Remote Sensing in Ecology and Conservation
Open Access
ZSL

RESEARCH ARTICLE

Aggregated time-series features boost species-specific differentiation of true and false positives in passive acoustic monitoring of bird assemblages

David Singer[1,2], Jonas Hagge[1,2], Johannes Kamp[3], Hermann Hondong[3] & Andreas Schuldt[1]

[1]Department of Forest Nature Conservation, University of Göttingen, Göttingen, Germany
[2]Department of Forest Nature Conservation, Northwest German Forest Research Institute, Hann. Münden, Germany
[3]Department of Conservation Biology, University of Göttingen, Göttingen, Germany

Keywords
AudioMoth, bird community monitoring, BirdNET performance, community bioacoustics, science-practice gap, threshold selection

Correspondence
David Singer, Department of Forest Nature Conservation, University of Göttingen, Büsgenweg 3, Göttingen 37077, Germany.
Tel: +49 551 69401245;
E-mail: d.singer@posteo.de

Editor: Vincent Lecours
Associate Editor: Jorge Ahumada

Received: 13 October 2023; Revised: 14 January 2024; Accepted: 9 February 2024

doi: 10.1002/rse2.385

Abstract

Passive acoustic monitoring (PAM) has gained increasing popularity to study behaviour, habitat preferences, distribution and community assembly of birds and other animals. Automated species classification algorithms like 'BirdNET' are capable of detecting and classifying avian vocalizations within extensive audio data, covering entire species assemblages. PAM reveals substantial potential for biodiversity monitoring that informs evidence-based conservation. Nevertheless, fully realizing this potential remains challenging, especially due to the issue of false-positive species detections. Here, we introduce an optimized thresholding framework, which incorporates contextual information extracted from the time-series of automated species detections (i.e. covariates on quality and quantity of species' detections measured at varying time intervals) to improve the differentiation of true and false positives. We verified a sample of BirdNET detections per species and modelled species-specific thresholds using conditional inference trees. These thresholds were designed to minimize false-positive detections while maximizing the preservation of true positives in the dataset. We tested this framework for a large dataset of BirdNET detections (5760 h of audio data, 60 sites) recorded over an entire breeding season. Our results revealed considerable interspecific variability of precision (percentage of true positives) within raw BirdNET data. Our optimized thresholding approach achieved high precision (≥0.9) for 70% of the 61 detected species, while species-specific thresholds solely relying on the BirdNET confidence scores achieved high precision for only 31% of the species. Conservative universal thresholds (not species-specific) reached high precision for 48% of the species. Our thresholding approach outperformed previous thresholding approaches and enhanced interspecific comparability for bird community analyses. By incorporating contextual information from the time-series of species detections, the differentiation of true and false positives was substantially improved. Our approach may enhance a straightforward application of PAM in biodiversity research, landscape planning and evidence-based conservation.

Introduction

Over the past decades, passive acoustic monitoring (PAM) has developed rapidly (e.g. Darras et al., 2019; Gibb et al., 2019; Sugai et al., 2019), especially driven by the availability of low-cost, energy efficient autonomous recording units (Hill et al., 2018). The ability to sample biodiversity at high temporal resolution is an outstanding feature of PAM, which results in extensive time-series data of detections of species vocalizations (Ross et al., 2023). Hence, compared to traditional observer-based methods, PAM enhances species detectability

(Darras et al., 2019; Metcalf, Barlow, Marsden, et al., 2022; Pérez-Granados et al., 2018) and thus reduces methodological issues regarding interspecific variation in detectability (Boulinier et al., 1998; Kéry & Schmidt, 2008) or observer biases (Kułaga & Budka, 2019; Schmidt et al., 2023). PAM is particularly valuable to detect rare and secretive species (Bota et al., 2023; Picciulin et al., 2019). Although not without limitations and imperfections, for example its restriction to vocalizing species (Darras et al., 2019), PAM can thus make a substantial contribution to biodiversity monitoring (Chhaya et al., 2021; Ross et al., 2023), which is needed to identify and understand threats for biodiversity and to provide evidence for conservation management (Lindenmayer et al., 2022; Sutherland et al., 2004).

Compared to cetaceans or bats, PAM of birds is a young field with specific challenges mainly due to noisy acoustic environments and the high complexity and variation of species' vocalizations (Kahl et al., 2021; Priyadarshani et al., 2018; Ross et al., 2023). Ecoacoustic indices correlate with species richness but lack information on species identities (Gasc et al., 2017). Therefore, numerous researchers have identified species from sound recordings by human listening and demonstrated advantages of PAM (Darras et al., 2018, 2019). Nevertheless, human-based species identification is limiting large-scale PAM applications (Gibb et al., 2019). Thus, various tools for automated detection of focal species have been developed (Florentin et al., 2020; Katz et al., 2016; Zwart et al., 2014). The deep artificial neural network 'BirdNET' is one of the first available algorithms covering entire bird communities. In its first release, BirdNET was able to detect vocalizations of 984 bird species from Europe and North America (Kahl et al., 2021). The algorithm was recently extended to cover 6000 bird species worldwide (https://github.com/kahst/BirdNET-Analyzer).

Nevertheless, false-positive species detections (i.e. misclassified sounds) are an inherent problem of classification algorithms like BirdNET (Clement et al., 2022; Rhinehart et al., 2022). False-positive detections are particularly problematic, since occupancy of threatened species may get overestimated and negative population trends may remain undetected (Rydell et al., 2017). As a result, decision-making in conservation management may be misguided, which demonstrates the need for the development of robust data post-processing methods in PAM to utilize its benefits for biodiversity research. Several studies provided valuable guidance on designing and optimizing PAM field studies (Froidevaux et al., 2014; Metcalf, Barlow, Marsden, et al., 2022; Sugai et al., 2020). However, a standardized workflow for PAM data post-processing has not been established yet and false-positive species detections hinder ecological analyses of PAM data (Barré et al., 2019; Clement et al., 2022; Rhinehart et al., 2022). To realize the full potential of PAM in large-scale and long-term applications, a minimization of false-positive detections of all species, not just selected or common ones, is required. Several approaches have been proposed to address false-positive detections, for example logistic regression (Barré et al., 2019; Bota et al., 2023), boosted regression trees (Knight et al., 2020), occupancy models (Clement et al., 2022; Rhinehart et al., 2022; Wright et al., 2020) and hierarchical modelling (Chambert et al., 2018; Cole et al., 2022). However, these approaches have only been demonstrated for a selected set of (common) species but not for entire species assemblages. Mitigation of false positives in large-scale bird monitoring applications comes with unique challenges due to extensive sources of noise and varying species assemblages (Lauha et al., 2022; Priyadarshani et al., 2018). Interspecific differences of false-positive rates have been identified as a challenge but remain unevaluated at the community scale (Knight et al., 2020; Pérez-Granados, 2023). Typically, automated species classification algorithms provide a continuous confidence score as a measure of the quality of a detection. In BirdNET applications, some authors applied universal confidence score thresholds (a minimum confidence score below which data are discarded) across species to reduce false positive (Sethi et al., 2021; Wood et al., 2021). However, interspecific differences of false-positive rates can be substantial, and false-positive rates remain unknown without human validation (Barré et al., 2019; Cole et al., 2022; Metcalf, Barlow, Bas, et al., 2022).

In this study, we present an optimized thresholding framework for data post-processing of automated species detection data. Our primary objective was to identify species-specific thresholds that maximize the precision [i.e., the number of correctly classified detections divided by the total number of correctly and incorrectly classified detections (Knight et al., 2017)]. We anticipated that including features derived from the time-series of automated species detections, that is variables related to the quality (confidence score) and quantity of detections calculated at varying time intervals, would improve the differentiation between true and false positives, as biologically meaningful information on detection probability is included (Chambert et al., 2018; Madhusudhana et al., 2021; Metcalf, Barlow, Bas, et al., 2022). Furthermore, we expected that these aggregated time-series features (ATF) could especially contribute to an optimization of the trade-off between precision and recall (Knight & Bayne, 2019). We expected that threshold models incorporating ATF would differentiate true- and false-positive detections with higher accuracy than a threshold solely based on the BirdNET confidence score.

As a result, fewer true positives would need to be discarded while false positives would be minimized compared to previous thresholding approaches, hence optimizing both precision and recall. We tested this optimized thresholding approach on a large dataset from a realistic bird monitoring application and compared it to previous thresholding approaches. This not only allowed us to evaluate the effectiveness of our thresholding approach but also provided general insights into the performance of the BirdNET algorithm in extended PAM studies of entire bird species assemblages with low-cost autonomous recording units like AudioMoth.

Materials and Methods

(1) Automated audio recording followed by (2) automated species classification are two basic and well-established steps of the PAM workflow (Gibb et al., 2019, Fig. 1). However, (3) data post-processing is an essential third step preceding ecological analyses with PAM data (Barré et al., 2019; Knight & Bayne, 2019). Our species-specific thresholding approach specifies the data post-processing in the PAM workflow. This approach is composed of three phases: (a) human validation of a sample of detections per species, (b) modelling of candidate threshold models and

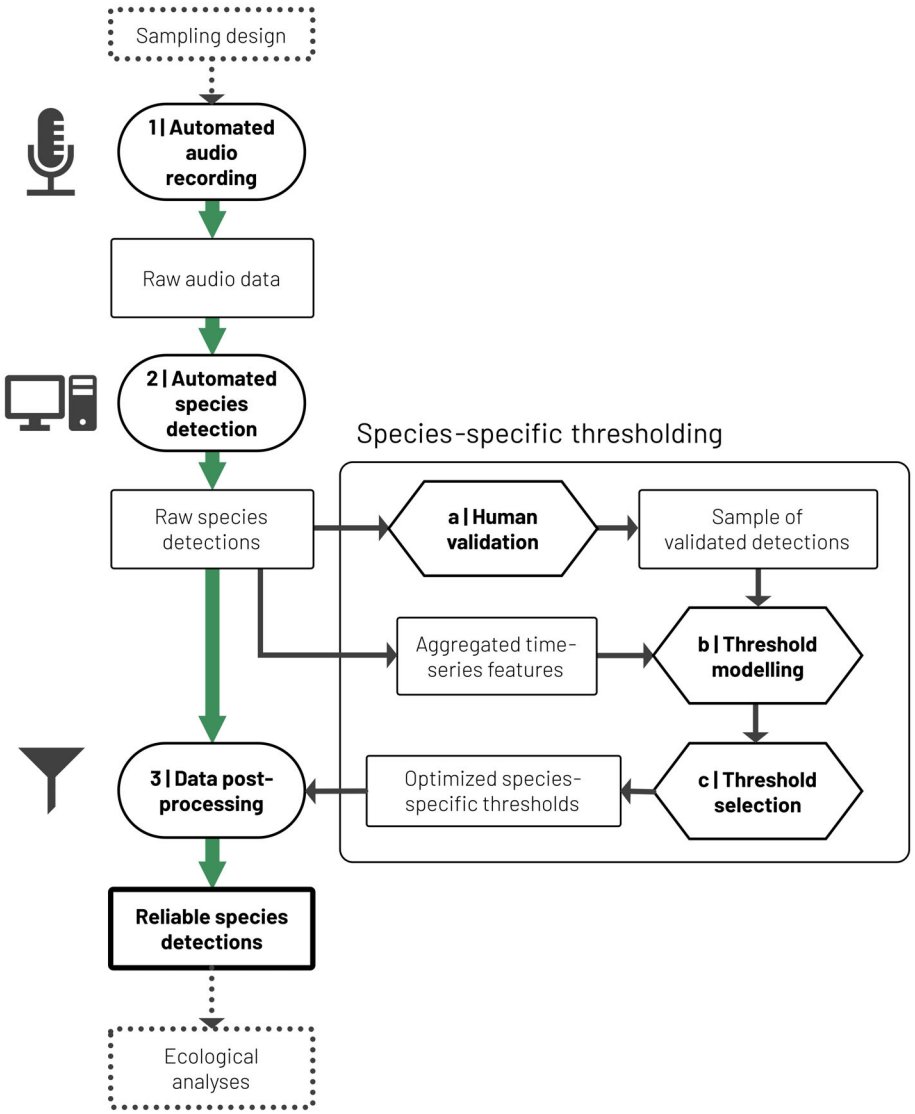

Figure 1. Workflow scheme of passive acoustic monitoring, including three basic steps (left: 1. Automated audio recording, 2. Automated species classification, 3. Data post-processing) and our species-specific thresholding approach, which consists of three phases (a. Human validation, b. Threshold modelling, c. Threshold selection). In the species-specific thresholding approach, a sample of the raw species detections is used to derive optimized thresholds for each species.

(c) final selection of optimized species-specific thresholds (Fig. 1). All analyses were conducted with R 4.2.2 (R Core Team, 2023), accessed via R Studio (Posit team, 2023).

Automated audio recording

The audio data used in our study were recorded at 60 beech forest sites in Hainich National Park in central Germany (51.09° N, 10.43° E). The forest of Hainich National Park is formed by a mixture of broadleaved tree species, dominated by European beech (*Fagus sylvatica*). We mounted one AudioMoth (versions 1.0.0 and 1.1.0, technically equivalent) in waterproof IPX7 cases at 1.5 m height to tree trunks at all sites. AudioMoth devices are capable of recording large amounts of audio data automatically (Hill et al., 2018). They were configured to record sound at a sampling rate of 32 kHz, with amplifier level (gain) set to medium. AudioMoth recordings covered the regional breeding season of forest birds from 18th March until 7th June 2021 (80 days). Audio data were recorded for 30 s at intervals of 10 min throughout day and night for the whole study period (5% of time), totalling 5760 h (96 h per site) of audio data. We chose this high-temporal-resolution sampling scheme because it records acoustic species assemblages more efficiently regarding completeness of species than sampling schemes with lower temporal resolution or limited to a certain daytime (Metcalf, Barlow, Marsden, et al., 2022; Wood et al., 2021).

Automated species classification

We employed the artificial neural network 'BirdNET' in its first version (Kahl et al., 2021) (available at https://github.com/kahst/BirdNET) to automatically detect and classify avian vocalizations in the audio recordings. We kept default settings (min_conf = 0.1, sensitivity = 1, spp = 1, overlap = 0) since effects of adjusting BirdNET parameters have not been evaluated yet (Pérez-Granados, 2023). BirdNET provided a list of species and corresponding confidence scores (continuous values from 0.1 to 0.99) for ten 3-s intervals for each 30 s audio file. We deactivated the use of eBird species distribution data (Sullivan et al., 2009) within the BirdNET processing to avoid a premature exclusion of species. However, before proceeding with subsequent analyses, we limited the species set to the breeding birds of the study region (Gedeon et al., 2015), resulting in a list of 110 species (Table S1).

Species-specific thresholding

Human validation

The first step of our species-specific thresholding approach (Fig. 1) was the validation of a sample of 225 BirdNET detections per species, adapted from Barré et al. (2019), Metcalf, Barlow, Bas, et al. (2022). For each species, 25 detections per 0.1 class of the confidence score were selected randomly from the entire pool of BirdNET detections. For some species, there were less than 25 available detections in certain classes, leading to reduced sample sizes. An experienced ornithologist (D.S.) validated the sample by listening to the 3-s audio snippets with studio headphones (AKG K701) and assigned whether they were true or false positives. Audio snippets were extracted using the *tuneR* package (Ligges et al., 2018). Detections that could not be identified at species level (e.g., unspecific calls or fragments of songs) were considered false positives. In cases where no true-positive detection was identified among the top 25 detections in the sample for a species, human validation was stopped. Only species with at least one validated true-positive detection were included in the subsequent analyses.

Threshold modelling

Secondly, we modelled species-specific thresholds to minimize false-positive detections (Fig. 1). Confidence scores are provided as a metric of the quality of a species detection, but initial validation checks of BirdNET detections indicated that there were considerable percentages of false positives also among detections with high confidence scores for several species (e.g. Marsh tit, Grey-headed woodpecker, Spotted flycatcher, Eurasian woodcock). Hence, we calculated ATF from the species detection time-series per site to be used as additional predictors. We basically assumed that a species detection that is embedded in a time interval with many high confidence score detections of that species at the sampling site is more likely to be a true-positive detection than a detection (with similar confidence score) from a time interval at the site with overall less and lower confidence score detections. We calculated simple statistical parameters that aggregate information on the quality (average, median, maximum and minimum of confidence scores) and the quantity (number of detections with different minimum confidence scores) at different time intervals from the time-series of BirdNET detections for each species per site (Table S2). The selection of the statistical parameters was guided by the intention to align them with interspecific differences in the distribution of true and false positives along the confidence score continuum (Rhinehart et al., 2022).

All ATF were calculated for 12 different time intervals centred on the detections' timestamp. We varied the time interval length at three temporal scales, meaning that we included information from temporally adjacent detections (±3, ±6, ±9 and ±12 s), files (±10, ±20, ±30 and ±40 min) and days (±12, ±24, ±36 and ±48 h). Extending the time interval is biologically meaningful in the way

that the predictors may better fit the varying song characteristics and temporal activity patterns among bird species than the original BirdNET confidence scores, which are based on 3 s (Kahl, 2020). By incorporating information from temporally adjacent detections, the predictors may capture different durations or repetition patterns of vocalizations (Benedict & Najar, 2019). Predictors integrating information from 20 to 80 min may capture vocalizations of species with distinct diel activity peaks [e.g. dawn or dusk chorus (Farina & Ceraulo, 2017)], while those calculated for time intervals of 24–96 h may capture distinct seasonal activity peaks, such as those of migratory species (Thompson et al., 2017). Together with the original BirdNET confidence score, we ended up with 169 predictor variables per species.

We employed conditional inference trees (Hothorn et al., 2006) to identify threshold values along the continuous scales of the predictor variables, which maximize the differentiation into true- and false-positive detections for each species (Fig. 2). Conditional inference trees partition a dataset into binary subsets recursively and select the most significant predictor variables based on statistical tests. They are robust to overfitting (Müller & Bütler, 2010). For methodological details on conditional inference trees, see Hothorn et al. (2006) and Müller and Bütler (2010). Conditional inference trees were fitted using the *ctree* function from the *partykit* package (Hothorn & Zeileis, 2015). Specific parameter settings are provided in the R code (03_threshold_modelling.R).

For each of the 61 investigated species, we fitted conditional inference trees of two different model types. Type 1 models were designed to assess the performance of the 169 predictor variables independently. Setting maximum tree depth in the *ctree* function to one allowed for either zero or one split based on the selected predictor variable (see, e.g., Fig. 2A). Consequently, we fitted a total of 169 type 1 models for each species. Type 2 models were designed to allow interactions between combinations of predictor variables. Hence, we used all 169 variables as single predictor variables as well as all possible combinations of them, respectively. To allow for up to two split levels, we set the maximum tree depth to two. This means that threshold rules can consist of two conditions, for example a minimum confidence score combined with a minimum number of detections per time interval (see, e.g., Fig. 2B). Accordingly, we fitted 14,365 type 2 models for each of the 61 species.

Threshold selection

Finally, we selected optimized species-specific threshold rules out of all available models (Fig. 3). Our primary objective was to minimize false-positive detections in the BirdNET data. Therefore, the precision defined as the ratio of true positives to the total of false and true positives (Knight et al., 2017) was the main selection criterion for an optimized threshold rule. We only considered threshold rules leading to the terminal node of a model that provided the maximum precision (Fig. 2).

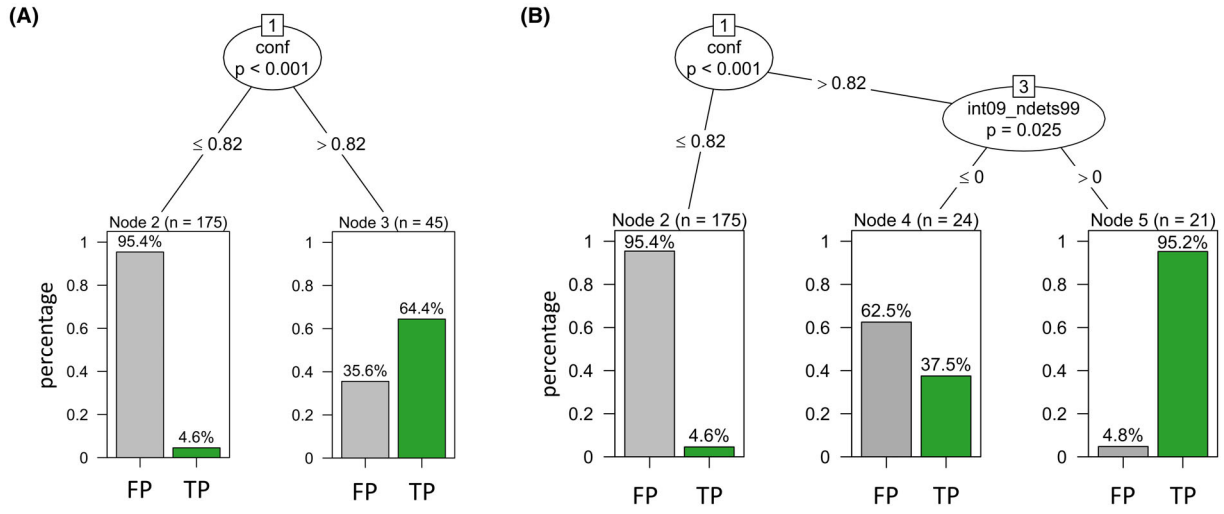

Figure 2. Examples of threshold selection with conditional inference trees for the Grey-headed woodpecker (*Picus canus*). FP, false-positive detections, TP, true-positive detections. Panel (A) shows an example of a type 1 model with maxdepth = 1, panel (B) shows an example of a type 2 model with maxdepth = 2 and two predictor variables (conf = original BirdNET confidence score, int09_ndets99 = number of detections with confidence ≥0.99 in a time interval of ±12 h). As an example, the right branch of the tree in panel (B) can be read as follows: The percentage of true-positive detections (=precision) increases up to 95.2% when their confidence score is >0.82 and they are embedded in a 24-h interval in which at least once a Grey-headed woodpecker was detected with confidence score of 0.99 at the recording site. With the type 1 model (A), percentage of true-positive detections can only be pushed to 64.4% when filtering confidence score >0.82.

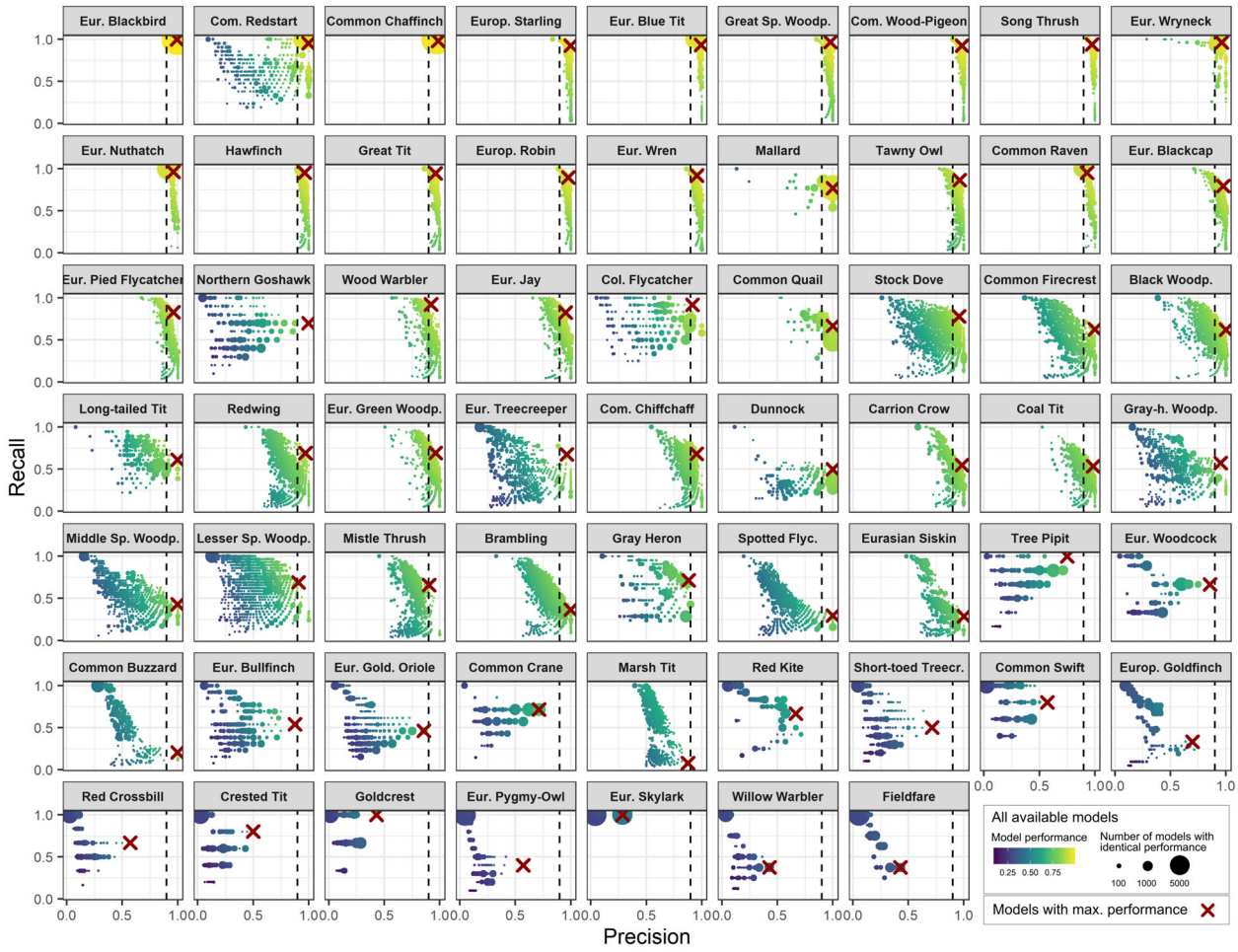

Figure 3. Performance of all available models per species as a metric of precision and recall. Models with maximum performance are marked by red crosses (=candidate models for optimized thresholds). Gradient colours indicate the model performance, calculated as the weighted sum of precision and recall. Point size is scaled by the number of models with identical performance. We regarded precision values ≥0.9 as 'high precision' (marked by a dashed line). Species are sorted by maximum model performance.

However, the recall, defined as the ratio of true positives included after thresholding to the total of all human-validated true positives (Knight et al., 2017), should also be considered within the selection process, as a model with maximum precision could (in a worst case) only include a minimum of the available true positives, meaning that many true positives are discarded. Accordingly, we assessed the model performance by calculating the weighted sum of the precision p and the recall r: model performance = $p \times w + r \times (1 - w)$. The weighting factor w as well as the model performance can take values from zero to one. As low precision is more likely to bias ecological inferences than lower recall (Metcalf, Barlow, Bas, et al., 2022), we chose $w = 0.75$. This means that precision was three times higher weighted than recall, effectively moderating the trade-off between minimal false-positive rates and maximum inclusion of true positives (Knight & Bayne, 2019).

Out of all available models, we selected those as candidates for optimized thresholding that exhibited the highest available model performance for each species (Fig. 3). Due to the multicollinearity of the aggregated time-series features, numerous candidate models with identical model performance were found for some species, while only one optimal candidate model was identified for others. Consequently, we obtained a matrix representing candidate models for each species, indicating that models exhibited the highest performance. Hence, we faced a classical set cover problem in optimization. In the set cover problem, the goal is to select a subset of elements from a given collection such that every element in the original collection is covered by at least one element in the subset. In our context, each species represents an element in the original collection, and the candidate models correspond to the elements in the subset we are trying to select. Hence, we

applied the set cover algorithm from the *lpSolve* package (Berkelaar & Csárdi, 2023) to find the minimum set of unique models required to ensure that at least one optimized model is retained for each species.

Performance evaluation

To evaluate the performance of our optimized species-specific thresholding approach, we compared the optimized thresholds to the type 1 models with BirdNET confidence score as predictor variable. We refer to these type 1 models as basic thresholds in the following since they have the simplest possible tree architecture and rely on the original measure of detection quality from BirdNET. Additionally, we fitted logistic regressions with confidence score as predictor to identify species-specific thresholds, as proposed by Barré et al. (2019). For species where no logistic threshold could be derived, we set performance metrics to raw values without any thresholding. To contrast the species-specific approaches with non-species-specific ones, we applied three universal confidence score thresholds (UNI10: confidence score ≥0.1, equals no thresholding when using BirdNET default settings; UNI50: ≥0.5; UNI90, ≥0.9) to the data. Universal confidence score thresholds were applied to BirdNET data previously (Sethi et al., 2021; Wood et al., 2021). To test whether the optimized thresholds improved the thresholding performance, we compared them to the five other threshold types by testing their effects on four performance metrics (precision, recall, model performance, number of species with precision ≥0.9) with Bonferroni-corrected Wilcoxon tests.

To optimize the number of predictor variables for future applications, we assessed the individual impact of the 169 predictors on the optimized thresholding using a backward selection approach. We iteratively removed predictors based on their length of time intervals from the pool of all candidate models. Similarly, we reduced the statistical parameters included in the models stepwise. Unlike time intervals, these parameters lack a straightforward ranking. Hence, we assessed their effects on the average model performance using a bootstrapping approach with 999 permutations. Finally, we removed the predictors with lowest effect on model performance first and selected the optimized thresholds out of the remaining candidates for each iteration.

Results

Interspecific variability of BirdNET performance

Overall, the BirdNET algorithm identified 10 848 360 bird detections within the 5760 h of audio material. We validated detections of 61 out of the 110 considered species as true positives. We found substantial interspecific variation of the distribution of confidence scores within the human-validated detections. While for some species (e.g. Blackbird *Turdus merula*, Chaffinch *Fringilla coelebs*) nearly all validated detections were true positives independent of confidence scores, for other species, only few true positives were found close to maximum confidence scores (e.g. Grey-headed woodpecker *Picus canus*) (Table S3).

Performance of optimized thresholds

Precision increased up to more than 0.9 (termed as *high precision* in the following) for 70% of the 61 species by optimized thresholding (Fig. 4). Basic thresholds reached high precision for 31% of the species. The thresholds from logistic regression, as proposed by Barré et al. (2019), reached high precision for 56% of the species. The logistic thresholds also performed better than our basic thresholds; however, they failed to identify thresholds for 31% of the species. When applying universal thresholds, precision was highly heterogeneous across species and reached high precision for 10% (UNI10), 26% (UNI50) or 48% (UNI90) of the species.

Our optimized thresholds significantly increased the precision compared to the three universal and two other species-specific thresholds (Figure S1). Furthermore, the optimized thresholds significantly increased the recall compared to the UNI90 threshold and no difference was found compared to the logistic thresholds; however, recall was lower compared to the UNI10, UNI50 and basic thresholds (Figure S2). The metric of model performance of the optimized thresholds was significantly higher compared to all but the basic thresholds (Figure S3). Number of species with high precision was significantly higher with the optimized thresholds compared to all other threshold types (Table S4).

After applying the set cover optimization to the candidate models with maximum performance per species, 41 unique type 2 models remained. Fifty-three of the 169 aggregated time-series features were included in the formulas of these models (Table S5). Predictor variables covering all 12 time intervals were included, but one of the 14 statistical parameters (number of detections with confidence ≥0.99) became redundant.

Post-hoc reduction of predictor variables

Overall, the performance of optimized thresholds decreased when ATF were removed iteratively (Fig. 5). Regarding the maximum length of time intervals, the percentage of species with high precision stayed at its maximum of 70% when predictors integrating more than ±12 h were

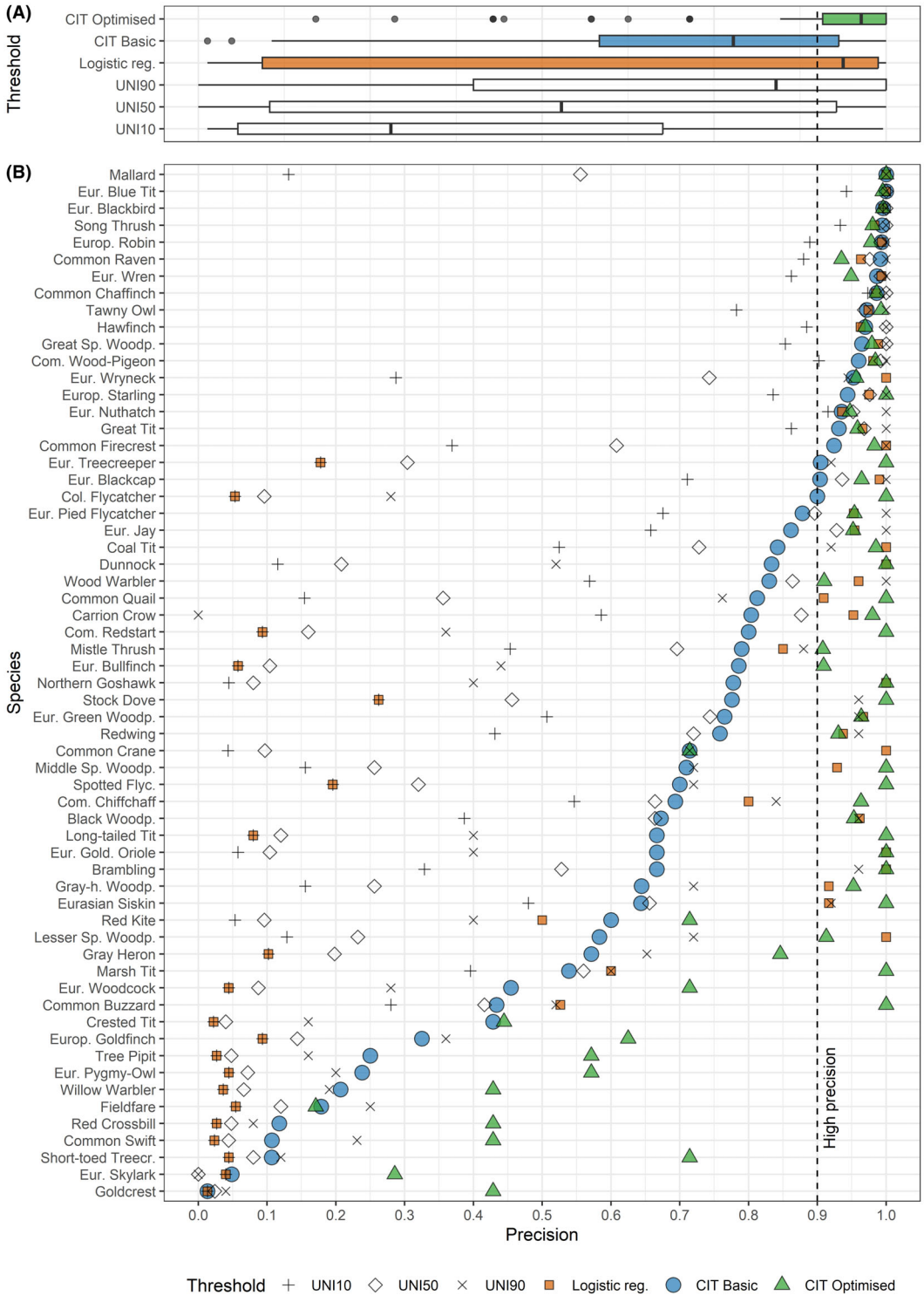

Figure 4. Comparison of precision between three universal thresholds (non-species-specific filtering above a certain confidence score: UNI10: confidence score ≥0.1, UNI50: confidence score ≥0.5, UNI90: confidence score ≥0.9), and three species-specific thresholds, derived from logistic regression (using script from Barré et al., 2019) and our own thresholds derived from conditional inference trees (CIT): basic thresholds (only original BirdNET confidence score was used as predictor) and the optimized thresholds (all 169 aggregated time-series features included as predictors). The dashed line marks a precision of 0.9, which we termed as precision. Plot (A) shows the distribution of precision as boxplots, and plot (B) shows the species-specific values.

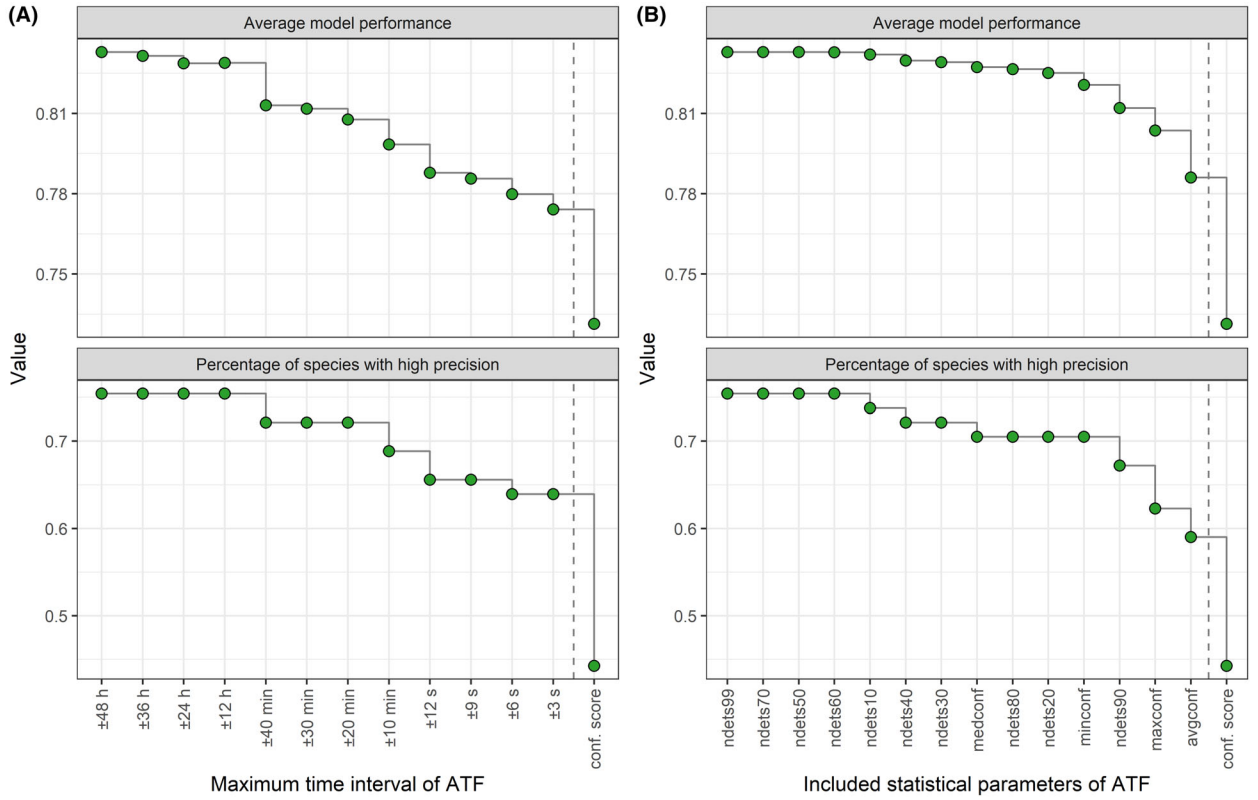

Figure 5. Effects of stepwise reduction of aggregated time-series features [(A) reduction of time interval length; (B) reduction of statistical parameters] on the performance of optimized species-specific threshold models. High precision = precision ≥0.9, performance measures are explained in detail in the methods section. Predictors were ordered according to (A) the time interval length and (B) the average effect on model performance, derived from bootstrapping with 999 permutations. All species: species with at least one validated true-positive detection from passive acoustic monitoring. The dashed line separates the models including simply the original BirdNET confidence score, which is based on a 3-s interval, from the aggregated time-series features.

included. The mean model performance just slightly increased when predictors integrating more than ±12 h were included. Similar patterns were observed for the reduction of statistical parameters. Removing the seven least informative parameters revealed no effect on the performance but removing the remaining eight parameters reduced model performance constantly. Removing all ATF (i.e. limiting the predictors to the confidence score) caused the most pronounced drop in model performance and percentage of species with high precision.

Discussion

Our study demonstrated high heterogeneity of precision in automated bird species detections among species. Hence, BirdNET confidence scores do not provide a uniform measure of detection quality across species but seem to depend on species abundance, at least in noisy real-world applications. It appears essential to implement species-specific post-processing of BirdNET data to ensure the comparability of classification errors among species. Otherwise, inferences regarding habitat preferences or population trends may be biased (Barré et al., 2019; Metcalf, Barlow, Bas, et al., 2022) or, as a worst case, harm threatened species through wrong conservation priorities (Russo & Voigt, 2016; Rydell et al., 2017).

The utilization of ATF substantially improved the differentiation of true and false species detections in our optimized thresholding approach. Optimized thresholds particularly outperformed non-species-specific universal confidence score thresholds, which were applied in previous studies (Sethi et al., 2021; Wood et al., 2021), emphasizing the importance of species-specific threshold adaptation (Cole et al., 2022; Pérez-Granados, 2023). The vast improvement of correct species classification achieved through the usage of ATF may be explained by their biological significance. By incorporating information from temporally adjacent detections, the ATF serve as proxies of species-specific temporal dynamics of detection probabilities and are fitted to the actual data. Hence, they

encapsulate study-specific information on the likelihood to observe a species at a specific time of the year or day in a certain habitat.

Several previous studies successfully utilized spatio-temporal information to improve the differentiation of true and false positives, for example for fin whales (Madhusudhana et al., 2021), two frog species (Chambert et al., 2018), two bat species (Clement et al., 2022) or five to six bird species (Metcalf, Barlow, Bas, et al., 2022; Rhinehart et al., 2022). However, none of them utilized information from the detection time-series comparable to the ATF in our thresholding approach. Hence, results are not directly comparable. Furthermore, such case studies including a limited set of species clearly differ from the extent of field studies (Rhinehart et al., 2022). As our species-specific thresholding approach has revealed high performance in an extended field study (240 days of audio data from 60 sites) including an entire assemblage of 61 species, it demonstrated applicability for comprehensive PAM studies.

Nevertheless, there is potential for future development of the utilization of ATFs. Since we validated a random sample of species detections, stratified by species and BirdNET confidence score but not by site or time, the derived species-specific threshold rules operate as an average threshold at the scale of the complete dataset but do not fully consider site- and time-specific differences of BirdNET performance within species. Hence, we implicitly assumed that performance of BirdNET is spatio-temporally consistent within the study area. This assumption may be challenged for some species, as intra-specific classifier performance can vary significantly across sites (Metcalf, Barlow, Bas, et al., 2022). Especially in studies that cover a broader environmental gradient, automated classifier performance is known to vary (Lauha et al., 2022). As our results demonstrate, BirdNET performance already varies largely between species. Hence, optimized thresholds based on ATF are a substantial contribution to improve the interspecific comparability of BirdNET results, even though future research should focus on spatio-temporal variation of the performance of optimized thresholds. When accounting for spatio-temporal variation, the performance for (locally) rare species may be enhanced. However, accounting for spatio-temporal variation of the thresholds would require a much larger effort of human validation when covering entire species assemblages and numerous sites. We hypothesize that sites could be pre-classified based on their raw BirdNET detection time-series; hence, human validation could be additionally stratified based on clusters of sites with similar data structure to improve the site specificity but reduce the required amount of human validation.

We are aware that our thresholding approach may be statistically more basic than approaches in the majority of the recent studies, which relied on more advanced statistical models [e.g. boosted regression trees (Knight et al., 2020), occupancy models (Clement et al., 2022; Rhinehart et al., 2022; Wright et al., 2020) and hierarchical modelling (Chambert et al., 2018; Cole et al., 2022)]. However, regarding the high performance we consider this as an advantage, as we intended to design a straightforward thresholding approach that is applicable by a wide audience. Our approach yields thresholds, which consist of up to two conditions for species-specific filtering of big data in extended PAM field studies. The threshold conditions are directly derived from the time-series of BirdNET detections and do not require external environmental information, as included by Metcalf, Barlow, Bas, et al. (2022). Furthermore, conditional inference trees are statistically robust and intuitively interpretable (Hothorn et al., 2006; Müller & Bütler, 2010). Compared to models relying on Bayesian statistics (Chambert et al., 2018; Clement et al., 2022; Cole et al., 2022; Rhinehart et al., 2022), conditional inference trees may be much more accessible to average PAM users in biodiversity research and conservation. Therefore, our thresholding approach may help to bridge the science-practice gap in conservation (Fabian et al., 2019).

Our thresholding approach simply requires a time-series of confidence scores from a classification algorithm and a rather small sample of 225 human-validated detections from the actual study area. A trained observer can validate the sample of audio snippets in near real-time, so the effort totals about 15 min per species. Therefore, it may be well adaptable to studies in other habitat types, regions and in acoustic monitoring of other taxonomic groups like bats. In principle, it is even adaptable to automated image classification in camera-trapping applications as this also provides time-series of confidence scores (Tabak et al., 2019). However, our chosen approach to validate BirdNET detections by listening to 3-s audio intervals may underestimate true-positive detections in some cases, as BirdNET may be able to correctly identify fragments of calls or songs of certain species that a human observer is not able to identify within a 3-s interval. Hence, it could be worth to prolong the time interval used for human validation in future applications.

As the backward selection of predictors demonstrated (Fig. 5), including statistical parameters calculated at time intervals of 9 s can substantially improve predictive power compared to the models that only include BirdNET confidence scores. Among the statistical parameters, the maximum and average of confidence scores yielded the most information. Consequently, it may already be highly informative to calculate maximum and average confidence

scores for two adjacent BirdNET detections. Average confidence scores at a time interval of 9 s also improved the performance in a previous study (Wood et al., 2021). Nevertheless, the set cover optimization revealed that 41 unique model formulas are required to identify at least one optimized threshold for all species. These 41 model formulas included 53 of the 169 ATF, aggregating confidence score information at all 12 time intervals and only one statistical parameter became redundant. Therefore, vast potential for reducing computational costs lies not in limiting the set of ATF, but rather in excluding uninformative combinations of ATF from the model formulas within in the thresholding modelling step.

It is likely that the selected optimized thresholds are not transferable one-on-one to other regions or habitat types with different species assemblages or other recording devices (Cole et al., 2022). Automated species recognition is generally known to perform differently across regions; therefore, site- or region-specific adaptations need to be made (Cole et al., 2022; Lauha et al., 2022; Metcalf, Barlow, Bas, et al., 2022). As the available BirdNET algorithm is trained on 'weak labels' that may include significant amounts of noise and non-target species (Kahl et al., 2021), researchers should make use of the new possibility to train BirdNET based on 'strong labels', for example own labelled data from regional audio collections, to improve its classification performance (Ghani et al., 2023; McGinn et al., 2023).

Our study unequivocally established the efficacy of aggregated time-series features in enhancing species-specific post-processing of data derived from passive acoustic bird monitoring. We designed a species-specific thresholding approach that minimizes false-positive species detections for an entire assemblage of bird species in an extended, realistic bird monitoring application. It also efficiently balances the trade-off between maximizing precision and recall. Due to the statistical simplicity of the underlying conditional inference trees, it may be straightforward to apply even by users not used to deal with complex statistical models. Hence, our optimized species-specific thresholding approach may enhance the application of PAM to inform evidence-based conservation efforts. Independent from the choice of conditional inference trees, we also encourage other researchers to consider the inclusion of aggregated time-series features into their models, as they turned out to vastly enhance the performance of classification models to distinguish true- and false-positive species detections.

Author contributions

DS developed the methodology, collected the data, conducted the analyses and drafted the paper; AS, HH, JH and JK made substantial contributions to methodology, analyses and interpretation of results; all authors contributed critically to the drafts and gave final approval for publication.

Acknowledgements

We thank the administration of the Hainich National Park for supporting the study. We thank A. Piter for getting the initial BirdNET version running on a Windows workstation. We thank P. Hansen and T. Leise for helpful ideas concerning the analyses. We also thank two anonymous reviewers for their valuable and constructive comments on an earlier version of this paper. We acknowledge support by the Open Access Publication Funds/transformative agreements of the Göttingen University. Open Access funding enabled and organized by Projekt DEAL.

Conflict of interest

The authors declare no conflict of interest.

Data availability statement

R code and data are available at GitHub: https://github.com/d-singer/BirdNET-thresholds.

References

Barré, K., Le Viol, I., Julliard, R., Pauwels, J., Newson, S.E., Julien, J. et al. (2019) Accounting for automated identification errors in acoustic surveys. *Methods in Ecology and Evolution*, **10**, 1171–1188. Available from: https://doi.org/10.1111/2041-210X.13198

Benedict, L. & Najar, N.A. (2019) Are commonly used metrics of bird song complexity concordant? *The Auk*, **136**(1), uky008. Available from: https://doi.org/10.1093/auk/uky008

Berkelaar, M. & Csárdi, G. (2023) lpSolve: Interface to 'Lp_solve' v. 5.5 to solve Linear/Integer programs, version 5.6.19. Available from: https://cran.r-project.org/web/packages/lpSolve/index.html

Bota, G., Manzano-Rubio, R., Catalán, L., Gómez-Catasús, J. & Pérez-Granados, C. (2023) Hearing to the unseen: AudioMoth and BirdNET as a cheap and easy method for monitoring cryptic bird species. *Sensors*, **23**, 16. Available from: https://doi.org/10.3390/s23167176

Boulinier, T., Nichols, J.D., Sauer, J.R., Hines, J.E. & Pollock, K.H. (1998) Estimating species richness: the importance of heterogeneity in species detectability. *Ecology*, **79**(3), 1018–1028. Available from: https://doi.org/10.1890/0012-9658(1998)079[1018:ESRTIO]2.0.CO;2

Chambert, T., Waddle, J.H., Miller, D.A.W., Walls, S.C. & Nichols, J.D. (2018) A new framework for analysing

automated acoustic species detection data: occupancy estimation and optimization of recordings post-processing. *Methods in Ecology and Evolution*, **9**(3), 560–570. Available from: https://doi.org/10.1111/2041-210X.12910

Chhaya, V., Lahiri, S., Jagan, M.A., Mohan, R., Pathaw, N.A. & Krishnan, A. (2021) Community bioacoustics: studying acoustic community structure for ecological and conservation insights. *Frontiers in Ecology and Evolution*, **9**, 706445. Available from: https://doi.org/10.3389/fevo.2021.706445

Clement, M.J., Royle, J.A. & Mixan, R.J. (2022) Estimating occupancy from autonomous recording unit data in the presence of misclassifications and detection heterogeneity. *Methods in Ecology and Evolution*, **13**(8), 1719–1729. Available from: https://doi.org/10.1111/2041-210X.13895

Cole, J.S., Michel, N.L., Emerson, S.A. & Siegel, R.B. (2022) Automated bird sound classifications of long-duration recordings produce occupancy model outputs similar to manually annotated data. *Ornithological Applications*, **124**, 1–15. Available from: https://doi.org/10.1093/ornithapp/duac003

Darras, K., Batáry, P., Furnas, B., Celis-Murillo, A., Van Wilgenburg, S.L., Mulyani, Y.A. et al. (2018) Comparing the sampling performance of sound recorders versus point counts in bird surveys: a meta-analysis. *Journal of Applied Ecology*, **55**(6), 2575–2586. Available from: https://doi.org/10.1111/1365-2664.13229

Darras, K., Batáry, P., Furnas, B.J., Grass, I., Mulyani, Y.A. & Tscharntke, T. (2019) Autonomous sound recording outperforms human observation for sampling birds: a systematic map and user guide. *Ecological Applications*, **29**, e01954. Available from: https://doi.org/10.1002/eap.1954

Fabian, Y., Bollmann, K., Brang, P., Heiri, C., Olschewski, R., Rigling, A. et al. (2019) How to close the science-practice gap in nature conservation? Information sources used by practitioners. *Biological Conservation*, **235**, 93–101. Available from: https://doi.org/10.1016/j.biocon.2019.04.011

Farina, A. & Ceraulo, M. (2017) The acoustic chorus and its ecological significance. In: Farina, A. & Gage, S.H. (Eds.) *Ecoacoustics: the ecological role of sounds*. Hoboken: Wiley, pp. 81–94. Available from: https://doi.org/10.1002/9781119230724.ch5

Florentin, J., Dutoit, T. & Verlinden, O. (2020) Detection and identification of European woodpeckers with deep convolutional neural networks. *Ecological Informatics*, **55**, 101023. Available from: https://doi.org/10.1016/j.ecoinf.2019.101023

Froidevaux, J.S.P., Zellweger, F., Bollmann, K. & Obrist, M.K. (2014) Optimizing passive acoustic sampling of bats in forests. *Ecology and Evolution*, **4**(24), 4690–4700. Available from: https://doi.org/10.1002/ece3.1296

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

Gedeon, K., Grüneberg, C., Mitschke, A., Sudfeldt, C., Eikhorst, W., Fischer, S. et al. (2015) *Atlas Deutscher Brutvogelarten*. Münster: Stiftung Vogelmonitoring Deutschland und Dachverband Deutscher Avifaunisten.

Ghani, B., Denton, T., Kahl, S. & Klinck, H. (2023) Feature embeddings from large-scale acoustic bird classifiers enable few-shot transfer learning. *arXiv*. 2307.06292. https://doi.org/10.1038/s41598-023-49989-z

Gibb, R., Browning, E., Glover-Kapfer, P. & Jones, K.E. (2019) Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. *Methods in Ecology and Evolution*, **10**(2), 169–185. Available from: https://doi.org/10.1111/2041-210X.13101

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

Hothorn, T., Hornik, K. & Zeileis, A. (2006) Unbiased recursive partitioning: a conditional inference framework. *Journal of Computational and Graphical Statistics*, **15**(3), 651–674. Available from: https://doi.org/10.1198/106186006X133933

Hothorn, T. & Zeileis, A. (2015) Partykit: a modular toolkit for recursive partytioning in R. *The Journal of Machine Learning Research*, **16**(1), 3905–3909.

Kahl, S. (2020) *Identifying birds by sound: large-scale acoustic event recognition for avian activity monitoring*. Chemnitz: Universitätsverlag Chemnitz.

Kahl, S., Wood, C.M., Eibl, M. & Klinck, H. (2021) BirdNET: a deep learning solution for avian diversity monitoring. *Ecological Informatics*, **16**, 101236. Available from: https://doi.org/10.1016/j.ecoinf.2021.101236

Katz, J., Hafner, S.D. & Donovan, T. (2016) Tools for automated acoustic monitoring within the R package monitoR. *Bioacoustics*, **25**(2), 197–210. Available from: https://doi.org/10.1080/09524622.2016.1138415

Kéry, M. & Schmidt, B. (2008) Imperfect detection and its consequences for monitoring for conservation. *Community Ecology*, **9**(2), 207–216. Available from: https://doi.org/10.1556/ComEc.9.2008.2.10

Knight, E.C. & Bayne, E.M. (2019) Classification threshold and training data affect the quality and utility of focal species data processed with automated audio-recognition software. *Bioacoustics*, **28**(6), 539–554. Available from: https://doi.org/10.1080/09524622.2018.1503971

Knight, E.C., Hannah, K.C., Foley, G.J., Scott, C.D., Brigham, R.M. & Bayne, E. (2017) Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs. *Avian*

Conservation and Ecology, **12**(2), 120214. Available from: https://doi.org/10.5751/ACE-01114-120214

Knight, E.C., Sòlymos, P., Scott, C. & Bayne, E.M. (2020) Validation prediction: a flexible protocol to increase efficiency of automated acoustic processing for wildlife research. *Ecological Applications*, **30**(7), e02140. Available from: https://doi.org/10.1002/eap.2140

Kułaga, K. & Budka, M. (2019) Bird species detection by an observer and an autonomous sound recorder in two different environments: forest and farmland. *PLoS One*, **14**(2), e0211970. Available from: https://doi.org/10.1371/journal.pone.0211970

Lauha, P., Somervuo, P., Lehikoinen, P., Geres, L., Richter, T., Seibold, S. et al. (2022) Domain-specific neural networks improve automated bird sound recognition already with small amount of local data. *Methods in Ecology and Evolution*, **13**, 2799–2810. Available from: https://doi.org/10.1111/2041-210X.14003

Ligges, U., Krey, S., Mersmann, O. & Schnackenberg, S. (2018) *tuneR: analysis of music and speech* [manual]. https://CRAN.R-project.org/package=tuneR

Lindenmayer, D.B., Lavery, T. & Scheele, B.C. (2022) Why we need to invest in large-scale, long-term monitoring programs in landscape ecology and conservation biology. *Current Landscape Ecology Reports*, **7**, 137–146. Available from: https://doi.org/10.1007/s40823-022-00079-2

Madhusudhana, S., Shiu, Y., Klinck, H., Fleishman, E., Liu, X., Nosal, E.-M. et al. (2021) Improve automatic detection of animal call sequences with temporal context. *Journal of the Royal Society Interface*, **18**, 20210297.

McGinn, K., Kahl, S., Peery, M.Z., Klinck, H. & Wood, C.M. (2023) Feature embeddings from the BirdNET algorithm provide insights into avian ecology. *Ecological Informatics*, **74**, 101995. Available from: https://doi.org/10.1016/j.ecoinf.2023.101995

Metcalf, O.C., Barlow, J., Bas, Y., Berenguer, E., Devenish, C., França, F. et al. (2022) Detecting and reducing heterogeneity of error in acoustic classification. *Methods in Ecology and Evolution*, **13**, 2559–2571. Available from: https://doi.org/10.1111/2041-210X.13967

Metcalf, O.C., Barlow, J., Marsden, S., Gomes de Moura, N., Berenguer, E., Ferreira, J. et al. (2022) Optimizing tropical forest bird surveys using passive acoustic monitoring and high temporal resolution sampling. *Remote Sensing in Ecology and Conservation*, **8**(1), 45–56. Available from: https://doi.org/10.1002/rse2.227

Müller, J. & Bütler, R. (2010) A review of habitat thresholds for dead wood: a baseline for management recommendations in European forests. *European Journal of Forest Research*, **129**(6), 981–992. Available from: https://doi.org/10.1007/s10342-010-0400-5

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

Pérez-Granados, C., la Rosa, D.B., Gómez-Catasús, J., Barrero, A., Abril-Colón, I. & Traba, J. (2018) Autonomous recording units as effective tool for monitoring of the rare and patchily distributed Dupont's lark *Chersophilus duponti*. *Ardea*, **106**(2), 139. Available from: https://doi.org/10.5253/arde.v106i2.a6

Picciulin, M., Kéver, L., Parmentier, E. & Bolgan, M. (2019) Listening to the unseen: passive acoustic monitoring reveals the presence of a cryptic fish species. *Aquatic Conservation: Marine and Freshwater Ecosystems*, **29**(2), 202–210. Available from: https://doi.org/10.1002/aqc.2973

Posit Team. (2023) *RStudio: integrated development environment for R* (2023.6.2.561) [computer software]. Posit Software, PBC. http://www.posit.co/

Priyadarshani, N., Marsland, S. & Castro, I. (2018) Automated birdsong recognition in complex acoustic environments: a review. *Journal of Avian Biology*, **49**(5), 01447. Available from: https://doi.org/10.1111/jav.01447

R Core Team. (2023) R: a language and environment for statistical computing (4.2.2) [computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rhinehart, T.A., Turek, D. & Kitzes, J. (2022) A continuous-score occupancy model that incorporates uncertain machine learning output from autonomous biodiversity surveys. *Methods in Ecology and Evolution*, **13**(8), 1778–1789. Available from: https://doi.org/10.1111/2041-210X.13905

Ross, S.R.P.-J., O'Connell, D.P., Deichmann, J.L., Desjonquères, C., Gasc, A., Phillips, J.N. et al. (2023) Passive acoustic monitoring provides a fresh perspective on fundamental ecological questions. *Functional Ecology*, **37**(4), 959–975. Available from: https://doi.org/10.1111/1365-2435.14275

Russo, D. & Voigt, C.C. (2016) The use of automated identification of bat echolocation calls in acoustic monitoring: a cautionary note for a sound analysis. *Ecological Indicators*, **66**, 598–602. Available from: https://doi.org/10.1016/j.ecolind.2016.02.036

Rydell, J., Nyman, S., Eklöf, J., Jones, G. & Russo, D. (2017) Testing the performances of automated identification of bat echolocation calls: a request for prudence. *Ecological Indicators*, **78**(Suppl C), 416–420. Available from: https://doi.org/10.1016/j.ecolind.2017.03.023

Schmidt, B.R., Cruickshank, S.S., Bühler, C. & Bergamini, A. (2023) Observers are a key source of detection heterogeneity and biased occupancy estimates in species monitoring. *Biological Conservation*, **283**, 110102. Available from: https://doi.org/10.1016/j.biocon.2023.110102

Sethi, S.S., Fossøy, F., Cretois, B. & Rosten, C.M. (2021) *Management relevant applications of acoustic monitoring for Norwegian nature – the sound of Norway* (2064; NINA Report). Norwegian Institute for Nature Research.

Sugai, L.S.M., Desjonquères, C., Silva, T.S.F. & Llusia, D. (2020) A roadmap for survey designs in terrestrial acoustic monitoring. *Remote Sensing in Ecology and Conservation*, **6**(3), 220–235. Available from: https://doi.org/10.1002/rse2.131

Sugai, L.S.M., Silva, T.S.F., Ribeiro, J.W. & Llusia, D. (2019) Terrestrial passive acoustic monitoring: review and

perspectives. *Bioscience*, **69**(1), 15–25. Available from: https://doi.org/10.1093/biosci/biy147

Sullivan, B.L., Wood, C.L., Iliff, M.J., Bonney, R.E., Fink, D. & Kelling, S. (2009) eBird: a citizen-based bird observation network in the biological sciences. *Biological Conservation*, **142**(10), 2282–2292. Available from: https://doi.org/10.1016/j.biocon.2009.05.006

Sutherland, W.J., Pullin, A.S., Dolman, P.M. & Knight, T.M. (2004) The need for evidence-based conservation. *Trends in Ecology & Evolution*, **19**(6), 305–308. Available from: https://doi.org/10.1016/j.tree.2004.03.018

Tabak, M.A., Norouzzadeh, M.S., Wolfson, D.W., Sweeney, S.J., Vercauteren, K.C., Snow, N.P. et al. (2019) Machine learning to classify animal species in camera trap images: applications in ecology. *Methods in Ecology and Evolution*, **10**, 585–590. Available from: https://doi.org/10.1111/2041-210X.13120

Thompson, S.J., Handel, C.M. & Mcnew, L.B. (2017) Autonomous acoustic recorders reveal complex patterns in avian detection probability. *The Journal of Wildlife Management*, **81**(7), 1228–1241. Available from: https://doi.org/10.1002/jwmg.21285

Wood, C.M., Kahl, S., Chaon, P., Peery, M.Z. & Klinck, H. (2021) Survey coverage, recording duration and community composition affect observed species richness in passive acoustic surveys. *Methods in Ecology and Evolution*, **12**(5), 885–896. Available from: https://doi.org/10.1111/2041-210X.13571

Wright, W.J., Irvine, K.M., Almberg, E.S. & Litt, A.R. (2020) Modelling misclassification in multi-species acoustic data when estimating occupancy and relative activity. *Methods in Ecology and Evolution*, **11**(1), 71–81. Available from: https://doi.org/10.1111/2041-210X.13315

Zwart, M.C., Baker, A., McGowan, P.J.K. & Whittingham, M.J. (2014) The use of automated bioacoustic recorders to replace human wildlife surveys: an example using nightjars. *PLoS One*, **9**(7), e102770. Available from: https://doi.org/10.1371/journal.pone.0102770

Supporting Information

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

Table S1. List of 110 species which were pre-selected for the BirdNET analysis based on their potential occurrence in the study region.

Table S2. Aggregated time-series features (ATF), included as predictor variables for threshold modelling with conditional inference trees.

Table S3. List of 61 species with at least one validated true positive detection during human validation of a sample of BirdNET detections.

Table S4. Comparison of the number of species with high precision (≥0.9) of our optimised thresholds, derived from conditional inference trees.

Table S5. List of 41 optimised threshold models that were identified with the set cover optimisation.

Table S6. Examples of species-specific threshold rules resulting from the selection of basic and optimised threshold models for 61 bird species.

Figure S1. Pairwise (per bird species) comparison of precision of our optimised thresholds, derived from conditional inference trees (CIT Optimised, all 169 aggregated time series features included as predictors) and three universal thresholds (non-species specific filtering above a certain confidence score: UNI10: confidence score ≥0.1, UNI50: confidence score ≥0.5, UNI90: confidence score ≥0.9) and two species-specific thresholds, derived from logistic regression (using script from Barré et al., 2019) and our own basic thresholds (only original BirdNET confidence score was used as predictor). Differences were tested by Bonferroni corrected paired Wilcox-tests.

Figure S2. Pairwise (per bird species) comparison of recall of our optimised thresholds, derived from conditional inference trees (CIT Optimised, all 169 aggregated time series features included as predictors) and three universal thresholds (non-species specific filtering above a certain confidence score: UNI10: confidence score ≥0.1, UNI50: confidence score ≥0.5, UNI90: confidence score ≥0.9) and two species-specific thresholds, derived from logistic regression (using script from Barré et al., 2019) and our own basic thresholds (only original BirdNET confidence score was used as predictor). Differences were tested by Bonferroni corrected paired Wilcox-tests.

Figure S3. Pairwise (per bird species) comparison of the model performance (calculated as weighted sum of precision and recall) of our optimised thresholds, derived from conditional inference trees (CIT Optimised, all 169 aggregated time series features included as predictors) and three universal thresholds (non-species specific filtering above a certain confidence score: UNI10: confidence score ≥0.1, UNI50: confidence score ≥0.5, UNI90: confidence score ≥0.9) and two species-specific thresholds, derived from logistic regression (using script from Barré et al., 2019) and our own basic thresholds (only original BirdNET confidence score was used as predictor). Differences were tested by Bonferroni corrected paired Wilcox-tests.