-based thresholds to define the start and end of an acoustic event. We defined an acoustic event as any temporal sequence that started when the probability of presence was above 0.9 and consistently remained above 0.5 for at least 400 milliseconds. These parameter values were determined based on preliminary sensitivity analysis. To gauge the model's performance, we ran the model on field data that was not annotated and selected 300 events identified by the model. We then listened to these samples to evaluate whether birds were actually present during each event.

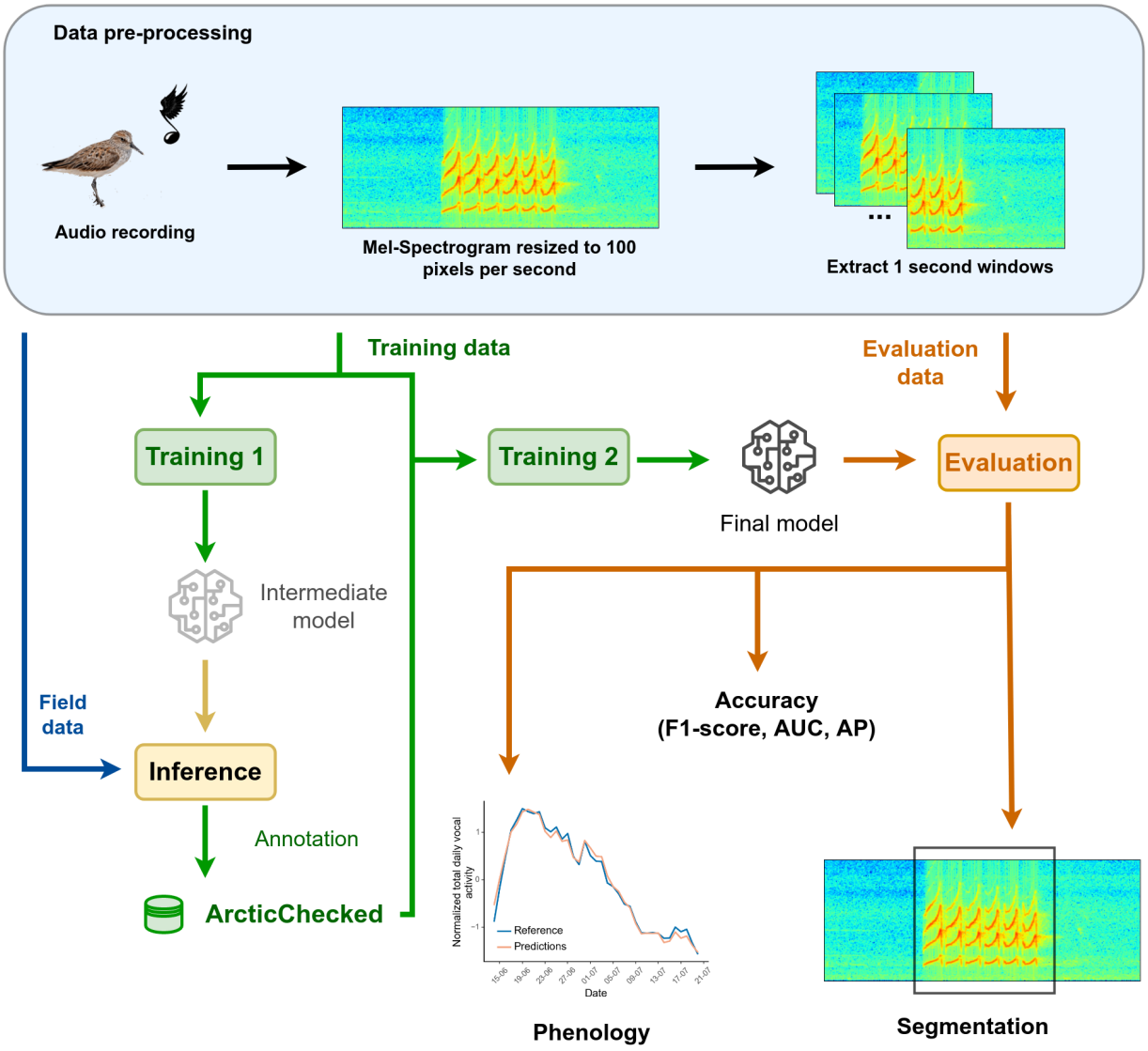

Figure 1: Workflow for the training and evaluation of the BioSoundNet model. In the blue panel on top, data pre-processing is applied to all audio samples. Green arrows and boxes represent steps of the training process. Orange arrows and boxes are steps of the evaluation process.

3 Results

3.1 Model accuracy

Our best model consistently achieved an AUC greater than 0.88 across all evaluations, with the highest performance observed on the Citynet dataset, where it achieved an AUC of 0.929 (Table 2). Comparable results were obtained on the target ArcticSummer dataset, with an AUC of 0.925. Furthermore, the average precision values were generally similar or higher for all datasets.

Table 2: Area under the curve (AUC), average precision, average recording time, average time needed to classify a recording, and average time needed to classify one minute of recording for all evaluation databases using the BioSoundNet model. All times given are in seconds. Performance evaluations were done on a middle range laptop with GPU acceleration

Dataset	AUC	Average Precision	Recording duration(s)	Time per file(s)	Time per minute(s)
Citynet	0.93	0.94	60	0.11	0.11
NIPS4B	0.9	0.94	5*	0.06	0.94
ArcticSummer	0.93	0.92	30	0.08	0.16
ENAB	0.88	0.97	300	0.45	0.09
ArcticChecked	0.92	0.87	0.9 to 272	0.08	0.43

3.2 Model performance

One of the primary objectives of our model was to have fast performance on a wide range of hardware. On a mid-range laptop with GPU acceleration (Intel(R) Core (TM) i5-6300 @2.30GHz with 4 cores and Nvidia GeForce GTX 960M) , it took less than 100 ms to classify one minute of audio on files with a 5-minute duration. However, the average processing time was slightly longer for files with a shorter duration. For instance, on the NIPS4B

dataset where files had an average duration time of 5 seconds, average processing time was around 60 milliseconds. Extrapolating this to a minute of audio, the estimated time would be approximately one second. This implies that our model is useful for analyzing large datasets of acoustic recordings and can be utilized with readily available technology by most researchers.

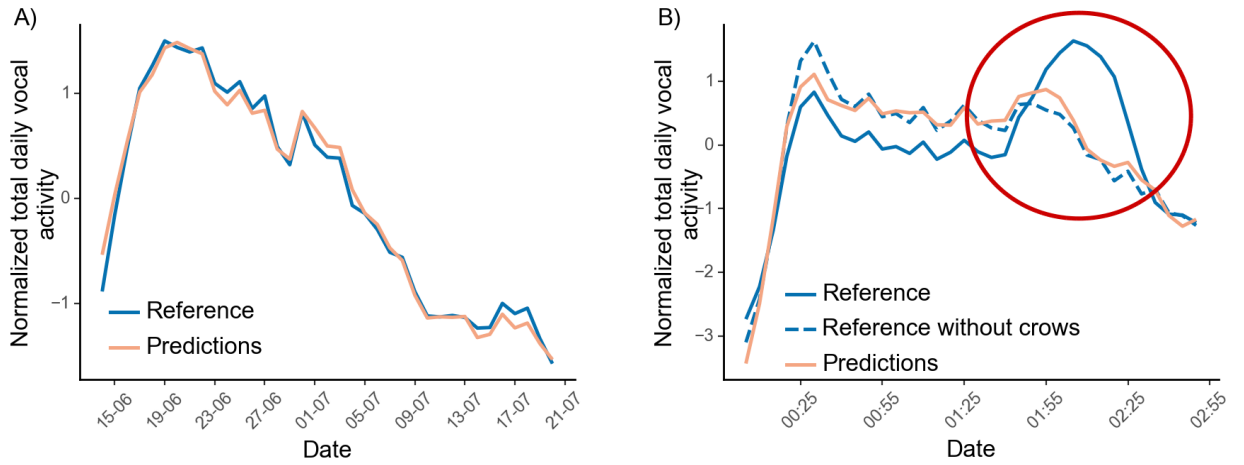

Figure 2: Evaluation of the ability of the BioSoundNet model to detect temporal variations of acoustic activity of avian communities. The two plots represent the normalized total vocal activity by recording. The reference curves are represented by the blue line, while the model predictions are shown in orange. A) Vocal activity patterns for the ArcticSummer dataset between 13 June and 20 July 2018. B) Vocal activity patterns for the 3-h long Recording 1 from the ENAB dataset collected on April 27, 2018 at Powdermill Nature Reserve, Rector, PA, USA. The activity pattern from the original reference data and the reference data without crows are in solid and dashed blue line, respectively. The red circle identifies the area of intense crow activity. Curves were obtained using the direct evaluation method with a threshold of 0.75.

3.3 Ability to detect phenological trends

Our model's predictions for the total daily vocal activity closely aligned with the reference data on the ArcticSummer dataset, as shown in Figure 2.A. On the shorter timescale of the ENAB dataset, the predicted trends closely matched the reference curve at the beginning

of the time series but deviated towards the end (Figure 2.B). Manual exploration revealed an increased activity of American crows (*Corvus brachyrhynchos*) during that period. To assess whether this discrepancy was due to the crows' vocalisations and our model inability to detect them, we excluded them from the reference dataset and this led to a closer match between the trends. In figure 2.B, the euclidean distance between the normalized predicted curve with crows and the normalized reference curve is 4.63, while it is 1.17 when crows are removed.

3.4 Influence of detection threshold

Selecting an appropriate detection threshold can pose challenges in finding the right balance between precision and recall. Here we show that, while different thresholds influenced the estimates of total vocal activity (Figure 3.A), this did not affect the normalized trends (Figure 3.B).

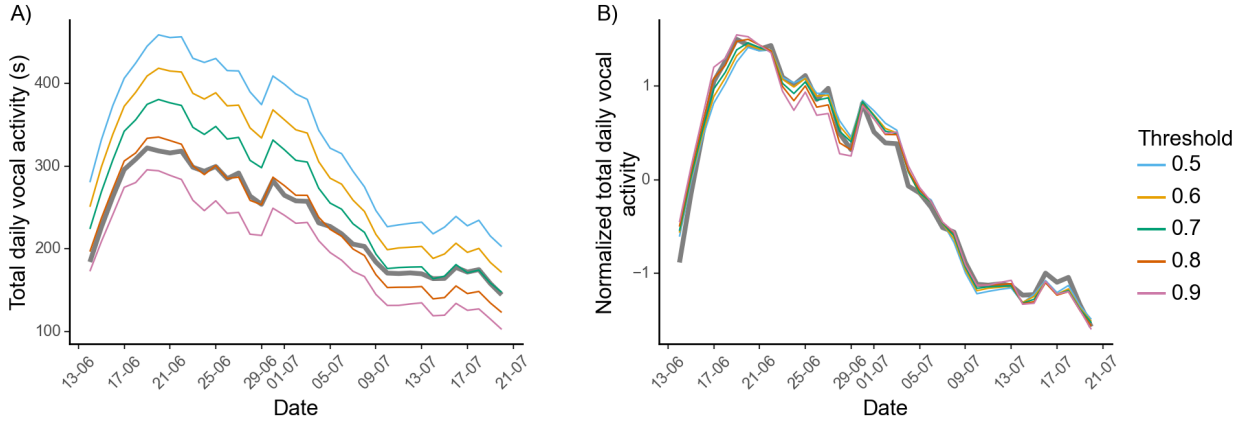

Figure 3: Influence of detection threshold on patterns of vocal activity using the BioSoundNet model. A) Total time spent vocalizing per day in seconds for various detection thresholds. B) Normalized daily vocal activity for the same thresholds. In both panels, the black line represents the vocal activity curve from the reference dataset.

3.5 Segmentation performance

Upon applying the event detection method to isolate events, we found that bird vocalizations were present in 262 (87.3%) of the 300 randomly selected events.

4 Discussion

The BioSoundNet model demonstrated good accuracy results, with AUC scores ranging from 0.88 to 0.93 and average precision between 0.87 and 0.97. These results hold true for both the target datasets with Arctic birds and general datasets, indicating BioSoundNet's potential use in different biomes, a much needed application for detecting trends in phenology. The model also performs well on datasets from new locations not included in the training data, underscoring its adaptability. Deep learning bird vocalization detectors similar to ours are scarce, but the average precision of our model on the CityNet test set is similar to the CityBioNet model described by Fairbrass et al. (2019).

Our assessment also goes beyond accuracy evaluation, highlighting the ability of the BioSoundNet model to be used in ecological applications. By comparing the model's predictions to acoustic time series at various temporal scales, we demonstrated that BioSoundNet can be reliably used in tracking temporal patterns of vocal activity for target species using automatic settings. To our knowledge, this is the first time such an evaluation has been conducted on a time series as extensive as an entire summer. This paves the way for tracking activity and presence at multiple spatial and temporal scales.

The predicted acoustic activity curve for the ArcticSummer dataset exhibits great similarity

to the reference curve, with the normalized curve closely aligned. Predicted curves are also quite robust to threshold selection temporally. Threshold selection can sometimes be an arduous task with detection models, as it involves finding a compromise between precision and recall. Changing the detection threshold can increase or decrease the number of events predicted, for instance by increasing the number of false positives. This could in turn impact activity patterns and modify the global signal. With BioSoundNet, the primary effect of altering the detection threshold was a shift in the estimated absolute duration of vocal activity, while overall temporal activity patterns remained consistent. Consequently, while threshold selection did lead to over- or under-estimation of vocal activity, it did not impact within season phenology of vocal activity, but rather affected all events in a similar way. This characteristic removes the constraint of finding the perfect threshold. This feature positions the BioSoundNet model as a valuable tool for applications that involve comparison of values, such as phenological studies.

Despite these successes, activity patterns were more challenging to detect in the ENAB dataset. Some discrepancies were observed, particularly towards the end of the time series, where an important increase in the activity of American crows was apparent. This issue likely arose from the model's limited ability to detect this type of vocalization, as confirmed by the improved matching of activity curves when crows were removed from the reference dataset. This can be explained by the fact that deep learning models are very sensitive to the type of data provided for training (Christin, Hervet, and Lecomte 2019). In this case, few examples of crow vocalisations were present in the training databases. Additionally, crows produce diverse vocalisations that can occur in different frequency bands. In such instances,

the low-frequency vocalisations of crows could have been mistaken for ambient noise, such as wind. Fortunately, addressing this problem with deep learning is relatively straightforward: it only requires re-training the model with relevant examples for the missing class. In our case, this issue seemed limited to this single species, as excluding it improved the match of activity curves. From a temporal standpoint, we thus believe that BioSoundNet can handle a diverse range of time scales relevant to ecological studies. However, it is important to keep in mind that, while trained to a wide range of species and vocalizations, BioSoundNet remains sensitive to vocalizations not present in the training set, especially when they dominate the soundscape like the crows did in our test set. This kind of problem can however be mitigated depending on the frequency of such undetected vocalizations and the temporal scope of the study. Indeed, while discrepancies appear in a three-hour sampling period, its impact might be less evident over the course of a full summer.

The crow detection problem encountered with the ENAB dataset highlights two observations when evaluating the performance of a model. First, the model should always be evaluated on a dataset representative of the specific question it aims to address and the specific vocal community it aims to study (Christin, Hervet, and Lecomte 2021; Chicco 2017). This is particularly crucial in deep learning, where performance is closely tied to the training data (Christin, Hervet, and Lecomte 2019). This ensures that all target species can be adequately detected. Secondly, model evaluation should extend beyond accuracy metrics and encompass the intended future use of the model. Relying solely on accuracy metrics would have led us to believe that our model was suitable for use on the ENAB dataset. Since North American bird samples were present in our training data and we achieved good accuracy results, we

might have assumed that activity curves on the ENAB dataset would be similar to those in the ArcticSummer dataset. However, the discovery of a key species going undetected reveals the potential for missing vocal activity patterns, which could have resulted in misleading conclusions. Note that while the BioSoundNet model is a detector, this also applies to any classifier.

The BioSoundNet model currently works at the avian soundscape level, providing a reliable alternative to acoustic indices (Farina, Pieretti, and Piccioli 2011; Bradfer-Lawrence et al. 2023). One significant advantage is the ease of quantifying the model's accuracy, which is often challenging with traditional indices, (e.g. Fairbrass et al. 2019). Moreover, by selecting sounds, BioSoundNet offers the capability to target species and actively filter out unwanted noise, such as anthropophony or geophony (Fairbrass et al. 2017; Fairbrass et al. 2019). Therefore, it holds potential for seasonal timing studies. However, reliable examination of breeding phenology through acoustic activities still requires additional work and should be corroborated by traditional monitoring methods like nest monitoring to evaluate the relationship between breeding stage and vocal activity patterns.

Beyond its direct ecological applications, we believe BioSoundNet could also serve as a valuable segmentation tool due to its high detection accuracy and low computational time. The model has modest hardware requirements and can efficiently run even on a middle-range laptop. Although the average prediction time may vary across datasets, this is primarily due to the differing lengths of audio files in each dataset. The computational costs of setting up the model remain relatively constant, and as audio file sizes decrease, the relative time for processing becomes longer. Therefore, for optimal performance, we recommend using

audio samples with a duration greater than one minute. This capability makes BioSoundNet particularly useful for filtering large datasets. For example, users can initially extract samples with a high probability of bird activity and use them as input for a more powerful species classifier. This approach would be especially beneficial in areas with low bird densities and minimal vocal activity, where processing a large number of audio files with a high-resolution classifier could be time and resource-intensive. For instance, in the ArcticBirdSound+ dataset, the total duration of vocalizations accounted for only around 15% of the total duration of the dataset. Pre-selecting segments with a high probability of a bird singing would reduce the amount of audio a more powerful but slower classifier has to process by a factor of five. Additionally, BioSoundNet could be used as an exploratory tool to extract audio samples for annotation purposes.

Finally, while BioSoundNet is currently targeted to bird detection, its core inputs are audio files, allowing great versatility across many vocal species, as long as the relevant training data is provided. The adaptable architecture of BioSoundNet facilitates straightforward retraining (Christin, Hervet, and Lecomte 2019; Christin and Lecomte 2022), rendering it a valuable resource for ecologists engaged in acoustic recording studies involving various taxa. And while we used a community approach here, it could also be used to quickly detect targeted species, depending on the examples provided in the training dataset.

Automated acoustic monitoring holds great potential in not only detecting specific vocal species but also in advancing our understanding of species phenology. Its potential benefits extend to diverse applications, including the enhanced detection of seasonal migrants arriving at and departing from their breeding territories. This temporal window often extends beyond

the regular field seasons, aligning seamlessly with the recording capabilities of contemporary acoustic devices. In the near future, we foresee the necessity for increased calibration and annotations of bioacoustic sounds to further streamline biodiversity tracking, particularly for rapidly declining species. In tandem with conventional monitoring methods, the expansion of acoustic networks across diverse biomes is imperative to address our ongoing biodiversity crisis. Fast-processing models like BioSoundnet are poised to effectively tackle the demands posed by this vast dataset challenge.

Acknowledgements

NL and SC were supported by the Canada Research Chair Program, NSERC, the New Brunswick Innovation Fund, and ArcticNet, a Network of Centres of Excellence Canada. We thank L. McKinnon for deployment of recorders in Churchill, MB, Canada, and the Wildlife Conservation Society for deployment of recorders in Prudhoe Bay, AK, USA and all field teams that helped deploy acoustic recorders across the Arctic. For a full list of people involved, please refer to the supplementary acknowledgement file. Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References

Aide, T. M., C. Corrada-Bravo, M. Campos-Cerqueira, C. Milan, G. Vega, and R. Alvarez. 2013. "Real-Time Bioacoustics Monitoring and Automated Species Identification." *PeerJ* 1 (July). https://doi.org/10.7717/peerj.103.

Bateman, H. L., S. B. Riddle, and E. S. Cubley. 2021. "Using Bioacoustics to Examine Vocal Phenology of Neotropical Migratory Birds on a Wild and Scenic River in Arizona." *Birds* 2, no. 3 (September): 261–274. https://doi.org/10.3390/birds2030019.

Bradfer-Lawrence, T., C. Desjonqueres, A. Eldridge, A. Johnston, and O. Metcalf. 2023. "Using Acoustic Indices in Ecology: Guidance on Study Design, Analyses and Interpretation." *Methods in Ecology and Evolution* 14 (9): 2192–2204. https://doi.org/10.1111/2041-210X.14194.

Buxton, R. T., E. Brown, L. Sharman, C. M. Gabriele, and M. F. McKenna. 2016. "Using Bioacoustics to Examine Shifts in Songbird Phenology." *Ecology and Evolution* (June). https://doi.org/10.1002/ece3.2242.

Buxton, R. T., M. F. McKenna, M. Clapp, E. Meyer, E. Stabenau, L. M. Angeloni, K. Crooks, and G. Wittemyer. 2018. "Efficacy of Extracting Indices from Large-Scale Acoustic Recordings to Monitor Biodiversity." *Conservation Biology* 32 (5): 1174–1184. https://doi.org/10.1111/cobi.13119.

Charmantier, A., and P. Gienapp. 2014. "Climate Change and Timing of Avian Breeding and Migration: Evolutionary versus Plastic Changes." *Evolutionary Applications* 7, no. 1 (January): 15–28. https://doi.org/10.1111/eva.12126.

Chicco, D. 2017. "Ten Quick Tips for Machine Learning in Computational Biology." *BioData Mining* 10, no. 1 (December): 35. https://doi.org/10.1186/s13040-017-0155-3.

Christin, S., C. Chicoine, T. O'Neill Sanger, M. F. Guigueno, J. Hansen, R. B. Lanctot, D. MacNearney, et al. 2023. "ArcticBirdSounds: An Open-Access, Multiyear, and Detailed Annotated Dataset of Bird Songs and Calls." *Ecology* 104 (6): e4047. https://doi.org/10.1002/ecy.4047.

Christin, S., É. Hervet, and N. Lecomte. 2019. "Applications for Deep Learning in Ecology." Edited by H. Ye. *Methods in Ecology and Evolution* 10, no. 10 (October): 1632–1644. https://doi.org/10.1111/2041-210X.13256.

———. 2021. "Going Further with Model Verification and Deep Learning." *Methods in Ecology and Evolution* 12 (1): 130–134. https://doi.org/10.1111/2041-210X.13494.

Christin, S., and N. Lecomte. 2022. *Introducing Mouffet, a Unified Framework to Make Model Creation Easier and More Reproducible.* bioRxiv. https://doi.org/10.1101/2022.07.06.498965.

———. 2023. "Taking the Pulse of Changing Phenologies and Biodiversity: The Acoustic Way." *The Bulletin of the Ecological Society of America* 104 (3): e2085. https://doi.org/10.1002/bes2.2085.

Chronister, L. M., T. A. Rhinehart, A. Place, and J. Kitzes. 2021. "An Annotated Set of Audio Recordings of Eastern North American Birds Containing Frequency, Time, and Species Information." *Ecology* 102 (6): e03329. https://doi.org/10.1002/ecy.3329.

Delplanque, A., S. Foucher, P. Lejeune, J. Linchant, and J. Théau. 2022. "Multispecies Detection and Identification of African Mammals in Aerial Imagery Using Convolutional Neural Networks." *Remote Sensing in Ecology and Conservation* 8 (2): 166–179. https://doi.org/10.1002/rse2.234.

Fairbrass, A. J., M. Firman, C. Williams, G. J. Brostow, H. Titheridge, and K. E. Jones. 2019. "CityNet—Deep Learning Tools for Urban Ecoacoustic Assessment." *Methods in Ecology and Evolution* 10 (2): 186–197. https://doi.org/10.1111/2041-210X.13114.

Fairbrass, A. J., P. Rennert, C. Williams, H. Titheridge, and K. E. Jones. 2017. "Biases of Acoustic Indices Measuring Biodiversity in Urban Areas." *Ecological Indicators* 83 (December): 169–177. https://doi.org/10.1016/j.ecolind.2017.07.064.

Farina, A., N. Pieretti, and L. Piccioli. 2011. "The Soundscape Methodology for Long-Term Bird Monitoring: A Mediterranean Europe Case-Study." *Ecological Informatics* 6, no. 6 (November): 354–363. https://doi.org/10.1016/j.ecoinf.2011.07.004.

Gibb, R., E. Browning, P. Glover-Kapfer, and K. E. Jones. 2019. "Emerging Opportunities and Challenges for Passive Acoustics in Ecological Assessment and Monitoring." *Methods in Ecology and Evolution* 10 (2): 169–185. https://doi.org/10.1111/2041-210X.13101.

Höchst, J., H. Bellafkir, P. Lampe, M. Vogelbacher, M. Mühling, D. Schneider, K. Lindner, et al. 2022. "Bird@Edge: Bird Species Recognition at the Edge." In *Networked Systems,* edited by M.-A. Koulali and M. Mezini, 69–86. Lecture Notes in Computer Science. Cham: Springer International Publishing. ISBN: 978-3-031-17436-0. https://doi.org/10.1007/978-3-031-17436-0_6.

Kahl, S., C. M. Wood, M. Eibl, and H. Klinck. 2021. "BirdNET: A Deep Learning Solution for Avian Diversity Monitoring." *Ecological Informatics* 61 (March): 101236. https://doi.org/10.1016/j.ecoinf.2021.101236.

Krause, B., and A. Farina. 2016. "Using Ecoacoustic Methods to Survey the Impacts of Climate Change on Biodiversity." *Biological Conservation* 195 (March): 245–254. https://doi.org/10.1016/j.biocon.2016.01.013.

Kwon, E., E. L. Weiser, R. B. Lanctot, S. C. Brown, H. R. Gates, G. Gilchrist, S. J. Kendall, et al. 2019. "Geographic Variation in the Intensity of Warming and Phenological Mismatch between Arctic Shorebirds and Invertebrates." *Ecological Monographs* 89 (4): e01383. https://doi.org/10.1002/ecm.1383.

LeCun, Y., Y. Bengio, and G. Hinton. 2015. "Deep Learning." *Nature* 521, no. 7553 (May): 436–444. https://doi.org/10.1038/nature14539.

Lostanlen, V., A. Cramer, J. Salamon, A. Farnsworth, B. M. V. Doren, S. Kelling, and J. P. Bello. 2022. *BirdVox: Machine Listening for Bird Migration Monitoring.* bioRxiv, May. https://doi.org/10.1101/2022.05.31.494155.

Mann, H. M. R., A. Iosifidis, J. U. Jepsen, J. M. Welker, M. J. J. E. Loonen, and T. T. Høye. 2022. "Automatic Flower Detection and Phenology Monitoring Using Time-Lapse Cameras and Deep Learning." *Remote Sensing in Ecology and Conservation* 8 (6): 765–777. https://doi.org/10.1002/rse2.275.

McFee, B., A. Metsai, M. McVicar, S. Balke, C. Thomé, C. Raffel, F. Zalkow, et al. 2022. *Librosa/Librosa: 0.9.2.* Zenodo, June. https://doi.org/10.5281/zenodo.6759664.

Menzel, A., Y. Yuan, M. Matiu, T. Sparks, H. Scheifinger, R. Gehrig, and N. Estrella. 2020. "Climate Change Fingerprints in Recent European Plant Phenology." *Global Change Biology* 26 (4): 2599–2612. https://doi.org/10.1111/gcb.15000.

Morfi, V., Y. Bas, H. Pamuła, H. Glotin, and D. Stowell. 2019. "NIPS4Bplus: A Richly Annotated Birdsong Audio Dataset." *PeerJ Computer Science* 5 (October): e223. https://doi.org/10.7717/peerj-cs.223.

Nakazawa, T., and H. Doi. 2012. "A Perspective on Match/Mismatch of Phenology in Community Contexts." *Oikos* 121 (4): 489–495. JSTOR: 41429317.

Park, D. S., W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le. 2019. "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition." In *Interspeech 2019,* 2613–2617. September. https://doi.org/10.21437/Interspeech.2019-2680. arXiv: 1904.08779 [cs, eess, stat].

Parmesan, C., and G. Yohe. 2003. "A Globally Coherent Fingerprint of Climate Change Impacts across Natural Systems." *Nature* 421, no. 6918 (January): 37.

Pecl, G. T., M. B. Araújo, J. D. Bell, J. Blanchard, T. C. Bonebrake, I.-C. Chen, T. D. Clark, et al. 2017. "Biodiversity Redistribution under Climate Change: Impacts on Ecosystems and Human Well-Being." *Science* 355, no. 6332 (March): eaai9214. https://doi.org/10.1126/science.aai9214.

Reback, J., jbrockmendel, W. McKinney, J. V. den Bossche, M. Roeschke, T. Augspurger, S. Hawkins, et al. 2022. *Pandas-Dev/Pandas: Pandas 1.4.3*. Zenodo, June. https://doi.org/10.5281/zenodo.6702671.

Ross, M. V., R. T. Alisauskas, D. C. Douglas, and D. K. Kellett. 2017. "Decadal Declines in Avian Herbivore Reproduction: Density-Dependent Nutrition and Phenological Mismatch in the Arctic." *Ecology* 98 (7): 1869–1883. https://doi.org/10.1002/ecy.1856.

Ross, S. R. P.-J., D. P. O'Connell, J. L. Deichmann, C. Desjonquères, A. Gasc, J. N. Phillips, S. S. Sethi, C. M. Wood, and Z. Burivalova. 2023. "Passive Acoustic Monitoring Provides a Fresh Perspective on Fundamental Ecological Questions." *Functional Ecology* n/a (n/a). https://doi.org/10.1111/1365-2435.14275.

Ruff, Z. J., D. B. Lesmeister, L. S. Duchac, B. K. Padmaraju, and C. M. Sullivan. 2020. "Automated Identification of Avian Vocalizations with Deep Convolutional Neural Networks." Edited by N. Pettorelli and V. Lecours. *Remote Sensing in Ecology and Conservation* 6, no. 1 (March): 79–92. https://doi.org/10.1002/rse2.125.

Saalfeld, S. T., B. L. Hill, C. M. Hunter, C. J. Frost, and R. B. Lanctot. 2021. "Warming Arctic Summers Unlikely to Increase Productivity of Shorebirds through Renesting." *Scientific Reports* 11, no. 1 (July): 15277. https://doi.org/10.1038/s41598-021-94788-z.

Samplonius, J. M., A. Atkinson, C. Hassall, K. Keogan, S. J. Thackeray, J. J. Assmann, M. D. Burgess, et al. 2021. "Strengthening the Evidence Base for Temperature-Mediated Phenological Asynchrony and Its Impacts." *Nature Ecology & Evolution* 5, no. 2 (February): 155–164. https://doi.org/10.1038/s41559-020-01357-0.

Schmidt, N. M., T. Kankaanpää, M. Tiusanen, J. Reneerkens, T. S. Versluijs, L. H. Hansen, J. Hansen, et al. 2023. "Little Directional Change in the Timing of Arctic Spring Phenology over the Past 25 Years." *Current Biology* 33, no. 15 (August): 3244–3249.e3. https://doi.org/10.1016/j.cub.2023.06.038.

Senner, N. R., M. Stager, and B. K. Sandercock. 2017. "Ecological Mismatches Are Moderated by Local Conditions for Two Populations of a Long-Distance Migratory Bird." *Oikos* 126, no. 1 (January): 61–72. https://doi.org/10.1111/oik.03325.

Shrivastava, A., A. Gupta, and R. Girshick. 2016. "Training Region-Based Object Detectors with Online Hard Example Mining." In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 761–769. June. https://doi.org/10.1109/CVPR.2016.89.

Slagsvold, T. 1977. "Bird Song Activity in Relation to Breeding Cycle, Spring Weather, and Environmental Phenology." *Ornis Scandinavica (Scandinavian Journal of Ornithology)* 8 (2): 197–222. https://doi.org/10.2307/3676105. JSTOR: 3676105.

Stenseth, N. C., and A. Mysterud. 2002. "Climate, Changing Phenology, and Other Life History Traits: Nonlinearity and Match–Mismatch to the Environment." *Proceedings of the National Academy of Sciences* 99, no. 21 (October): 13379–13381. https://doi.org/10.1073/pnas.212519399.

Thackeray, S. J., P. A. Henrys, D. Hemming, J. R. Bell, M. S. Botham, S. Burthe, P. Helaouet, et al. 2016. "Phenological Sensitivity to Climate across Taxa and Trophic Levels." *Nature* 535, no. 7611 (July): 241–245. https://doi.org/10.1038/nature18608.

Van Doren, B. M., V. Lostanlen, A. Cramer, J. Salamon, A. Dokter, S. Kelling, J. P. Bello, and A. Farnsworth. 2023. "Automated Acoustic Monitoring Captures Timing and Intensity of Bird Migration." *Journal of Applied Ecology* 60 (3): 433–444. https://doi.org/10.1111/1365-2664.14342.

Varpe, Ø. 2017. "Life History Adaptations to Seasonality." *Integrative and Comparative Biology* 57, no. 5 (November): 943–960. https://doi.org/10.1093/icb/icx123.

Visser, M. E., and C. Both. 2005. "Shifts in Phenology Due to Global Climate Change: The Need for a Yardstick." *Proceedings: Biological Sciences* 272 (1581): 2561–2569. https://doi.org/10.2307/30047868. JSTOR: 30047868.

Visser, M. E., and P. Gienapp. 2019. "Evolutionary and Demographic Consequences of Phenological Mismatches." *Nature Ecology & Evolution* 3, no. 6 (June): 879–885. https://doi.org/10.1038/s41559-019-0880-8.

Visser, M. E., L. J. M. Holleman, and P. Gienapp. 2006. "Shifts in Caterpillar Biomass Phenology Due to Climate Change and Its Impact on the Breeding Biology of an Insectivorous Bird." *Oecologia* 147 (1): 164–172. JSTOR: 20445808.

Zhao, Z. 2018. "North American Bird Species" (May). https://doi.org/10.5281/zenodo.1250690.