Ecological Indicators 182 (2026) 114550

Contents lists available at ScienceDirect

Ecological Indicators

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

A global assessment of BirdNET performance: Differences among continents, biomes, and species

David Funosas [a,b,*,1], Esther Sebastián-González [c,d,1], Jon Morant [c], Oscar H. Marín Gómez [e], Irene Mendoza [f,g], Miguel A. Mohedano-Muñoz [h], Eduardo Santamaría [f], Giulia Bastianelli [i], Alba Márquez-Rodríguez [j], Michał Budka [k], Gerard Bota [l], Cristina D. Alonso-Moya [l], José M. de la Peña-Rubio [m], Eladio L. García de la Morena [n], Manu Santa-Cruz [o], Pablo de la Nava [p], Mario Fernández-Tizón [q], Hugo Sánchez-Mateos [r], Adrián Barrero [s,t], Juan Traba [s,t], Tomasz S. Osiejuk [k], Patrick J. Hart [u], Amanda K. Navine [u], Andrés F. Montoya Muñoz [v], Carlos B. de Araújo [w], Gabriel L.M. Rosa [x,y], Ingrid M.D. Torres [y], Ana L.C. Catalano [y], Cassio Rachid Simões [y], Diego Llusia [s,t,w], Manuel B. Morales [s,t], Pablo Acebes [s,t], Juan A. Medina [z], Nicholas Brown [s,aa], Christos Astaras [ab], Ilias Karmiris [ab], Elizabeth Navarrete [ac], Maxime Cauchoix [a], Luc Barbaro [ac], Dominik Arend [ad], Sandra Müeller [ad], Fernando González-García [ae], Alberto González-Romero [ae], Christos Mammides [af], Michaelangelo Pontikis [ag], Giordano Jacuzzi [ah], Julian D. Olden [ah], Sara P. Bombaci [ai], Gabriel Marcacci [aj], Alain Jacot [aj], Juan P. Zurano [w,ak], Elena Gangenova [w,ak], Diego Varela [w,ak], Facundo Di Sallo [w,ak], Gustavo A. Zurita [w,ak], Andrey Atemasov [al], Junior A. Tremblay [am], Vincent Lamarre [am], Anja Hutschenreiter [an], Alan Monroy-Ojeda [ao], Mauricio Díaz-Vallejo [ap], Sergio Chaparro-Herrera [aq], Robert A. Briers [ar], Renata Sousa-Lima [as], Thiago Pinheiro [as], Wigna C. Da Silva [as], Alice Calvente [at], Raiane V. Paz [au,av], Carlos Salustio-Gomes [au,av], Dorgival D. Oliveira-Júnior [au,av], Cicero S. lima-Santos [au,av], Mauro Pichorim [au,av], Anamaria Dal Molin [aw], Alexandre Antonelli [ax,ay,az], Svetlana Gogoleva [ba,bb], Igor Palko [bb], Hiếu V. Trong [bb], Marina H.L. Duarte [bc], Natalia dos Santos Saturnino [bd], Samuel R. Silva [bd], Ana Rainho [be], Paula Lopes [be,bf], Karl-L. Schuchmann [bg,bh], Marinêz I. Marques [bg], Ana S. de Oliverira Tissiani [bg], Nick A. Littlewood [bi], Mao-Ning Tuanmu [bj], Sebastian Kepfer-Rojas [bk], Andrea L. Aguilera [bl], Lluís Brotons [l,bm,bn], Mariano J. Feldman [l], Louis Imbeau [bo], Pooja Panwar [bp], Aaron S. Weed [bq], Anant Dehwal [br], Alfredo Attisano [bs], Jörn Theuerkauf [bs], Eben Goodale [bt], Kevin F.A. Darras [bu], Cristian Pérez-Granados [l,bv]

[a] *Station d'Écologie Théorique et Expérimentale, SETE, CNRS. 09200, Moulis, France*
[b] *Université de Toulouse. 31077, Toulouse, France*
[c] *Department of Ecology, University of Alicante, 03690 San Vicente del Raspeig, Alicante, Spain*
[d] *Instituto Multidisciplinar para el Estudio del Medio "Ramón Margalef", University of Alicante, 03690 San Vicente del Raspeig, Alicante, Spain*
[e] *Fundación Soy Conservación, Caicedonia, Colombia. Programa de Biología, Grupo de Investigación en Biodiversidad y Biotecnología (GIBUQ), Universidad del Quindío, Armenia, Quindío, Colombia*
[f] *Estación Biológica de Doñana (CSIC), Department of Ecology and Evolution, Avda. Américo Vespucio, 26, 41092 Sevilla, Spain*
[g] *University of Sevilla, Department of Plant Biology and Ecology, Faculty of Biology, Avda Reina Mercedes s/n, 41012 Sevilla, Spain*

* Corresponding author.
E-mail address: davidfunosas@gmail.com (D. Funosas).

https://doi.org/10.1016/j.ecolind.2025.114550
Received 15 October 2025; Received in revised form 2 December 2025; Accepted 13 December 2025
Available online 6 January 2026

[h] Escuela Técnica Superior de Ingeniería Informática, Universidad Rey Juan Carlos, 28933, Móstoles, Madrid, Spain
[i] ICTS-Doñana, Estación Biológica de Doñana (EBD), CSIC, C/ Américo Vespucio 26, 41092 Sevilla, Spain
[j] Institute of Marine Research (INMAR), International Campus of Excellence in Marine Science (CEIMAR), University of Cádiz, 11510, Puerto Real, Cádiz, Spain
[k] Department of Behavioural Ecology, Institute of Environmental Biology, Faculty of Biology, Adam Mickiewicz University Poznań, Poland
[l] Biodiversity Conservation and Management Programme, Forest Science and Technology Center of Catalonia (CTFC), 25280 Lleida, Spain
[m] Blue Nature Birding and Nature Tours SL, Calle Rufino Blanco 17, 28028 Madrid, Spain
[n] Biodiversity Node, Sector Foresta, 17. 1°B, 28760, Tres Cantos, Madrid, Spain
[o] Eurofins MITOX B.V, Science Park 408, 1098 XH Amsterdam, Netherlands
[p] SEO/BirdLife, 28052 Madrid, Spain
[q] Instituto de Investigación en Recursos Cinegéticos, IREC-CSIC-UCLM-JCCM, Ciudad Real, Spain
[r] Iduna Tours, 10720, Villar de Plasencia, Cáceres, Spain
[s] Terrestrial Ecology Group (TEG-UAM), Departamento de Ecología, Universidad Autónoma de Madrid, 28049 Madrid, Spain
[t] Centro de Investigación en Biodiversidad y Cambio Global (CIBC-UAM), Universidad Autónoma de Madrid, 28049 Madrid, Spain
[u] Department of Biology, University of Hawai'i at Hilo. 200 W. Kawili St, Hilo, Hawai'i, USA, 96720
[v] Colección de ornitología (COUQ), Universidad del Quindío, Armenia, Quindío, Colombia
[w] Atlantic Forest Biodiversity Observatory, Instituto de Biología Subtropical, CONICET-Universidad Nacional de Misiones, Puerto Iguazú, Argentina
[x] Laboratório de Biodiversidade, Universidade Estadual de Londrina, Londrina, Brazil
[y] ConservaSom, João Pessoa, Paraíba, Brazil
[z] BUTEO Environmental Initiatives, Mojados, Valladolid, Spain
[aa] Department of Life Sciences, Imperial College London, Exhibition Road, South Kensington, London, SW7 2AZ, United Kingdom.
[ab] Forest Research Institute, ELGO-DIMITRA, Loutra Thermis, Thessaloniki 57006, Greece
[ac] Dynafor, INRAE-INPT, University of Toulouse. 31326, Castanet-Tolosan, France
[ad] Geobotany, Faculty of Biology, University of Freiburg, Freiburg, Germany
[ae] Red Biología y Conservación de Vertebrados, Instituto de Ecología, A.C. Carretera Antigua a Coatepec No. 351, El Haya, Xalapa, Veracruz, Mexico
[af] Nature Conservation Unit, Frederick University, Gianni Freiderikou 7, Nicosia 1036, Cyprus
[ag] Department of Biological Sciences, University of Cyprus, Nicosia, Cyprus
[ah] School of Aquatic and Fishery Sciences, University of Washington, Seattle, WA, 98195, USA
[ai] Department of Fish, Wildlife, and Conservation Biology, Colorado State University, 1474 Campus Delivery, Fort Collins, CO 80521, USA
[aj] Swiss Ornithological Institute, Seerose 1, 6204 Sempach, Switzerland
[ak] Asociación Civil Centro de Investigaciones del Bosque Atlántico (CeIBA), Bertoni 85, Puerto Iguazú, Misiones, Argentina
[al] V.N.Karazin Kharkiv National University, Kharkiv, Ukraine
[am] Environment and Climate Change Canada, 801-1550 Ave d'Estimauville, Québec, Canada, G1J 0C3
[an] Instituto de Investigaciones en Ecosistemas y Sustentabilidad, Universidad Autónoma de México, Morelia, Michoacán, Mexico
[ao] Centro de Investigaciones Tropicales, Universidad Veracruzana, Xalapa, Veracruz, Mexico
[ap] Laboratorio de Bioclimatología. Red Biología Evolutiva, Instituto de Ecología, A.C. Carretera Antigua a Coatepec No. 351, El Haya, Xalapa, Veracruz, Mexico
[aq] Proyecto Atlapetes, Antioquia, Colombia; Laboratorio de Ecología Evolutiva y Urbana, Universidad del Norte, Barranquilla, Colombia
[ar] Centre for Conservation and Restoration Science, School of Applied Sciences, Edinburgh Napier University, Sighthill Campus, Edinburgh, EH11 4BN, UK
[as] Laboratory of Bioacoustics / EcoAcoustic Research Hub. Biosciences Center. Universidade Federal do Rio Grande do Norte. Campus Universitário, Lagoa Nova, Natal, RN 59078-970, Brazil
[at] Departamento de Botânica e Zoologia, Universidade Federal do Rio Grande do Norte, Campus Universitário, 59078-970 - Lagoa Nova, Natal, RN, Brazil
[au] Laboratório de Ornitologia, Departamento de Botânica e Zoologia, Universidade Federal do Rio Grande do Norte, Campus Universitário, Lagoa Nova, Natal, RN, 59078-970, Brazil
[av] Programa de Pós-Graduação em Ecologia, Universidade Federal do Rio Grande do Norte, Campus Universitário, Lagoa Nova, Natal, RN, 59078-970, Brazil
[aw] Department of Microbiology and Parasitology, Biosciences Center, Universidade Federal do Rio Grande do Norte, 59078-970 Natal, RN, Brazil
[ax] Royal Botanic Gardens, Kew, Richmond, Surrey, TW9 3AE, UK
[ay] Department of Biology, University of Oxford, South Parks Road, Oxford, OX1 3RB, 96, United Kingdom
[az] Gothenburg Global Biodiversity Centre, Department of Biological and Environmental Sciences, University of Gothenburg, Box 463, 405 30 Göteborg, Sweden
[ba] A.N. Severtsov Institute of Ecology and Evolution, Russian Academy of Sciences, Leninsky Ave. 33, Moscow, 119071, Russia
[bb] Southern Branch of the Joint Vietnam-Russia Tropical Science and Technology Research Center, HoChiMinh City, Viet Nam
[bc] Environmental Research and Innovation Centre (ERIC), School of Science, Engineering and Environment. University of Salford, Manchester, United Kingdom
[bd] Graduate Program in Biodiversity and Environment. Pontifical Catholic University of Minas Gerais, Belo Horizonte, Brazil
[be] CE3C-Centre for Ecology, Evolution and Environmental Changes, CHANGE-Global Change and Sustainability Institute & Departamento de Biologia Animal, Faculdade de Ciências, Universidade de Lisboa, Lisbon 1749-016, Portugal
[bf] SPEA - Portuguese Society for the Study of Birds, Lisbon 1700-031, Portugal
[bg] Computational Bioacoustics Research Unit (CO.BRA), Institute for Science and Technology in Wetlands (INAU), Federal University of Mato Grosso (UFMT), Cuiabá 78060-900, Brazil
[bh] Ornithology, Zoological Research Museum A. Koenig (ZFMK), 53113 Bonn, Germany
[bi] Scotland's Rural College, Craibstone Estate, Bucksburn, Aberdeen, AB21 9YA, UK
[bj] Biodiversity Research Center, Academia Sinica, Taipei, Taiwan
[bk] Department of Geosciences and Natural Resource Management, University of Copenhagen. Rolighedsvej 23, Frederiksberg C. Copenhagen, Denmark
[bl] Centro de Datos para la Conservación, Centro de Estudios Conservacionistas, Universidad de San Carlos de Guatemala, Avenida La Reforma 0-63, zona 10, Ciudad de Guatemala, Guatemala
[bm] CREAF, 08193 Cerdanyola del Vallès, Spain
[bn] CSIC, 08193 Cerdanyola del Vallès, Spain
[bo] Institut de recherche sur les forêts, Université du Québec en Abitibi-Témiscamingue, 445, boul de l'Université, Rouyn-Noranda, Québec J9X 5E4, Canada
[bp] Ecology, Evolution, Environment, and Society, Dartmouth College, 78 College St, Hanover, NH, USA
[bq] US Department of Interior, National Park Service, Northeast Temperate Network, 54 Elm St, Woodstock, VT, USA
[br] Biology Department, Bradley University, Peoria, IL, 61625, USA
[bs] Museum and Institute of Zoology, Polish Academy of Sciences, Warsaw, Poland
[bt] Department of Health and Environmental Sciences, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China
[bu] EFNO, ECODIV, INRAE, Nogent-sur-Vernission 45290, France
[bv] IUCN SSC Species Monitoring Specialist Group. Gland, Switzerland

[1] David Funosas and Esther Sebastián-González contributed equally to the study.

ARTICLE INFO

Keywords:
Passive acoustic monitoring
Bird communities
BirdNET
Deep learning
Automated detection
Confidence threshold

ABSTRACT

Recent advances in machine learning have accelerated automated species detection across diverse ecological domains, enabling large-scale, non-invasive monitoring of biodiversity. In ornithological research, the combination of passive acoustic monitoring (PAM) and rapidly-developing novel identification tools such as BirdNET—a deep learning–based sound recognition algorithm—offers new opportunities for surveying vocally active bird communities. Here, we present the first worldwide evaluation of BirdNET using 4224 one-minute recordings from 67 sites across all continents annotated by local experts. More specifically, we assessed the capacity of BirdNET to accurately identify individual vocalizations and characterize bird communities based on the automated analysis of passively collected soundscapes. We further analyzed how its performance varies across continents, biomes, species, and minimum confidence thresholds. The proportion of correct BirdNET predictions (precision) was generally high and consistent across continents (range: 0.57–0.71) and biomes (range: 0.55–0.76). In contrast, the proportion of vocalizations successfully detected (recall) was generally lower and more heterogeneous across continents (range: 0.24–0.52) and biomes (range: 0.34–0.72), reflecting differences in species coverage and local ecological context. BirdNET predictive power, as measured by the Precision-Recall Area Under the Curve (PR AUC; higher values indicating better performance), was highest in North America, Oceania, and Europe (range: 0.16–0.23), moderate in Central/South America (0.13), and lowest in Africa and Asia (range: 0.03–0.04). Species-specific analyses revealed substantial heterogeneity in detection accuracy, with optimal confidence thresholds varying widely by species and analytical goal. Our results establish a global reference point for BirdNET reliability and highlight where algorithmic refinement and expanded acoustic sampling are most needed.

1. Introduction

In recent decades, various automated and non-invasive approaches have become standard in biodiversity monitoring (Lahoz-Monfort and Magrath, 2021). Among these, passive acoustic monitoring (PAM) has proven particularly effective for surveying diverse taxa, including anurans, bats, birds, cetaceans, and soniferous insects (Sugai et al., 2019; Hoefer et al., 2023; Darras et al., 2025). Among terrestrial taxa, birds have been primary targets of PAM (Shonfield and Bayne, 2017; Sugai et al., 2019), and soundscape ecology has historically drawn on acoustic ornithology (Gasc et al., 2017). The documented capacity of this approach to characterize avian communities and estimate population densities from passively collected soundscapes (Darras et al., 2019, Pérez-Granados and Traba, 2021) has prompted the development of dedicated methods and analytical pipelines for automated or semi-automated bird monitoring. In recent years, the development of effective low-cost audio recorders, such as the AudioMoth (Hill et al., 2018) and Song Meter Micro (Wildlife Acoustics) devices, along with essential advances in automated signal recognition software, have significantly expanded the use of PAM in biodiversity research.

While PAM offers numerous possibilities, it also presents challenges. One of its primary strengths—the ability to easily scale acoustic monitoring both spatially and temporally—also results in acoustic datasets far too large for manual analysis. To address this, most current projects rely on deep learning (DL) algorithms for the automated analysis of passively collected data (Stowell, 2022, Xie et al., 2023). Unfortunately, many state-of-the-art DL algorithms remain largely inaccessible to ecologists, land managers, and non-specialists lacking computational training. To bridge this gap, a new generation of user-friendly, ready-to-use sound recognition tools has emerged, helping to further streamline PAM workflows. Notably, many of these applications focus on birds as a model group, including BirdNET (Kahl et al., 2021), Perch (Ghani et al., 2023), HawkEars (Huus et al., 2025), Nighthawk (Van Doren et al., 2024), Chirpity (Kirkland, 2024), and the British Trust for Ornithology's Acoustic Pipeline (British Trust for Ornithology, 2023). Among them, BirdNET stands out for its broad taxonomic and geographic coverage, as well as its high predictive accuracy (Kahl et al., 2021; Pérez-Granados, 2023).

BirdNET is a free bird vocalization recognition software based on a convolutional neural network (Kahl et al., 2021). Its latest version (v.2.4) can identify more than 6000 bird species worldwide, as well as a smaller set of mammals and anurans (Wood et al., 2023a, 2023b; Pérez-Granados et al., 2023; Bota et al., 2024). The model analyzes recordings in 3-s windows, predicting zero, one, or multiple species per segment. Each prediction is assigned a confidence score ranging from 0.01 (very low model certainty in the prediction) to 1 (very high certainty), enabling users to filter results by using a customizable minimum confidence threshold. Low confidence thresholds favor high detection rates but increase the risk of false positives, while high confidence thresholds reduce errors at the cost of missed detections (Wood and Kahl, 2024). Two additional parameters also shape BirdNET performance: *Overlap* (ranging from 0 to 3 s), which determines the degree of temporal overlap between consecutive 3-s windows, and *Sensitivity* (ranging from 0.5 to 1.5), which adjusts how confidence scores are distributed across predictions. Lower *Sensitivity* values increase model certainty in its top predictions, while higher values make confidence scores more uniform across predictions (Pérez-Granados et al., 2025). Recent studies have explored how these parameters influence BirdNET performance, with the optimal configuration differing among species, regions, and research goals (Funosas et al., 2024; Pérez-Granados et al., 2025).

In recent years, multiple studies have used BirdNET to automatically classify bird vocalizations and derive ecological or conservation insights from the data collected (e.g., Funosas et al., 2024; wa Maina and Njoroge, 2025; Winiarska et al., 2025). Recent research suggests that BirdNET can provide a reliable characterization of bird communities in Europe, often yielding higher species richness estimates than traditional on-site surveys due to its ability to detect nocturnal and cryptic species (Funosas et al., 2024, Winiarska et al., 2025). However, most prior studies evaluating BirdNET have been limited in scope, focusing either on individual species at local scales (Manzano-Rubio et al., 2022) or on datasets restricted to Europe and North America (Pérez-Granados, 2023). Only a handful of studies have evaluated BirdNET performance in other regions (Amorós-Ausina et al., 2024; Pérez-Granados et al., 2025). This geographic bias stems in part from the initial restriction of the BirdNET training set to European and North American species (Kahl et al., 2021). Given the recent expansion of BirdNET species coverage and its increasing adoption worldwide, there is a growing need for a rigorous assessment of its performance across diverse ecological, acoustic, and biogeographic contexts.

The main goal of this study is to fill this gap by providing the first global evaluation of BirdNET. Using 4224 one-minute soundscapes recorded at 67 sites across 28 administrative regions worldwide, we aim to assess how accurately BirdNET identifies bird vocalizations and how effectively it characterizes bird communities from passively collected recordings. Our study has three objectives: (1) to examine geographic and biome-level variation in BirdNET capacity to identify individual

vocalizations and characterize bird communities, (2) to provide species-level performance metrics by reporting mean precision (the proportion of correct predictions) and recall (the proportion of vocalizations successfully detected) values for each of the 1102 bird species included in the study, and (3) to analyze how BirdNET performance varies across confidence thresholds. Given the rapidly growing use of BirdNET for automated bird monitoring, we hope that our findings will help identify its strengths and limitations across diverse ecological contexts, thus guiding future use of the software and contributing to its continued improvement.

2. Methods

2.1. Acoustic data

The recordings used in this study were sourced from the World Annotated Bird Acoustic Dataset (WABAD, version 2.0, Pérez Granados et al., 2025b). WABAD comprises expert-annotated recordings collected from 72 recording sites across 28 administrative regions (mainly representing countries). To facilitate further comparisons across datasets and ensure methodological consistency, we focused our study on a

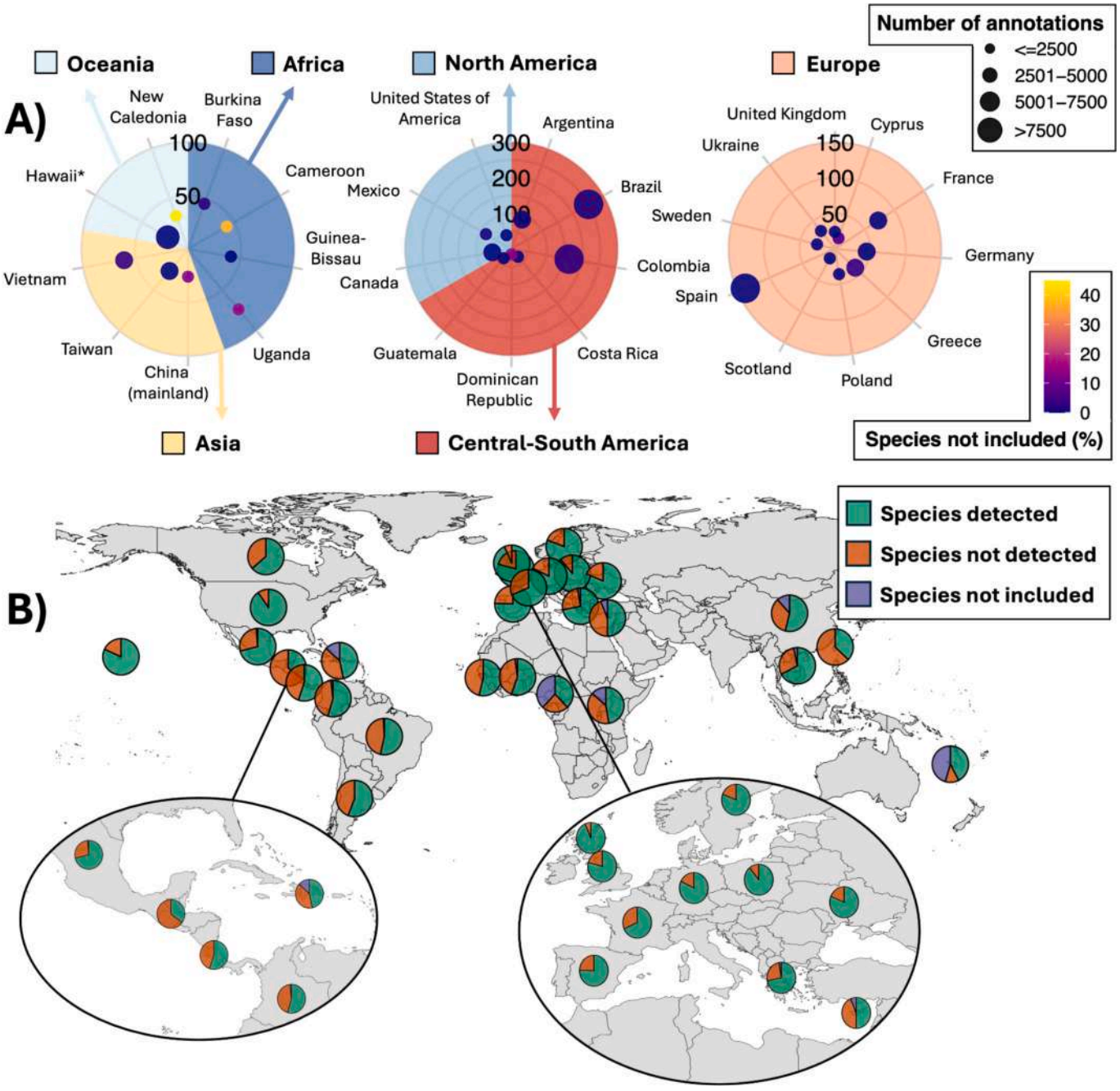

Fig. 1. (A) Number of annotations (dot size), number of annotated species (radial distance from the center, with scales varying among continents), and proportion of annotated species not covered by BirdNET-Analyzer v2.4 (color scale) in each administrative region. Administrative regions (mostly countries) are grouped by continent*, each represented within a separate circle or distinguished by a different background color. (B) Global mapping of the proportion of species 1) correctly detected by BirdNET (green), 2) not detected by BirdNET (orange), and 3) not covered by BirdNET (purple), using a minimum confidence threshold of 0.1. * Hawai'i and New Caledonia were both classified in Oceania following biogeographical criteria. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

subset of 4224 one-minute soundscapes from 67 distinct recording sites (henceforth referred to as datasets) in all 28 administrative regions, having retained only those with "strong labels" (i.e., providing the precise start and end times of each bird vocalization). The spatial distribution of the acoustic datasets used is shown in Fig. 1, and metadata for each site (minutes annotated, recording device, sampling frequency, geographic region, biome, and coordinates) are provided in Supplementary Table S1. Further details about the datasets analyzed in the study are available in Pérez Granados et al. (2025c).

Data coverage in this study varied substantially across geographic regions, generally aligning with continental boundaries (Fig. 1). The only exception is the Americas, which we divided into North America (Canada, USA, and Mexico) and Central/South America. This division allows us to evaluate whether BirdNET performance in North America—where public training data are more abundant—differs from that in the rest of the continent. For simplicity, we refer to geographic regions as "continents" hereafter. In Europe and Central/South America we gathered data from ≥20 datasets comprising >1100 recordings and > 27,000 annotated vocalizations each. In contrast, the other continents have far fewer annotated datasets (range: 3–9), recordings (range: 181–498), and annotated vocalizations (>6000–8000; Fig. 1). BirdNET species coverage, estimated as the proportion of bird species annotated by humans that was included in the algorithm, was near-complete in Europe and the Americas (≥99 %). Still, substantial gaps remained in Asia (95 %), Africa (84 %), and especially Oceania (72 %). Table 1 summarizes, for each continent, 1) the number of species and vocalizations annotated, and 2) the percentage of annotated species included in the BirdNET algorithm.

The ecological distribution of the recordings annotated was also uneven across biomes, categorized as in Olson et al. (2001). Tropical/subtropical and temperate broadleaf forests, as well as Mediterranean forests & shrublands, have 9000–35,000 vocalizations annotated. Boreal and temperate coniferous forests, tropical/subtropical grasslands, and wetlands are moderately represented (>3000–6000 vocalizations annotated), while deserts & xeric shrublands, as well as temperate and montane grasslands & savannahs, have >500–2000 annotated vocalizations each. BirdNET species coverage was near-complete (≥98 %) in all biomes except for tropical/subtropical broadleaf forests (91 %). Table 2 summarizes, for each biome, 1) the number of annotated species and vocalizations, and 2) the percentage of annotated species included in the BirdNET algorithm.

2.2. Annotation procedure

Expert ornithologists with in-depth knowledge of the local bird communities annotated all recordings. Audio spectrograms were analyzed in Raven Pro (v1.6), with experts being allowed to adjust the software parameters at their convenience and listen and visualize the recordings as many times as needed. Each vocalization was labeled at the species level following the Clements Checklist nomenclature (Clements et al., 2021), the taxonomy used in BirdNET-Analyzer v.2.4 (Kahl et al., 2021), thus ensuring direct comparability between human annotations and BirdNET output. Each recording site was annotated by a single observer, who delineated vocalizations using bounding boxes encompassing their temporal and frequency ranges and exported the annotations as .txt files named after the corresponding recordings. Multiple vocalizations of the same species were grouped into a single annotation box if they occurred within one second of each other; otherwise, separate annotations were created. A more thorough description of the annotation workflow is available in Pérez Granados et al. (2025b), and both the recordings and annotations used in this study can be downloaded from Pérez Granados et al. (2025c).

2.3. Audio analysis

We processed the acoustic recordings using BirdNET-Analyzer v2.2.0 (model v2.4: BirdNET_GLOBAL_6K_V2.4_Model_FP32.tflite) via a Linux shell script interfaced with the algorithm's Python backend, following the approach described in Funosas et al. (2024). BirdNET includes four adjustable settings that influence detection performance: *Minimum occurrence frequency threshold*, *Confidence threshold*, *Overlap*, and *Sensitivity* (for detailed descriptions of these settings and their impact, see Kahl et al., 2021, Funosas et al., 2024, Wood and Kahl, 2024, or Pérez Granados et al., 2025c). Based on prior research assessing nine combinations of *Overlap* and *Sensitivity* values (Pérez Granados et al., 2025c), we selected an *Overlap* of 2 s, maximizing performance at both vocalization and dataset levels, and a *Sensitivity* of 1 (default value). High *Overlap* values enhance recall by increasing the probability that an entire vocalization is contained within a single prediction window, longer within-window durations consistently yielding higher recall (Funosas et al., 2024). Although the strongest gains occur at the vocalization level—substantially higher recall without precision loss—higher *Overlap* also confers modest improvements at the dataset level (moderately higher recall with a minimal precision penalty; Pérez-Granados et al., 2025). Because the effects of *Sensitivity* on BirdNET performance vary with continent and analytical goal (e.g., identifying individual vocalizations versus characterizing bird communities), we opted for the default, balanced setting.

The *Minimum occurrence frequency threshold* defines the lowest regional and temporal occurrence frequency a species must have to be included in the list of potentially detectable species generated by BirdNET (range: 0.01–0.99). BirdNET-Analyzer v2.4 uses eBird checklist frequency data to estimate species ranges and probabilities of occurrence given geographic coordinates and week of the year (see https://github.com/birdnet-team/BirdNET-Analyzer/discussions/234). A low threshold broadens the list of potentially detectable species by including those with low likelihoods of occurrence, whereas a higher threshold restricts the list to species with the highest expected occurrence based on eBird data. We used a *Minimum occurrence frequency threshold* of 0.02—i.e., requiring a 2 % probability of species presence for inclusion in BirdNET-generated species lists—following Funosas et al. (2024), and retained the default *Confidence threshold* of 0.1 to minimize the risk of missed detections and to evaluate how detection performance varies across different thresholds.

2.4. BirdNET performance assessment

We evaluated BirdNET performance by comparing its predictions to expert annotations using a suite of custom R scripts (v4.2.2; R Core Team 2025) adapted, and available, from Funosas et al. (2024). Performance was assessed at two hierarchical levels: 1) the vocalization level, providing detailed insight into BirdNET capacity to correctly identify individual songs or calls, and 2) the dataset level, reflecting its capacity to characterize bird community composition based on multiple recordings from a single recording site. BirdNET predictions were classified into four categories:

- **True Positives (TP):** At the vocalization level, a BirdNET prediction was classified as a TP when an expert labeled the same species at the same 3-s time interval. At the dataset level, a bird species was considered a TP if there was at least one correct identification of that species by BirdNET in any of the recordings from the same study site (i.e., dataset).
- **False Positives (FP):** At the vocalization level, a BirdNET prediction was classified as a FP when an expert did not detect the same species at the same time. At the dataset level, a bird species was considered a FP when all BirdNET predictions of that species in the dataset were incorrect.
- **True Negatives (TN):** At both levels of analysis, a species was classified as a TN when it was covered by BirdNET but was neither identified by the expert nor predicted by the algorithm.

Table 1
BirdNET performance across continents when using a minimum confidence threshold of 0.1. For each continent, the table reports the number of annotated vocalizations and species, the number and proportion of annotated species 1) correctly detected, 2) in BirdNET auto-generated lists but not detected, 3) covered but not in BirdNET auto-generated lists, and 4) not covered by BirdNET, as well as precision, recall, and FPR (reported as mean ± standard deviation across datasets, followed by the bootstrapped 95 % confidence interval) calculated at both vocalization (voc_precision, voc_recall, voc_FPR) and dataset (ds_precision, ds_recall, ds_FPR) levels.

Continent	Vocalizations annotated	Species annotated	Species correctly detected	Species in auto-generated lists but not detected	Species covered but not included in auto-generated lists	Species not covered	voc_precision	ds_precision	voc_recall	ds_recall	voc_FPR	ds_FPR
Africa	6455	169	82 (49 %)	55 (33 %)	5 (3 %)	27 (16 %)	0.568 ± 0.156 (0.445–0.715)	0.272 ± 0.095 (0.185–0.348)	0.236 ± 0.084 (0.156–0.299)	0.487 ± 0.079 (0.42–0.546)	4e-05 ± 3e-05 (2.04e-05–6.93e-05)	0.011 ± 0.004 (0.00706–0.0133)
Asia	8410	106	62 (58 %)	28 (26 %)	11 (10 %)	5 (5 %)	0.68 ± 0.22 (0.444–0.878)	0.274 ± 0.096 (0.163–0.336)	0.257 ± 0.123 (0.183–0.338)	0.527 ± 0.151 (0.37–0.672)	5e-05 ± 2e-05 (1.84e-05–6.38e-05)	0.009 ± 0.005 (0.00323–0.0125)
Central and South America	27,582	482	269 (56 %)	178 (37 %)	28 (6 %)	7 (1 %)	0.707 ± 0.118 (0.654–0.758)	0.344 ± 0.132 (0.288–0.407)	0.458 ± 0.179 (0.383–0.536)	0.59 ± 0.161 (0.524–0.659)	3e-05 ± 2e-05 (2.55e-05–4.06e-05)	0.007 ± 0.005 (0.00503–0.00909)
Europe	29,781	182	149 (82 %)	20 (11 %)	11 (6 %)	2 (1 %)	0.605 ± 0.167 (0.543–0.669)	0.282 ± 0.1 (0.243–0.32)	0.512 ± 0.165 (0.443–0.572)	0.692 ± 0.159 (0.634–0.751)	6e-05 ± 3e-05 (4.87e-05–7.22e-05)	0.008 ± 0.003 (0.00653–0.00861)
North America	8748	164	124 (76 %)	37 (23 %)	2 (1 %)	1 (1 %)	0.647 ± 0.164 (0.54–0.745)	0.371 ± 0.124 (0.297–0.446)	0.517 ± 0.255 (0.353–0.657)	0.696 ± 0.174 (0.587–0.809)	5e-05 ± 5e-05 (2.36e-05–8.49e-05)	0.005 ± 0.003 (0.00352–0.00649)
Oceania	8085	54	32 (59 %)	5 (9 %)	2 (4 %)	15 (28 %)	0.692 ± 0.131 (0.572–0.78)	0.624 ± 0.23 (0.426–0.823)	0.448 ± 0.358 (0.166–0.73)	0.611 ± 0.292 (0.341–0.81)	5e-05 ± 5e-05 (1.02e-05–9.17e-05)	0.001 ± 0.001 (0.000384–0.00161)

Table 2
BirdNET performance across biomes when using a minimum confidence threshold of 0.1. For each biome, the table reports the number of annotated vocalizations and species, the number and proportion of annotated species 1) correctly detected, 2) in BirdNET auto-generated lists but not detected, 3) covered by BirdNET but not included in BirdNET auto-generated lists, and 3) not covered by BirdNET, as well as precision, recall, and FPR (reported as mean ± standard deviation across datasets, followed by the bootstrapped 95 % confidence interval) calculated at both vocalization (voc_precision, voc_recall, voc_FPR) and dataset (ds_precision, ds_recall, ds_FPR) levels. Standard deviations and confidence intervals are not provided for biomes being represented by a single dataset.

Biome	Vocalizations annotated	Species annotated	Species correctly detected	Species in auto-generated lists but not detected	Species covered but not included in auto-generated lists	Species not covered	voc_precision	ds_precision	voc_recall	ds_recall	voc_FPR	ds_FPR
Boreal Forest/ Taiga	5604	79	56 (71 %)	23 (29 %)	0 (0 %)	0 (0 %)	0.586 ± 0.191 (0.445–0.748)	0.347 ± 0.088 (0.281–0.42)	0.411 ± 0.149 (0.275–0.521)	0.636 ± 0.117 (0.545–0.73)	5e-05 ± 4e-05 (2.72e-05–8.01e-05)	0.005 ± 0.003 (0.00268–0.00767)
Deserts & Xeric Shrublands	623	13	11 (85 %)	2 (15 %)	0 (0 %)	0 (0 %)	0.675	0.143	0.72	0.846	4e-05	0.01
Mediterranean Forests & Shrublands	10,572	116	79 (68 %)	27 (23 %)	8 (7 %)	2 (2 %)	0.641 ± 0.138 (0.555–0.717)	0.218 ± 0.104 (0.165–0.29)	0.528 ± 0.153 (0.434–0.611)	0.633 ± 0.165 (0.536–0.729)	5e-05 ± 2e-05 (3.85e-05–5.85e-05)	0.009 ± 0.003 (0.00739–0.0103)
Montane Grasslands & Savannahs	529	32	19 (59 %)	12 (38 %)	1 (3 %)	0 (0 %)	0.761 ± 0.014 (0.751–0.771)	0.357 ± 0.017 (0.345–0.368)	0.452 ± 0.086 (0.392–0.513)	0.507 ± 0.044 (0.476–0.538)	3e-05 ± 0 (3.13e-05–3.65e-05)	0.003 ± 0.001 (0.00292–0.00369)
Temperate Broadleaf & Mixed Forest	10,099	118	96 (81 %)	21 (18 %)	1 (1 %)	0 (0 %)	0.598 ± 0.166 (0.476–0.705)	0.373 ± 0.064 (0.331–0.419)	0.547 ± 0.269 (0.346–0.712)	0.855 ± 0.099 (0.788–0.928)	9e-05 ± 5e-05 (5.07e-05–0.000125)	0.006 ± 0.002 (0.00539–0.00758)
Temperate Coniferous Forest	3611	50	43 (86 %)	5 (10 %)	2 (4 %)	0 (0 %)	0.644 ± 0.15 (0.539–0.75)	0.401 ± 0.034 (0.377–0.425)	0.657 ± 0.149 (0.552–0.762)	0.878 ± 0.093 (0.812–0.944)	8e-05 ± 6e-05 (4.18e-05–0.00012)	0.005 ± 0.002 (0.00431–0.00662)
Temperate Grasslands	2107	34	24 (71 %)	8 (24 %)	2 (6 %)	0 (0 %)	0.672	0.245	0.533	0.706	6e-05	0.011
Tropical/ Subtropical Grasslands	4281	91	57 (63 %)	31 (34 %)	1 (1 %)	2 (2 %)	0.676 ± 0.056 (0.636–0.716)	0.221 ± 0.005 (0.217–0.224)	0.441 ± 0.172 (0.319–0.562)	0.626 ± 0.099 (0.556–0.696)	6e-05 ± 1e-05 (5.33e-05–7.26e-05)	0.016 ± 0.002 (0.0139–0.0171)
Tropical/ Subtropical Dry Broadleaf Forest	9334	191	118 (62 %)	50 (26 %)	5 (3 %)	18 (9 %)	0.753 ± 0.072 (0.704–0.802)	0.415 ± 0.199 (0.291–0.556)	0.387 ± 0.269 (0.212–0.575)	0.575 ± 0.198 (0.44–0.701)	3e-05 ± 2e-05 (1.52e-05–4.03e-05)	0.004 ± 0.003 (0.00216–0.00574)
Tropical/ Subtropical Moist Broadleaf Forest	35,687	605	331 (55 %)	197 (33 %)	42 (7 %)	35 (6 %)	0.659 ± 0.155 (0.594–0.724)	0.37 ± 0.165 (0.306–0.439)	0.409 ± 0.208 (0.324–0.496)	0.577 ± 0.169 (0.509–0.641)	4e-05 ± 3e-05 (2.66e-05–4.9e-05)	0.006 ± 0.004 (0.00484–0.00813)
Wetland	6614	183	106 (58 %)	72 (39 %)	5 (3 %)	0 (0 %)	0.547 ± 0.226 (0.394–0.711)	0.271 ± 0.068 (0.227–0.318)	0.505 ± 0.14 (0.415–0.604)	0.62 ± 0.127 (0.536–0.7)	6e-05 ± 2e-05 (4.12e-05–7.34e-05)	0.009 ± 0.005 (0.00556–0.0117)

- **False Negatives (FN):** At both levels of analysis, a species was classified as a FN when it was identified by the expert but not predicted by BirdNET.

Based on these categorizations, we evaluated BirdNET precision, recall and False Positive Rate (FPR) at both levels of analysis. Precision quantifies the proportion of correct identifications among all BirdNET predictions, whereas recall measures the proportion of expert-identified vocalizations or species that were correctly detected by BirdNET (Pérez-Granados, 2023). FPR complements these metrics by estimating the probability that BirdNET falsely detects a species absent from the acoustic sample—defined as a 3-s prediction window at the vocalization level and as the entire dataset at the dataset level—and is computed as the number of spurious detections divided by the number of species absent from the sample but covered by BirdNET. Importantly, recall calculations include species not covered by BirdNET: any annotated species that went undetected was counted as a FN, whether because BirdNET failed to predict it despite coverage or because the species was outside the taxonomic scope of the algorithm. This choice was deliberate, as our concern is not BirdNET recall conditional on its species list, but the practical outcome for ecological applications—i.e., if BirdNET is deployed in a region, what fraction of vocalizations or species it will detect successfully.

Precision, recall and FPR were calculated across 90 confidence thresholds ranging from 0.10 to 0.99 in 0.01 increments, following Funosas et al. (2024). At the vocalization level, recall was computed by aggregating all BirdNET predictions that overlapped a given vocalization (i.e., those starting before its end and ending after its onset). Precision was calculated by pooling all expert annotations that overlapped each prediction. At the dataset level, comparisons were based on species lists: a species was marked as correctly predicted if it appeared in both BirdNET predictions and expert annotations, as long as they both coincided in time at some point. Note that under this criterion, a single correct detection of a species suffices for it to be classified as a TP at the dataset level, thus favoring higher recall values in longer datasets. The formulas used were the following:

Precision = TP / (TP + FP).

Recall = TP / (TP + FN).

FPR = FP / (FP + TN).

To visually represent the variation of these metrics across confidence thresholds, we used the Precision-Recall (PR) curve, accompanied by its corresponding Area Under the Curve (AUC; Davis and Goadrich, 2006). PR curves plot precision against recall across all confidence thresholds, capturing the trade-off between these two metrics, with higher AUC values (range: 0–1) being indicative of higher predictive power. Because PR AUC integrates precision across the entire recall range, broader recall ranges—even those including lower recall values—can yield higher AUCs. Thus, to enable fair comparisons across continents with varying recall ranges, we adjusted PR AUC scores to account for recall ranges using the following formula:

$$\text{adj\_PR AUC} = \frac{PRAUC}{max(recall) - min(recall)}$$

Finally, we computed the F-score, which integrates precision and recall into a single performance metric:

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

An F-score with $\beta = 1$ gives equal weight to precision and recall, whereas $\beta > 1$ emphasizes recall and $\beta < 1$ emphasizes precision. We computed F-scores using three β values: $\beta = 1$ as a standard value to enable comparison with previous studies, $\beta = 0.25$ to prioritize precision over recall, and $\beta = 4$ to prioritize recall over precision. A β value <1 was chosen because high precision is typically paramount in biodiversity research, where failing to detect a present species is generally less problematic than falsely detecting an absent one (Tolkova et al., 2021). Furthermore, specific population models (e.g., occupancy models) explicitly account for imperfect detection (Brunk et al., 2023; Bielski et al., 2024), further reducing the relative cost of FNs compared with FPs. A β value >1, on the other hand, might be particularly relevant to research teams targeting rare or cryptic species and able to manually validate large numbers of BirdNET predictions. Moreover, some classification–occupancy models can explicitly incorporate classification errors, making FPs less problematic (Ogawa et al., 2025).

The metrics described above were used to capture complementary aspects of BirdNET performance across continents, biomes, and species. PR AUC scores provided a global measure of predictive power, with PR curves offering a fine-grained view of how precision and recall shift with confidence threshold choice in each continent. F-score curves were used to identify, for every continent, the threshold that maximizes the trade-off between precision and recall under three weighting schemes: equal weighting, marked emphasis on precision, and marked emphasis on recall. To enable consistent cross-continent and cross-biome comparisons for different purposes—whether identifying individual vocalizations (vocalization level) or characterizing bird communities (dataset level)—, we calculated recall and precision at both levels of analysis using BirdNET's default confidence threshold of 0.1. Two additional thresholds (0.5 and 0.75) were also used to evaluate how optimal threshold choice varies across species. We quantified uncertainty in continent- and biome-specific performance metrics using bootstrapped 95 % confidence intervals across datasets within each continent and biome.

3. Results

3.1. Performance across continents

We found BirdNET performance to be moderately heterogeneous across continents. When using the default confidence threshold of 0.1, precision was relatively consistent across continents at the vocalization level (range = 0.57–0.71, SD = 0.053), and more variable at the dataset level (range = 0.27–0.62, SD = 0.134; Fig. 2), with Africa performing worst at both levels of analysis and Oceania performing best at the dataset level and second-to-best at the vocalization level (Table 1). Recall, in turn, varied more widely across continents at the vocalization level (range = 0.24–0.52, SD = 0.125) but was more uniform at the dataset level (range = 0.49–0.70, SD = 0.084), with Europe and North America performing consistently better and Africa and Asia performing consistently worse at both levels (Table 1, Fig. 2, Supplementary Fig. S1). Finally, FPR varied widely across continents at the vocalization level (range = 3e-05–6e-05, SD = 1.03e-05) and even more so at the dataset level (range = 0.001–0.011, SD = 0.0034; Table 1), with high discrepancies in continent ranking between the two levels of analysis. Central and South America performed best and Europe performed worst at the vocalization level, while Oceania performed best and Africa performed worst at the dataset level.

At the vocalization level, adjusted PR AUC values highlight more evident differences: Asia (0.03) and Africa (0.04) exhibit very low scores at the vocalization level, with Central/South America (0.13) and Europe (0.16) having intermediate scores, and Oceania (0.20) and North America (0.23) performing best. Asia shows the lowest adjusted PR AUC score because, even though the precision scores obtained with high confidence thresholds are very high, these drop precipitously when the threshold is lowered (Fig. 3E). At the same time, the maximum recall score for this continent is already low (0.26), implying that any confidence threshold yielding reasonable precision will necessarily drive recall to very low levels. In Africa, both maximum precision and recall are lower than in Asia, but its precision demonstrates greater robustness, decreasing more slowly with reduced confidence thresholds (Fig. 3A). At the dataset level, adjusted PR AUC values follow similar trends, but with a lower degree of cross-continent heterogeneity. Africa exhibits the lowest score (0.16), while Europe, Asia, Central/South America, and North America have intermediate scores (range = 0.22–0.29) and Oceania stands out (0.35) due to consistently high precision ($\geq$0.62) across

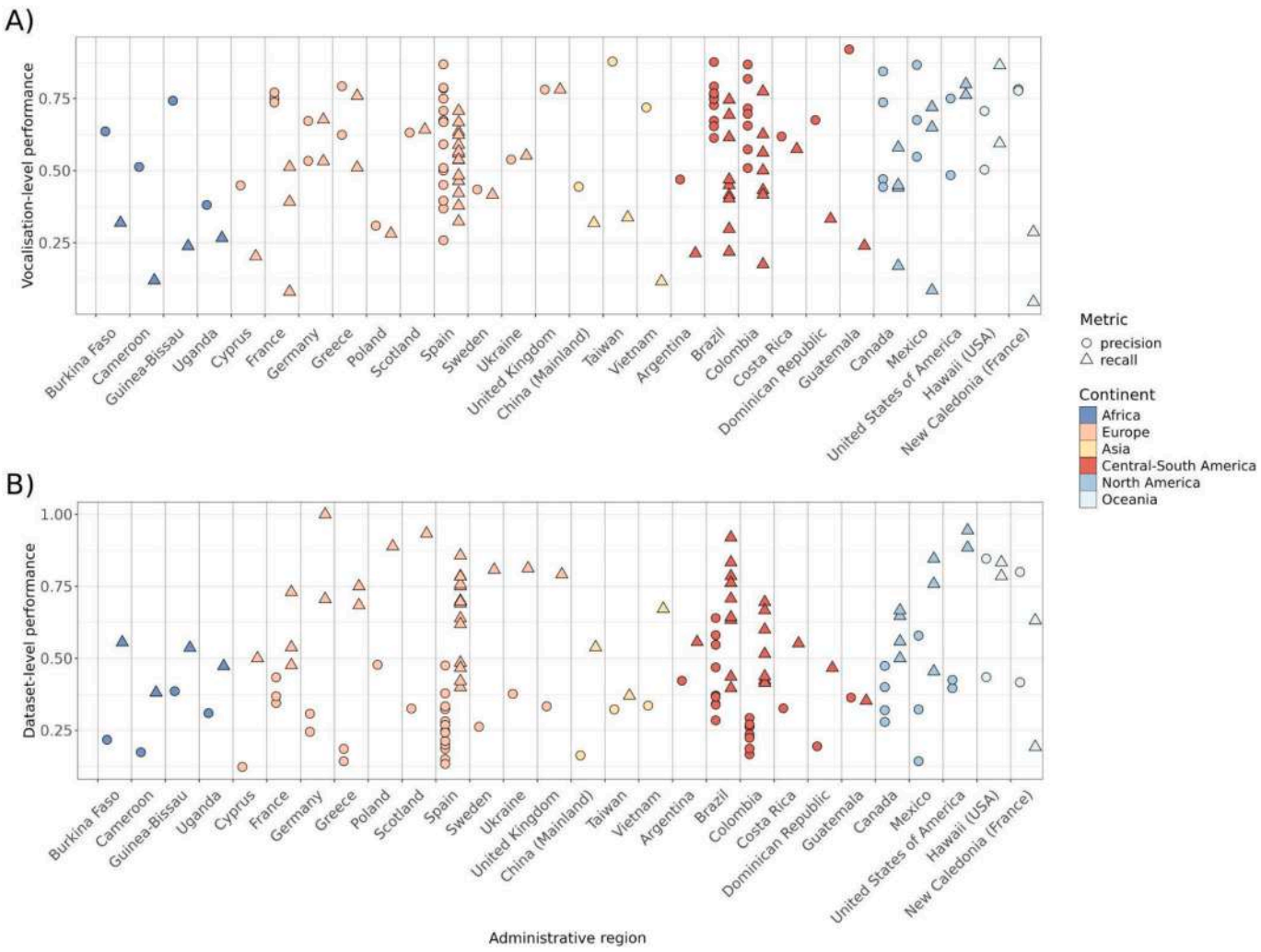

Fig. 2. Mean BirdNET precision and recall by administrative region using a minimum confidence threshold of 0.1. Results are shown separately at the (A) vocalization and (B) dataset levels.

confidence thresholds (Fig. 3).

F-score curves show broadly similar shapes across continents. F1-scores peak at a confidence threshold of 0.1 at the vocalization level and plateau at a mostly flat maximum between 0.2 and 0.6 at the dataset level. F0.25-scores, for the most part, plateau at a relatively flat maximum between 0.3 and 0.8 at the vocalization level and reach a peak at a confidence threshold of around 0.9 at the dataset level. F4-scores steadily decrease along with higher confidence thresholds at both levels of analysis, with declines being particularly pronounced at the dataset level (Supplementary Fig. S2).

3.2. Performance across biomes

Consistent with continent-scale results, we found that, when using the default confidence threshold of 0.1, precision was relatively uniform across biomes at the vocalization level (range = 0.55–0.76, SD = 0.065) but more heterogeneous at the dataset level (range = 0.14–0.41, SD = 0.090; Table 2). Best- and worst-performing biomes are not entirely consistent across the two levels: BirdNET performed especially poorly in wetlands at the vocalization level and in deserts & xeric shrublands at the dataset level. In contrast, classifications in recordings from tropical/subtropical dry broadleaf forests and montane grasslands & savannahs exhibited the highest precision, maintaining strong performance at both levels of analysis.

Recall varied more widely across biomes, with values of 0.39–0.72 (SD = 0.105) at the vocalization level and 0.51–0.85 (SD = 0.127) at the dataset level (Table 2). Scores were consistently highest in deserts & xeric shrublands, whereas tropical/subtropical broadleaf forests ranked lowest at the vocalization level and near the bottom at the dataset level, only surpassing montane grasslands & savannahs (Table 2, Supplementary Fig. S3). Finally, FPR varied widely across biomes at the vocalization level (range = 3e-05–9e-05, SD = 1.91e-05) and even more so at the dataset level (range = 0.003–0.016, SD = 0.0038), with broad similarities in biome ranking between the two levels of analysis (Table 2). Montane grasslands & savannahs performed best at both levels, with temperate broadleaf & mixed forest and tropical/subtropical grasslands performing worst at the vocalization and dataset levels, respectively. Confidence intervals for mean performance across both continents and biomes overlapped partially across all three metrics and both levels of analysis (Table 1), indicating moderate heterogeneity rather than clear-cut regional contrasts. Some of this overlap likely reflects uneven sample sizes across continents and biomes, as regions with fewer datasets exhibited broader confidence intervals and hence greater uncertainty in estimated means.

3.3. Performance across species

Our species-specific analyses, focused on the species annotated in more than 50 recordings and conducted across three confidence thresholds (0.1, 0.5, 0.75), show that, as the minimum confidence

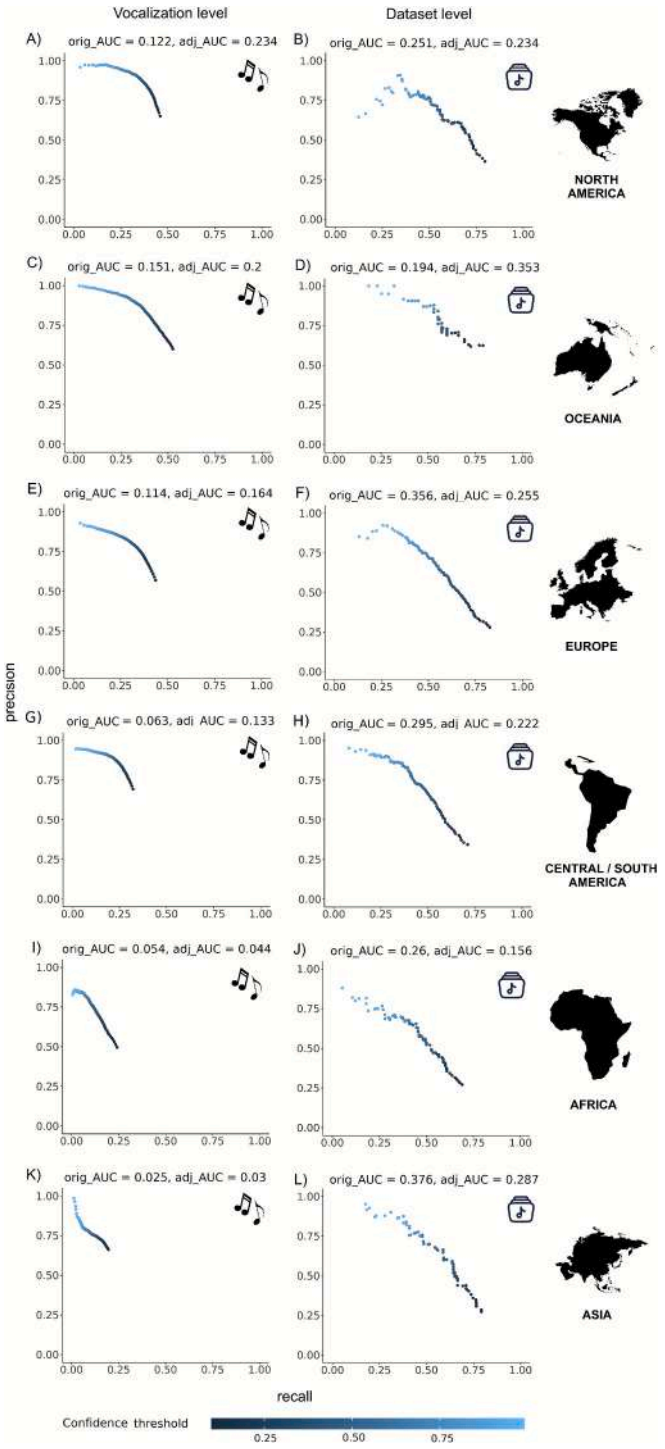

Fig. 3. BirdNET Precision-Recall (PR) curves for each continent and level of analysis. Continents are ordered by decreasing vocalization-level adj_AUC. Original Area Under the Curve (AUC) scores (orig_AUC) and AUC scores adjusted to account for recall range (adj_AUC) are shown on top of each curve.

threshold increases, both TPs and FPs decline, with FNs following the opposite trend (Fig. 4). Consequently, the numbers of correctly and mistakenly detected species vary substantially with threshold choice. Because higher-confidence predictions are more likely to be correct, raising the threshold reduces FPs far more than TPs. This asymmetry, however, is highly species-dependent (Fig. 4, Supplementary Table S2). Some species (e.g., *Arremon brunneinucha*) retain most TPs while eliminating nearly all FPs when increasing the confidence threshold from 0.1 to 0.75, achieving consistently high F1-scores (>0.7) across thresholds. However, other species (e.g., *Acrocephalus arundinaceus*) lose ≥90 % of TPs with the same confidence threshold increase, maintaining persistently low F1-scores (<0.3, Fig. 4). Vocalization-level precision and recall scores, along with the numbers of species-specific annotated vocalizations and BirdNET predictions for all 1102 species in the WABAD dataset, are reported in Supplementary Table S2.

4. Discussion

The use of deep learning algorithms for automated wildlife detection from passively collected data has expanded rapidly in recent years (Stowell, 2022, Xie et al. 2022). Here, we provide the first comprehensive evaluation of BirdNET (Kahl et al., 2021) performance across continents and biomes, assessing both its ability to detect bird vocalizations and to characterize bird communities. We analyzed an extensive acoustic dataset composed of 4224 one-minute soundscapes annotated by local experts from 67 recording sites worldwide (Pérez-Granados et al., 2025). Our results suggest that BirdNET cross-continent and cross-biome performances are highly heterogenous, with recall being more variable than precision. Much of this variation can be attributed to gaps in BirdNET species coverage, highlighting the importance of consulting the most up-to-date BirdNET species list when interpreting results. The fact that precision remains imperfect even at the highest confidence thresholds, and that relaxing these thresholds is unavoidable to obtain reasonable recall, makes expert validation indispensable for achieving accurate and comprehensive ecological inference.

Precision exhibited substantial variation across continents and biomes, with vocalization-level estimates being more consistent than those at the dataset level. Dataset-specific factors, including recording duration—which tends to increase the number of FPs at fixed thresholds (Funosas et al., 2024)—and species richness—where lower values magnify the relative impact of FPs—likely contributed to this discrepancy. At the biome scale, recordings collected in wetland habitats exhibited the weakest performance at the vocalization level. Since many wetland species are widespread and well represented in reference libraries, low performance possibly reflects ecological complexity driven by high species richness and abundance (Goëau et al., 2018), as well as the presence of species with relatively uncharacteristic or poorly differentiated calls (e.g., many waterfowl). Unlike passerines, where strong sexual selection and territoriality have driven the evolution of distinctive songs, waterfowl typically depend more on visual courtship displays, reducing pressure for acoustically distinctive signals (Johnsgard, 1971, Ten Cate 2021) and thereby complicating discrimination by BirdNET. Deserts, by contrast, exhibited the lowest dataset-level precision despite strong recall: with few species annotated per site, even a small number of incorrect detections can drastically reduce precision. At the continental scale, precision was found to be the weakest in Africa and Asia at both levels of analysis, possibly due to the relatively small amount of training data available for species occurring in these continents (Funosas et al., 2024). Oceania, however, achieved the highest dataset-level precision and the second-highest vocalization-level precision despite having the lowest degree of species coverage. This result may be partly explained by the partnership established between the Cornell Lab of Ornithology, developer of BirdNET, and the Listening Observatory for Hawaiian Ecosystems, which contributed extensive annotated material (both full soundscapes and short species-specific clips) for the 2022 BirdCLEF competition and for further improvement of the BirdNET algorithm (Kahl et al., 2022; Navine et al., 2022). This intensive training corpus may have improved BirdNET's ability to discriminate Hawaiian taxa.

FPR broadly tracked precision, with vocalization-level estimates more consistent than dataset-level ones. At the dataset level, continental rankings were nearly identical: Africa and Asia performed worst, Oceania and North America best. At the vocalization level, however, Central and South America showed the lowest FPR, while Europe had the

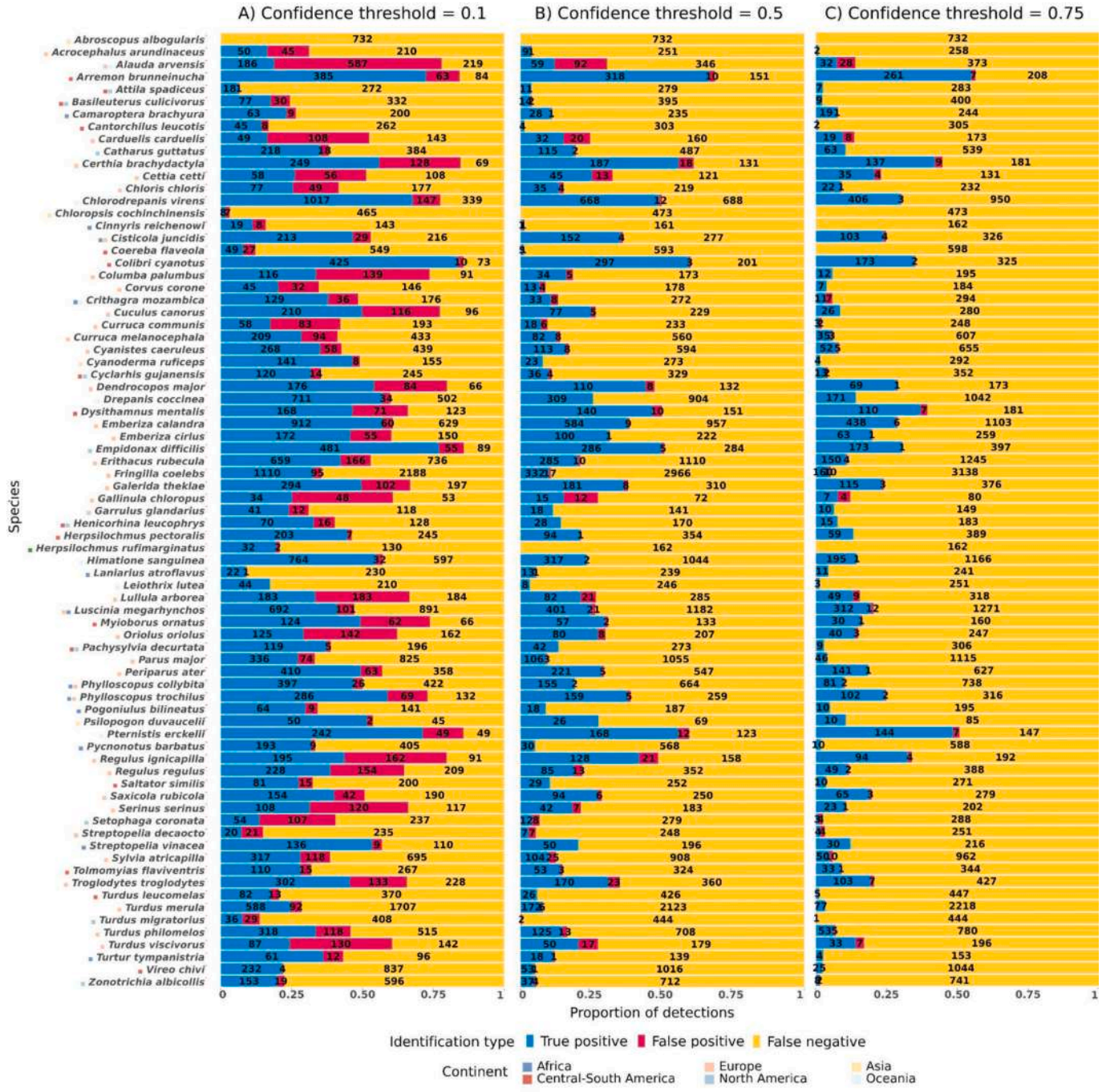

Fig. 4. Proportion of true positives (TP, blue), false positives (FP, red), and false negatives (FN, yellow) by bird species in BirdNET output with minimum confidence scores of (A) 0.1, (B) 0.5, and (C) 0.75. Results are calculated at the vocalization level, so the combined total of TPs and FNs per species is fixed (equal to the total number of vocalizations annotated), while FPs vary with confidence threshold, with lower thresholds yielding more predictions and thus more FPs. Only species present in at least 50 recordings are shown, listed in alphabetical order. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

highest—indicating that European precision is penalised mainly by its high number of FPs, whereas Central and South American precision is mainly driven down by its low number of TPs. At the biome scale, the biomes with the highest precision (tropical/subtropical dry broadleaf forests and montane grasslands & savannahs) also had the lowest FPR, while those with the highest FPR (temperate broadleaf & mixed forests and temperate coniferous forests at the vocalization level; tropical/subtropical and temperate grasslands at the dataset level) also exhibited the lowest precision. This tight correspondence highlights the central role of FPs in shaping both metrics.

Recall also displayed broadly similar patterns to precision: continents and biomes with the most significant proportion of uncovered species (e.g., Africa, Asia, tropical/subtropical broadleaf forests) generally exhibited the lowest scores, with Oceania again standing out as an exception. Oceanian recall was surpassed only by Europe and North America and clearly outperformed Africa and Asia at both levels of analysis (Table 1). Beyond the potential emphasis on Hawaiian avifauna during BirdNET training, the low species richness in the geographically isolated Hawaiian bird communities—accounting for half of our Oceanian data—may also have contributed to higher

detection rates by increasing the likelihood of detecting a large share of the present species. A similar pattern was observed in cross-biome analyses: deserts & xeric shrublands, which have the lowest species richness, achieved the highest recall at both levels of analysis (Table 2). Taken together, these results underscore the combined influence of species coverage, dataset composition, and local context on BirdNET performance. However, given the limited number of datasets and uneven spatial coverage for certain continents and biomes—most notably Oceania and deserts—, these apparent patterns may partly reflect sampling artifacts rather than generalisable performance differences.

When integrating across the F1-score and PR AUC metrics, BirdNET performed best in North America, Europe, and Oceania, moderately in Central/South America, and relatively poorly in Asia and Africa, with these cross-continent patterns largely consistent between both levels of analysis. Geographic differences likely reflect disparities in training data: online acoustic libraries are richer in recordings from North America and Europe, whereas African and Asian species remain underrepresented (Macaulay, 2025; Xeno-canto, 2025). The paucity of species-specific training data in these regions may predispose the algorithm to learn narrow or context-specific acoustic cues that fail to generalise more broadly. This could undermine the capacity of BirdNET to recognize species whose vocal signatures vary geographically or ecologically (Knight et al., 2024, Sebastián-González and Pérez-Granados, 2025), as well as species with very wide acoustic repertoires.

Confidence threshold analyses, in contrast, highlight a high degree of consistency across continents but a large difference between scales of evaluation. At the vocalization level, F1-scores peak at a confidence threshold of 0.1 and decline steadily at higher thresholds, as recall decreases more sharply than precision improves. F0.25-scores plateau at intermediate thresholds because the greater weighting of precision over recall means that small precision gains can compensate for moderate recall losses. F4-scores, driven mainly by recall and only weakly by precision, decline steadily with higher confidence thresholds (Supplementary Fig. S2). At the dataset level, since only one correct detection is required for a species to be considered a TP, higher thresholds can improve results by favoring precision without as steep a recall penalty. This may explain why both recall rates and optimum thresholds are higher at the dataset level. These differences highlight that there is no universally optimal confidence threshold: the best choice is inherently context-dependent (Wood and Kahl, 2024, Tueng et al. ., 2025). Lower thresholds enhance recall and are therefore better suited to the monitoring of rare or elusive species, whereas higher thresholds are preferable for biodiversity monitoring, where minimizing FPs is essential to prevent inflated richness estimates. The instability of precision-focused F-scores at high thresholds also cautions against overconfidence, highlighting the need for manual validation—especially in regions with low training coverage. More broadly, threshold selection is not just a technical choice but a reflection of ecological or conservation priorities. Explicitly reporting and justifying threshold choices could therefore enhance reproducibility and interpretability in future BirdNET applications.

Our species-level analyses also revealed substantial heterogeneity in BirdNET vocalization-level performance across species (Fig. 4, Supplementary Table S2), a result that aligns with prior research suggesting strong variation across species, including within the same family (Amorós et al. 2024). Strong cross-context intraspecific variation in BirdNET performance has also been reported, with mean precision for the Common Raven (*Corvus corax*) ranging from 0.29 (Cole et al., 2022) to 0.66 (Kahl, 2020) and 0.94 (Sethi et al., 2021). These findings underscore the difficulty of inferring species-specific performance from global or aggregate metrics, highlighting the need for assessments at the species level (see Tseng et al., 2025). In our study, interspecific variability in vocalization-level precision and recall diminished with increasing confidence thresholds, with standard deviations shrinking from 0.23 and 0.19 at a threshold of 0.1 to 0.18 and 0.13 at a threshold of 0.75. As expected, increasing thresholds improved precision but reduced recall, though the magnitude of this trade-off varied widely across species. In some cases, TPs declined far more slowly than FPs, making these species better suited to higher thresholds. In others, TPs and FPs declined at similar rates (Fig. 4), favoring lower thresholds. While the use of uniform thresholds for all species remains a practical option, these results underscore that species-specific adjustments are likely to yield more reliable outcomes (Wood and Kahl, 2024, Tseng et al. ., 2025).

Despite having limited our species-level analysis to those present in at least 50 recordings, there was substantial heterogeneity in the number of annotated vocalizations per species. Some species (e.g., *Fringilla coelebs*, *Turdus merula*, *Erithacus rubecula*, *Sylvia atricapilla*) had more than 1000 annotated vocalizations across over 300 recordings, while others (e.g., *Herpsilochmus rufimarginatus*, *Curruca undata*, *Phasianus colchicus*, *Eurillas curvirostris*) had fewer than 200 annotations drawn from only 50 recordings. Results for species with sparse representation—whether in annotated vocalizations or BirdNET predictions—should therefore be interpreted with particular caution. Moreover, several species, despite being frequent in the dataset, were largely undetected by BirdNET because they were excluded from many auto-generated species lists. A notable case is *Abroscopus albogularis*, present in 969 recordings but included in only 111 lists and predicted in none. This result underlines the importance of revising the species lists automatically generated by BirdNET for each location and week of the year to detect potentially missing species. Since BirdNET generates these lists based on eBird data, this consideration will be especially important in undersampled regions, where lists are more likely to erroneously exclude locally present species (Table 1, Table 2). Further, prior work suggests that the performance of DL acoustic classifiers is not only species- but also context-dependent, influenced by factors such as background noise and the presence of acoustically similar taxa (Ventura et al., 2024, Tseng et al., 2025). This underscores the need for prudence when extrapolating our species-specific results to datasets collected under different recording settings, background noise profiles, or bird community compositions (Wood and Kahl, 2024).

Interpretation of our cross-continent and cross-biome results should also be undertaken cautiously, as the acoustic datasets analyzed exhibit uneven representation across continents and biomes. Europe and Central/South America account for more than half of the data, whereas Africa, Asia, and Oceania are sparsely represented. Within continents, datasets are often concentrated in a few countries (e.g., Brazil and Colombia in Central/South America and Spain in Europe), so our findings cannot be assumed to be fully representative of entire world regions. This bias is particularly pronounced for Oceania, as our study covers only two regions in a continent with exceptional endemic diversity. Similarly, biome-level analyses reveal significant imbalances: tropical/subtropical, temperate, and Mediterranean forests are well represented, while deserts & xeric shrublands and montane grasslands & savannahs are poorly sampled in WABAD (Pérez Granados et al., 2025c).

A further limitation arises from the fact that continent and biome categories are not fully independent, since many biomes occur only in specific continents, complicating attribution. The partial overlap of cross-continent and cross-biome confidence intervals further suggests that the observed performance differences, while directionally consistent, may partly reflect these unequal sampling efforts rather than intrinsic model limitations. Consequently, apparent performance gaps should be interpreted with caution. In addition, the absence of common habitats such as croplands and urban areas, now central to biodiversity monitoring and research, limits the applicability of our findings to these environments. Additional sources of bias stem from the datasets themselves. Recording equipment and settings, annotation effort, and local conditions vary across locations, potentially influencing BirdNET performance (Leroy et al., 2018, Duc et al., 2021, Wood and Kahl, 2024, Pérez-Granados, 2025). Although all recordings were annotated by local experts following standardized protocols, differences in annotation

quality remain possible, potentially biasing our results. Finally, we would like to acknowledge that although incorporating temporal and spatial filters (e.g., week and location) into the species list potentially detected by BirdNET is conceptually sound and improves ecological plausibility, it may inadvertently exclude species that are present but not expected according to reference databases, such as eBird. Such mismatches can affect detection performance and introduce seasonal or regional biases, affecting to a larger extent those areas with smaller reference databases. The impact of applying temporal and spatial filters on BirdNET performance should be further assessed to ensure robust and unbiased classification across contexts.

5. Conclusion

Overall, our study shows that BirdNET represents a powerful tool to assist in the assessment of bird community composition through the automated analysis of ecoacoustic data. However, its performance exhibits substantial regional and species-specific heterogeneity. While precision remains relatively consistent across regions, recall drops sharply in continents and biomes with limited acoustic data availability. Responsible BirdNET use requires objective- and, ideally, species-specific confidence thresholds, coupled with manual validation to ensure a rigorous and thorough characterization of recorded bird communities. Our findings set the first global reference for these calibration decisions and emphasize the need to fill acoustic data gaps to ensure reliable, globally consistent model performance. In particular, future global studies should prioritize increased sampling in Africa, Asia, and underrepresented biomes to improve both coverage and model reliability.

CRediT authorship contribution statement

David Funosas: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Esther Sebastián-González:** Writing – review & editing, Supervision, Resources, Methodology, Data curation, Conceptualization. **Jon Morant:** Writing – review & editing. **Oscar H. Marín Gómez:** Writing – review & editing. **Irene Mendoza:** Writing – review & editing. **Miguel A. Mohedano-Muñoz:** Writing – review & editing. **Eduardo Santamaría:** Writing – review & editing. **Giulia Bastianelli:** Writing – review & editing. **Alba Márquez-Rodríguez:** Writing – review & editing. **Michał Budka:** Writing – review & editing. **Gerard Bota:** Writing – review & editing. **Cristina D. Alonso-Moya:** Writing – review & editing. **José M. de la Peña-Rubio:** Writing – review & editing. **Eladio L. García de la Morena:** Writing – review & editing. **Manu Santa-Cruz:** Writing – review & editing. **Pablo de la Nava:** Writing – review & editing. **Mario Fernández-Tizón:** Writing – review & editing. **Hugo Sánchez-Mateos:** Writing – review & editing. **Adrián Barrero:** Writing – review & editing. **Juan Traba:** Writing – review & editing. **Tomasz S. Osiejuk:** Writing – review & editing. **Patrick J. Hart:** Writing – review & editing. **Amanda K. Navine:** Writing – review & editing. **Andrés F. Montoya Muñoz:** Writing – review & editing. **Carlos B. de Araújo:** Writing – review & editing. **Gabriel L.M. Rosa:** Writing – review & editing. **Ingrid M.D. Torres:** Writing – review & editing. **Ana L.C. Catalano:** Writing – review & editing. **Cassio Rachid Simões:** Writing – review & editing. **Diego Llusia:** Writing – review & editing. **Manuel B. Morales:** Writing – review & editing. **Pablo Acebes:** Writing – review & editing. **Juan A. Medina:** Writing – review & editing. **Nicholas Brown:** Writing – review & editing. **Christos Astaras:** Writing – review & editing. **Ilias Karmiris:** Writing – review & editing. **Elizabeth Navarrete:** Writing – review & editing. **Maxime Cauchoix:** Writing – review & editing. **Luc Barbaro:** Writing – review & editing. **Dominik Arend:** Writing – review & editing. **Sandra Müeller:** Writing – review & editing. **Fernando González-García:** Writing – review & editing. **Alberto González-Romero:** Writing – review & editing. **Christos Mammides:** Writing – review & editing. **Michaelangelo Pontikis:** Writing – review & editing. **Giordano Jacuzzi:** Writing – review & editing. **Julian D. Olden:** Writing – review & editing. **Sara P. Bombaci:** Writing – review & editing. **Gabriel Marcacci:** Writing – review & editing. **Alain Jacot:** Writing – review & editing. **Juan P. Zurano:** Writing – review & editing. **Elena Gangenova:** Writing – review & editing. **Diego Varela:** Writing – review & editing. **Facundo Di Sallo:** Writing – review & editing. **Gustavo A. Zurita:** Writing – review & editing. **Andrey Atemasov:** Writing – review & editing. **Junior A. Tremblay:** Writing – review & editing. **Vincent Lamarre:** Writing – review & editing. **Anja Hutschenreiter:** Writing – review & editing. **Alan Monroy-Ojeda:** Writing – review & editing. **Mauricio Díaz-Vallejo:** Writing – review & editing. **Sergio Chaparro-Herrera:** Writing – review & editing. **Robert A. Briers:** Writing – review & editing. **Renata Sousa-Lima:** Writing – review & editing. **Thiago Pinheiro:** Writing – review & editing. **Wigna C. Da Silva:** Writing – review & editing. **Alice Calvente:** Writing – review & editing. **Raiane V. Paz:** Writing – review & editing. **Carlos Salustio-Gomes:** Writing – review & editing. **Dorgival D. Oliveira-Júnior:** Writing – review & editing. **Cicero S. lima-Santos:** Writing – review & editing. **Mauro Pichorim:** Writing – review & editing. **Anamaria Dal Molin:** Writing – review & editing. **Alexandre Antonelli:** Writing – review & editing. **Svetlana Gogoleva:** Writing – review & editing. **Igor Palko:** Writing – review & editing. **Hiếu V. Trong:** Writing – review & editing. **Marina H.L. Duarte:** Writing – review & editing. **Natalia dos Santos Saturnino:** Writing – review & editing. **Samuel R. Silva:** Writing – review & editing. **Ana Rainho:** Writing – review & editing. **Paula Lopes:** Writing – review & editing. **Karl-L. Schuchmann:** Writing – review & editing. **Marinêz I. Marques:** Writing – review & editing. **Ana S. de Oliverira Tissiani:** Writing – review & editing. **Nick A. Littlewood:** Writing – review & editing. **Mao-Ning Tuanmu:** Writing – review & editing. **Sebastian Kepfer-Rojas:** Writing – review & editing. **Andrea L. Aguilera:** Writing – review & editing. **Lluís Brotons:** Writing – review & editing. **Mariano J. Feldman:** Writing – review & editing. **Louis Imbeau:** Writing – review & editing. **Pooja Panwar:** Writing – review & editing. **Aaron S. Weed:** Writing – review & editing. **Anant Dehwal:** Writing – review & editing. **Alfredo Attisano:** Writing – review & editing. **Jörn Theuerkauf:** Writing – review & editing. **Eben Goodale:** Writing – review & editing. **Kevin F.A. Darras:** Writing – review & editing. **Cristian Pérez-Granados:** Writing – original draft, Visualization, Supervision, Resources, Project administration, Methodology, Data curation, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: David Funosas reports financial support was provided by University of Toulouse. Cristian Perez-Granados reports a relationship with Departament de Recerca i Universitats de la Generalitat de Catalunya that includes: funding grants. Esther Sebastian-Gonzalez reports a relationship with Spanish Ministry of Science, Innovation and Universities that includes: funding grants. Esther Sebastian-Gonzalez reports a relationship with ESF Investing in your future that includes: funding grants. Esther Sebastian-Gonzalez reports a relationship with HORIZONMSCA-2021-SE-0 that includes: funding grants. Esther Sebastian-Gonzalez reports a relationship with NextGenerationEU that includes: funding grants. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

DF was funded by the University of Toulouse; CP-G was funded by Departament de Recerca i Universitats de la Generalitat de Catalunya through the 2021-SGR 00302 project; and ES-G was funded by the Spanish Ministry of Science, Innovation and Universities (MCIN/AEI/10.13039/501100011033), NextGenerationEU/PRTR, ESF Investing in

your future (TED2021-130890B-C21 and RYC2019-027216-I), and HORIZONMSCA-2021-SE-0 (action number: 101086387). We are also grateful to the Forest Science and Technology Center of Catalonia (CTFC) IT team for their help and support with the server-based analyses, especially Daniel Macedo and Albert Sanahuja.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecolind.2025.114550.

Data availability

All recordings and annotations used in the study are available on Zenodo at https://doi.org/10.5281/zenodo.14191524. This is explicitly stated in the Data Availability section.

References

Amorós-Ausina, D., Schuchmann, K.L., Marques, M.I., Pérez-Granados, C., 2024. Living together, singing together: revealing similar patterns of vocal activity in two tropical songbirds applying BirdNET. Sensors 24 (17), 5780. https://doi.org/10.3390/s24175780.

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 (1), 4. https://doi.org/10.5751/JFO-00390-950104.

Bota, G., Manzano-Rubio, R., Fanlo, H., Franch, N., Brotons, L., Villero, D., Pérez-Granados, C., 2024. Passive acoustic monitoring and automated detection of the American bullfrog. Biol. Invasions 26 (4), 1269–1279. https://doi.org/10.1007/s10530-023-03244-8.

British Trust for Ornithology, 2023. BTO Acoustic Pipeline: Detection and identification of birds, bats and wildlife from audio recordings. Available at: http://bto.org/pipeline.

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 Ecol. 19 (1), 1–13. https://doi.org/10.1186/s42408-023-00180-9.

Clements, J.F., Schulenberg, T.S., Iliff, M.J., Billerman, S.M., Fredericks, T.A., Gerbracht, J.A., Lepage, D., Sullivan, B.L., Wood, C.L., 2021. The eBird/Clements checklist of birds of the. World v2021.

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 Applications 124 (2), duac003. https://doi.org/10.1093/ornithapp/duac003.

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 (6), e01954. https://doi.org/10.1002/eap.1954.

Darras, K.F., Rountree, R., Van Wilgenburg, S., Cord, A.F., Chen, Y., Dong, L., Wanger, T. C., 2025. Worldwide soundscapes: a synthesis of passive acoustic monitoring across realms. Glob. Ecol. Biogeogr. 34 (5), e70021. https://doi.org/10.1111/geb.70021.

Duc, P.N.H., Torterotot, M., Samaran, F., White, P.R., Gérard, O., Adam, O., Cazau, D., 2021. Assessing inter-annotator agreement from collaborative annotation campaign in marine bioacoustics. Eco. Inform. 61, 101185. https://doi.org/10.1016/j.ecoinf.2020.101185.

Funosas, D., Barbaro, L., Schillé, L., Elger, A., Castagneyrol, B., Cauchoix, M., 2024. Assessing the potential of BirdNET to infer European bird communities from large-scale ecoacoustic data. Ecol. Indic. 164, 112146. https://doi.org/10.1016/j.ecolind.2024.112146.

Gasc, A., Francomano, D., Dunning, J.B., Pijanowski, B.C., 2017. Future directions for soundscape ecology: the importance of ornithological contributions. Auk 134 (1), 215–228. https://doi.org/10.1642/AUK-16-124.1.

Ghani, B., Denton, T., Kahl, S., Klinck, H., 2023. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. Sci. Rep. 13 (1), 22876. https://doi.org/10.1038/s41598-023-49989-z.

Goëau, H., Kahl, S., Glotin, H., Planqué, R., Vellinga, W.-P., Joly, A., 2018. Overview of BirdCLEF 2018: Monospecies vs. soundscape bird identification. CEUR Workshops Proceedings 2125 (9). http://ceur-ws.org/Vol-2125/invited_paper_9.pdf.

Hill, A.P., Prince, P., Piña Covarrubias, E., Doncaster, C.P., Snaddon, J.L., Rogers, A., 2018. AudioMoth: evaluation of a smart open acoustic device for monitoring biodiversity and the environment. Methods in Ecology and Evolution 9 (5), 1199–1211. https://doi.org/10.1111/2041-210X.12955.

Hoefer, S., McKnight, D.T., Allen-Ankins, S., Nordberg, E.J., Schwarzkopf, L., 2023. Passive acoustic monitoring in terrestrial vertebrates: a review. Bioacoustics 32 (5), 506–531. https://doi.org/10.1080/09524622.2023.2209052.

Huus, J., Kelly, K.G., Bayne, E.M., Knight, E.C., 2025. HawkEars: A regional, high-performance avian acoustic classifier. Ecological Informatics 87, 103122. https://doi.org/10.1016/j.ecoinf.2025.103122.

Johnsgard, P.A., 1971. Observations on sound production in the Anatidae. Wildfowl 22, 46–59. https://digitalcommons.unl.edu/context/johnsgard/article/1013/viewcontent/WILDFOWL_1971_Observations_on_sound_Anatidae.pdf.

Kahl, S., 2020. Identifying Birds by Sound: Large-Scale Acoustic Event Recognition for Avian Activity Monitoring. Unpublished doctoral dissertation, Chemnitz University of Technology.

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.

Kahl, S., Navine, A., Denton, T., Klinck, H., Hart, P., Glotin, H., Goëau, H., Vellinga, W.-P., Planqué, R., Joly, A., 2022. Overview of BirdCLEF 2022: endangered bird species recognition in soundscape recordings, 3180(154), p. 1929. https://hal.inrae.fr/hal-03791428.

Kirkland, M., 2024. Chirpity-Electron: AI-powered audio analyzer for bird call visualization, detection and cataloguing. Available at: https://github.com/Mattk70/Chirpity-Electron.

Knight, E., Rhinehart, T., de Zwaan, D.R., Weldy, M.J., Cartwright, M., Hawley, S.H., Kitzes, J., 2024. Individual identification in acoustic recordings. Trends Ecol. Evol. 39 (10), 947–960. https://doi.org/10.1016/j.tree.2024.05.007.

Lahoz-Monfort, J.J., Magrath, M.J., 2021. A comprehensive overview of technologies for species and habitat monitoring and conservation. BioScience 71 (10), 1038–1062. https://doi.org/10.1093/biosci/biab073.

Leroy, E.C., Thomisch, K., Royer, J.Y., Boebel, O., Van Opzeeland, I., 2018. On the reliability of acoustic annotations and automatic detections of Antarctic blue whale calls under different acoustic conditions. J. Acoust. Soc. Am. 144 (2), 740–754. https://doi.org/10.1121/1.5049803.

Macaulay, 2025. The World's Premier Scientific Archive of Natural History Audio, Video, and Photographs. https://www.macaulaylibrary.org/about/.

Manzano-Rubio, R., Bota, G., Brotons, L., Soto-Largo, E., Pérez-Granados, C., 2022. Low-cost open-source recorders and ready-to-use machine learning approaches provide effective monitoring of threatened species. Eco. Inform. 72, 101910. https://doi.org/10.1016/j.ecoinf.2022.101910.

Navine, A., Kahl, S., Tanimoto-Johnson, A., Klinck, H., Hart, P., 2022. A collection of fully-annotated soundscape recordings from the Island of Hawai'i. https://doi.org/10.5281/zenodo.7078499.

Ogawa, R., Gosselin, F., Darras, K.F.A., Roilo, S., Cord, A.F., 2025. A classification-occupancy model based on automatically identified species data. Ecology 106 (5), e70086. https://doi.org/10.1002/ecy.70086.

Olson, D.M., Dinerstein, E., Wikramanayake, E.D., Burgess, N.D., Powell, G.V., Underwood, E.C., Kassem, K.R., 2001. Terrestrial ecoregions of the world: a new map of life on earth: a new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. BioScience 51 (11), 933–938. https://doi.org/10.1641/0006-3568(2001)051[0933:TEOTWA]2.0.CO;2.

Pérez Granados, C., Morant, J., Darras, K.F.A., Gómez, O.H.M., Mendoza, I., Mohedano-Muñoz, M.A., Sebastián-González, E., 2025b. WABAD: A World Annotated Bird Acoustic Dataset for Passive Acoustic Monitoring. Ecology. Accepted.

Pérez Granados, C., Morant, J., Darras, K.F.A., Gómez, O.H.M., Mendoza, I., Mohedano-Muñoz, M.A., Sebastián-González, E., 2025c. WABAD: a world annotated bird acoustic dataset for passive acoustic monitoring. Zenodo. https://doi.org/10.5281/zenodo.15629388.

Pérez-Granados, C., 2023. BirdNET: applications, performance, pitfalls and future opportunities. Ibis 165 (3), 1068–1075. https://doi.org/10.1111/ibi.13193.

Pérez-Granados, C., 2025. Birdnet's confidence scores decrease with bird distance to the recorder: revisiting Pérez-Granados (2023). Ardeola 72 (2), 149–159. https://doi.org/10.13157/arla.72.2.2025.fo1.

Pérez-Granados, C., Traba, J., 2021. Estimating bird density using passive acoustic monitoring: a review of methods and suggestions for further research. Ibis 163 (3), 765–783. https://doi.org/10.1111/ibi.12944.

Pérez-Granados, C., Feldman, M.J., Mazerolle, M.J., 2023. Combining two user-friendly machine learning tools increases species detection from acoustic recordings. Can. J. Zool. 102 (4), 403–409. https://doi.org/10.1139/cjz-2023-0154.

Pérez-Granados, C., Funosas, D., Morant, J., Gómez, O.H.M., Mendoza, I., Sebastián-González, E., 2025. Optimisation of passive acoustic bird surveys: a global assessment of BirdNET settings. Ibis. https://doi.org/10.1111/ibi.70013.

Sebastián-González, E., Pérez-Granados, C., 2025. Geographic Variation in Acoustic Signals in Wildlife: A Systematic Review. Journal of Biogeography 52 (6), e15116. https://doi.org/10.1111/jbi.15116.

Sethi, S.S., Fossøy, F., Cretois, B., Rosten, C.M., 2021. Management relevant applications of acoustic monitoring for Norwegian nature – the sound of Norway. NINA report 2064. In 31. In: Norsk institutt for naturforskning (NINA). https://brage.nina.no/nina-xmlui/bitstream/handle/11250/2832294/ninarapport2064.pdf.

Shonfield, J., Bayne, E.M., 2017. Autonomous recording units in avian ecological research: current use and future applications. Avian Conservation & Ecology 12 (1), 14. https://doi.org/10.5751/ACE-00974-120114.

Stowell, D., 2022. Computational bioacoustics with deep learning: a review and roadmap. PeerJ 10, e13152. https://doi.org/10.7717/peerj.13152.

Sugai, L.S.M., Silva, T.S.F., Ribeiro Jr., J.W., Llusia, D., 2019. Terrestrial passive acoustic monitoring: review and perspectives. BioScience 69 (1), 15–25. https://doi.org/10.1093/biosci/biy147.

Tolkova, I., Chu, B., Hedman, M., Kahl, S., & Klinck, H. (2021). Parsing birdsong with deep audio embeddings. *Preprint. Available at. ArXiv*. Doi:10.48550/arXiv.2108.09203.

Tseng, S., Hodder, D.P., Otter, K.A., 2025. Setting BirdNET confidence thresholds: species-specific vs. universal approaches. Journal of Ornithology 1–13. https://doi.org/10.1007/s10336-025-02260-w.

Van Doren, B.M., Farnsworth, A., Stone, K., Osterhaus, D.M., Drucker, J., Van Horn, G., 2024. Nighthawk: acoustic monitoring of nocturnal bird migration in the Americas. Methods Ecol. Evol. 15 (2), 329–344. https://doi.org/10.1111/2041-210X.14272.

Ventura, T.M., Ganchev, T.D., Pérez-Granados, C., De Oliveira, A.G., de Pedroso, S.G.G., Marques, M.I., Schuchmann, K.L., 2024. The importance of acoustic background modelling in CNN-based detection of the neotropical White-lored Spinetail (Aves, Passeriformes, Furnaridae). Bioacoustics 33 (2), 103–121. https://doi.org/10.1080/09524622.2024.2309362.

Wa Maina, C., Njoroge, P., 2025. Comparing point counts, passive acoustic monitoring, citizen science and machine learning for bird species monitoring in the Mount Kenya ecosystem. Philosophical Transactions B 380 (1928), 20240057. https://doi.org/10.1098/rstb.2024.0057.

Winiarska, D., Neubauer, G., Budka, M., Szymański, P., Barczyk, J., Cholewa, M., Osiejuk, T.S., 2025. BirdNET provides superior diversity estimates compared to observer-based surveys in long-term monitoring. Ecol. Indic. 177, 113747. https://doi.org/10.1016/j.ecolind.2025.113747.

Wood, C.M., Kahl, S., 2024. Guidelines for appropriate use of BirdNET scores and other detector outputs. J. Ornithol. 165, 777–782. https://doi.org/10.1007/s10336-024-02144-5.

Wood, C.M., Kahl, S., Barnes, S., Van Horne, R., Brown, C., 2023a. Passive acoustic surveys and the BirdNET algorithm reveal detailed spatiotemporal variation in the vocal activity of two anurans. Bioacoustics 32 (5), 532–543. https://doi.org/10.1080/09524622.2023.2211544.

Wood, C.M., Barceinas Cruz, A., Kahl, S., 2023b. Pairing a user-friendly machine-learning animal sound detector with passive acoustic surveys for occupancy modeling of an endangered primate. Am. J. Primatol. 85 (8), e23507. https://doi.org/10.1002/ajp.23507.

Xeno-canto, 2025. Sharing Bird Sounds from Around the World. https://www.xeno-canto.org/about/xeno-canto.

Xie, J., Zhong, Y., Zhang, J., Liu, S., Ding, C., Triantafyllopoulos, A., 2023. A review of automatic recognition technology for bird vocalizations in the deep learning era. Eco. Inform. 73, 101927. https://doi.org/10.1016/j.ecoinf.2022.101927.