Assessing the potential of BirdNET to infer European bird communities from large-scale ecoacoustic data

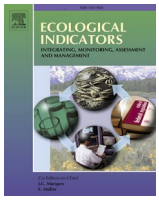
David Funosas, Luc Barbaro, Laura Schillé, Arnaud Elger, Bastien Castagneyrol, Maxime Cauchoix

Ecological Indicators 164 (2024) 112146

Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind



Original Articles

Assessing the potential of BirdNET to infer European bird communities from large-scale ecoacoustic data

David Funosas [a,b,g,*], Luc Barbaro [b,c,d], Laura Schillé [e], Arnaud Elger [a,b], Bastien Castagneyrol [e], Maxime Cauchoix [a,b,f]

[a] Laboratoire Écologie Fonctionnelle et Environnement, Université de Toulouse, CNRS, INPT, UPS, Toulouse, France
[b] LTSER Zone Atelier 'Pyrénées Garonne', ZA PYGAR, Auzeville-Tolosane, France
[c] Dynafor, INRAE-INPT, University of Toulouse, Castanet-Tolosan, France
[d] CESCO, Museum national d'Histoire naturelle, CNRS, Sorbonne University, Paris, France
[e] University of Bordeaux, INRAE, BIOGECO, Cestas, France
[f] Station d'Écologie Théorique et Expérimentale, SETE, CNRS, Moulis, France
[g] Université de Caen Normandie, UNICAEN, Caen, France

ARTICLE INFO

Keywords:
Automated species identification
Bird communities
BirdNET
Confidence threshold
Passive Acoustic Monitoring
Precision and recall
Soundscape

ABSTRACT

1. Passive acoustic monitoring has become increasingly popular as a practical and cost-effective way of obtaining highly reliable acoustic data in ecological research projects. Increased ease of collecting these data means that, currently, the main bottleneck in ecoacoustic monitoring projects is often the time required for the manual analysis of passively collected recordings. In this study we evaluate the potential and current limitations of BirdNET-Analyzer v2.4, the most advanced and generic deep learning algorithm for bird recognition to date, as a tool to assess bird community composition through the automated analysis of large-scale ecoacoustic data.
2. To this end, we study 3 acoustic datasets comprising a total of 629 environmental soundscapes collected in 194 different sites spread across a 19° latitude span in Europe. We analyze these soundscapes using both BirdNET and manual listening by local expert birders, and we then compare the results obtained through the two methods to evaluate the performance of the algorithm both at the level of each single vocalization and for entire recording sequences (1, 5 or 10 min).
3. Since BirdNET provides a confidence score for each identification, minimum confidence thresholds can be used to filter out identifications with low scores, thus retaining only the most reliable ones. The volume of ecoacoustic data used in this study did not allow us to estimate species-specific minimum confidence thresholds for most taxa, so we instead used and evaluated global confidence thresholds selected for optimized results when consistently applied across all species.
4. Our analyses reveal that BirdNET identifications can be highly reliable if a sufficiently high minimum confidence threshold is used. However, the inevitable trade-off between precision and recall does not allow to obtain satisfactory results for both metrics at the same time. We found that F1-scores remain moderate (<0.5) for all datasets and confidence thresholds studied, and that acoustic datasets of extended duration seem to be currently necessary for BirdNET to provide a reliable and minimally comprehensive picture of the target bird community. We estimate, however, that the usage of species- and context-specific minimum confidence thresholds would allow to substantially improve the global performance benchmarks obtained in this study.

* Corresponding author at: Laboratoire Écologie Fonctionnelle et Environnement, Université de Toulouse, CNRS, INPT, UPS, Toulouse, France.
E-mail address: davidfunosas@gmail.com (D. Funosas).

https://doi.org/10.1016/j.ecolind.2024.112146
Received 18 December 2023; Received in revised form 25 April 2024; Accepted 15 May 2024
Available online 20 May 2024

5. We conclude that a judicious use of AI-based identifications provided by BirdNET can represent a powerful method to assist in the assessment of bird community composition through the automated analysis of ecoacoustic data, especially when applied to acoustic datasets of extended duration.

1. Introduction

In recent years, passive acoustic monitoring (PAM) based on autonomous recording units (ARUs) has become an increasingly popular option for monitoring avian populations and communities (Shonfield & Bayne, 2017). Compared to active acoustic monitoring (AAM) methods such as point counts or active searches, ARUs could provide a more cost-effective way of obtaining data (Melo et al., 2021; Darras et al., 2018) and a higher degree of standardization in the data collection process (Darras et al., 2019). Moreover, ARUs offer the possibility of listening to distant, noisy or dubious animal sounds multiple times, potentially facilitating their identification (Sugai et al., 2019). At the same time, ARUs make it possible to collect substantial amounts of ecoacoustic data in a non-invasive way for local wildlife.

ARUs can be especially useful for biodiversity monitoring projects in hard-to-reach areas, where difficult access conditions might hinder frequent data collection through AAM methods, and might improve the detectability of species whose vocal activity patterns do not coincide with the dates and times at which active monitoring is usually conducted (Sebastián-González et al., 2018). This greater facility to collect ecoacoustic data at regular time intervals in a standardized way makes it possible to study the variations in vocal activity patterns of one or more species along a given time gradient (e.g., time of day, week of the year) (Gasc et al., 2013; Towsey et al., 2014). Additionally, the analysis of soundscape variation across time and space allows researchers to study potential correlations between the presence of anthropogenic sounds and the acoustic activity of sound-producing animals (Blumstein et al., 2011). This means that PAM could serve as a reliable way to obtain complementary information to cover the gaps left by traditional monitoring methods (Bradfer-Lawrence et al., 2023). Indeed, recent studies suggest that the combined use of PAM and AAM may provide a more accurate picture of the presence and abundance of target species than the use of only one of the two methods (Bobay et al., 2018; Shaw et al., 2021).

The increased ease of obtaining large amounts of ecoacoustic data –easily within the range of hundreds to tens of thousands of hours for multi-site projects (Dufourq et al., 2021)– through PAM means that, currently, the main bottleneck in ecoacoustic monitoring is often the time required for the analysis of passively collected data (Symes et al., 2022). Manual analysis of recordings by experts requires time-intensive dedication and therefore severely limits the amount of data that can be processed. Thus, the automated analysis of recordings could represent a potential solution to the current disparity between the efforts required to obtain data through PAM and the efforts necessary to manually analyze these data. In recent years, deep learning techniques have allowed for significant advances in the development of different tools for the automated recognition of biological sounds (Stowell, 2021). However, this field is still in a developing phase, and the use of these algorithms often results in significant numbers of false positives or false negatives, thus requiring subsequent manual validation (Knight et al., 2017) or the use of sophisticated population models to filter out false positives (Clare et al., 2021). Hence, the development of ecoacoustic data processing algorithms that can reliably identify the species recorded without manual supervision could allow for a substantial transformation of the methods through which biodiversity monitoring projects are conducted (Liu et al., 2022).

In this study we evaluate BirdNET-Analyzer v2.4, a deep neural network algorithm developed by the Cornell Lab of Ornithology which is capable of identifying $>$ 6000 bird, mammal and amphibian species worldwide based on their sounds (Kahl et al., 2021). BirdNET, in addition to detecting and identifying animal sounds present in a given recording, assigns a confidence score between 0 and 1 to each identification, thus providing information about the reliability of its results. If considered sufficiently reliable, BirdNET could partially or even fully automate the analysis of ample amounts of ecoacoustic data that can already be efficiently obtained with ARUs (Stowell, 2021; Borowiec et al., 2022). Moreover, the recently released BirdNET App (Wood et al., 2022) might further contribute to the collection of ornithological data by helping non-experts identify bird vocalizations heard on the field.

Several recent studies have tested the potential and limitations of BirdNET as a tool to assess bird community composition through the automated analysis of environmental soundscapes (Kahl et al., 2021; Arif et al., 2020; Brüggemann et al., 2021; Tolkova et al., 2021; Sethi et al., 2021; Toenies & Rich, 2021; Cole et al., 2022; Höchst et al., 2022; Malamut, 2022; Wood et al., 2021). However, as highlighted in a recent review of these previous studies (Pérez-Granados, 2023), multiple analyses of potential relevance are still lacking. More concretely, the impact of (i) the number of species vocalizing simultaneously, (ii) the recording environment (i.e., the soundscape or acoustic habitat type), (iii) the type of recorder employed, (iv) the volume of species-specific acoustic data available online, and (v) the input values for the overlap and detection sensitivity parameters of the algorithm on BirdNET performance remains to be evaluated. Moreover, the review suggests that BirdNET performance should be studied not only at the entire recording level but also at the level of each single vocalization. Last but not least, BirdNET detection capacity (i.e., recall rate) across a large range of European soundscapes has yet to be investigated.

To help fill this knowledge gap we analyze, both automatically with BirdNET and manually by expert birders, 629 environmental soundscapes collected in 194 different sites spread across a 19° latitude gradient in Europe (Fig. 1) and amounting to a total of 3137 min of recording (Table 1). We then compare the identifications obtained with BirdNET against those made by experts in order to assess the performance of the algorithm in terms of precision and recall (see subsection 2.4.1). Subsequently, we examine how to deal with the existing trade-off between these two metrics and how different features of the acoustic dataset analyzed might influence BirdNET performance. The focus of this study being on the use of BirdNET for the broad characterization of bird communities, our recommendations do not apply to other widespread applications (e.g., occupancy modeling (Cole et al., 2022), bioacoustic tracking (Verreycken et al., 2021)) of the algorithm. We also suggest caution in the extrapolation of our results to regions with different bird community compositions or geophonic or anthrophonic profiles.

2. Methods

2.1. Study area and soundscape collection

The environmental soundscapes analyzed in this study originated from three independent acoustic datasets (Table 1, Appendix S1). Our first dataset, hereafter referred to as ZA-Pygar, consists of a total of 237 1-minute recordings collected from 2019 to 2022 in 79 different sampling sites in the French region of Occitanie (Barbaro et al., 2022). More precisely, the region sampled is part of the Zone Atelier Pyrénées Garonne (Ouin et al., 2021), considered a long-term socio-ecological research site by French research institute CNRS since 2017. Out of the 79 sampling sites from which recordings were obtained, 60 are found in the Aurignac canton, 16 are found either in alpine meadows or at the edge between alpine meadows and subalpine forests at altitudes between

1400 and 2100 m in the Ariège department, and the remaining 3 are found in lowland agricultural areas in South-West Occitanie (Fig. 1a).

Our second dataset, hereafter referred to as Rambouillet, consists of a total of 204 5-minute recordings collected during the COVID-19 lockdown of 2020 in 68 different sampling sites in the public forest of Rambouillet (Fig. 1b). This lowland sub-Atlantic broadleaf forest, covering 220 km^2, is located in South-West Île-de-France (France), about 70 km from Paris (Barbaro et al., 2023). Finally, our third dataset, hereafter referred to as TreeBodyguards, was inherited from a pan-European citizen science project and consists of a total of 188 10-minute

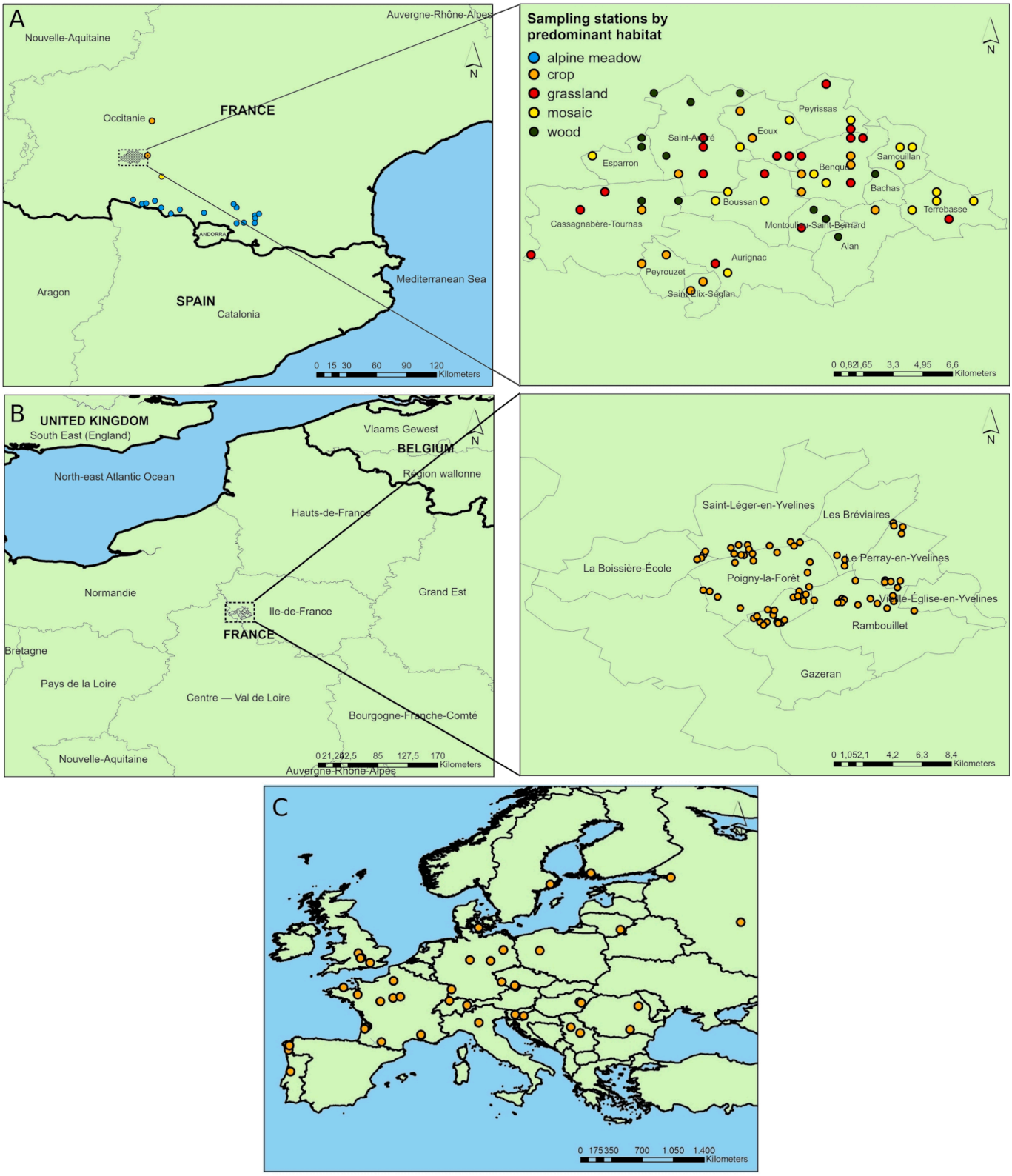


Fig. 1. Maps showing (a) the 79 sampling stations in the ZA-Pygar dataset, (b) the 68 sampling stations in the Rambouillet dataset, and (c) the 47 sampling stations in the TreeBodyguards dataset. The predominant habitat in each station is only specified for the ZA-Pygar dataset.

Table 1
Global description of the three datasets studied along with multiple indicators of BirdNET performance for each dataset.

Dataset Name	ZA-Pygar	Rambouillet	TreeBodyguards
Number of sites	79	68	47
Number of recordings	237	204	188
Recording duration (min)	1	5	10
Total duration (min)	237	1020	1880
Unit of identification	Bird vocalization	Bird presence in the recording	Bird presence in the recording
Number of species	74	44	86
Number of identifications	4101[1]	2004	1630
BirdNET results at the vocalization level			
Maximal F1-score	0.339	–	–
Optimal confidence threshold[2]	0.4	–	–
Precision with optimal conf. threshold	0.584	–	–
Recall with optimal conf. threshold	0.238	–	–
PR AUC	0.432	–	–
ROC AUC	0.054	–	–
BirdNET results at the recording level			
Maximal F1-score	0.401	0.477	0.521
Optimal confidence threshold[2]	0.3	0.3	0.45
Precision with optimal conf. threshold	0.59	0.684	0.592
Recall with optimal conf. threshold	0.30	0.312	0.506
PR AUC	0.416	0.37	0.451
ROC AUC	0.157	0.185	0.153

[1] 3340 songs, 754 calls and 7 drummings.
[2] Confidence threshold with the highest F1-score.

recordings collected in 47 different sampling sites from 17 European countries ranging from Spain to North-West Russia (Fig. 1c) (Schillé et al., 2024).

2.2. Manual bird identification by experts

Expert birders identified bird vocalizations in the three datasets. In the ZA-Pygar dataset, all bird sounds found within the first minute of each of the 237 recordings were manually identified and annotated by a single trained birder (DF) by drawing time–frequency bounding boxes around every single vocalization on recording spectrograms (Appendix S2, Fig. S1). In the two other datasets, a list of species present in each recording was generated after the listening, the specific vocalization times of each species remaining unknown. In the Rambouillet dataset, each 5-minute recording selected was manually listened to once, all species detected being listed regardless of their total vocalization time by a single trained birder (LB). As for the TreeBodyguards dataset, given its wide geographical range, we distributed the recordings among 21 expert birders (see Acknowledgements), each of them manually listening to and identifying bird sounds from between 4 (one site) to 52 recordings (13 sites).

2.3. BirdNET-assisted analysis of recordings

The BirdNET version chosen for this study is BirdNET-Analyzer v2.4, the most recent release of the algorithm at the time of publication. BirdNET takes a sound file as input, which it splits into 3-second segments before providing a list of species detected in each segment along with a confidence score assigned to each identification, with values ranging from 0 to 1. By default there is no overlap between consecutive prediction segments, i.e., each segment begins where the previous segment ends, but BirdNET offers the possibility of allowing a certain degree of overlap (Fig. S2). Another configurable parameter is the detection sensitivity, with potential input values ranging from 0.5 to 1.5. Higher values of this parameter result in a lower threshold for the detection of bird vocalizations in the input recording, thus translating into a greater expected number of bird identifications by BirdNET. In this study we initially performed the same analyses with allowed overlaps of 0 s, 1 s and 2 s between consecutive segments and with detection sensitivity values of 0.5, 1 and 1.5. Despite the increase in BirdNET processing time associated with the use of higher overlaps, the superior performances (as measured by Area Under the Curve scores for both Receiver Operating Characteristic and Precision-Recall curves) obtained with an overlap of 2 s and a detection sensitivity of 1.5 on the ZA-Pygar dataset made us select them as the input values to evaluate BirdNET with in all datasets (Table 2, section 3.1). Finally, we configured BirdNET to filter its list of potentially detectable species based on the approximate date (in the form of week of the year) and geographic coordinates of each recording (Appendix S3).

2.4. Comparing BirdNET and expert identifications

We compared the results obtained with BirdNET against the identifications made by the expert birders in order to evaluate the performance of the algorithm. BirdNET results corresponding to the ZA-Pygar dataset could be analyzed at a higher level of detail (at the vocalization level) than those obtained for the other two datasets, since its temporally annotated identifications make it possible to confront human and BirdNET identifications on each 3 s segment of recording analyzed. For our evaluation we defined 4 possible categories –True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN)– for each species and acoustic sample, the categorization criteria being slightly variable between the three acoustic datasets analyzed (see detailed explanation in Table S1).

Based on the categorization described, we calculated the precision, recall and False Positive Rate (FPR) of the algorithm for each minimum confidence threshold studied, going from 0.10 to 0.99 with a step of 0.01 (e.g., a minimum confidence threshold of 0.5 implies that all BirdNET identifications with confidence levels lower than 0.5 were filtered out). Each of these thresholds was applied consistently across all species present in the dataset or identified by BirdNET, rather than using variable thresholds on a species-by-species basis.

Precision is a measure of the reliability of BirdNET identifications, i.e., it estimates the probability that a given BirdNET identification will be correct, and is calculated by dividing the number of species correctly detected by the total number of species detected by the algorithm in the acoustic sample analyzed. Recall estimates the probability that a species present in the acoustic sample will be correctly detected by BirdNET, and is calculated by dividing the number of species correctly detected by the total number of species actually present in the acoustic sample analyzed. Finally, FPR estimates the probability that a species absent from the acoustic sample will be detected by BirdNET, and is calculated by dividing the number of species mistakenly detected by the total number of species absent from the acoustic sample analyzed (only considering those featuring in the species lists generated following the procedure explained in Appendix S3). The specific formulas used are the following:

- Precision = TP/(TP + FP)
- Recall = TP/(TP + FN)
- FPR = FP/(FP + TN)

Different conclusions can be drawn from these metrics at different levels of analysis. In the ZA-Pygar dataset, measuring these metrics at the vocalization level provides us with a fine-grained picture of BirdNET performance. However, since the total vocalization time of any given species varies significantly across different recordings, evaluating

Table 2
Area Under the Curve (AUC) scores for both Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves depending on the detection sensitivity used and the levels of overlap allowed between consecutive prediction segments by BirdNET. The *voc_*, *rec_* and *ds_* prefixes correspond to precision, recall and FPR being calculated at the vocalization level, recording level and dataset level, respectively. Best results are shown in bold and worst in italics. Only results from the ZA-Pygar dataset are included.

Overlap	Detection sensitivity	voc_PR AUC	voc_ROC AUC	rec_PR AUC	rec_ROC AUC	ds_PR AUC	ds_ROC AUC
0 s	0.5	*0.312*	*0.011*	*0.177*	*0.049*	*0.255*	0.084
0 s	1.0	0.384	0.012	0.316	0.051	0.360	0.076
0 s	1.5	0.382	0.023	0.420	0.136	**0.449**	0.130
1 s	0.5	0.374	0.014	0.213	0.051	0.363	*0.038*
1 s	1.0	0.408	0.018	0.336	0.060	0.409	0.062
1 s	1.5	0.388	0.031	**0.423**	0.143	0.428	0.148
2 s	0.5	0.363	0.036	0.246	0.058	0.354	0.057
2 s	1.0	0.428	0.036	0.380	0.062	0.445	0.058
2 s	1.5	**0.432**	**0.054**	0.416	**0.157**	0.446	**0.152**

BirdNET based on the correctness of each individual identification could result in a small number of recordings with long vocalization times having a disproportionate weight over a greater number of recordings with shorter vocalization times. Thus, calculating precision and recall at the recording level can be an appropriate way of homogenizing the weight of each recording in the final results. Finally, calculating precision and recall at the level of the whole acoustic dataset can inform us about the reliability and exhaustiveness of BirdNET results when using them to provide us with a broad picture of the composition of a bird community recorded over multiple hours or days.

We therefore calculated precision, recall and FPR at all three different levels of analysis: (i) at the vocalization level (only for the ZA-Pygar dataset), (ii) at the recording level and (iii) at the dataset level. The metrics calculated at the first level will be referred to as voc_precision, voc_recall and voc_FPR, the metrics calculated at the second level will be referred to as rec_precision, rec_recall and rec_FPR and the metrics calculated at the third level will be referred to as ds_precision, ds_recall and ds_FPR (Table 3). It is important to note that, since our categorization criteria imply that one correct identification can override any number of incorrect identifications of the same species within the same acoustic sample, the values obtained for these metrics will be highly influenced by the level of analysis chosen. Longer acoustic samples provide more opportunities for BirdNET to detect a given species, either correctly or incorrectly, but the asymmetric weight given to correct identifications over incorrect ones implies that the number of TPs will scale more rapidly with time than the number of FPs. This results in a bias towards more positive results when longer acoustic samples are used.

Once calculated, we used these values to plot the Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves, both with an estimation of the Area Under the Curve (AUC) (Davis & Goadrich, 2006). The ROC curve consists in plotting recall against FPR for each minimum confidence threshold studied in order to show the trade-off between the two metrics: as the minimum confidence score required to accept BirdNET identifications increases, the number of FPs will decrease, but so will the number of TPs. The PR curve, on the other hand, plots precision against recall for each minimum confidence threshold used, showing the trade-off between these two metrics as well. The AUC is, in both cases, a measure of the predictive power of the algorithm with a range of possible values going from 0 to 1, higher values being indicative of a higher predictive power.

Another metric that we used to evaluate the performance of the algorithm for each minimum confidence threshold studied is the F-score. The F-score measures the overall predictive power of the algorithm by considering both precision and recall scores, the weight assigned to each variable depending on the β coefficient:

$$F-score = (\beta^2+1)*precision*recall(\beta^2*precision+recall)$$

An F-score with $\beta = 1$ assigns the same importance to precision and recall, whereas an F-score with $\beta > 1$ weighs recall more heavily than precision, and vice versa for $\beta < 1$. We performed F-score analyses using three different β values: $\beta = 1$ as a standard value to facilitate the comparison of our results with those of other studies, $\beta = 0.25$ to reflect a clear prioritization of precision over recall and $\beta = 0.1$ to reflect an even more marked asymmetry between the importance assigned to these two metrics. We chose such asymmetrical coefficients because we estimate that ensuring high precision levels is paramount in most biodiversity research projects, i.e., it is usually considered preferable not to detect present species rather than mistakenly detecting non-present ones (Tolkova et al., 2021). Moreover, some types of population models (e.g., occupancy models) can account for the imperfect detection of species in the target study sites (Brunk et al., 2023; Bielski et al., 2024), further underscoring the importance of false negatives with respect to false positives.

2.5. Factors influencing BirdNET performance

Using only the Za-Pygar dataset, we further analyzed how BirdNET performance might be influenced by factors related to the acoustic dataset analyzed (total duration recorded, habitat recorded, passive recorder used, preponderance of biophony over anthropophony and geophony as measured by NDSI (Bradfer-Lawrence et al., 2023; Kasten et al., 2012; see Appendix S6), and number of species vocalizing at the same time) as well as the number of recordings available online for each species as a proxy for the size of the training dataset used by BirdNET.

To estimate the influence of the size of the dataset (i.e., total recording duration) analyzed on BirdNET performance, we calculated recall and FPR at the dataset level (ds_recall and ds_FPR) for 23 different subset sizes going from 10 to 230 recordings using a step increase of 10. For each subset size, we performed 200 random selections of recordings from the whole set of 237 recordings making up the ZA-Pygar dataset and then averaged out their results (i.e., we generated 200 random subsets of 10 recordings and then averaged out the ds_recall and ds_FPR corresponding to each of these subsets; likewise for all other subset sizes until 230). This number of random selections proved to be sufficient to perfectly replicate the ordinality of the ds_recall and ds_FPR averages by subset size across 2 iterations of the same analysis.

As for the variability of the predictive power of BirdNET across species, we distinguished between two different types of causal factors: (i) factors related to the difficulty inherent in identifying certain vocalizations, such as some acoustic patterns being easier to identify than others, or bird species with vocalizations highly similar to those of other species being harder to identify than species with more idiosyncratic sounds, and (ii) the size and quality of the acoustic data available for each species on the online platforms used as sources of data for the training of the algorithm. The first type of factor being considered out of scope for this study, we analyzed the influence of species-specific data availability on BirdNET performance. Despite not having direct access to the training datasets used by BirdNET, we know that the Xeno-canto platform (Xeno-canto, 2023) and the Macaulay Library of Natural

Table 3
Description of the different metrics used to evaluate BirdNET performance.

Evaluation Metric	Acoustic sample	Description	Aggregation method
voc_precision[1]	3-second audio segment	Proportion of BirdNET identifications that are correct (match in time and species with an annotation by an expert).	No aggregation: each BirdNET identification is either correct (TP) or incorrect (FP)
voc_recall[1]	3-second audio segment	Proportion of manually annotated bird vocalizations that have been correctly identified by BirdNET	No aggregation: each bird vocalization has either been correctly identified (TP) or not (FN)
voc_FPR[1]	3-second audio segment	Number of species mistakenly detected by BirdNET as a fraction of all species actually absent from the prediction segment[2]	No aggregation: each species absent from the 3-second audio segment has either been mistakenly detected (FP) or not detected (TN) by BirdNET
rec_precision	Entire recording[3]	Proportion of species detected by BirdNET that are actually present in the recording	Aggregation at the recording level: each species detected by BirdNET has either been correctly detected at least once (TP) or not (FP) in the recording
rec_recall	Entire recording[3]	Proportion of species present in the recording that have been correctly detected by BirdNET	Aggregation at the recording level: each species present in the recording has either been correctly identified at least once (TP) or missed (FN) by BirdNET
rec_FPR	Entire recording[3]	Number of species mistakenly detected by BirdNET as a fraction of all species actually absent from the recording[2]	Aggregation at the recording level: each species absent from the recording has either been mistakenly detected (FP) or not detected (TN) by BirdNET
ds_precision	Acoustic dataset (compilation of recordings)	Proportion of species detected by BirdNET that are actually present in the acoustic dataset	Aggregation at the dataset level: each species detected by BirdNET has either been correctly detected at least once (TP) or not (FP) in the dataset
ds_recall	Acoustic dataset	Proportion of species present in the acoustic dataset that have been correctly detected by BirdNET	Aggregation at the dataset level: each species present in the dataset has either been correctly identified at least once (TP) or missed (FN) by BirdNET
ds_FPR	Acoustic dataset	Number of species mistakenly detected by BirdNET as a fraction of all species actually absent from the acoustic dataset[2]	Aggregation at the dataset level: each species absent from the dataset has either been mistakenly detected (FP) or not detected (TN) by BirdNET

[1] Only calculated for the ZA-Pygar dataset.
[2] Only considering those featuring in the species lists generated following the procedure described in Appendix S3.
[3] Corresponding to 1 min in the ZA-Pygar dataset, 5 min in the Rambouillet dataset and 10 min in the TreeBodyguards dataset.

Sounds (Macaulay, 2023) were used as sources of audio data for this purpose (Kahl et al., 2021). Even though recordings from other sources were included in the training of the algorithm and the sampling procedures used to convert archival recordings to training data were not simply 1:1 (personal communication, January 31, 2024), we used the total number of foreground recordings available on both platforms for each species analyzed as a rough proxy for the species-specific availability of recordings during the training phase of the algorithm. More specifically, we examined the correlation between the number of recordings available on these platforms and BirdNET precision, recall and F1-scores across species. This allowed us to roughly estimate the explanatory power of species-specific recording availability over the variability of BirdNET performance across species. Species detected fewer than 10 times were filtered out from the analysis, their sample sizes being too low for results to be minimally reliable. Finally, it is important to note that, unless explicitly specified, all figures and statistical results correspond to the analysis of recordings with BirdNET-Analyzer v2.4, using a detection sensitivity of 1.5 and an overlap window of 2 s between consecutive prediction segments.

3. Results

3.1. Optimizing BirdNET parameters

Our evaluation of BirdNET performance under different input parameter configurations suggests that 1 s overlap windows improve AUC scores on both ROC and PR curves with respect to the default 0-second overlap, and that 2 s overlaps improve these scores even further (Table 2). This improvement seems to arise from the fact that high degrees of overlap facilitate the capture of a substantial part of any given bird vocalization within a single prediction segment (Fig. S1). Since longer vocalization times within a given prediction segment are predictive of higher voc_recall levels (Fig. S3; Spearman, $r_s = 0.956$ and $p < 0.001$), an increase in overlap between consecutive prediction segments should correspondingly result in higher recall scores. Likewise, the predictive power of BirdNET as measured by the ROC and PR AUC scores seems to improve along with higher values of the detection sensitivity parameter (Table 2), suggesting that a detection sensitivity of 1.5 (from a range of possible values going from 0.5 to 1.5) might be the optimal value for overall performance in the specific context of this study.

3.2. Assessing BirdNET performance

We found the confidence scores provided by the algorithm to be positively correlated with the actual probability that the corresponding identification is correct (Figs S4 and S5; Kruskal-Wallis, $H(1) = 5977.7$ and $p < 0.001$). This means that a minimum confidence threshold can be established so that identifications with low confidence scores are filtered out, thus retaining only the most reliable ones. In accordance with this premise, the ROC and PR curves (Fig. 2) reveal a clear trade-off between the precision and recall of the algorithm, as well as between recall and FPR: the higher the minimum confidence score required to accept BirdNET identifications, the higher the precision and the lower the FPR will be. Nonetheless, improvements in these two metrics have to be weighed against the lower recall associated with a higher level of selectiveness. Average precision scores by species improve considerably by setting higher minimum confidence thresholds (Fig. S6), recall scores by species following the opposite trend (Fig. S7).

When analyzing BirdNET results at the recording level for the ZA-Pygar dataset, we found that the optimal minimum confidence thresholds for maximizing global F-scores for β-values of 1, 0.25 and 0.1 are 0.3, 0.55 and 0.65, respectively (Fig. 3). The results obtained for each of these thresholds show that, as the minimum confidence score required to accept BirdNET identifications becomes stricter, both TPs and FPs decrease while FNs rise sharply (Fig. 4). This causes both the number of species correctly detected and the number of species mistakenly

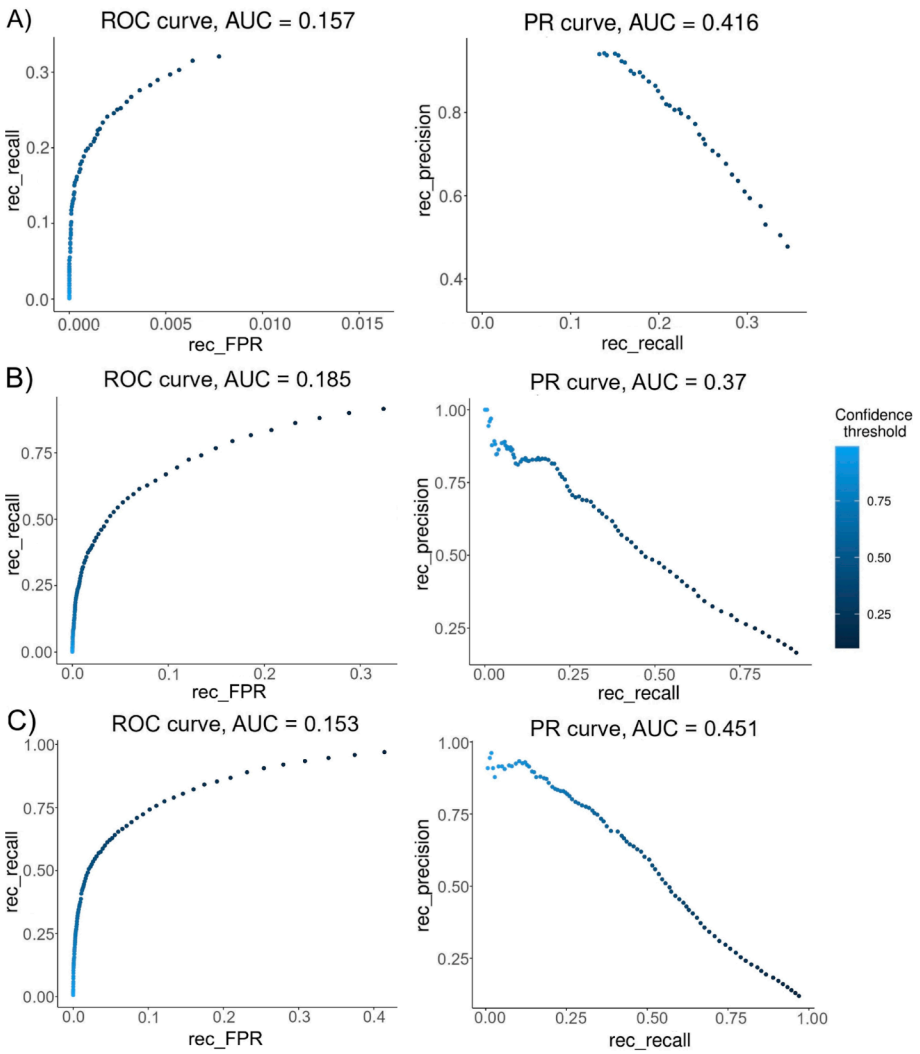


Fig. 2. Receiver Operating Characteristic (left) and Precision-Recall (right) curves of BirdNET-Analyzer. The three plots correspond to results obtained at the recording level for the (A) ZA-Pygar, (B) Rambouillet and (C) TreeBodyguards datasets.

detected to vary considerably depending on the minimum confidence threshold used. While a minimum confidence threshold of 0.3 allows for a ds_recall of 0.65 (i.e., the correct detection of 65 % of the species present in the dataset), a threshold of 0.55 reduces this figure to 0.55 and a threshold of 0.65 reduces it further to 0.5. On the other hand, ds_FPR (i.e., the proportion of species absent from the dataset that are mistakenly detected by the algorithm) varies even more strongly along with the confidence threshold used, with respective values of 0.41, 0.07 and 0.03. In absolute terms, this translates into 48, 41 and 37 species correctly detected and 85, 15 and 6 species mistakenly detected, respectively.

It is important to note that a non-negligible number of species have only been manually detected in fewer than 10 recordings per dataset (47, 15 and 53 species in the ZA-Pygar, Rambouillet and TreeBodyguards datasets, respectively) (Fig. S8). Hence, we deem our estimations of the predictive power of the algorithm for these species to be highly unreliable. Limiting the analysis to species detected –either by BirdNET or by the expert birder– in a minimum of 10 recordings, while filtering out less common species, results in a substantially higher overall performance of the algorithm. More specifically, when analyzing the ZA-Pygar dataset with a confidence threshold of 0.3, this filtering procedure improves rec_precision from 0.59 to 0.71 without compromising rec_recall, which actually remains stable at 0.32. The results obtained after applying this filter (Fig. 4d) also suggest that the variability in BirdNET predictive power across species is not as pronounced as what we might infer from the analysis of unfiltered results (Fig. 4a), since the low number of data points available for many species leads to more variable outcomes.

Overall, our analyses yielded highly consistent results across the three datasets studied, with maximal F1-scores (corresponding to minimum confidence thresholds between 0.3 and 0.45) ranging from 0.4 to 0.52 at the recording level (Table 1). This roughly means that, in the best case, we can expect two errors (FP or FN) for each correct BirdNET prediction. BirdNET performance is lower (maximal F1-score of 0.33 for a minimum confidence threshold of 0.4) when results are analyzed at the vocalization level (i.e., 1 correct prediction for every 3 errors).

3.3. Factors influencing BirdNET performance

In order to properly isolate the variables of interest and ensure that the heterogeneous identification procedures and recording durations used in the three datasets do not have a confounding effect, only recordings from the ZA-Pygar dataset are included in the following analyses.

Regarding the influence of sample size on BirdNET performance, there appears to be a positive linear correlation between ds_F1-scores and the logarithm of the number of recordings sampled (Fig. 5; linear regression, adjusted $R^2 = 0.539$, $p < 0.001$). The same seems to be true

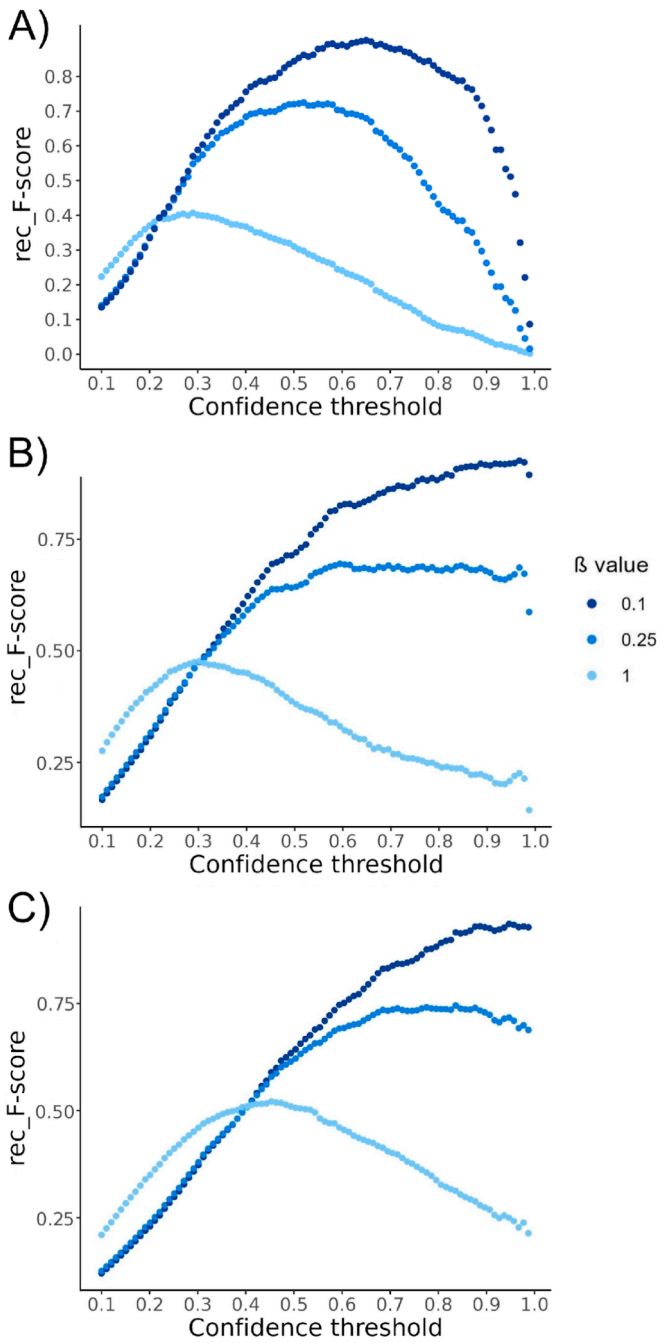


Fig. 3. F-score curves of BirdNET-Analyzer for β values of 0.1, 0.25 and 1. The three plots correspond to results obtained at the recording level for the (A) ZA-Pygar, (B) Rambouillet and (C) TreeBodyguards datasets.

for ds_TPR (adjusted $R^2 = 0.376$, $p < 0.001$) and ds_recall (adjusted $R^2 = 0.095$, $p < 0.001$) as well, in all cases after controlling for the minimum confidence threshold used. ds_FPR shows a pronounced decrease when higher confidence scores are required whereas ds_recall, while following the same trend, shows differences of more moderate proportions as a response to increased minimum confidence thresholds (Fig. 5). This suggests that a given global recall target can be reached either by setting a low enough minimum confidence threshold or by setting a high minimum confidence threshold and procuring a large enough acoustic dataset, the latter approach resulting in substantially lower ds_FPR values.

The correlation tests performed between the number of different bird species vocalizing simultaneously and BirdNET performance metrics suggest that vocalization superposition has a significant influence on voc_precision (Fig. S9; Spearman, $r_s = 0.122$ and $p < 0.001$) but not on voc_recall and voc_F1-scores (Spearman, $r_s = -0.051$ and $p = 0.285$ for voc_recall, $r_s = 0.017$ and $p = 0.72$ for voc_F1-score). In contrast, neither the recorder used nor the predominant habitat sampled seem to affect BirdNET performance (Fig. S10), since neither rec_precision, rec_recall nor rec_F1-scores show significant differences across different habitats (Kruskal-Wallis, $H_4 = 1.11$ and $p = 0.775$ for rec_precision, $H_4 = 0.79$ and $p = 0.852$ for rec_recall and $H_4 = 1.24$ and $p = 0.744$ for rec_F1_score) or recorders (Kruskal-Wallis, $H_3 = 4.08$ and $p = 0.130$ for rec_precision, $H_3 = 8.92$ and $p = 0.116$ for rec_recall and $H_3 = 2.04$ and $p = 0.36$ for rec_F1_score). Finally, the number of foreground recordings available on Xeno-canto and the Macaulay Library for each species studied appears to have a significant influence on species-specific rec_precision, rec_recall and rec_F1-scores. More concretely, the logarithm of the number of recordings available seems to correlate positively with rec_precision and rec_F1-score but negatively with rec_recall (Fig. 6; Spearman, $r_s = 0.506$ and $p < 0.001$ for rec_precision, $r_s = -0.484$ and $p = 0.002$ for rec_recall and $r_s = 0.599$ and $p < 0.001$ for rec_F1-score).

4. Discussion

BirdNET appears as a promising tool to assist in the assessment of bird community composition through the automated processing of large-scale ecoacoustic data. Our analyses reveal that BirdNET can provide us with reasonably high levels of precision or recall, at least in the regions studied, but the inevitable trade-off between these two metrics prevents satisfactory results for both at the same time (maximal F1-score < 0.5). Despite BirdNET still having considerable room for improvement in both precision and recall scores, it is important to note that identifications by field observers are not exempt from errors either. Recent studies suggest that acoustic identifications by highly experienced birders present precision scores no higher than 0.94 and recall scores no higher than 0.89 (Farmer et al., 2012; Campbell & Francis, 2011), implying that BirdNET identifications with exceptionally high confidence scores might prove to be as accurate or more as those made by reasonably experienced birders.

Regarding precision scores in particular, our results are in line with previous studies (Sethi et al., 2021; Cole et al., 2022) showing that BirdNET identifications can be highly reliable, especially for common species (Fig. 6), provided that a sufficiently high minimum confidence threshold is used (Fig. S6). Indeed, confidence scores assigned to each identification by the algorithm are positively correlated with the actual probability (species-agnostic id_precision) that the identification in question is correct (Fig. S4). It is important to note, though, that the correspondence between confidence scores and precision is not of 1 to 1, i.e., a confidence score of 0.5 is not equivalent to a 50 % probability that the corresponding identification is correct. This correspondence can also vary across different species and recording settings, i.e., a given confidence score can translate into different levels of precision depending on the species identified and acoustic features such as the sampling rate of the recording (Wood & Kahl, 2024). Moreover, recent findings suggest that thresholds based on the use of contextual information in combination with BirdNET confidence scores can yield higher precision results than thresholds based on BirdNET confidence scores alone. More specifically, the precision of a given BirdNET identification might be less reliably correlated with its own confidence score than with the aggregate quality (average, median, minimum and maximum confidence scores) and quantity of BirdNET identifications obtained for the same species over a certain recording duration within the same site (Singer et al., 2024).

Our results further suggest that BirdNET precision can be improved not only by raising the minimum confidence threshold used, but also by filtering out species not having reached a minimum number of BirdNET detections across the whole dataset (Fig. S6). We found, however, that

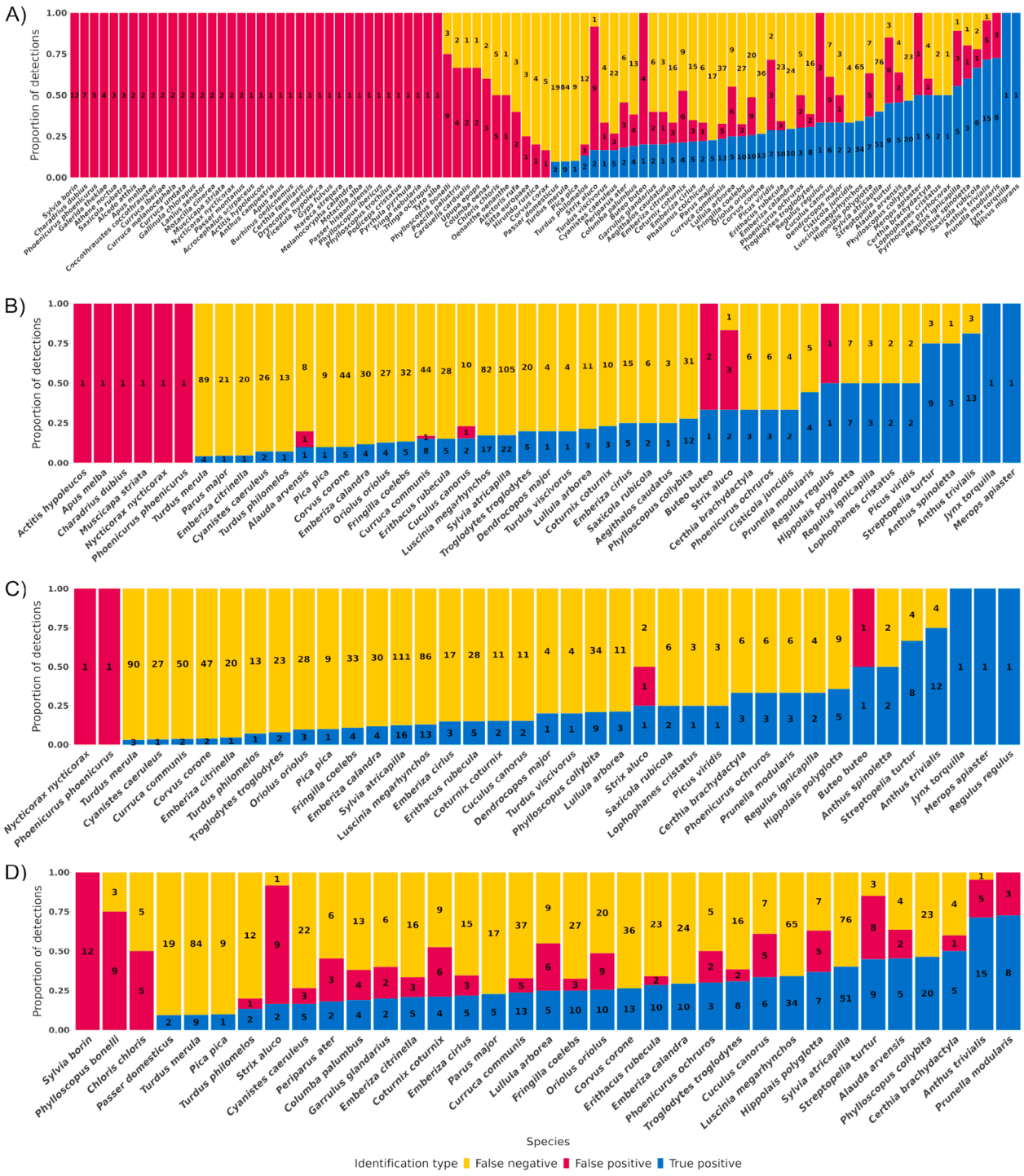


Fig. 4. Proportion of true positives, false positives and false negatives by bird species in BirdNET results with confidence scores (A) ≥ 0.3, (B) ≥ 0.55, (C) ≥ 0.65 and (D) ≥ 0.3, the latter only including species with ≥ 10 detections (either by BirdNET or by an expert). Results are calculated at the recording level and only recordings from the ZA-Pygar dataset are included.

species such as garden warblers (*Sylvia borin*) and Western Bonelli's Warblers (*Phylloscopus bonelli*) were identified by BirdNET on a high number of occasions, always incorrectly, and exceeded our minimum occurrence frequency threshold (Fig. 4d). Non-negligible numbers of FPs for other non-present species suggest caution around their identifications as well. On the positive side, these misidentifications seem to be effectively filtered out by establishing a strict enough global minimum confidence threshold (Fig. 4c), and could probably be filtered out even more effectively with the usage of species-specific thresholds. It is important to note, though, that background sounds can vary considerably between different locations and times of the year, so the list of problematic species can turn out to be substantially different in other spatio-temporal contexts. We therefore recommend that, prior to its use, a preliminary assessment of BirdNET be conducted with soundscapes from the target study site. This would allow researchers to preemptively identify the species appearing most often as FPs and to pay special attention to identifications of these species when examining BirdNET results. Another possible method to reduce FPs, as suggested by a recent

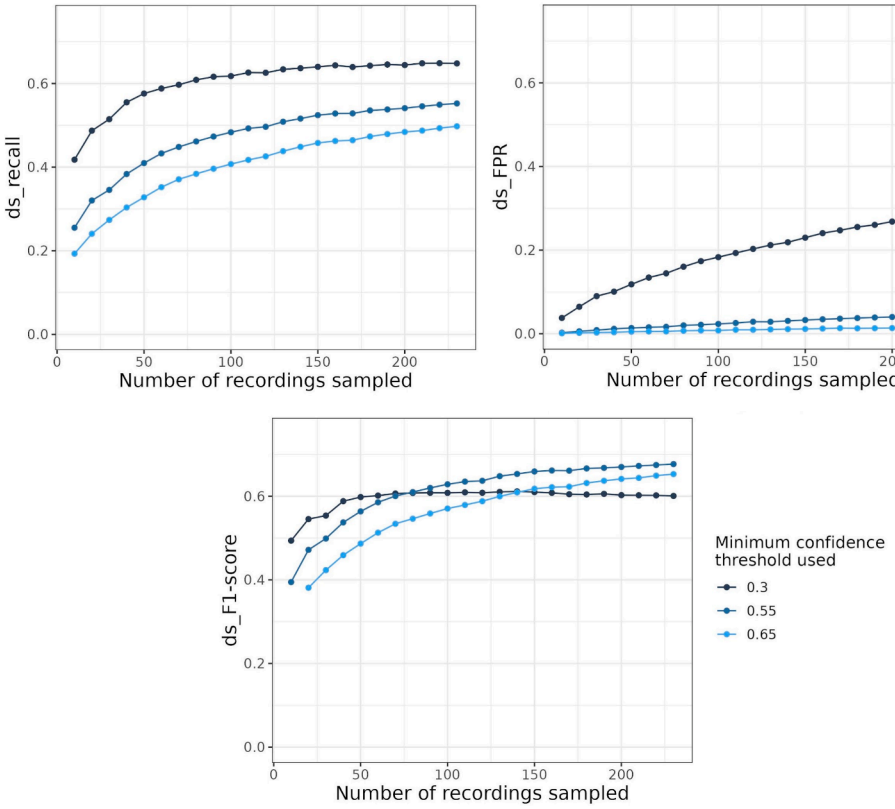


Fig. 5. Dataset performance depending on the number of recordings included in the analysis. ds_recall (top-left), ds_FPR (top-right) and ds_F1 scores (bottom) are shown for virtual datasets of increasing size depending on the minimum confidence threshold required to validate BirdNET identifications. Only recordings from the ZA-Pygar dataset are included.

study (Toenies & Rich, 2021), would be to filter out BirdNET identifications of diurnal species when analyzing nocturnal recordings, thus effectively preventing most dog barks and frog croaks from being misclassified as diurnal waterbird calls.

Our results (Fig. 6, S6, S7), along with those of previous studies (Kelly et al., 2023; Wood et al., 2023; Singer et al., 2024), make it clear that BirdNET precision and recall results can be highly variable across species, even when controlling for the minimum confidence threshold used. We thus recommend that, when using BirdNET for species-specific occupancy modeling, researchers conduct preliminary tests to determine the optimal minimum confidence threshold for the species of interest rather than relying on a threshold optimized for global performance when consistently used across a wide range of species. We estimate that this approach could allow BirdNET users targeting specific species to readily exceed the global performance benchmarks presented in this study. While we expect the use of species-specific minimum confidence thresholds to improve BirdNET performance when applied to the characterization of bird communities as well, the estimation of optimal thresholds for each species potentially present in a given study area represents a much greater challenge for this approach. It is also important to bear in mind that the optimal minimum confidence threshold to be used for any species will always be conditional on recording settings and on the importance that each user or research group assigns to minimizing FPs vs. minimizing FNs.

Despite our study being based on $>$ 600 recordings and $>$ 4000 manually annotated bird vocalizations (Table 1), the number of BirdNET identifications obtained for any given species is quite low, especially in the high confidence range. More precisely, no species has been identified by BirdNET in more than 29 recordings across all datasets (Table S2) or in more than 32 prediction segments in the ZA-Pygar dataset (Table S3) with a confidence score greater or equal than 0.9. This is partly due to the wide bird diversity covered by our acoustic datasets and the short duration of each recording analyzed, and partly because the vast majority of BirdNET identifications obtained have low confidence scores (Figure S5). More specifically, only 1692 out of 1,047,794 BirdNET identifications obtained across the three datasets have a confidence score $>=$ 0.9, corresponding to 0.16 % of all identifications. Hence, the limited per-species amount of high-confidence BirdNET identifications did not allow for the analysis of BirdNET performance based on species-specific minimum confidence thresholds. We therefore limited our analyses to the estimation of the best possible global threshold for optimized results at the community level. These values, while suboptimal for any given species, might provide researchers seeking to use BirdNET for the general characterization of bird community compositions (Hartig et al., 2023) with a basic picture of the performance that can be expected from the algorithm at different levels of precision exigency.

Another question that arises from our results is the degree to which the cross-species variability in BirdNET performance is due to (i) cross-species differences in the amount or quality of recordings available to train the algorithm with, (ii) the inherent difficulty in identifying vocalizations with certain acoustic patterns, or (iii) the identification difficulty arising from two or more species emitting highly similar sounds. Assuming that the first obstacle is the most tractable among the three mentioned, the higher the percentage of cross-species variation in BirdNET performance that is explained by recording availability, the more optimistic we should be regarding the improvement of the performance of the algorithm over time. In this regard, our results suggesting that 19 % of the variance in F1-scores between species can be explained by online recording availability (Fig. 6) provide ground for optimism about the future reliability of BirdNET. More concretely, this could mean that the diagnostic capacity of BirdNET for less common species could be considerably improved by enlarging the pool of available acoustic data for these species.

We presume that a higher availability of acoustic data could facilitate

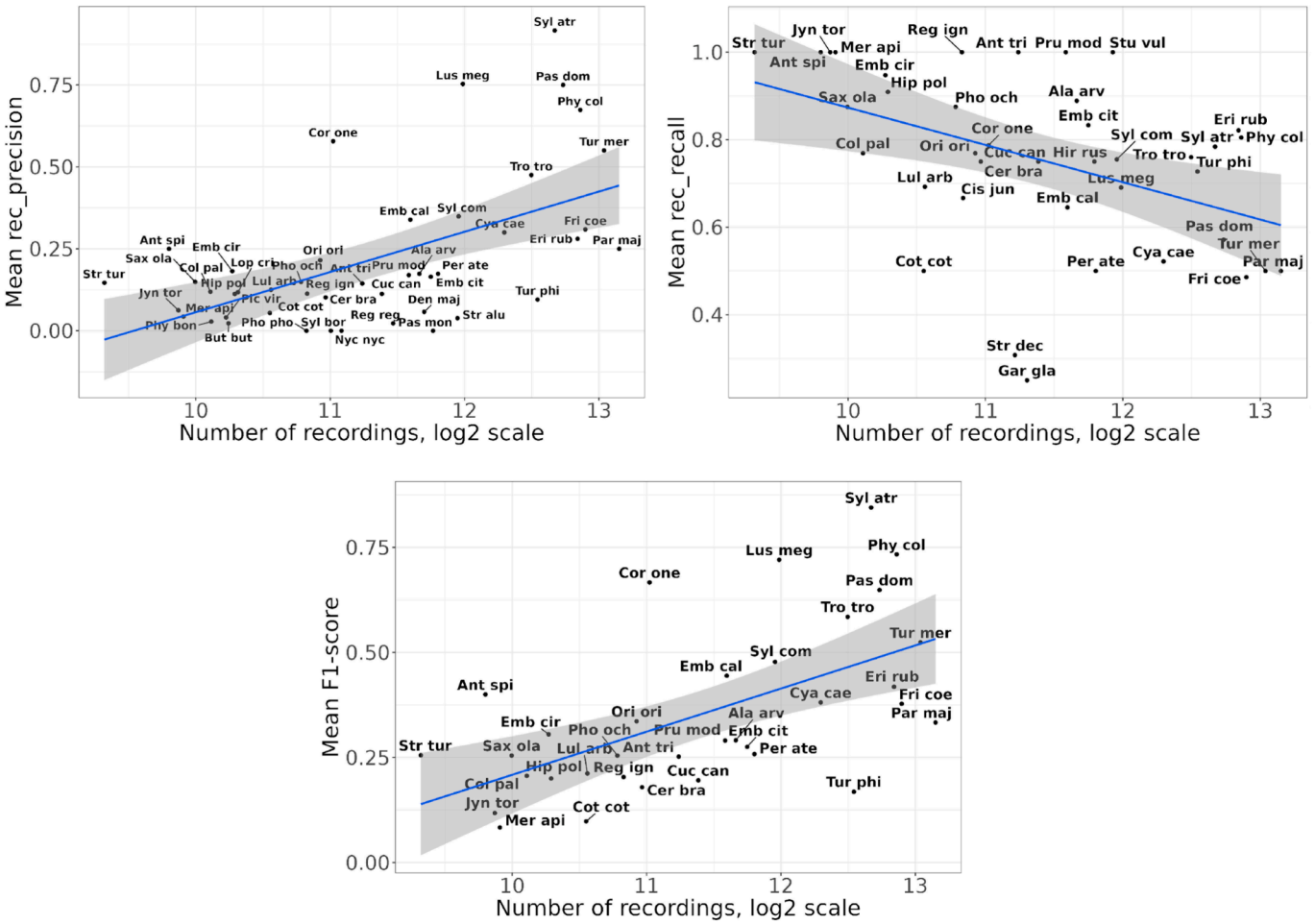


Fig. 6. Relationship between BirdNET performance and the number of foreground recordings available for each species on the Xeno-canto and Macaulay Library platforms. More specifically, the mean rec_precision (top-left), rec_recall (top-right) and rec_F1 (bottom) scores obtained for each species are plotted against the base-2 logarithm of the number of recordings available online for the species in question. Only species having been detected 10 or more times by BirdNET are included in the precision analysis, only species with at least 30 s of (manually annotated) total vocalization time are included in the recall analysis and only species meeting both criteria are included in the F1-score analysis. The minimum confidence threshold used is 0.1, and only recordings from the ZA-Pygar dataset are included.

BirdNET performance through two different mechanisms: (i) by enabling BirdNET to enlarge the size of its training dataset, and (ii) by improving the quality of the recordings comprising its training dataset. Acoustic quality scores are included in the metadata of recordings on Xeno-canto and the Macaulay Library, and higher-quality recordings were prioritized in the selection of training data for the BirdNET algorithm (Kahl et al., 2021). Hence, we would expect the average acoustic quality to be higher in the training datasets of species with larger pools of available recordings to choose from. Insomuch as both higher-quality recordings and larger training datasets could facilitate a better adjustment of the algorithm to the target bird vocalizations, these factors could provide a plausible explanation to the positive correlation observed between species-specific acoustic data availability and BirdNET predictive power.

We further found a significant linear relationship between the logarithm of the recording time analyzed and the proportion of recorded species having been correctly detected by BirdNET (Fig. 5). A possible explanation to this correlation would be that the short total vocalization time recorded for the least frequent species often results in no correct detections of these species by BirdNET. Thus, insofar as longer recording times contain a greater number of vocalizations per species in expectation, we should expect larger acoustic datasets to provide more chances for the algorithm to correctly detect less frequent species. We therefore hypothesize that a sufficiently high number of recording hours analyzed could partially compensate for the low recall scores obtained when high minimum confidence thresholds are used. Hence, were the necessary conditions to be met, BirdNET might prove successful in capturing most bird species present in a study site while also ensuring a relatively low number of FPs. These results, in line with recent findings (Toenies & Rich, 2021; Cole et al., 2022; Wood et al., 2021), are encouraging with respect to the potential of BirdNET to reliably and comprehensively describe bird communities in research projects with large amounts of ecoacoustic data at their disposal. Similar results have been obtained in studies targeting specific species and having used highly conservative minimum confidence thresholds (Brunk et al., 2023; Bielski et al., 2024), providing further evidence that ecologically meaningful results can be obtained even with a very strong emphasis on precision over recall when analyzing acoustic datasets of sufficient duration.

Our results also show that the predominance of biophony over anthropophony and geophony in recordings, as measured by NDSI (Kasten et al., 2012), appears to be positively correlated with recall scores (see Appendix S6, Fig. S11). We propose three possible hypotheses explaining this correlation: (i) anthropogenic and geological sounds might hinder the performance of the algorithm by concealing or blurring bird vocalizations, (ii) BirdNET might have mostly been trained with recordings with relatively low levels of background noise, thus being less well calibrated to identify bird vocalizations in noisy recordings, and (iii) high NDSI scores can be indicative of high-amplitude bird vocalizations, which might to be easier to identify than low-amplitude bird vocalizations (Pérez Granados, 2023). If the second hypothesis were

correct, we consider it plausible that the application of a noise reduction filter to recordings prior to their analysis with BirdNET could result in an enhanced performance of the algorithm.

Overall, the results of this study may provide valuable information for research groups considering the possibility of using BirdNET for the automated analysis of passively collected environmental soundscapes. However, these results provide an overview of the current potential of BirdNET in a very specific context, and therefore cannot be reliably extrapolated to other regions with different habitats, levels of anthropogenic pressure or bird community compositions. Care must also be taken when extrapolating our results to other areas with similar but geographically distant bird communities, as bird songs and calls can present subtle variations across different regions (Slabbekoorn & Smith, 2002). Likewise, the analyses performed to evaluate the effect of dataset size on BirdNET performance have been conducted with subset sizes of very short durations (on the order of minutes or hours), so our findings might fail to generalize to acoustic datasets of larger scale. The same caveat applies to the optimal values for BirdNET configuration parameters found in this study as well. Higher overlap values between consecutive prediction segments might improve BirdNET performance by facilitating the capture of a substantial part of any given bird vocalization within a single prediction segment (Fig. S1), but they increase processing times as well. In small datasets with few vocalizations of any given species, the improvement in detection capacity might largely compensate for the potential increase in processing time, but the compromise can look very different in large datasets, which provide more opportunities for BirdNET to detect any given species. In these cases, the increased processing time can become a substantial burden without providing much benefit in terms of recall.

Another point that should be borne in mind is that this study starts from the fundamental assumption that all our expert identifications are correct, which is why we used them as the reference to compare BirdNET against. As already mentioned, even highly experienced birders seem to fall short of infallibility (Farmer et al., 2012; Campbell & Francis, 2011), so the possibility that some of the expert identifications included in this study are incorrect cannot be entirely ruled out. As a concluding note, in light of the improvement detected between the recall scores of BirdNET-Analyzer v2.4 and those of older versions of the algorithm –BirdNET-Lite and BirdNET-Analyzer v2.1 to 2.3– that we previously tested with the same data, we consider it plausible that future releases will be capable of identifying bird vocalizations even more reliably and exhaustively. This finding, coupled with the fact that the usage of species-specific minimum confidence thresholds could readily improve the performance benchmarks presented in this study, implies that our results should not be assumed to accurately estimate the best-case performance of BirdNET, but they should rather be interpreted as a lower bound for its potential.

CRediT authorship contribution statement

David Funosas: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Luc Barbaro:** Writing – review & editing, Validation, Supervision, Resources, Investigation. **Laura Schillé:** Writing – review & editing, Resources, Investigation, Formal analysis, Data curation. **Arnaud Elger:** Writing – review & editing. **Bastien Castagneyrol:** Writing – review & editing, Resources. **Maxime Cauchoix:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The code used for the analyses performed in this study, as well as both BirdNET and manual identifications/annotations for the three acoustic datasets studied, can be found in the following GitHub repository: https://anonymous.4open.science/r/BirdNET_study_Occitanie_2023-8D8F/. The raw acoustic data analyzed is available upon request. We share our code through a link in the "Data availability statement" section

Acknowledgements

We thank Arndt Hampe, Maarten de Groot, Thomas Boivin, Vojtech Lanta, Lyudmila Pukinskaya, Frédéric Archaux, Thomas Perot, Márton Molnár, Nickolay Sedikhin, Valentin Moser, Elina Mäntylä, Jan Grünwald, Karthik Thrikkadeeri, Anders Mårell, Mona C Bjørn, Becky Thomas, Andreas Prinzing and Lucian Grosu for manual bird identifications of the TreeBodyguards European dataset. We are grateful to Michel Beal, Jean-Luc Témoin , Valérie Delage and Laurent Tillon from Office National des Forêts for their help while recording the Rambouillet dataset. We also thank Thierry Feuillet, Amandine Gasc, Jérémy Froidevaux, Kevin Darras, Yves Bas, Elena Valdés-Correcher, Thomas Delattre for helpful advice and support. We are also indebted to the CESAB Acoucene working group for fruitful discussions.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecolind.2024.112146.

References

Arif, M., Hedley, R., Bayne, E., 2020. Testing the Accuracy of a birdNET, Automatic bird song Classifier [WWW Document]. ERA. https://doi.org/10.7939/r3-6khb-kz18.

Barbaro, L., Sourdril, A., Froidevaux, J.S.P., Cauchoix, M., Calatayud, F., Deconchat, M., Gasc, A., 2022. Linking acoustic diversity to compositional and configurational heterogeneity in mosaic landscapes. Landsc Ecol. 37, 1125–1143. https://doi.org/10.1007/s10980-021-01391-8.

Barbaro, L., Froidevaux, J.S.P., Valdés-Correcher, E., Calatayud, F., Tillon, L., Sourdril, A., 2023. COVID-19 shutdown revealed higher acoustic diversity and vocal activity of flagship birds in old-growth than in production forests. Sci. Total Environ. 901, 166328 https://doi.org/10.1016/j.scitotenv.2023.166328.

Bielski, L., Cansler, C.A., McGinn, K., Peery, M.Z., Wood, C.M., 2024. Can the Hermit Warbler (Setophaga occidentalis) serve as an old-forest indicator species in the Sierra Nevada? J. Field Ornithol. 95 https://doi.org/10.5751/JFO-00390-950104.

Blumstein, D.T., Mennill, D.J., Clemins, P., Girod, L., Yao, K., Patricelli, G., Deppe, J.L., Krakauer, A.H., Clark, C., Cortopassi, K.A., Hanser, S.F., McCowan, B., Ali, A.M., Kirschel, A.N.G., 2011. Acoustic monitoring in terrestrial environments using microphone arrays: applications, technological considerations and prospectus. J. Appl. Ecol. 48, 758–767. https://doi.org/10.1111/j.1365-2664.2011.01993.x.

Bobay, L.R., Taillie, P.J., Moorman, C.E., 2018. Use of autonomous recording units increased detection of a secretive marsh bird. J. Field Ornithol. 89, 384–392. https://doi.org/10.1111/jofo.12274.

Borowiec, M.L., Dikow, R.B., Frandsen, P.B., McKeeken, A., Valentini, G., White, A.E., 2022. Deep learning as a tool for ecology and evolution. Methods Ecol. Evol. 13, 1640–1660. https://doi.org/10.1111/2041-210X.13901.

Bradfer-Lawrence, T., Desjonqueres, C., Eldridge, A., Johnston, A., Metcalf, O., 2023. Using acoustic indices in ecology: Guidance on study design, analyses and interpretation. Methods Ecol. Evol. 14, 2192–2204. https://doi.org/10.1111/2041-210X.14194.

Brüggemann, L., Schütz, B., Aschenbruck, N., 2021. Ornithology meets the IoT: Automatic Bird Identification, Census, and Localization. In: 2021 IEEE 7th World Forum on Internet of Things (WF-IoT). Presented at the 2021 IEEE 7th World Forum on Internet of Things (WF-IoT), pp. 765–770. https://doi.org/10.1109/WF-IoT51360.2021.9595401.

Brunk, K.M., Gutiérrez, R.J., Peery, M.Z., Cansler, C.A., Kahl, S., Wood, C.M., 2023. Quail on fire: changing fire regimes may benefit mountain quail in fire-adapted forests. Fire Ecology 19, 19. https://doi.org/10.1186/s42408-023-00180-9.

Campbell, M., Francis, C.M., 2011. Using stereo-microphones to evaluate observer variation in north american breeding bird survey point counts. Auk 128, 303–312. https://doi.org/10.1525/auk.2011.10005.

Clare, J.D.J., Townsend, P.A., Zuckerberg, B., 2021. Generalized model-based solutions to false-positive error in species detection/nondetection data. Ecology 102, e03241.

Cole, J.S., Michel, N.L., Emerson, S.A., Siegel, R.B., 2022. Automated bird sound classifications of long-duration recordings produce occupancy model outputs similar

to manually annotated data. Ornithological Appl. 124, duac003. https://doi.org/10.1093/ornithapp/duac003.
Darras, K., Batáry, P., Furnas, B., Celis-Murillo, A., Van Wilgenburg, S.L., Mulyani, Y.A., Tscharntke, T., 2018. Comparing the sampling performance of sound recorders versus point counts in bird surveys: A meta-analysis. J. Appl. Ecol. 55, 2575–2586. https://doi.org/10.1111/1365-2664.13229.
Darras, K., Batáry, P., Furnas, B.J., Grass, I., Mulyani, Y.A., Tscharntke, T., 2019. Autonomous sound recording outperforms human observation for sampling birds: a systematic map and user guide. Ecol. Appl. 29, e01954.
Davis, J., Goadrich, M., 2006. The relationship between Precision-Recall and ROC curves, in: Proceedings of the 23rd International Conference on Machine Learning, ICML '06. Association for Computing Machinery, New York, NY, USA, pp. 233–240. https://doi.org/10.1145/1143844.1143874.
Dufourq, E., Durbach, I., Hansford, J.P., Hoepfner, A., Ma, H., Bryant, J.V., Stender, C.S., Li, W., Liu, Z., Chen, Q., Zhou, Z., Turvey, S.T., 2021. Automated detection of Hainan gibbon calls for passive acoustic monitoring. Remote Sens. Ecol. Conserv. 7, 475–487. https://doi.org/10.1002/rse2.201.
Farmer, R.G., Leonard, M.L., Horn, A.G., 2012. Observer Effects and Avian-Call-Count Survey Quality: Rare-Species Biases and Overconfidence. Auk 129, 76–86. https://doi.org/10.1525/auk.2012.11129.
Gasc, A., Sueur, J., Pavoine, S., Pellens, R., Grandcolas, P., 2013. Biodiversity sampling using a global acoustic approach: contrasting sites with microendemics in New Caledonia. PLoS One 8, e65311.
Hartig, F., Abrego, N., Bush, A., Chase, J.M., Guillera-Arroita, G., Leibold, M.A., Ovaskainen, O., Pellissier, L., Pichler, M., Poggiato, G., Pollock, L., Si-Moussi, S., Thuiller, W., Viana, D.S., Warton, D.I., Zurell, D., Yu, D.W., 2023. Novel community data in ecology-properties and prospects. Trends Ecol. Evol. https://doi.org/10.1016/j.tree.2023.09.017.
Höchst, J., Bellafkir, H., Lampe, P., Vogelbacher, M., Mühling, M., Schneider, D., Lindner, K., Rösner, S., Schabo, D.G., Farwig, N., Freisleben, B., 2022. Bird@Edge: Bird Species Recognition at the Edge, in: Koulali, M.-A., Mezini, M. (Eds.), Networked Systems, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 69–86. https://doi.org/10.1007/978-3-031-17436-0_6.
Kahl, S., Wood, C.M., Eibl, M., Klinck, H., 2021. BirdNET: A deep learning solution for avian diversity monitoring. Eco. Inform. 61, 101236 https://doi.org/10.1016/j.ecoinf.2021.101236.
Kasten, E.P., Gage, S.H., Fox, J., Joo, W., 2012. The remote environmental assessment laboratory's acoustic library: An archive for studying soundscape ecology. Eco. Inform. 12, 50–67. https://doi.org/10.1016/j.ecoinf.2012.08.001.
Kelly, K.G., Wood, C.M., McGinn, K., Kramer, H.A., Sawyer, S.C., Whitmore, S., Reid, D., Kahl, S., Reiss, A., Eiseman, J., Berigan, W., Keane, J.J., Shaklee, P., Gallagher, L., Munton, T.E., Klinck, H., Gutiérrez, R.J., Peery, M.Z., 2023. Estimating population size for California spotted owls and barred owls across the Sierra Nevada ecosystem with bioacoustics. Ecol. Ind. 154, 110851 https://doi.org/10.1016/j.ecolind.2023.110851.
Knight, E., Hannah, K., Foley, G., Scott, C., Brigham, R., Bayne, E., 2017. Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs. Avian Conservation Ecol. 12 https://doi.org/10.5751/ACE-01114-120214.
Liu, M., Sun, Q., Brewer, D.E., Gehring, T.M., Eickholt, J., 2022. An Ornithologist's guide for including machine learning in a workflow to identify a secretive focal species from recorded audio. Remote Sens. (Basel) 14, 3816. https://doi.org/10.3390/rs14153816.
Macaulay, 2023. The World's Premier Scientific Archive of Natural History Audio, Video, and Photographs. https://www.macaulaylibrary. org/about/.
Malamut, E.J., 2022. Using Autonomous Recording Units and Image Processing to Investigate Patterns in Avian Singing Activity and Nesting Phenology. UCLA. https://escholarship.org/uc/item/92p9z0gp.
Melo, I., Llusia, D., Bastos, R.P., Signorelli, L., 2021. Active or passive acoustic monitoring? Assessing methods to track anuran communities in tropical savanna wetlands. Ecol. Ind. 132, 108305 https://doi.org/10.1016/j.ecolind.2021.108305.
Ouin, A., Andrieu, E., Vialatte, A., Balent, G., Barbaro, L., Blanco, J., Ceschia, E., Clement, F., Fauvel, M., Gallai, N., Hewison, A.J.M., Jean-François, D., Kephaliacos, C., Macary, F., Probst, A., Probst, J.-L., Ryschawy, J., Sheeren, D., Sourdril, A., Tallec, T., Verheyden, H., Sirami, C., 2021. Chapter Two - Building a shared vision of the future for multifunctional agricultural landscapes. Lessons from a long term socio-ecological research site in south-western France. In: Bohan, D.A., Dumbrell, A.J., Vanbergen, A.J. (Eds.), Advances in Ecological Research, the Future of Agricultural Landscapes, Part III. Academic Press, pp. 57–106. https://doi.org/10.1016/bs.aecr.2021.05.001.
Pérez Granados, C., 2023. A First Assessment of Birdnet Performance at Varying Distances: A Playback Experiment 70, 257–269. https://doi.org/10.13157/arla.70.2.2023.sc1.
Pérez-Granados, C., 2023. BirdNET: applications, performance, pitfalls and future opportunities. Ibis 165, 1068–1075. https://doi.org/10.1111/ibi.13193.
Schillé, L., Valdés-Correcher, E., Archaux, F., Bălăcenoiu, F., Bjørn, M.C., Bogdziewicz, M., Boivin, T., Branco, M., Damestoy, T., de Groot, M., Dobrosavljević, J., Duduman, M.-L., Dulaurent, A.-M., Green, S., Grünwald, J., Eötvös, C.B., Faticov, M., Fernandez-Conradi, P., Flury, E., Funosas, D., Galmán, A., Gossner, M.M., Gripenberg, S., Grosu, L., Hagge, J., Hampe, A., Harvey, D., Houston, R., Isenmann, R., Kavčič, A., Kozlov, M.V., Lanta, V., Le Tilly, B., Lopez-Vaamonde, C., Mallick, S., Mäntylä, E., Mårell, A., Milanović, S., Molnár, M., Moreira, X., Moser, V., Mrazova, A., Musolin, D.L., Perot, T., Piotti, A., Popova, A.V., Prinzing, A., Pukinskaya, L., Sallé, A., Sam, K., Sedikhin, N.V., Shabarova, T., Tack, A.J.M., Thomas, R., Thrikkadeeri, K., Toma, D., Vaicaityte, G., van Halder, I., Varela, Z., Barbaro, L., Castagneyrol, B., n.d. Decomposing drivers in avian insectivory: Large-scale effects of climate, habitat and bird diversity. Journal of Biogeography. https://doi.org/10.1111/jbi.14808.
Sebastián-González, E., Camp, R., Tanimoto, A., de Oliveira, P., Lima, B., Marques, T., Hart, P., 2018. Density estimation of sound-producing terrestrial animals using single automatic acoustic recorders and distance sampling. Avian Conserv. Ecol. 13 https://doi.org/10.5751/ACE-01224-130207.
Sethi, S.S., Fossøy, F., Cretois, B., Rosten, C.M., 2021. Management relevant applications of acoustic monitoring for Norwegian nature – The Sound of Norway, 31. Norsk institutt for naturforskning (NINA). https://brage.nina.no/nina-xmlui/handle/11250/2832294.
Shaw, T., Hedes, R., Sandstrom, A., Ruete, A., Hiron, M., Hedblom, M., Eggers, S., Mikusiński, G., 2021. Hybrid bioacoustic and ecoacoustic analyses provide new links between bird assemblages and habitat quality in a winter boreal forest. Environ. Sustain. Indicators 11, 100141. https://doi.org/10.1016/j.indic.2021.100141.
Shonfield, J., Bayne, E., 2017. Autonomous recording units in avian ecological research: Current use and future applications. Avian Conservation Ecol. 12, 14. https://doi.org/10.5751/ACE-00974-120114.
Singer, D., Hagge, J., Kamp, J., Hondong, H., Schuldt, A., 2024. Aggregated time-series features boost species-specific differentiation of true and false positives in passive acoustic monitoring of bird assemblages. Remote Sens. Ecol. Conserv. https://doi.org/10.1002/rse2.385.
Slabbekoorn, H., Smith, T.B., 2002. Bird song, ecology and speciation. Philosophical Transactions of the Royal Society of London. Series b: Biological Sciences 357, 493–503. https://doi.org/10.1098/rstb.2001.1056.
Stowell, D., 2021. Computational bioacoustics with deep learning: a review and roadmap. https://doi.org/10.48550/arXiv.2112.06725.
Sugai, L.S.M., Silva, T.S.F., Ribeiro Jr, J.W., Llusia, D., 2019. Terrestrial Passive Acoustic Monitoring: Review and Perspectives. Bioscience 69, 15–25. https://doi.org/10.1093/biosci/biy147.
Symes, L.B., Kittelberger, K.D., Stone, S.M., Holmes, R.T., Jones, J.S., Castaneda Ruvalcaba, I.P., Webster, M.S., Ayres, M.P., 2022. Analytical approaches for evaluating passive acoustic monitoring data: A case study of avian vocalizations. Ecol. Evol. 12, e8797.
Toenies, M., Rich, L., 2021. Advancing bird survey efforts through novel recorder technology and automated species identification. California Fish Wildlife J. 107, 56–70. https://doi.org/10.51492/cfwj.107.5.
Tolkova, I., Chu, B., Hedman, M., Kahl, S., Klinck, H., 2021. Parsing Birdsong with Deep Audio Embeddings. https://doi.org/10.48550/arXiv.2108.09203.
Towsey, M., Zhang, L., Cottman-Fields, M., Wimmer, J., Zhang, J., Roe, P., 2014. Visualization of Long-duration Acoustic Recordings of the Environment. Procedia Computer Science, 2014 International Conference on Computational Science 29, 703–712. https://doi.org/10.1016/j.procs.2014.05.063.
Verreycken, E., Simon, R., Quirk-Royal, B., Daems, W., Barber, J., Steckel, J., 2021. Bio-acoustic tracking and localization using heterogeneous, scalable microphone arrays. Commun. Biol. 4, 1–11. https://doi.org/10.1038/s42003-021-02746-2.
Wood, C.M., Kahl, S., Chaon, P., Peery, M.Z., Klinck, H., 2021. Survey coverage, recording duration and community composition affect observed species richness in passive acoustic surveys. Methods Ecol. Evol. 12, 885–896. https://doi.org/10.1111/2041-210X.13571.
Wood, C.M., Kahl, S., Rahaman, A., Klinck, H., 2022. The machine learning–powered BirdNET App reduces barriers to global bird research by enabling citizen science participation. PLoS Biol. 20, e3001670.
Wood, C.M., Kahl, S., Barnes, S., Van Horne, R., Brown, C., 2023. Passive acoustic surveys and the BirdNET algorithm reveal detailed spatiotemporal variation in the vocal activity of two anurans. Bioacoustics 32, 532–543. https://doi.org/10.1080/09524622.2023.2211544.
Wood, C.M., Kahl, S., 2024. Guidelines for appropriate use of BirdNET scores and other detector outputs. J. Ornithol. https://doi.org/10.1007/s10336-024-02144-5.
Xeno-canto, 2023. Sharing Bird Sounds from Around the World. https://www.xeno-canto.org/about/xeno-canto.