IBIS
international journal of avian science
Ibis (2025)
doi: 10.1111/ibi.70013

Short Communication

Optimization of passive acoustic bird surveys: a global assessment of BirdNET settings

CRISTIAN PÉREZ-GRANADOS,*[1,2] DAVID FUNOSAS,[3,4] JON MORANT,[5] OSCAR H. MARÍN GÓMEZ,[6] IRENE MENDOZA,[7,8] MIGUEL A. MOHEDANO-MUNOZ,[9] EDUARDO SANTAMARÍA,[7] GIULIA BASTIANELLI,[10] ALBA MÁRQUEZ-RODRÍGUEZ,[11] MICHAŁ BUDKA,[12] GERARD BOTA,[1] JOSÉ M. DE LA PEÑA-RUBIO,[13] ELADIO GARCÍA DE LA MORENA,[14] MANU SANTA-CRUZ,[15] PABLO DE LA NAVA,[16] MARIO FERNÁNDEZ-TIZÓN,[17] HUGO SÁNCHEZ-MATEOS,[18] ADRIÁN BARRERO,[19,20] JUAN TRABA,[19,20] TOMASZ S. OSIEJUK,[12] PATRICK J. HART,[21] AMANDA K. NAVINE,[21] ANDRÉS F. MONTOYA MUÑOZ,[6] CARLOS B. DE ARAÚJO,[22] GABRIEL L. M. ROSA,[23,24] INGRID M. D. TORRES,[24] ANA L. C. CATALANO,[24] CASSIO RACHID SIMÕES,[24,25] DIEGO LLUSIA,[19,20,23] MANUEL B. MORALES,[19,20] PABLO ACEBES,[19,20] JUAN A. MEDINA,[26] NICHOLAS BROWN,[19,27] CHRISTOS ASTARAS,[28] ILIAS KARMIRIS,[28] ELIZABETH NAVARRETE,[29] MAXIME CAUCHOIX,[29] LUC BARBARO,[29] DOMINIK AREND,[30] SANDRA MÜELLER,[30] FERNANDO GONZÁLEZ-GARCÍA,[31] ALBERTO GONZÁLEZ-ROMERO,[31] CHRISTOS MAMMIDES,[32] MICHAELANGELO PONTIKIS,[33] GIORDANO JACUZZI,[34] JULIAN D. OLDEN,[34] SARA P. BOMBACI,[35] GABRIEL MARCACCI,[36] ALAIN JACOT,[36] JUAN P. ZURANO,[22,37] ELENA GANGENOVA,[22,37] DIEGO VARELA,[22,37] FACUNDO DI SALLO,[22,37] GUSTAVO A. ZURITA,[22,37] ANDREY ATEMASOV,[38] JUNIOR A. TREMBLAY,[39] ANJA HUTSCHENREITER,[40] ALAN MONROY-OJEDA,[41] MAURICIO DÍAZ-VALLEJO,[42] SERGIO CHAPARRO-HERRERA,[43,44] ROBERT A. BRIERS,[45] RENATA SOUSA-LIMA,[46] THIAGO PINHEIRO,[46] WIGNA C. DA SILVA,[46] ALICE CALVENTE,[47] ANAMARIA DAL MOLIN,[48] ALEXANDRE ANTONELLI,[49,50,51] SVETLANA GOGOLEVA,[51,52] IGOR PALKO,[52] HI U V. TRONG,[52] MARINA H. L. DUARTE,[53] NATALIA DOS SANTOS SATURNINO,[54] SAMUEL R. SILVA,[54] ANA RAINHO,[55] PAULA LOPES,[55,56] KARL-L. SCHUCHMANN,[57,58] MARINÊZ I. MARQUES,[57] ANA S. DE OLIVERIRA TISSIANI,[57] NICK A. LITTLEWOOD,[59] MAO-NING TUANMU,[60] YI-RU CHENG,[60] HSUAN CHAO,[60] SEBASTIAN KEPFER-ROJAS,[61] ANDREA L. AGUILERA,[62] LLUÍS BROTONS,[1,63,64] MARIANO J. FELDMAN,[1] LOUIS IMBEAU,[65] POOJA PANWAR,[66] AARON S. WEED,[67] ANANT DEHWAL,[68] ALFREDO ATTISANO,[69] JÖRN THEUERKAUF,[69] DORGIVAL D. OLIVEIRA-JÚNIOR,[47,70] CICERO S. LIMA-SANTOS,[47,70] CARLOS SALUSTIO-GOMES,[47,70] RAIANE V. PAZ,[47,70] MAURO PICHORIM,[47,70] EBEN GOODALE[71] & ESTHER SEBASTIÁN-GONZÁLEZ[5,72]

[1]*Biodiversity Conservation and Management Programme, Forest Science and Technology Center of Catalonia (CTFC), 25280, Lleida, Spain*
[2]*IUCN SSC Species Monitoring Specialist Group, Gland, Switzerland*
[3]*Station d'Écologie Théorique et Expérimentale, SETE, CNRS, 09200, Moulis, France*
[4]*Université Paul Sabatier – Toulouse III, 31077, Toulouse, France*
[5]*Department of Ecology, University of Alicante, 03690 San Vicente del Raspeig, Alicante, Spain*
[6]*Programa de Biología, Grupo de Investigación en Biodiversidad y Biotecnología (GIBUQ), Universidad del Quindío, Armenia, Quindío, Colombia*
[7]*Department of Ecology and Evolution, Estación Biológica de Doñana (CSIC), Avda. Américo Vespucio, 26, 41092, Sevilla, Spain*
[8]*Department of Plant Biology and Ecology, Faculty of Biology, University of Sevilla, Avda Reina Mercedes s/n, 41012, Sevilla, Spain*
[9]*Grupo de Sistemas Complejos, Departamento de Ingeniería Agroforestal, ETSIAAB, Universidad Politécnica de Madrid, 28040, Madrid, Spain*
[10]*ICTS-Doñana, Estación Biológica de Doñana (EBD), CSIC, C/Américo Vespucio 26, 41092, Sevilla, Spain*
[11]*Institute of Marine Research (INMAR), International Campus of Excellence in Marine Science (CEIMAR), University of Cádiz, 11510, Puerto Real, Cádiz, Spain*
[12]*Department of Behavioural Ecology, Institute of Environmental Biology, Faculty of Biology, Adam Mickiewicz University in Poznań, Poznań, Poland*

[13]Blue Nature Birding and Nature Tours SL, Calle Rufino Blanco 17, 28028, Madrid, Spain
[14]Biodiversity Node, Sector Foresta, 17. 1°B, 28760, Tres Cantos, Madrid, Spain
[15]Eurofins MITOX B.V, Science Park 408, 1098 XH, Amsterdam, the Netherlands
[16]SEO/BirdLife, 28052, Madrid, Spain
[17]Instituto de Investigación en Recursos Cinegéticos, IREC-CSIC-UCLM-JCCM, Ciudad Real, Spain
[18]Iduna Tours, 10720, Villar de Plasencia, Cáceres, Spain
[19]Terrestrial Ecology Group (TEG-UAM), Departamento de Ecología, Universidad Autónoma de Madrid, 28049, Madrid, Spain
[20]Centro de Investigación en Biodiversidad y Cambio Global (CIBC-UAM), Universidad Autónoma de Madrid, 28049, Madrid, Spain
[21]Department of Biology, University of Hawai'i at Hilo, 200 W. Kawili St, Hilo, Hawaii, 96720, USA
[22]Atlantic Forest Biodiversity Observatory, Instituto de Biología Subtropical, CONICET-Universidad Nacional de Misiones, Puerto Iguazú, Argentina
[23]Laboratório de Ornitologia e Bioacústica, Universidade Estadual de Londrina, Londrina, Brazil
[24]ConservaSom, João Pessoa, Paraíba, Brazil
[25]Programa de Pós-graduação em Biodiversidade e Evolução, Instituto de Biologia, Universidade Federal da Bahia, Salvador, Brazil
[26]BUTEO Environmental Initiatives, Mojados, Valladolid, Spain
[27]Department of Life Sciences, Imperial College London, Exhibition Road, South Kensington, London, SW7 2AZ, UK
[28]Forest Research Institute, ELGO-DIMITRA, Loutra Thermis, Thessaloniki, 57006, Greece
[29]Dynafor, INRAE-INPT, University of Toulouse, 31326, Castanet-Tolosan, France
[30]Geobotany, Faculty of Biology, University of Freiburg, Freiburg, Germany
[31]Red Biología y Conservación de Vertebrados, Instituto de Ecología, A.C. Carretera Antigua a Coatepec No. 351, El Haya, Xalapa, Veracruz, Mexico
[32]Nature Conservation Unit, Frederick University, Gianni Freiderikou 7, Nicosia, 1036, Cyprus
[33]Department of Biological Sciences, University of Cyprus, Nicosia, Cyprus
[34]School of Aquatic and Fishery Sciences, University of Washington, Seattle, WA, 98195, USA
[35]Department of Fish, Wildlife, and Conservation Biology, Colorado State University, 1474 Campus Delivery, Fort Collins, Colorado, 80521, USA
[36]Swiss Ornithological Institute, Seerose 1, 6204, Sempach, Switzerland
[37]Asociación Civil Centro de Investigaciones del Bosque Atlántico (CeIBA), Bertoni 85, Puerto Iguazú, Misiones, Argentina
[38]V.N.Karazin Kharkiv National University, Kharkiv, Ukraine
[39]Environment and Climate Change Canada, 801-1550 Ave d'Estimauville, Québec, Canada,G1J 0C3
[40]Instituto de Investigaciones en Ecosistemas y Sustentabilidad, Universidad Autónoma de México, Morelia, Michoacan, Mexico
[41]Kiekari Bird Observatory, KiekariTerra A.C, Xico, Veracruz, Mexico
[42]Laboratorio de Bioclimatología, Red Biología Evolutiva, Instituto de Ecología, A.C. Carretera Antigua a Coatepec No. 351, El Haya, Xalapa, Veracruz, México
[43]Proyecto Atlapetes, Antioquia, Colombia
[44]Laboratorio de Ecología Evolutiva y Urbana, Universidad del Norte, Barranquilla, Colombia
[45]Centre for Conservation and Restoration Science, School of Applied Sciences, Edinburgh Napier University, Sighthill Campus, Edinburgh, EH11 4BN, UK
[46]Laboratory of Bioacoustics/EcoAcoustic Research Hub. Biosciences Center, Universidade Federal do Rio Grande do Norte, Campus Universitário, Lagoa Nova, Natal, RN, 59078-970, Brazil
[47]Departamento de Botânica e Zoologia, Universidade Federal do Rio Grande do Norte, Campus Universitário, Lagoa Nova, Natal, RN, 59078-970, Brazil
[48]Department of Microbiology and Parasitology, Biosciences Center, Universidade Federal do Rio Grande do Norte, 59078-970, Natal, RN, Brazil
[49]Royal Botanic Gardens, Kew, Richmond, Surrey, TW9 3AE, UK
[50]Department of Biology, University of Oxford, South Parks Road, Oxford, OX1 3RB, 96, UK
[51]Gothenburg Global Biodiversity Centre, Department of Biological and Environmental Sciences, University of Gothenburg, Box 463, 405 30, Göteborg, Sweden
[52]Southern Branch of the Joint Vietnam-Russia Tropical Science and Technology Research Center, HoChi Minh City, Vietnam
[53]Environmental Research and Innovation Centre (ERIC), School of Science, Engineering and Environment, University of Salford, Manchester, UK
[54]Graduate Program in Biodiversity and Environment, Pontifical Catholic University of Minas Gerais, Belo Horizonte, Brazil
[55]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
[56]SPEA – Portuguese Society for the Study of Birds, Lisbon, 1700-031, Portugal
[57]Computational Bioacoustics Research Unit (CO.BRA), Institute for Science and Technology in

Wetlands (INAU), Federal University of Mato Grosso (UFMT), Cuiabá, 78060-900, Brazil
[58]*Ornithology, Zoological Research Museum A. Koenig (ZFMK), 53113, Bonn, Germany*
[59]*Scotland's Rural College, Craibstone Estate, Bucksburn, Aberdeen, AB21 9YA, UK*
[60]*Biodiversity Research Center, Academia Sinica, Taipei, Taiwan*
[61]*Department of Geosciences and Natural Resource Management, University of Copenhagen, Rolighedsvej 23, Frederiksberg C. Copenhagen, Denmark*
[62]*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*
[63]*CREAF, 08193, Cerdanyola del Vallès, Spain*
[64]*CSIC, 08193, Cerdanyola del Vallès, Spain*
[65]*Institut de recherche sur les forêts, Université du Québec en Abitibi-Témiscamingue, 445, boul de l'Université, Rouyn-Noranda, Québec, Canada,J9X 5E4*
[66]*Ecology, Evolution, Environment, and Society, Dartmouth College, 78 College St, Hanover, New Hampshire, USA*
[67]*US Department of Interior, National Park Service, Northeast Temperate Network, 54 Elm St, Woodstock, Vermont, USA*
[68]*Biology Department, Bradley University, Peoria, Illinois, 61625, USA*
[69]*Museum and Institute of Zoology, Polish Academy of Sciences, Warsaw, Poland*
[70]*Programa de Pós-Graduação em Ecologia, Universidade Federal do Rio Grande do Norte, Campus Universitário, Lagoa Nova, Natal, RN, 59078-970, Brazil*
[71]*Department of Health and Environmental Sciences, Xi'an Jiaotong-Liverpool University, Suzhou, 215123, China*
[72]*Instituto Multidisciplinar para el Estudio del Medio "Ramón Margalef", University of Alicante, 03690 San Vicente del Raspeig, Alicante, Spain*

BirdNET is a popular machine learning tool for automated recognition of bird sounds. However, evidence on how to optimize its settings for accurate bird monitoring remains limited. Here, we evaluate how BirdNET settings influence model performance in identifying bird vocalizations and characterizing bird communities, using 4224 1-min recordings from 67 recording locations worldwide. Giving equal importance to recall and precision, a low confidence score threshold (0.1–0.3) appears optimal for detecting bird vocalizations, whereas higher thresholds (around 0.5) are more suitable for characterizing bird communities. Based on our findings, we recommend increasing the *Overlap* parameter from its default value of 0 to 2 s, as this consistently improves BirdNET performance in detecting both bird vocalizations and species presence. The effect of the *Sensitivity* parameter varied across regions. However, a value of 0.5 maximizes global performance for community-level analyses across all confidence thresholds, and a value of 1.5 generally yields better results for vocalization-level studies, particularly at low confidence thresholds. Our findings offer practical guidance for selecting BirdNET settings in passive acoustic bird surveys, enhancing both the identification of bird vocalizations and the characterization of bird communities.

Keywords: automated detection, bird monitoring, convolutional neural networks, machine learning, novel communities, passive acoustic monitoring.

Passive acoustic monitoring (PAM) is a non-invasive, automated method extensively used for bird monitoring (Darras *et al.* 2019, Pérez-Granados & Traba 2021). A key advancement in this field has been the development of machine learning and deep learning algorithms for the automated identification of bird vocalizations (Stowell 2022, Xie *et al.* 2023), with BirdNET being among the most widely used software (Kahl *et al.* 2021, Pérez-Granados 2023). BirdNET is based on a convolutional neural network, capable of identifying over 6500 bird species worldwide (https://github.com/birdnet-team/BirdNET-Analyzer). BirdNET divides recordings into 3-s segments and generates multispecies predictions of species presence for each segment. Each prediction is assigned a quantitative Confidence score from 0.01 (low model certainty in the identification) to 1 (very high model certainty), allowing users to filter BirdNET outputs based on a confidence score threshold. Setting a low confidence score threshold minimizes the risk of false negatives (i.e. missed detections) but increases the likelihood of false positives (i.e. mislabelled detections), and vice versa for a high confidence threshold (Wood & Kahl 2024).

In addition to the Confidence score threshold, BirdNET allows users to adjust two other parameters: (1) Overlap (range: 0–3 s), which controls the degree of overlap between consecutive 3-s segments, and (2) Sensitivity (range: 0.5–1.5), which modulates the spread of Confidence scores: Sensitivity values less than 1 increase the model's certainty in its top predictions and decrease

*Corresponding author.
Email: cristian.perez@ctfc.cat
Twitter id: @MorantJon

Cristian Pérez-Granados and David Funosas contributed equally to the study.

its certainty in the bottom predictions, whereas values greater than 1 make confidence scores more uniform across predictions. In summary, low Confidence score thresholds combined with high Overlap and Sensitivity values maximize the recall rate, i.e. the proportion of vocalizations detected among those present in a recording, to the detriment of precision, i.e. the proportion of vocalizations correctly identified by BirdNET. As a result, an inherent trade-off emerges between recall and precision (Funosas *et al.* 2024).

Although previous research has explored the impact of adjusting the input values of BirdNET parameters (Wood *et al.* 2023, Funosas *et al.* 2024), evidence on optimal settings for automated bird monitoring remains scarce. BirdNET performance varies significantly across species and environmental contexts (Funosas *et al.* 2024, Pérez-Granados 2025). Large-scale research is therefore needed to define parameter settings that optimize monitoring outcomes for bird monitoring using BirdNET. To address this gap, we provide a comprehensive evaluation to identify the best set of settings – at both vocalization (i.e. the best settings to correctly classify a specific vocalization within a recording) and dataset (i.e. the best settings to correctly identify the species appearing in a collection of recordings from the same study area) levels – to optimize the balance between BirdNET precision and recall. Data at the vocalization level are helpful when it is important to know the exact moment within a recording when a species is vocally active (e.g. for studies on bird activity patterns), whereas data at the dataset level may be useful for studies aiming to obtain a list of species from a study area. To achieve this, we analysed 4224 1-min audio recordings collected from 67 recording locations across six continents, comprising a total of 89 061 bird vocalizations – all identified and annotated by expert ornithologists. We hope our results will guide future studies in determining optimal parameter settings and support the continued refinement of BirdNET, both for ecological monitoring of bird species and for the characterization of novel bird acoustic communities (*sensu* Hartig *et al.* 2024).

METHODS

Soundscape collection

The analysed soundscapes are part of the World Annotated Bird Acoustic Dataset (WABAD, Pérez-Granados *et al.* 2025). These recordings were annotated at the vocalization level by local experts. For consistency across datasets, our analyses included all recordings from the 67 recording locations in WABAD with annotations providing the exact start and end times for each bird vocalization present in the recording (see Audio annotations section). The five recording locations with annotations lacking exact start and end times were excluded from this study, as such labels are unsuitable for vocalization-level analyses. Most of the data were collected in the northern hemisphere (mainly from Europe and North America). Nonetheless, the database includes data from six continents, with several recording locations in Central and South America but low representation from Africa, Asia and Oceania (Fig. 1). In total, we analysed 4224 1-min recordings collected at 67 recording locations (Fig. 1). We provide detailed information (e.g. recording location, minutes annotated per location, geographical coordinates, biome, recorder used) for each recording site in Table S1. The recordings and annotations used in this study are publicly available. For access and further details, see Pérez-Granados *et al.* (2025).

Audio annotations

Expert ornithologists familiar with the local avifauna examined each 1-min audio recording spectrogram and identified every single bird vocalization at the species level. All annotations followed the Clements Checklist (Clements *et al.* 2021), which guarantees taxonomic alignment with the nomenclature used in BirdNET. The experts annotated each vocalization using bounding boxes: the start and end points of the box (x-axis) mark the duration of the sound and the top and bottom boundaries (y-axis) indicate its frequency range (lowest to highest). Two vocalizations from the same species could be included in the same box when they were separated by less than 1 s; otherwise, a separate annotation was made. The coordinator of each recording location ensured that audio annotations met the criteria specified, with a subset of the files of each recording location (c.30%) being double-checked by the WABAD coordinators. All labels included in that study were reviewed by a single observer, with no formal inter-observer validation. A detailed description of the annotation process, along with all audio annotations, can be found in Pérez-Granados *et al.* (2025).

BirdNET settings

We analysed the recordings by running BirdNET-Analyser v2.4.0 (model BirdNET_GLOBAL_6K_V2.4_Model_FP32.tflite) with varying input parameter values via a Linux shell script interfacing with the algorithm's Python codebase, following Funosas *et al.* (2024). We processed the data with the default minimum Confidence score threshold of 0.1 and nine value combinations of Overlap (0, 1 or 2 s) and Sensitivity (0.5, 1.0 or 1.5), spanning the respective parameter ranges of 0–3 s and 0.5–1.5. We configured BirdNET to filter the list of potentially detectable species based on

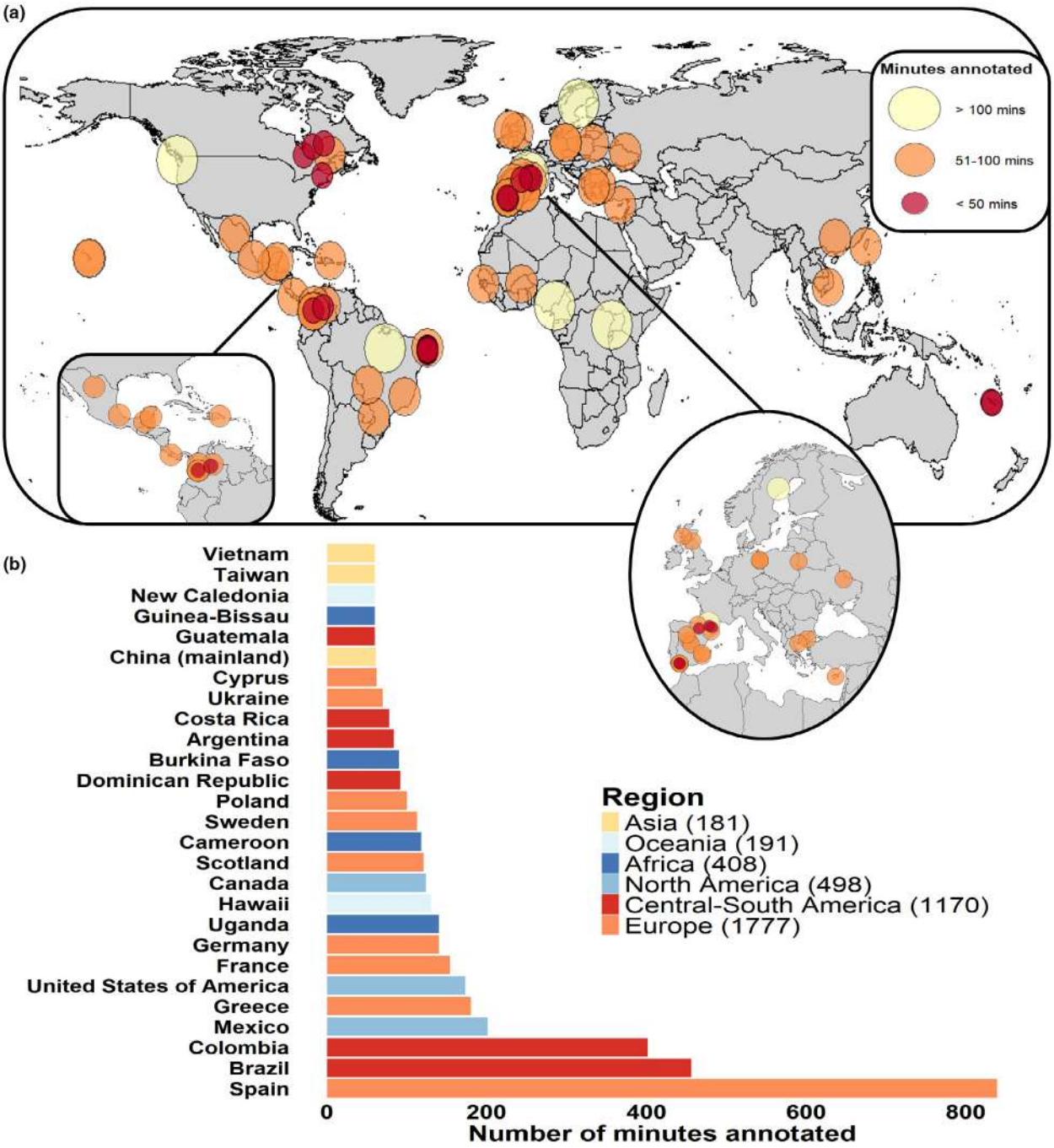

Figure 1. (a) Global mapping of 67 recording locations considered in the study. Colours and sizes of circles refer to the number of minutes annotated per recording site. The small circles show the location of recording locations in Europe and Central America. (b) Number of minutes annotated per location. Colours of the recording locations in this panel refer to different regions, with the total number of minutes annotated per location provided in parentheses. *Although Hawai'i is part of the USA, we classified it separately within the Oceania region based on biogeographical criteria.

the following criteria: (1) recording site location (Table S1), (2) recording date (week of the year) and (3) a minimum occurrence frequency threshold of 0.02 (following Funosas *et al.* 2024), which defines the lowest regional and temporal occurrence frequency a species must have to be included in BirdNET's list of potentially

detectable species (range 0.01–0.99). BirdNET-Analyser v2.4 uses eBird checklist frequency data to estimate the range of bird species and the probability of their occurrence given coordinates and week of the year (see https://github.com/birdnet-team/BirdNET-Analyzer/discussions/234). A low threshold (as in our study) broadens the list of potentially detectable species, as it includes those with low likelihoods of occurrence, whereas a higher threshold limits the list to species with the highest expected occurrence based on eBird data.

BirdNET assessment

We assessed BirdNET performance across the nine combinations of settings by comparing model predictions with the annotations made by experts through a series of custom R scripts (version 4.2.2; R Core Team 2025) that (1) categorize BirdNET predictions according to their correctness, (2) compute performance metrics based on this categorization, and (3) generate corresponding summary tables and plots. The scripts have been adapted from and can be accessed at Funosas *et al.* (2024). The assessments were conducted at two levels: (1) vocalization level, providing a fine-grained picture of BirdNET's ability to correctly detect and identify individual bird vocalizations; and (2) dataset level, offering insight into BirdNET's ability to characterize the composition of bird communities based on a collection of recordings from the same location. BirdNET predictions were categorized into four possible outcomes (Fig. S1):

- True positives (TP): At the vocalization level, a BirdNET prediction was classified as a TP when an expert labelled the same species at the same time (see definition below). At the dataset level, a bird species was considered a TP when there was at least one correct identification of that species by BirdNET in any of the recordings from the same dataset (i.e. recording location).
- False positives (FP): At the vocalization level, a BirdNET prediction was classified as an FP when an expert did not detect the same species at the same time. At the dataset level, a bird species was considered an FP when all BirdNET predictions of that species in the dataset were incorrect.
- True negatives (TN): A prediction was classified as a TN when a vocalization or species not identified by the expert was also not predicted by BirdNET, at either the dataset or vocalization level at the same time.
- False negatives (FN): A prediction was classified as an FN when a vocalization or species identified by the expert was not predicted by BirdNET at the same time.

Following the above categorization criteria, we evaluated BirdNET precision, recall and false-positive rate (FPR) at both vocalization and dataset levels. Precision is defined as the proportion of species or vocalizations correctly predicted relative to the total number of species or vocalizations predicted by BirdNET. The recall rate measures the proportion of species or vocalizations correctly predicted relative to the total number of species or vocalizations present in the recording (Pérez-Granados 2023). The FPR measures the likelihood of BirdNET falsely identifying an absent species as present. These three metrics were estimated using 90 different minimum confidence thresholds (from 0.1 to 0.99 with a step of 0.01; Funosas *et al.* 2024). Analyses at the vocalization level compare BirdNET predictions within 3-s segments to expert annotations, whereas dataset-level assessments match expert-annotated species lists to BirdNET-predicted species, counting only correct matches (i.e. instances where a BirdNET prediction temporally overlaps with an expert annotation of the same species). At the vocalization level, recall and FPR were calculated by pooling all BirdNET predictions that overlapped with a given vocalization (i.e. those ending after its onset or beginning before its offset), and precision was calculated by pooling all manual annotations that overlapped with each prediction segment using the same criterion. The specific formulae used to calculate precision, recall and FPR are the following:

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP})$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN})$$

$$\text{FPR} = \text{FP}/(\text{FP} + \text{TN})$$

The three metrics were calculated at each level of analysis: vocalization and dataset. It is essential to note that, according to our categorization criteria, a single correct prediction of a species by BirdNET was sufficient for the species to be considered a true positive at the dataset level, thereby favouring higher recall results in datasets of longer duration. The values obtained for the precision, recall and FPR metrics were used to plot the Precision–Recall (PR) and receiver operating characteristic (ROC) curves, both accompanied by an estimation of the area under the curve (AUC; Davis & Goadrich 2006, Knight *et al.* 2017). The PR curve plots precision against recall for each minimum confidence threshold considered, illustrating the trade-off between these two metrics. Similarly, the ROC curve plots recall against the FPR for each minimum confidence threshold, revealing the trade-off between these two metrics as well. For both curves, the AUC serves as a measure of the algorithm's predictive power, with values ranging from 0 to 1, where higher values indicate greater predictive power.

The AUC of PR integrates precision across the entire recall range, meaning that extending this range – even toward lower recall values – can increase the total area under the curve. Consequently, a PR curve with a broader recall range can have a higher AUC than one with a narrower range, even if the latter maintains higher precision at every overlapping recall level. The same principle applies to AUC of ROC scores and FPR ranges. Higher Sensitivity values are associated with greater variability in both recall and FPR scores across confidence levels, resulting in a broader range of recall and FPR values compared with those obtained with lower Sensitivity values (Fig. S2). Hence, to ensure comparability across different Sensitivity values, PR AUC was adjusted for the recall range and ROC AUC for FPR range using the following formulae:

$$adj\_PR\_AUC = \frac{PR\ AUC}{max(recall) - min(recall)}$$

$$adj\_ROC\_AUC = \frac{ROC\ AUC}{max(FPR) - min(FPR)}$$

Additionally, we estimated the F1-score, which evaluates an algorithm's predictive power by integrating both precision and recall (Knight *et al.* 2017), across 90 Confidence score thresholds (from 0.10 to 0.99). The formula used was the following:

$$F\text{-score} = (1 + \beta^2) \times \text{precision} \times \text{recall} / (\beta^2 \times \text{precision} + \text{recall})$$

For consistency and to facilitate comparisons with other studies, we computed the F1-score, i.e. the F-score with a β equal to 1, assigning equal importance to precision and recall. F1-score values range from 0 to 1, with higher F1-score values indicating a better model performance (i.e. a value of 1 represents perfect precision and recall).

RESULTS

Optimizing BirdNET parameters at the vocalization level

Both the Overlap and the Sensitivity values impacted BirdNET performance at the vocalization level (Table 1). The AUC scores for the PR curves evaluated globally across all datasets consistently increased with higher Overlap values at each specific Sensitivity score. In fact, the improved BirdNET performance at the vocalization level when using an Overlap of 2 was consistent within and among regions (Table 2) as well as among biomes (Table S2). We also found that a Sensitivity value of 0.5 yielded the highest PR AUC scores across the three Overlap values considered. However, the influence of Sensitivity on PR AUC scores seems to be substantially less robust than that of Overlap. Our results also show that, at the vocalization level, the PR AUC score was maximized with an Overlap of 2 and a Sensitivity of 0.5 (Table 1, see results at recording location level in Table S3). However, the optimal Sensitivity value for maximizing PR AUC scores varied across biogeographical regions, being 0.5 in three regions and 1.5 in the other three (Table 2). Regarding the AUC scores for the ROC curves, the largest differences appeared between the Sensitivity values of 1.0 and 1.5, with the latter yielding the lowest ROC AUC scores for the three Overlap values analysed (Table 1). The highest ROC AUC score was obtained with an Overlap of 2 and a Sensitivity of 1 or 0.5 (Table 1).

Optimizing BirdNET parameters at the dataset level

At the dataset level, Sensitivity had the strongest influence on BirdNET performance. Under all Overlap values considered, the highest PR AUC scores were obtained using a Sensitivity of 0.5, with large differences across Sensitivity values (Table 1). This result is consistent across different geographical regions, with all regions reaching their highest PR AUC scores at a Sensitivity of 0.5 (see also the high degree of consistency within regions and among biomes of the optimal Sensitivity value in Table 2 and Tables S2 and S3). The impact of Overlap on the PR curve was small, but higher PR AUC scores were obtained at the dataset level when using higher Overlap values. The combination of settings maximizing PR AUC consisted of an Overlap of 2 and a Sensitivity of 0.5 (Table 1), which was consistent in four of the six regions analysed (Table 2).

Regarding the ROC AUC scores, we found small differences between the different groups of settings tested. Nonetheless, the lowest ROC AUC scores corresponded to a Sensitivity of 0.5 at any given Overlap value. Differences in ROC AUC scores across Overlap values were small and variable. However, the highest ROC AUC score across all regions was achieved with an Overlap of 0 and a Sensitivity of 1.5.

F1-score curves: impact of confidence score threshold

The F1-score curves showed that BirdNET performance remained relatively consistent across the three Overlap settings at both vocalization and dataset levels (Fig. 2). However, at the vocalization level, performance showed a slight overall improvement as Overlap increased. In contrast, Sensitivity had a substantial impact on BirdNET performance at both levels. The effect of the

Overlap	Sensitivity	AUC values			
		Vocal_PR	Vocal_ROC	Dataset_PR	Dataset_ROC
0	0.5	0.102	0.083	0.342	0.119
0	1	0.092	0.085	0.238	0.135
0	1.5	0.099	0.070	0.138	**0.156**
1	0.5	0.120	0.085	0.369	0.124
1	1	0.109	0.089	0.260	0.136
1	1.5	0.118	0.069	0.141	0.152
2	0.5	**0.155**	0.090	**0.380**	0.130
2	1	0.138	**0.091**	0.262	0.148
2	1.5	0.151	0.065	0.153	0.150

Table 1. Area under the curve (AUC) scores for both Precision–Recall (PR) and receiver operating characteristic (ROC) curves using nine combinations of values for the Overlap and Sensitivity settings. The results shown have been obtained with the default minimum Confidence score threshold (0.1). Results are presented at the vocalization and dataset levels. The best results are highlighted in bold.

Table 2. Continent-specific optimal Overlap and Sensitivity settings for BirdNET-Analyser to maximize area under the curve (AUC) scores for the Precision–Recall (PR) curve. Nine combinations of settings – three levels of Overlap between consecutive predictions (0, 1 and 2 s) and three Sensitivity values (0.5, 1 and 1.5) – were evaluated using a minimum Confidence score threshold of 0.1. To measure model improvement, we report the variation (Δ) in AUC scores for both PR and receiver operating characteristic (ROC) curves between the best-performing settings for PR AUC optimization and the default settings (Overlap = 0, Sensitivity = 1). Finally, the number and percentage of regional datasets converging on the same optimal settings are presented for each region. Results are presented separately for vocalization-level and dataset-level analyses.

Analysis level	Region	Overlap (s)	Sensitivity	Δ PR_AUC	Δ ROC_AUC	Cross-dataset convergence on Overlap	Cross-dataset convergence on Sensitivity
Vocalization	Africa	2	0.5	0.020	0.029	4/4 (100%)	1/4 (25%)
	Asia	2	1.5	0.026	0.005	3/3 (100%)	2/3 (67%)
	Central-South America	2	0.5	0.057	0.029	18/20 (90%)	7/20 (35%)
	Europe	2	1.5	0.084	−0.053	27/27 (100%)	15/27 (56%)
	North America	2	0.5	0.122	0.009	9/9 (100%)	3/9 (33%)
	Oceania	2	1.5	0.139	−0.038	3/4 (75%)	2/4 (50%)
Dataset	Africa	2	0.5	0.111	−0.005	2/4 (50%)	4/4 (100%)
	Asia	2	0.5	0.105	−0.016	0/3 (0%)	3/3 (100%)
	Central-South America	2	0.5	0.124	−0.004	10/20 (50%)	19/20 (95%)
	Europe	1	0.5	0.140	−0.024	8/27 (30%)	27/27 (100%)
	North America	2	0.5	0.118	0.007	6/9 (67%)	8/9 (89%)
	Oceania	0	0.5	0.211	−0.023	2/4 (50%)	3/4 (75%)

Sensitivity setting varied between the two levels of analysis. At the vocalization level, when using Sensitivity values of 0.5 and 1, the F1-score declined almost linearly as the minimum Confidence score thresholds increased. However, with a Sensitivity of 1.5, the F1-score increased until it reached its maximum around a Confidence score threshold of 0.3. Interestingly, the F1-score curve with a Sensitivity of 1.5 showed better performance than the F1-score curves obtained with the other two Sensitivity settings between confidence thresholds of 0.15 and 0.6, while also showing poorer performance at both very low (<0.15) and very high (>0.75) confidence thresholds.

The highest F1-scores at the vocalization level were obtained with an Overlap of 2, a Sensitivity of 1.5 and a Confidence score threshold around 0.3. At the dataset level, the highest F1-scores were consistently achieved with a Confidence score threshold of around 0.5 across all settings. The best overall BirdNET performance was achieved with a Sensitivity of 0.5, followed by 1.0, while a Sensitivity of 1.5 yielded the lowest performance. Under all settings, the highest and nearly identical F1-scores were obtained with Confidence score thresholds around 0.5. The largest differences in F1-scores appeared between Sensitivity values at the lowest and highest minimum Confidence score thresholds, particularly at the higher end.

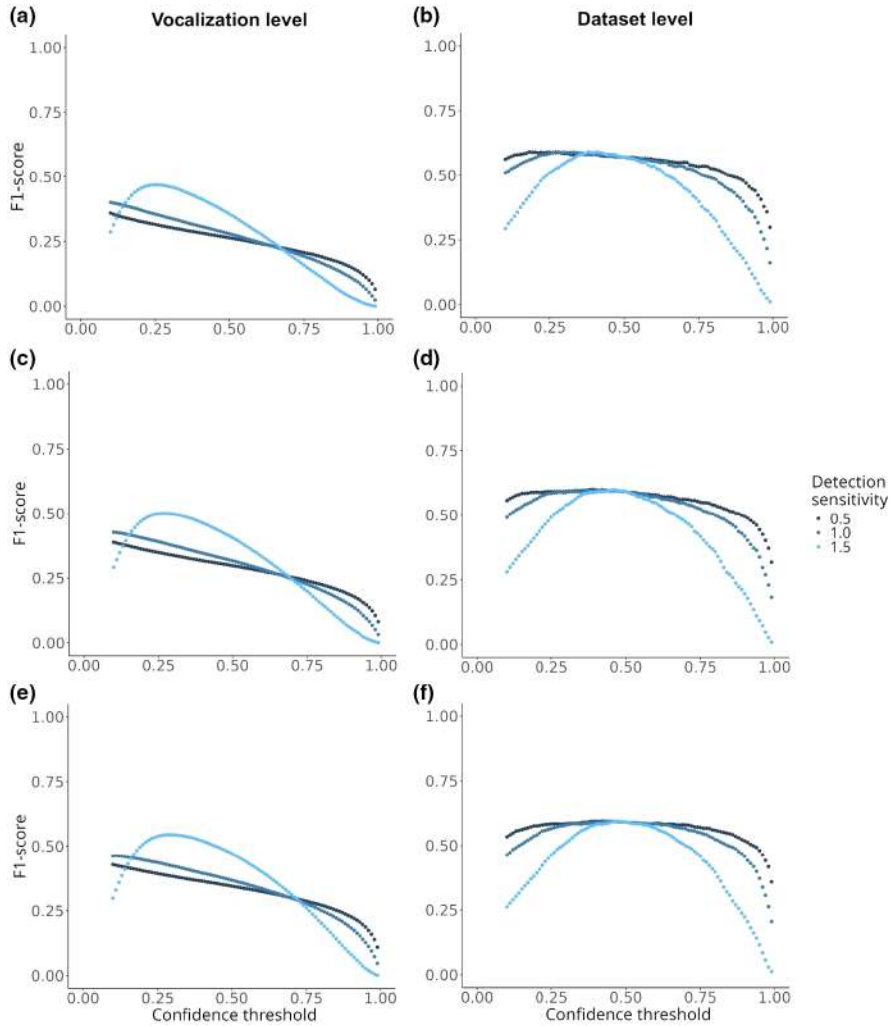

Figure 2. BirdNET-Analyser F1-score curves for nine combinations of settings. The three panels on the left (a, c and e) present results at the vocalization level, while the three panels on the right (b, d and f) show results at the dataset level. The panels are organized by Overlap settings: the top panels (a, b) correspond to Overlap = 0, the middle panels (c, d) correspond to Overlap = 1, and the bottom panels (e, f) correspond to Overlap = 2. Within each panel, the three different Sensitivity values (0.5, 1 and 1.5) are represented by three distinct colours.

DISCUSSION

BirdNET has become a widely adopted tool for automated bird sound recognition, yet the majority of past studies have relied heavily on its default settings, with minimal parameter adjustments – reviewed by Pérez-Granados (2023); see also Funosas *et al.* (2024). Here, we demonstrated that parameter tuning can substantially improve performance, with optimal settings varying according to the monitoring goal – whether focused on identifying individual vocalizations or detecting species presence in acoustic datasets. The large variability observed in BirdNET outputs across different parameter configurations highlights the need for standardized parameter guidelines. Such standards would improve cross-study comparisons, ensure temporal and spatial reproducibility, and facilitate the integration of acoustic data into broader biodiversity monitoring platforms.

Our findings provide strong evidence that increasing the Overlap parameter from its default value of 0 to 2 consistently improves BirdNET performance at the vocalization level, and moderate evidence that it improves performance at the dataset level as well. The improvement probably reflects increased temporal capture: greater overlap increases the chance that most of a bird vocalization falls within a single prediction segment,

and longer within-segment durations are associated with higher recall (Funosas *et al.* 2024). Although the benefits were most evident at the vocalization level, higher Overlap also led to performance gains at the dataset level, albeit to a lesser extent. Importantly, this improvement in recall appears not to come at a general cost of reduced precision, as shown by consistently higher PR and ROC AUC scores at both levels when using an Overlap of 2 (Table 1). Although higher Overlap values increase processing times, this limitation can be offset by using computing systems and server-based analyses. With our dataset, and in comparison with an Overlap of 0 (default setting), using an Overlap of 1 increased execution time by 25%, while an Overlap of 2 nearly doubled it (+96%). Overall, given the clear advantages in output quality using higher degrees of Overlap, BirdNET capabilities may be limited by the conservative default setting of zero Overlap.

Notably, BirdNET performance – at both the vocalization and dataset levels – varied more across Sensitivity values than across Overlap values, particularly at the vocalization level, where the most effective setting varied greatly between regions. As expected, assigning a high Sensitivity value in BirdNET increased the number of predictions, especially those with lower confidence scores (Fig. S3). However, it remains unclear why, in half of the regions, the best performance at the vocalization level was obtained with a value of 0.5, whereas in the other half it was achieved with a value of 1.5. Further research should aim to evaluate whether such differences among regions might be related to different bird diversity, bird song parameters, local vegetation structure or environmental noise.

Our results suggest that high Sensitivity values may not be optimal for maximizing PR AUC scores at the dataset level. This is because they amplify the asymmetry in how precision and recall respond to the Confidence score threshold: at low thresholds, precision drops sharply while recall increases more gradually; at high thresholds, recall drops sharply while precision increases more gradually. As PR AUC weights precision and recall equally, this imbalance reduces overall performance. Low Sensitivity values moderate these effects, producing a more balanced precision–recall trade-off across the confidence threshold range. As a result, F1-scores are consistently higher for low Sensitivity values when using either low or high confidence thresholds. In the mid-range of confidence thresholds (0.35–0.6), where the asymmetry is less pronounced, F1-scores are relatively unaffected by the Sensitivity setting. However, because higher Sensitivity values strengthen the positive correlation between Confidence scores and precision and the negative correlation between Confidence scores and recall (Fig. S4), they enable targeted optimization: combining high Sensitivity with a high confidence threshold maximizes precision, while pairing with a low confidence threshold heavily boosts recall. Therefore, despite low Sensitivity supporting balanced metrics like PR AUC and F1-scores, high Sensitivity values coupled with extreme confidence thresholds appear to be the most appropriate choice for users who strongly prioritize either precision or recall.

Our analyses also reveal how BirdNET performance varies depending on the minimum Confidence score threshold used. At the vocalization level, the best performance (i.e. the highest F1-score) was achieved at low confidence thresholds – around 0.1 for Sensitivity values of 0.5 and 1.0, and around 0.3 for a Sensitivity of 1.5. In contrast, at the dataset level, optimal performance was consistently achieved with minimum confidence thresholds around 0.5, regardless of the Sensitivity setting. This elevated performance at the dataset level probably stems from the greater number of opportunities for correctly predicting a species across the dataset duration (i.e. only a single correct prediction is required for the species to be classified as a true positive), such that raising the minimum confidence threshold at the dataset level – up to a certain point – improves precision more than it decreases recall.

The results of our study must be interpreted in light of the following four primary limitations: (1) the limited amount of data available for certain regions (Fig. 1), (2) the different recording equipment used (Pérez-Granados 2025, see recorder type of each site in Table S1), (3) the assumption that the expert human annotations – used as the benchmark to compare BirdNET against – are always correct (see Campbell & Francis 2011) and (4) the classification of a species as correctly identified when one prediction was correct, regardless of the number of incorrect predictions for that species within the dataset. Although our datasets were annotated by local experts following a strict protocol (Pérez-Granados *et al.* 2025), differences in the annotation effort among recording locations are still possible, potentially biasing results. Further research should aim to develop reference annotation catalogues in which acoustic samples are annotated in agreement by at least two expert observers – to reduce biases – and, whenever possible, to collect a similar number of samples at each site to avoid positive biases toward recording locations or regions where longer acoustic samples are used. Furthermore, we decided to give equal importance to recall and precision to evaluate BirdNET performance; however, future research could explore the impact of variable settings on BirdNET output depending on whether higher recall or precision is prioritized.

Our results provide practical guidance for future studies that employ BirdNET for the automated identification of bird vocalizations and the detection of species presence in audio recordings. The broad spatial scope of our study, combined with consistent performance trends across different setting values, suggests that our findings,

particularly the performance benefits of high Overlap values, can serve as a reliable starting point for BirdNET usage in other regions. Nonetheless, it would be advisable to assess the impact of BirdNET settings before applying them in regions that were underrepresented in our acoustic dataset, such as Africa, Asia and Oceania. It is also worth noting that BirdNET performance improves with the development of updated versions (Funosas *et al.* 2024). Therefore, the annotated acoustic dataset used in this study, which is freely available, may serve as a valuable benchmark for evaluating the comparative performance of future versions of BirdNET, as well as for comparative studies between BirdNET and other machine learning tools (e.g. Morfi *et al.* 2019, Ghani *et al.* 2023).

We are grateful to the CTFC IT team for their help and support during the analyses, especially Daniel Macedo and Albert Sanahuja for their assistance.

AUTHOR CONTRIBUTIONS

Cristian Pérez-Granados: Data curation; supervision; resources; project administration; software; formal analysis; methodology; validation; visualization; writing – review and editing; writing – original draft; funding acquisition; investigation; conceptualization. **David Funosas:** Conceptualization; investigation; writing – review and editing; visualization; validation; methodology; software; formal analysis; supervision; data curation; resources. **Jon Morant:** Data curation; formal analysis; writing – review and editing; visualization; validation; methodology; investigation. **Oscar H. Marín Gómez:** Writing – review and editing; validation; investigation. **Irene Mendoza:** Writing – review and editing; validation; investigation. **Miguel A. Mohedano-Munoz:** Investigation; validation; writing – review and editing. **Eduardo Santamaría:** Writing – review and editing; validation; investigation. **Giulia Bastianelli:** Writing – review and editing; validation; investigation. **Alba Márquez-Rodríguez:** Writing – review and editing; validation; investigation. **Michał Budka:** Writing – review and editing; validation; investigation. **Gerard Bota:** Writing – review and editing; validation; investigation. **José M. De la Peña-Rubio:** Writing – review and editing; validation; investigation. **Eladio García De La Morena:** Writing – review and editing; validation; investigation. **Manu Santa-Cruz:** Investigation; validation; writing – review and editing. **Pablo De la Nava:** Writing – review and editing; validation; investigation. **Mario Fernández-Tizón:** Writing – review and editing; validation; investigation. **Hugo Sánchez-Mateos:** Writing – review and editing; validation; investigation. **Adrián Barrero:** Writing – review and editing; validation; investigation. **Juan Traba:** Writing – review and editing; validation; investigation. **Tomasz S. Osiejuk:** Writing – review and editing; validation; investigation. **Patrick J. Hart:** Writing – review and editing; validation; investigation. **Amanda K. Navine:** Writing – review and editing; validation; investigation. **Andrés F. Montoya Muñoz:** Writing – review and editing; validation; investigation. **Carlos B. De Araújo:** Investigation; validation; writing – review and editing. **Gabriel L. M. Rosa:** Writing – review and editing; validation; investigation. **Ingrid M. D. Torres:** Writing – review and editing; validation; investigation. **Ana L. C. Catalano:** Writing – review and editing; validation; investigation. **Cassio Rachid Simões:** Investigation; validation; writing – review and editing. **Diego Llusia:** Writing – review and editing; validation; investigation. **Manuel B. Morales:** Writing – review and editing; validation; investigation. **Pablo Acebes:** Writing – review and editing; validation; investigation. **Juan A. Medina:** Writing – review and editing; validation; investigation. **Nicholas Brown:** Writing – review and editing; validation; investigation. **Christos Astaras:** Writing – review and editing; validation; investigation. **Ilias Karmiris:** Writing – review and editing; validation; investigation. **Elizabeth Navarrete:** Writing – review and editing; validation; investigation. **Maxime Cauchoix:** Writing – review and editing; validation; investigation. **Luc Barbaro:** Writing – review and editing; validation; investigation. **Dominik Arend:** Writing – review and editing; validation; investigation. **Sandra Müeller:** Writing – review and editing; validation; investigation. **Fernando González-García:** Writing – review and editing; validation; investigation. **Alberto González-Romero:** Writing – review and editing; validation; investigation. **Christos Mammides:** Writing – review and editing; validation; investigation. **Michaelangelo Pontikis:** Writing – review and editing; validation; investigation. **Giordano Jacuzzi:** Writing – review and editing; validation; investigation. **Julian D. Olden:** Writing – review and editing; validation; investigation. **Sara P. Bombaci:** Writing – review and editing; validation; investigation. **Gabriel Marcacci:** Writing – review and editing; validation; investigation. **Alain Jacot:** Writing – review and editing; validation; investigation. **Juan P. Zurano:** Writing – review and editing; validation; investigation. **Elena Gangenova:** Writing – review and editing; validation; investigation. **Diego Varela:** Writing – review and editing; validation; investigation. **Facundo Di Sallo:** Writing – review and editing; validation; investigation. **Gustavo A. Zurita:** Writing – review and editing; validation; investigation. **Andrey Atemasov:** Writing – review and editing; validation; investigation. **Junior A. Tremblay:** Writing – review and editing; validation; investigation. **Anja Hutschenreiter:** Writing – review and editing; validation; investigation. **Alan Monroy-Ojeda:** Writing –

review and editing; validation; investigation. **Mauricio Díaz-Vallejo:** Writing – review and editing; validation; investigation. **Sergio Chaparro-Herrera:** Writing – review and editing; validation; investigation. **Robert A. Briers:** Writing – review and editing; validation; investigation. **Renata Sousa-Lima:** Writing – review and editing; validation; investigation. **Thiago Pinheiro:** Writing – review and editing; validation; investigation. **Wigna C. Da Silva:** Writing – review and editing; validation; investigation. **Alice Calvente:** Writing – review and editing; validation; investigation. **Anamaria Dal Molin:** Writing – review and editing; validation; investigation. **Alexandre Antonelli:** Writing – review and editing; validation; investigation. **Svetlana Gogoleva:** Writing – review and editing; validation; investigation. **Igor Palko:** Writing – review and editing; validation; investigation. **Hi u V. Trong:** Writing – review and editing; validation; investigation. **Marina H. L. Duarte:** Writing – review and editing; validation; investigation. **Natalia Dos Santos Saturnino:** Writing – review and editing; validation; investigation. **Samuel R. Silva:** Writing – review and editing; validation; investigation. **Ana Rainho:** Writing – review and editing; validation; investigation. **Paula Lopes:** Writing – review and editing; validation; investigation. **Karl-L. Schuchmann:** Writing – review and editing; validation; investigation. **Marinêz I. Marques:** Writing – review and editing; validation; investigation. **Ana S. De Oliverira Tissiani:** Writing – review and editing; validation; investigation. **Nick A. Littlewood:** Writing – review and editing; validation; investigation. **Mao-Ning Tuanmu:** Writing – review and editing; validation; investigation. **Yi-Ru Cheng:** Writing – review and editing; validation; investigation. **Hsuan Chao:** Writing – review and editing; validation; investigation. **Sebastian Kepfer-Rojas:** Writing – review and editing; validation; investigation. **Andrea L. Aguilera:** Writing – review and editing; validation; investigation. **Lluís Brotons:** Writing – review and editing; validation; investigation. **Mariano J. Feldman:** Investigation; validation; writing – review and editing. **Louis Imbeau:** Writing – review and editing; validation; investigation. **Pooja Panwar:** Writing – review and editing; validation; investigation. **Aaron S. Weed:** Writing – review and editing; validation; investigation. **Anant Dehwal:** Writing – review and editing; validation; investigation. **Alfredo Attisano:** Writing – review and editing; validation; investigation. **Jörn Theuerkauf:** Writing – review and editing; validation; investigation. **Dorgival D. Oliveira-Júnior:** Investigation; validation; writing – review and editing. **Cicero S. Lima-Santos:** Writing – review and editing; validation; investigation. **Carlos Salustio-Gomes:** Investigation; validation; writing – review and editing. **Raiane V. Paz:** Investigation; validation; writing – review and editing. **Mauro Pichorim:** Writing – review and editing; investigation; validation. **Eben Goodale:** Investigation; validation; writing – review and editing. **Esther Sebastián-González:** Funding acquisition; investigation; conceptualization; methodology; validation; visualization; writing – review and editing; project administration; formal analysis; software; data curation; supervision; resources.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

FUNDING

The authors have nothing to report.

ETHICAL NOTE

None.

DATA AVAILABILITY STATEMENT

Audio recordings (WAV format) and audio annotations used in this study are freely available in Zenodo:https://zenodo.org/records/17293588.

REFERENCES

Campbell, M. & Francis, C.M. 2011. Using stereo-microphones to evaluate observer variation in North American breeding bird survey point counts. *The Auk* **128**: 303–312.

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.

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*: 233–240. New York, NY, USA: Association for Computing Machinery.

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.

Ghani, B., Denton, T., Kahl, S. & Klinck, H. 2023. Global birdsong embeddings enable superior transfer learning for bioacoustics classification. *Sci. Rep.* **13**: 22876.

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. 2024. Novel community data in ecology-properties and prospects. *Trends Ecol. Evol.* **39**: 280–293.

Kahl, S., Wood, C.M., Eibl, M. & Klinck, H. 2021. BirdNET: A deep learning solution for avian diversity monitoring. *Eco. Inform.* **61**: 101236.

Knight, E.C., Hannah, K.C., Foley, G., Scott, C., Mark Brigham, R. & Bayne, E. 2017. Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs. *Avian Conserv. Ecol.* **12**: 14.

Morfi, V., Bas, Y., Pamuła, H., Glotin, H. & Stowell, D. 2019. NIPS4Bplus: A richly annotated birdsong audio dataset. *PeerJ Comput. Sci.* **5**: e223.

Pérez-Granados, C. 2023. BirdNET: Applications, performance, pitfalls and future opportunities. *Ibis* **165**: 1068–1075.

Pérez-Granados, C. 2025. BirdNET's confidence scores decrease with bird distance to the recorder: Revisiting Pérez-Granados (2023). *Ardeola* **72**: 149–159.

Pérez-Granados, C. & Traba, J. 2021. Estimating bird density using passive acoustic monitoring: A review of methods and suggestions for further research. *Ibis* **163**: 765–783.

Pérez-Granados, C., Morant, J., Darras, K.F.A., MarínGómez, O.H., Mendoza, I., Mohedano-Muñoz, M.A., Santamaría, E., Bastianelli, G., Márquez-Rodríguez, A., Budka, M., Bota, G., de la Peña Rubio, J.M., Garcíade la Morena, E.L., Santa-Cruz, M., de la Nava, P., Fernández-Tizón, M., Sánchez-Mateos, H., Barrero, A., Traba, J., Osiejuk, T.S., Hart, P.J., Navine, A., MontoyaMuñoz, A.F., de Araujo, C.B., Medina Rosa, G.L., Denóbile Torres, I.M., CamargoCatalano, I.L., de Almeida Simões, C.R.M., Llusia, D., Morales, M.B., Acebes, P., Medina, J.A., Brown, N., Astaras, C., Karmiris, I., Navarrete, E., Cauchoix, M., Barbaro, B., Funosas, D., Arend, D., Müller, S., González-García, F., González-Romero, A., Mammides, C., Pontikis, M., Jacuzzi, G., Olden, J.D., Bombaci, S.P., Marcacci, G., Jacot, A., Zurano, J.P., Gangenova, E., Varela, D., Di Sallo, F., Zurita, G.A., Atemasov, A., Tremblay, J.A., Lamarre, V., Hutschenreiter, A., Monroy-Ojeda, A., Díaz-Vallejo, M., Chaparro-Herrera, S., Briers, R.A., Sousa-Lima, R.S., Pinheiro, T., Carla da Silva, W., Calvente, A., Dal Molin, A., Antonelli, A., Gogoleva, S., Palko, I., Tr ng, H.V., Duarte, M.H.L., Falcão Saturnino, N.d.S., Silva, S.R., Rainho, A., Lopes, P., Schuchmann, K.-L., Marques, M.I., de Oliveira, A.S., Littlewood, N.A., Tuanmu, M.-N., Cheng, Y.-R., Chao, C., Kepfer-Rojas, S., Aguilera, A.L., Brotons, L., Feldman, M.J., Imbeau, L., Panwar, P., Weed, A.S., Dehwal, A. & Sebastián-González, E. 2025. WABAD: A World Annotated Bird Acoustic Dataset for passive acoustic monitoring. *Preprint at Research Square* https://doi.org/10.21203/rs.3.rs-5729784/v1

R Core Team 2025. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Stowell, D. 2022. Computational bioacoustics with deep learning: A review and roadmap. *PeerJ* **10**: e13152.

Wood, C.M. & Kahl, S. 2024. Guidelines for appropriate use of BirdNET scores and other detector outputs. *J. Ornithol.* **165**: 777–782.

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.

Xie, J., Zhong, Y., Zhang, J., Liu, S., Ding, C. & Triantafyllopoulos, A. 2023. A review of automatic recognition technology for bird vocalisations in the deep learning era. *Eco. Inform.* **73**: 101927.

Received 15 May 2025;
Revision 23 July 2025;
revision accepted 10 November 2025.
Associate Editor: Stuart Sharp

SUPPORTING INFORMATION

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

Figure S1. Definitions of true positives, false positives, true negatives and false negatives are used to evaluate BirdNET performance at the (A) vocalization and (B) dataset levels.

Figure S2. Mean precision, along with the standard error bar, of BirdNET vocalizations predicted for different ranges of confidence scores and each of the Sensitivity values tested, having used an Overlap of 2 s.

Figure S3. Average number of BirdNET predictions made per minute for different ranges of confidence scores and for each of the Sensitivity values tested, having used an Overlap of 2 s.

Figure S4. BirdNET-analyser Precision–Recall (PR) and receiver operating characteristic (ROC) curves at the dataset level for Overlap = 0 and different Sensitivity values. The three panels on the left (a, c, e) display PR curves, while the three panels on the right (b, d, f) display ROC curves. The panels are organized by DS settings, the top panels (a, b) corresponding to DS = 0.5, the middle panels (c, d) to DS = 1.0, and the bottom panels (e, f) to DS = 1.5. Original area under the curve (AUC) scores (orig_AUC) and AUC scores divided by the recall range in the PR curve and by the false-positive rate range in the ROC curve (adj_AUC) are shown on top of each curve.

Table S1. Description of the 67 recording locations where the recordings were done, including the recording site ID, study area name, region, main biome, latitude, longitude (geographical coordinates in decimal degrees), recorder time (including microphone brand for handheld recorders) and number of recording minutes included in the dataset.

Table S2. Biome-specific optimal Overlap and Sensitivity settings for BirdNET-Analyser to maximize area under the curve (AUC) scores for the Precision–Recall (PR) curve. Nine combinations of settings – three levels of Overlap between consecutive predictions (0, 1 and 2 s) and three Sensitivity values (0.5, 1 and 1.5) – were evaluated using a minimum Confidence score threshold

of 0.1. To measure model improvement, we report the variation (Δ) in AUC scores for both PR and receiver operating characteristic (ROC) curves between the best-performing settings for PR AUC optimization and the default settings (Overlap = 0, Sensitivity = 1). Results are presented separately for vocalization-level and dataset-level analyses. The number and percentage of datasets converging on the same optimal settings are presented for each biome.

Table S3. Dataset-specific optimal Overlap and Sensitivity settings for BirdNET-Analyser to maximize area under the curve (AUC) scores for the Precision–Recall (PR) curve. Nine combinations of settings – three levels of Overlap between consecutive predictions (0, 1 and 2 s) and three Sensitivity values (0.5, 1 and 1.5) – were evaluated using a minimum Confidence score threshold of 0.1. To measure model improvement, we report the variation (Δ) in AUC scores for both PR and receiver operating characteristic (ROC) curves between the best-performing settings for PR AUC optimization and the default settings (Overlap = 0, Sensitivity = 1). Results are presented separately for vocalization-level and dataset-level analyses.