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

Cristian Pérez-Granados

cristian.perez@ctfc.cat

CTFC https://orcid.org/0000-0003-3247-4182

David Funosas

Station d'Écologie Théorique et Expérimentale, SETE, CNRS https://orcid.org/0000-0001-9964-1120

Jon Morant
Oscar H. Marín
Irene Mendoza
Miguel A. Mohedano-Muñoz
Eduardo Santamaría
Giulia Bastianelli
Alba Márquez-Rodríguez
Michał Budka
Gerard Bota
José M. De la Peña-Rubio
Eladio García de la Morena
Manuel Snata-Cruz
Pablo de la Nava
Mario Fernández-Tizón
Hugo Sánchez.Mateos
Adrián Barrero
Juan Traba
Tomasz S. Osiejuk
Patrick J Hart
Amanda K. Navine
Andrés F. Montoya Muñoz
Carlos B. de Araujo
Gabriel L. M. Rosa
Ingrid M. Torres
Ana L. Catalano
Cassio de Alameida Simões
Diego Llusia
Manuel B. Morales

Pablo Acebes
Juan A. Medina
Nicholas Brown
Christos Astaras
Ilias Kamiris
Eliza Navarrete
Maxime Cauchoix
Luc Barbaro
Dominik Arend
Sandra Müeller
Fernando González-García
Alberto González-Romero
Christos Mammides
Michaelangelo Pontikis
Giordano Jacuzzi
Julian D. Olden
Sara P. Bombaci
Gabriel Marcacci
Alain Jacot
Juan P. Zurano
Elena Gangenova
Diego Varela
Facundo di Sallo
Gustavo A. Zurita
Andrey Atemasov
Junior A. Tremblay
Anja Jutschrenteiter
Alan Monroy-Ojeda
Mauricio Díaz-Vallejo
Sergio Chaparro-Herrera
Robert A. Briers
Renata Sousa-Lima
Thiago Pinheiro
Wigna C. da Silva
Alice Calvente
Anamaria del Molin
Alexandre Antonelli
Svetlana Gogoleva
Igo Palko
Hiếu V. Trong
Marina H. L. Duarte
Natalia dos Santos Saturnino

Samuel R. Silva
Ana Rainho
Karl -L. Schuchmann
Marinez I. Marques
Ana S. de Oliveira Tissiani
Nick A. Littlewood
Mao-Ning Tuanmu
Yi-Ru Cheng
Hsuan Chao
Sebastian Kepfer-Rojas
Andrea L. Aguilera
Lluís Brotons
Mariano L. Feldman
Louis Imbeau
Pooja Panwar
Aaron S. Weed
Anant Dehwal
Alfredo Attisano
Jörn Theuerkauf
Dorgival D. Oliveira-Junior
Cicero S. Lima-Santos
Carlos Salustio-Gomes
Raiane C. da Paz
Mauro Pichorim
Eben Goodale
Esther Sebsatián-González
Alicante University https://orcid.org/0000-0001-7229-1845

Research Article

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

Posted Date: May 15th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-6633549/v1

Additional Declarations: The authors declare no competing interests.

Abstract

BirdNET is a popular machine learning tool for automated recognition of bird sounds. Here we evaluate how BirdNET settings affect the model performance both at vocalization and species levels, using 4,225 one-minute 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 vocalisations, whereas higher thresholds (around 0.5) are more suitable for characterising bird communities.

Based on our findings, we recommend increasing the *Overlap* parameter from its default value of 0 seconds to 2 seconds, as this consistently improves BirdNET performance in detecting both individual vocalisations and species presence.

The effect of the *Sensitivity* parameter varied across regions. However, a value of 0.5 maximises global performance for community-level analyses across all confidence thresholds, while a value of 1.5 generally yields better results for vocalisation-level studies, particularly at low confidence thresholds.

INTRODUCTION

Passive acoustic monitoring (PAM) is a non-invasive, automated method extensively used for bird monitoring (e.g., Darras et al., 2019; Pérez-Granados and Traba, 2021). A key advancement in this field has been the development of machine learning (ML) and deep learning algorithms for the automated identification of bird vocalisations (Stowell, 2022; Xie et al., 2023), with BirdNET being amongst the most widely used software (Kahl et al., 2021; Pérez-Granados, 2023). BirdNET is based on a convolutional neural network, capable of identifying over 6,500 bird species worldwide (Kahl et al. 2021). BirdNET divides recordings into 3-second 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 minimises 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 and 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-second segments, and 2) *Sensitivity* (range: 0.5–1.5), which adjusts how sharply BirdNET separates confidence scores for each species: values < 1 of *Sensitivity* increase the model's certainty in its top predictions, while values > 1 make confidence scores more uniform across predictions. In summary, low *Confidence score* thresholds combined with high *Overlap* and *Sensitivity* values maximise the recall rate, i.e., the proportion of vocalisations detected among those in a recording, to the detriment of precision, i.e., the proportion of vocalisations correctly identified by BirdNET. As a result, an inherent trade-off emerges between recall and precision (Funosas et al., 2024).

Although prior research has explored the impact of adjusting the input values of BirdNET parameters (e.g., 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 optimise 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 vocalisation (i.e., the best settings to correctly classify a vocalisation) and dataset levels (i.e., the best settings to correctly identify the species appearing in a collection of recordings)— to optimise the balance between BirdNET precision and recall. To achieve this, we analysed 4,225 one-minute audio recordings collected from 67 recording locations, with 89,061 fine-grained annotations of all bird vocalisations (strong labels) made by expert ornithologists worldwide. We hope our results will guide future studies in determining optimal parameter settings and support the continued refinement of BirdNET for both ecological monitoring of bird species and for the characterisation 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 vocalisation level by local experts. For consistency across datasets, our analyses only include sites with 1-minute audio recordings and strong labels (i.e., annotations with exact start and end times for each bird vocalisation present in the recording). In total, we analysed 4,225 1-minute recordings collected at 67 recording locations (Fig. 1; see Supplementary Table S1 for details on the biome, location, and specific coordinates of the surveyed sites). For more detailed information, including the particular recordings and annotations used in this study, refer to Pérez-Granados et al. (2025).

AUDIO ANNOTATIONS

Expert ornithologists familiar with the local avifauna examined each 1-minute audio recording spectrogram and identified every single bird vocalisation at the species level. All annotations followed the Clements Checklist (Clements et al., 2021), which guarantees taxonomic alignment with nomenclature used in BirdNET. The experts annotated each vocalisation 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 vocalisations from the same species could be included in the same box whether they were separated by less than one second; otherwise, a separate annotation was made. 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-Analyzer 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 for the *Overlap* and *Sensitivity* settings (i.e., a mixture between 0 s, 1 s, and 2 s *Overlap* and 0.5, 1, and 1.5 *Sensitivity*). We configured BirdNET to filter the list of potentially detectable species based on the following criteria: 1) recording site location (Supplementary Table S1), 2) recording date (week of the year), and 3) a minimum occurrence threshold of 0.02 (probability threshold of a species occurrence at a site and week of the year; see Funosas et al. 2024).

BIRDNET ASSESSMENT

We assessed BirdNET performance across the nine combinations of settings by comparing model predictions to the annotations made by experts through a series of custom R scripts (version 4.2.2; R Core Team 2025). The assessments were conducted at two levels: i) vocalisation level, which provides a fine-grained assessment of BirdNET performance for bird monitoring, and ii) dataset level, which offers insights into BirdNET ability to characterise the composition of bird communities in a collection of recordings from the same location. BirdNET predictions were categorised into four possible outcomes (Supplementary Fig. S1):

- **True Positives (TP)**: At the vocalisation level, a BirdNET prediction was classified as a TP when an expert labelled the same species at the same time. At the dataset level, a bird species was considered a TP whether there was at least one correct identification of that species by BirdNET in any of the audio recordings from the same study site (i.e. dataset).
- **False Positives (FP)**: At the vocalisation 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)**: A prediction was classified as a TN when a vocalisation or species not identified by the expert was also not predicted by BirdNET either at the dataset level or the vocalisation level.
- **False Negatives (FN)**: A prediction was classified as a FN when a vocalisation or species identified by the expert was not predicted by BirdNET .

Following the above categorisation criteria, we evaluated BirdNET precision, recall, and False Positive Rate (FPR) at both vocalisation and dataset levels. Precision is defined as the proportion of species or vocalisations correctly predicted relative to the total number of species or vocalisations predicted by BirdNET. The recall rate measures the proportion of species or vocalisations correctly predicted relative to the total number of species or vocalisations 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). We controlled for the possible double counting of the same bird vocalisation detected in two overlapping segments. Analyses at the vocalisation level compare BirdNET predictions within 3-second segments to expert annotations, while dataset-level assessments match expert-annotated species lists to BirdNET-predicted species, counting only correct matches. The specific formulas used to calculate precision, recall, and FPR are the following:

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

FPR = FP / (FP + TN)

The three metrics were calculated at each level of analysis: vocalisation and dataset. It is essential to note that, according to our categorisation 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 and Goadrich, 2006). The PR curve plots precision against recall for each minimum confidence threshold considered, illustrating these two metrics' trade-offs. Similarly, the ROC curve plots recall against the FPR for each minimum confidence threshold, revealing the trade-off between these two metrics. For both curves, the AUC serves as a measure of the algorithm's predictive power, with values ranging from 0 to 1; higher values indicate greater predictive power.

The AUC of Precision-Recall (PR AUC) 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 Receiver Operating Characteristic (ROC AUC) 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 than with lower *Sensitivity* values (Supplementary Fig. S3). 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 formulas:

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

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

Additionally, we estimated the F-score across confidence score thresholds, which evaluates an algorithm's predictive power by integrating both precision and recall, calculated using the following formula:

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

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

RESULTS

Optimising BirdNET parameters at vocalisation level

Both the *Overlap* and the *Sensitivity* values impacted BirdNET performance at the vocalisation 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. 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 appeared substantially less than that of *Overlap*. Our results also show that, at the vocalisation level, the PR AUC score is maximised with an *Overlap* of 2 and a *Sensitivity* of 0.5 (Table 1). However, the optimal *Sensitivity* value for maximising PR AUC scores varied across biogeographic 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).

Optimising BirdNET parameters at 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 geographic regions, with all regions reaching their highest PR AUC scores at a *Sensitivity* of 0.5 (Table 2). 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 set of settings maximising PR AUC was 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.

F-score curves: impact of confidence score threshold

The F-score curves showed that BirdNET performance remained relatively consistent across the three *Overlap* settings at both vocalisation and dataset levels (Figure 2). However, at the vocalisation 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 *Sensitivity* setting varied between the two levels of analysis. At the vocalisation level, when using *Sensitivity* values of 0.5 and 1, the F-score declined almost linearly as the minimum confidence thresholds increased. However, with a *Sensitivity* of 1.5, the F-score increased until reaching its maximum around a confidence score threshold of 0.3. Interestingly, the F-score curve with a *Sensitivity* of 1.5 showed better performance than the F-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 F-scores at the vocalisation 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 F-scores were consistently achieved with a confidence 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 F-scores were obtained with confidence score thresholds around 0.5. The largest differences in F-scores appeared between *Sensitivity* values at the lowest and highest minimum confidence score thresholds, particularly at the higher end.

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 (e.g., 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 vocalisations or detecting species presence in acoustic datasets. The large variability observed in BirdNET outputs across different parameter configurations highlights the need for optimised standardised parameter guidelines. Such standards would improve cross-study comparisons, ensure temporal and spatial reproducibility, and enhance 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 both the vocalisation and dataset levels. This improvement is likely due to the increased ability for BirdNET to detect short, split or faint bird vocalisations when higher degrees of overlap are used —vocalisations that might otherwise be missed between non-overlapping segments (Funosas et al. 2024). While the benefits were most evident at the vocalisation 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. 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 —both at the vocalisation and dataset levels— varied more across *Sensitivity* values than across *Overlap* values, particularly at the vocalisation 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 (Supplementary Fig. S2). However, it remains unclear why, in half of the regions, the best performance at the vocalisation level was obtained with a value of 0.5, while 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 maximising PR AUC scores at the dataset level, primarily due to very low precision at low confidence thresholds and very low recall at high confidence thresholds. In contrast, low *Sensitivity* values yield a more balanced trade-off between precision and recall across all confidence thresholds. This results in similar F-score values and sensitivity levels across *Sensitivity* values for minimum confidence thresholds ranging from 0.35 to 0.6, and with comparatively poor results for high *Sensitivity* values when either lowering or increasing the minimum confidence thresholds beyond this range. However, because higher *Sensitivity* values strengthen the correlation between confidence scores and precision (Supplementary Fig. S4), they enable targeted optimisation: combining high *Sensitivity* with a high confidence threshold maximises precision, while pairing with a low confidence threshold boosts recall. Therefore, despite their low PR AUC scores, high *Sensitivity* values appear to be the most appropriate choice for users who strongly prioritise either precision or recall.

Our analyses also reveal how BirdNET performance varies depending on the minimum confidence score threshold used. At the vocalisation level, the best performance (i.e., the highest F-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 likely 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 three primary limitations: i) the limited amount of data available for certain regions (Fig. 1); ii) the assumption that the expert human annotations —used as the benchmark to compare BirdNET against— are always correct (see Campbell

and Francis 2011); and iii) the consideration 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 sites may still occur, potentially biasing the results for certain locations. 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 sites or regions where longer acoustic samples are used. Furthermore, we gave 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 prioritised.

Our results provide practical guidance for future studies that employ BirdNET for the automated identification of bird vocalisations 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 the optimal BirdNET configurations, particularly when setting higher *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 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 serving as a basis for comparative studies between BirdNET and other machine learning tools (e.g. Ghani et al. 2023, Morfi et al. 2019).

Declarations

Open Research statement:

All recordings and annotations used in the study are available in Zenodo at https://doi.org/10.5281/zenodo.14191524.

ACKNOWLEDGEMENTS

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.

References

1. Campbell, M., & Francis, C. M. (2011). Using stereo-microphones to evaluate observer variation in North American Breeding Bird Survey point counts. The Auk, 128(2), 303-312.

2. 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.
3. 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. *Ecological Applications*, *29*(6), e01954.
4. Davis, J., Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves, in: Proceedings of the 23rd International Conference on Machine Learning, ICML '06. Association for Computing Machinery, New York, NY, USA, pp. 233–240.
5. 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. *Ecological Indicators*, *164*, 112146.
6. Ghani, B., Denton, T., Kahl, S., & Klinck, H. (2023). Global birdsong embeddings enable superior transfer learning for bioacoustics classification. Scientific Reports, 13(1), 22876.
7. Hartig, F., Abrego, N., Bush, A., Chase, J. M., Guillera-Arroita, G., Leibold, M. A., ... & Yu, D. W. (2024). Novel community data in ecology-properties and prospects. Trends in Ecology & Evolution, 39(3), 280-293.
8. Kahl, S., Wood, C. M., Eibl, M., & Klinck, H. (2021). BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, *61*, 101236.
9. Morfi, V., Bas, Y., Pamuła, H., Glotin, H., & Stowell, D. (2019). NIPS4Bplus: a richly annotated birdsong audio dataset. PeerJ Computer Science, 5, e223.
10. Pérez-Granados, C. (2023). BirdNET: applications, performance, pitfalls and future opportunities. *Ibis*, *165*(3), 1068-1075.
11. Pérez-Granados, C. (2025). BirdNET's confidence scores decrease with bird distance to the recorder: revisiting Pérez-Granados (2023). Ardeola, 72(2): 149-159.
12. 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.
13. Pérez-Granados, C., Morant, J., Darras, K., et al. (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]
14. R Core Team (2025). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
15. Stowell, D. (2022). Computational bioacoustics with deep learning: a review and roadmap. *PeerJ*, *10*, e13152.
16. Wood, C. M., & Kahl, S. (2024). Guidelines for appropriate use of BirdNET scores and other detector outputs. *Journal of Ornithology*, 165: 777-782.

17. 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*(5), 532-543.
18. 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. *Ecological Informatics*, *73*, 101927.

Tables

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 vocalisation and dataset levels. The best results are highlighted in bold.

		AUC VALUES			
Overlap	*Sensitivity*	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 2: Continent-specific optimal *Overlap* and *Sensitivity* settings for BirdNET-analyser to maximise Area Under the Curve (AUC) scores for the Precision-Recall (PR) curve. Nine combinations of settings — 3 levels of *Overlap* between consecutive predictions (0 s, 1 s and 2 s) and 3 *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 optimisation and the default settings (*Overlap* = 0, *Sensitivity* = 1). Results are presented separately for vocalisation-level and dataset-level analyses.

Analysis level	Region	Overlap (s)	Sensitivity	Δ PR_AUC	Δ ROC_AUC
Vocalisation	Africa	2	0.5	0.020	0.029
	Asia	2	1.5	0.026	0.005
	Central-South America	2	0.5	0.057	0.029
	Europe	2	1.5	0.084	-0.053
	North America	2	0.5	0.122	0.009
	Oceania	2	1.5	0.139	-0.038
Dataset	Africa	2	0.5	0.111	-0.005
	Asia	2	0.5	0.105	-0.016
	Central-South America	2	0.5	0.124	-0.004
	Europe	1	0.5	0.140	-0.024
	North America	2	0.5	0.118	0.007
	Oceania	0	0.5	0.211	-0.023

Supplementary Material

Supplementary Table S1 not available with this version

Figures

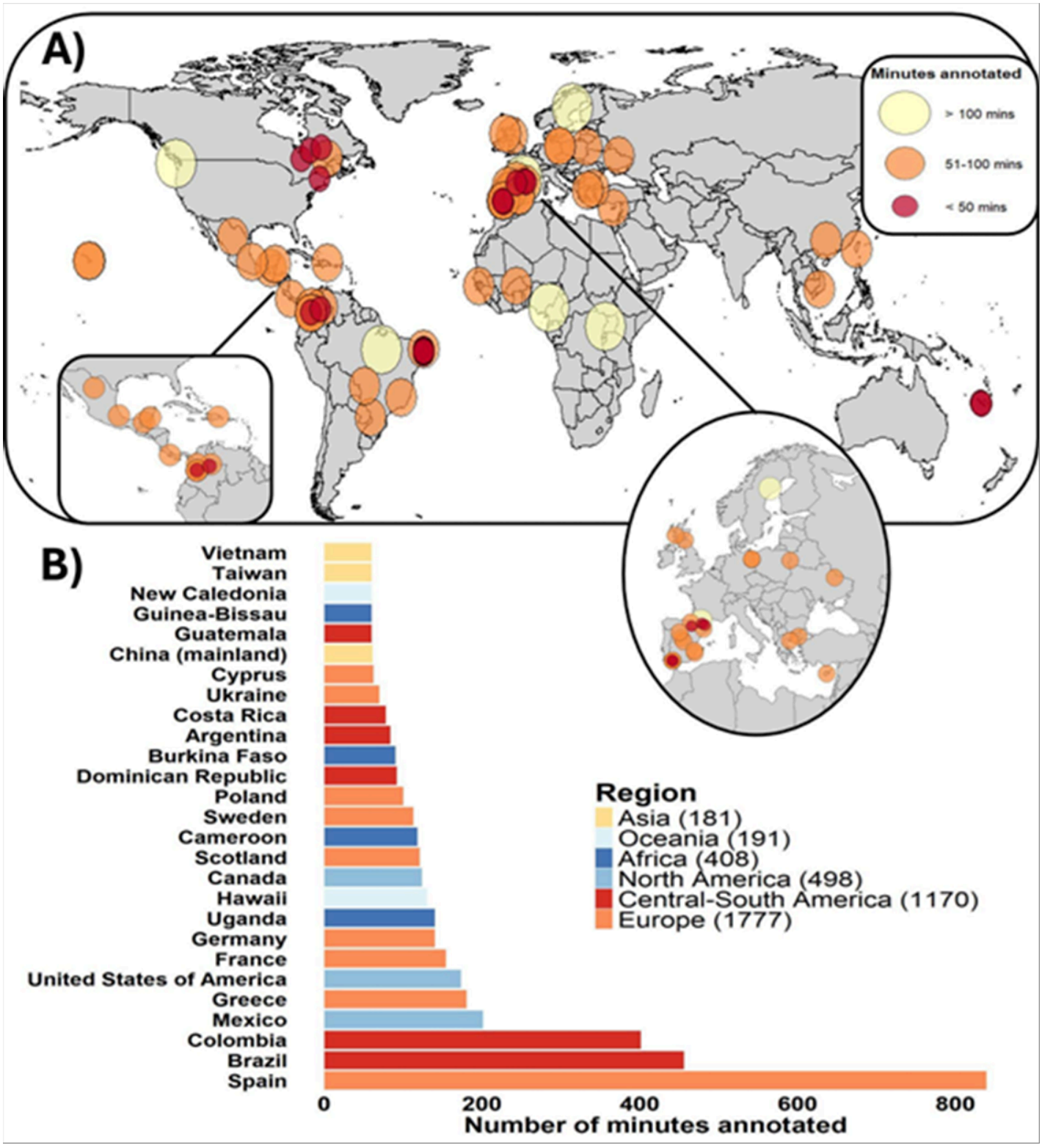

Figure 1

(A) Global mapping of 67 recording sites considered in the study. Colours and size of circles refer to the number of minutes annotated per recording site. The small circles show the location of recording sites 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 brackets. * Although Hawai'i is part of the United States, we classified it separately within the Oceania region based on biogeographical criteria.

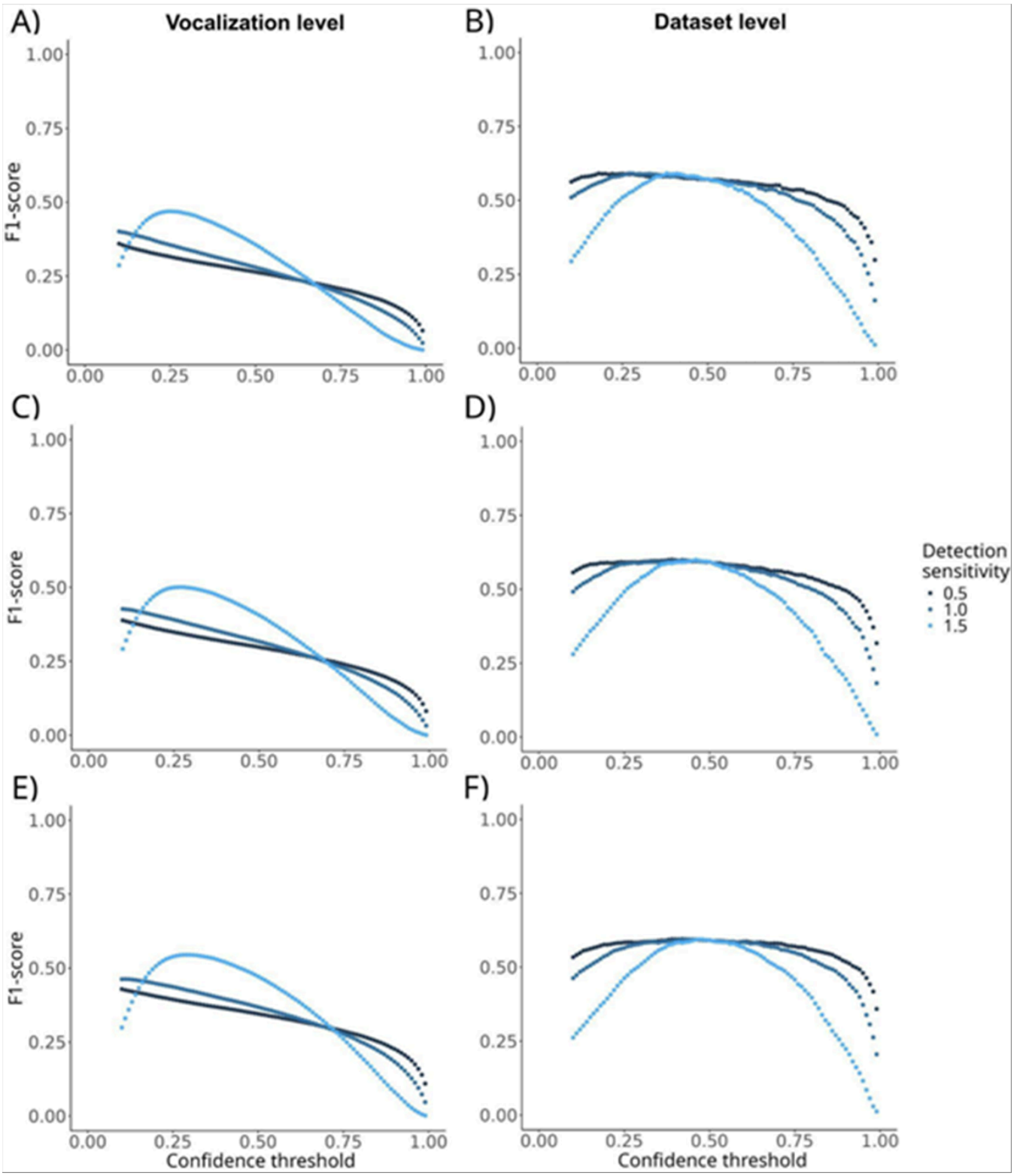

Figure 2

BirdNET-analyser F1-score curves for nine combinations of settings. The three panels on the left (A, C, and E) present results at the vocalisation level, while the three panels on the right (B, D, and F) show results at the dataset level. The panels are organised by *Overlap* settings: the top panels (A and B)

correspond to *Overlap* = 0, the middle panels (C and D) to *Overlap* = 1, and the bottom panels (E and F) to *Overlap* = 2. Within each panel, the three different S*ensitivity* values (0.5, 1, and 1.5) are represented by three distinct colours.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial.docx