doi:10.26443/seismica.v3i2.1344

An Open Source Hydroacoustic Benchmarking Framework for Geophonic Signal Detection

Pierre-Yves Raumer * [1,2], **Sara Bazin** [1], **Dorian Cazau** [2], **Vaibhav Vijay Ingale** [1,3], **Aude Lavayssière** [1], **Jean-Yves Royer** [1]

[1]Geo-Ocean, Univ Brest, Centre National de la Recherche Scientifique (CNRS), Ifremer, UMR6538, F-29280 Plouzane, France, [2]Laboratoire des Sciences et Techniques de l'Information, de la Communication et de la Connaissance (Lab-STICC), École Nationale Supérieure de Techniques Avancées (ENSTA) - Bretagne, UMR6285, F-29200, Brest, France, [3]Institute of Geophysics and Planetary Physics, Scripps Institution of Oceanography, University of California, San Diego, CA, USA

Author contributions: *Conceptualization*: Raumer. *Methodology*: Raumer, Bazin, Cazau, Lavayssière. *Software*: Raumer. *Resources*: Royer. *Writing - Original draft*: Raumer. *Writing - Review & Editing*: Bazin, Cazau, Ingale, Lavayssière, Royer. *Supervision*: Bazin, Cazau.

Abstract Passive hydroacoustic studies have underscored the efficiency and relevance of deploying autonomous hydrophones for the surveillance of underwater geophony. In particular, monitoring networks have been deployed for detecting SOFAR-propagating hydroacoustic waves generated by seismic events and locating their sources. The technique has been extended to study other hydroacoustic signals, such as P-waves from teleseismic events or impulsive waves generated by seawater-lava interactions. A significant challenge in this endeavor lies in the time required for the manual detection and annotation of these signals in long-term records. To address this issue, we tested the feasibility of implementing automated algorithms based on machine learning to detect and identify these various signals, and obtained satisfying classification and time picking accuracies. We incorporated those models in a benchmarking framework, proposing a training dataset, two evaluation datasets, two tasks to solve and the evaluations of the mentioned models on them. The goal of this framework is to foster the development of new models in the community, as it gives a clear way to evaluate them.

Production Editor: Carmine Gelasso
Handling Editor: Stephen Hicks
Copy & Layout Editor: Sarah Jaye Oliva

Signed reviewer(s): Matthias Pilot

Received: March 28, 2024
Accepted: September 20, 2024
Published: October 14, 2024

1 Introduction

T-waves are hydroacoustic signals generated by the conversion of emerging seismic waves at the ocean bottom (Tolstoy and Ewing, 1950) and arrive after P and S-waves. Since the 90s, their systematic analysis has helped build hydroacoustic catalogs in remote parts of the world ocean, poorly covered by land-based seismological networks (*e.g.,* Fox et al., 1995; Ingale et al., 2023). This purpose led to the development of specific recording instruments and software to analyze these waves.

Studying T-waves proved to be a cost-effective and efficient approach for monitoring the seismic activities in the open ocean, particularly along mid-ocean ridges (*e.g.,* Fox et al., 2001; Dziak et al., 2004; Giusti et al., 2018; Ingale et al., 2021) or underwater volcanoes (Tepp and Dziak, 2021; Bazin et al., 2022; Saurel et al., 2022). The cylindrical propagation of T-waves within the SOund Fixing And Ranging (SOFAR) channel in the ocean undergoes little attenuation over long distances (>1000 km), compared with the rapid spherical attenuation losses of seismic P- or S-waves traveling through the solid earth (Okal, 2008). Consequently, hydrophones can detect seismic events with smaller magnitudes compared to land-based seismometers (e.g., body-wave magnitude (m_b) of completeness is 3.3 for hydroacoustic detection versus 4.1 for terrestrial arrays (Ingale et al., 2023)).

Hydrophones can also detect other geophonic signals, such as regional (Pn) or teleseismic P-waves in the water column coming from distant events (thousands of kilometers away) (Dziak et al., 2004; de Melo et al., 2021), thereafter both termed as hydroacoustic P-waves. In addition, lava-seawater interactions generate impulsive (<10 s) signals directly in the water column (Bazin et al., 2022), thereafter termed as H-waves. For clarity, the term *geophony* will refer to earthquake- and volcano-generated sounds, while ice-generated sounds will be called *cryogenic* events.

For monitoring such geological events, hydrophones are moored in networks of three or more stations around an area of interest. When a particular signal is detected by three or more stations, trilateration using the Times of Arrival (ToA) can yield the source location. In the case of T-waves or hydroacoustic P-waves, this location corresponds to the area of conversion from seismic to hydroacoustic waves, often referred to as the hydroacoustic radiator (Fox et al., 2001). For shallow earthquakes, such hydroacoustic radiators generally match the epicenter (Williams et al., 2006).

Locating hydroacoustic events thus requires an initial process of selecting and annotating continuous records of several hydrophones. Until now, this stage has mainly relied on "manual" detection, with softwares such as *Seas* (Fox et al., 2001). This task is

*Corresponding author: pierre-yves.raumer@univ-brest.fr

time-consuming and user-dependent, resulting in incomplete and/or imprecise processing of long-term datasets. Consequently, automatic detection of these signals appears to be a relevant tool. However, such endeavors are not trivial. In other communities, open datasets are selected and used to benchmark and compare new techniques. For example, in the successful image recognition community, ImageNet was used in the context of the ILSVRC (ImageNet Large Scale Visual Recognition Challenge), which led to the emergence of nowadays well-known classification techniques (Russakovsky et al., 2015). Closer to our field, the DCASE (Detection and Classification of Acoustic Scenes and Events) challenge regularly proposes acoustic datasets and formalizes difficult tasks benchmarked with baseline methods to foster developments in these fields (Kong et al., 2016; Mesaros et al., 2018). Such benchmarking frameworks thus appear to ease the emergence of efficient techniques. We propose a similar approach in the hydroacoustic geophony context, with the following contributions:

- Three datasets: one covering 9 months intended for model training, and two of shorter periods (6 days) in different geographical and temporal contexts and intended for model evaluation. The latter two datasets have been exhaustively annotated.
- Two formalized tasks, intended to detect geophony sounds, together with relevant evaluation metrics to compare models.
- Reference models for detecting geophony sounds in hydroacoustic datasets, including state-of-the-art models from neighboring communities, and a new model to challenge those reference models.

The proposed framework is intended to serve as a comparison baseline, to test future models against these open datasets and to compare them with the proposed metrics.

2 Related works

While the automatic detection of T-waves has not yet been extensively explored, some studies started addressing this problem. For example, some studies focused on embedded systems for detecting some relevant events in time series data (*e.g.,* MERMAID project, Simons et al., 2009). A widely used technique for this purpose involves computing the Short-Term-Average/Long-Term-Average (STA/LTA) ratio, which compares the energy in a short time-window to that in a long time-window and considers detection when this ratio exceeds a predefined threshold. Sukhovich et al. (2014) improved this technique by incorporating spectral features derived from wavelets to classify STA/LTA detections with the objective of distinguishing T-waves from other signals (iceberg cracks, ship noise, whale communication). The classification of signals is then derived by statistically comparing scales from Discrete Wavelet Transform (DWT) (Sukhovich et al., 2011) or by employing a machine learning model, like Gradient Boosted Decision Trees (Sukhovich et al., 2014). In contrast, Matsumoto et al. (2006) proposed a threshold method that compares frequency bands. This approach dynamically adapts the detection threshold to account for the noise level.

Other related fields have sparked a renewed interest. Notably, earthquake monitoring using seismic arrays led to significant progress in automatic detection with highly effective techniques. We can cite PhaseNet (Zhu and Beroza, 2019), a U-Net-like model made of 1D convolutions, or EQTransformer (Mousavi et al., 2020), a model taking advantage of the recent Transformer architectures. Both use raw waveforms as inputs and work on a single seismic station at a time, considering each recorded component (N–S, E–W, vertical movement) as a different channel, and aim at picking P- and S-phases. These tools have been, for example, used for real-time monitoring with seismometers (*e.g.,* Retailleau et al., 2022) or to study data from the emerging Distributed Acoustic Sensing (DAS) technology (Zhu et al., 2023).

In marine biology, passive hydroacoustics have become a common tool for detecting sounds in the "biophony" domain. The various techniques used in this community include the classification of spectrograms derived from Short-Term Fourier Transforms (STFT) using Convolutional Neural Networks (CNN) (Zhong et al., 2020; Rasmussen and Širović, 2021; Zhong et al., 2021). Additionally, classification efforts extend to scalograms obtained with DWT (Ibrahim et al., 2018), and, albeit less frequently, to waveforms (Luo et al., 2019). Studies also explored data augmentation to enhance the performance of these models (Luo et al., 2019; Rasmussen and Širović, 2021) or non-parametric methods using features such as calls periodicities, matched filters or stochastic matched filters to classify the signals (Bouffaut et al., 2018). In bioacoustics, most detection works focus on the presence / absence classification task, whereas the along-time segmentation task may sometimes be required (Bermant et al., 2022). A key component in most approaches is the ground truth dataset to compare automatic detections with or to train machine learning models. The quality of such dataset, generally resulting from manual annotations, is often overlooked and can bias the comparisons or the training. For example, several papers dealing with marine bioacoustics (e.g., Leroy et al., 2018b; Duc et al., 2021; Dubus et al., 2023b,a) have quantitatively measured the inter-annotator variability, showing how this variability propagates to machine learning results. Dubus et al. (2023a), for example, showed an increased variability and a global worsening of model performance when using annotations produced by novices compared to experts, while techniques such as soft labelling aggregations mitigate this problem. On a similar topic, Leroy et al. (2018b) highlighted the influence of the personality of the annotator on the produced annotations, especially in terms of conservativeness, demonstrating the potential subjectivity of this task.

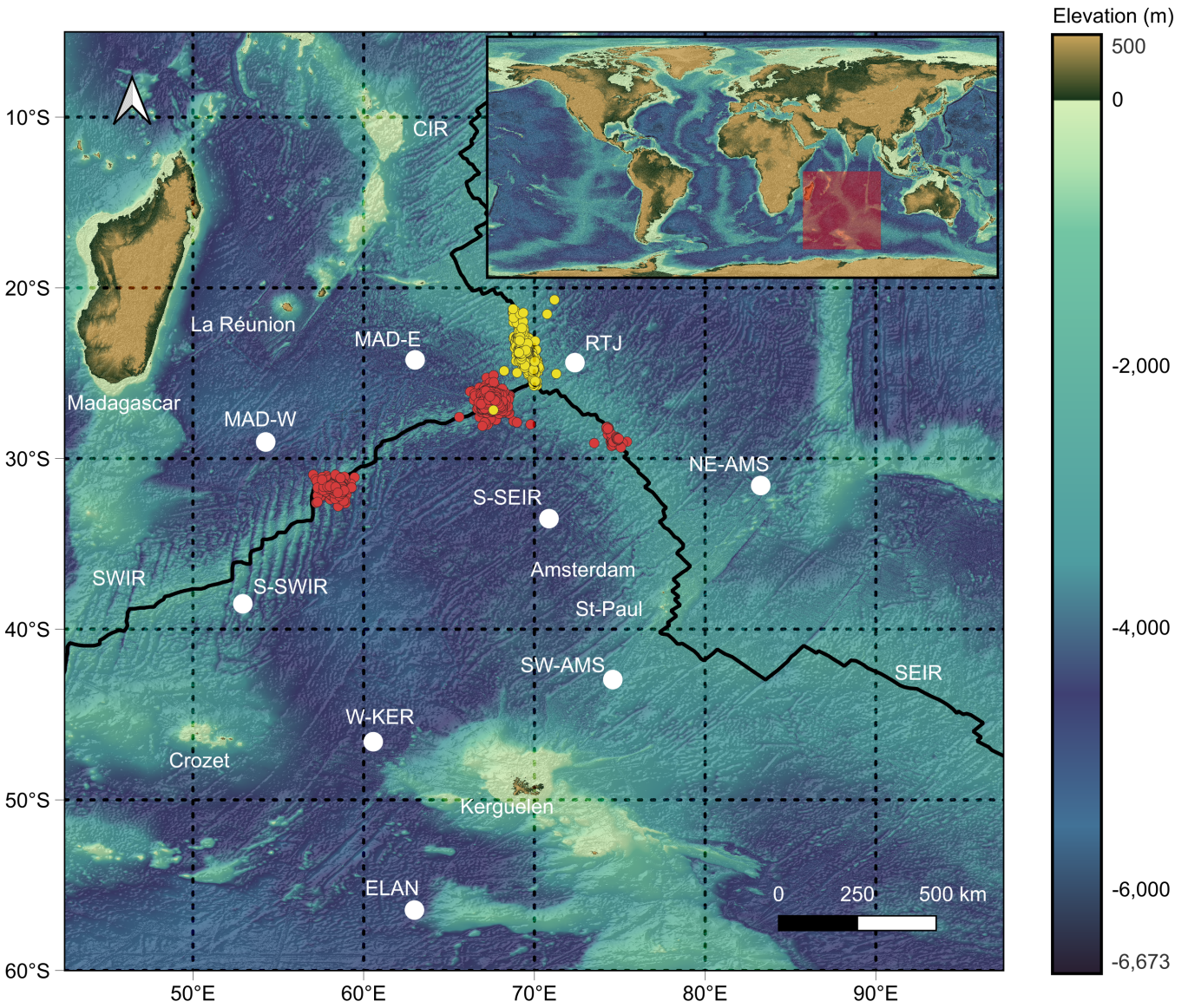

Figure 1 Location map of the OHASISBIO hydrophone network (white circles) in the Southern Indian Ocean. The OHASISBIO-2018 dataset includes recordings from all the stations shown in the map, while OHASISBIO-2020 only covers ELAN, MAD-W, NE-AMS, RTJ, SW-AMS and W-KER. Continuous black lines mark the three mid-oceanic ridges, which are responsible for most of the geophonic signals recorded. SWIR: Southwest Indian Ridge; CIR: Central Indian Ridge; SEIR: Southeast Indian Ridge. The red dots correspond to events belonging to the three swarms analysed by Ingale et al. (2021), used as a starting point to establish the OHASISBIO-2018 dataset. The yellow dots correspond to seismic events belonging to a 2020 swarm, used to select the optimal period of annotation for the OHASISBIO-2020 dataset. Bathymetry were taken from GEBCO grids (Kapoor, 1981).

3 Data collection

3.1 Hydroacoustic data

Our datasets are continuous recordings of pressure data from two remote and autonomous hydrophone networks: the OHASISBIO array, deployed in the southern Indian Ocean in 2009 and the HYDROMOMAR array, deployed in the central Atlantic Ocean in 2010. All hydrophones were moored in the SOFAR channel at an average depth of ∼1100 m, were recording at 240 Hz, and stored data with a 24-bit resolution.

The OHASISBIO network, operated in the Indian Ocean between 2009 and 2023 (Royer, 2009), consisted of up to 9 stations, spaced thousands of kilometers apart (Figure 1). Its primary purpose was to monitor seismic and volcanic activity of the three Indian mid-oceanic ridges and the biophony in the southern Indian Ocean. The first dataset provided in this paper, referred to as OHASISBIO-2018, covers about 9 months from 14 February 2018 to 3 November 2018. It was chosen because of the annotations already available from Ingale et al. (2021). The dataset includes recordings from 9 stations: ELAN, MAD-E, MAD-W, NE-AMS, RTJ, S-SEIR, S-SWIR, SW-AMS and W-KER. Recordings at MAD-E and W-KER ended on 21 July 2018 due to battery failure. The second dataset, named OHASISBIO-2020, includes 6 days from 4 to 10 June 2020, with recordings from 6 sites: ELAN, MAD-W, NE-AMS, RTJ, SW-AMS and W-KER. This period has been chosen because of its dense seismic activity

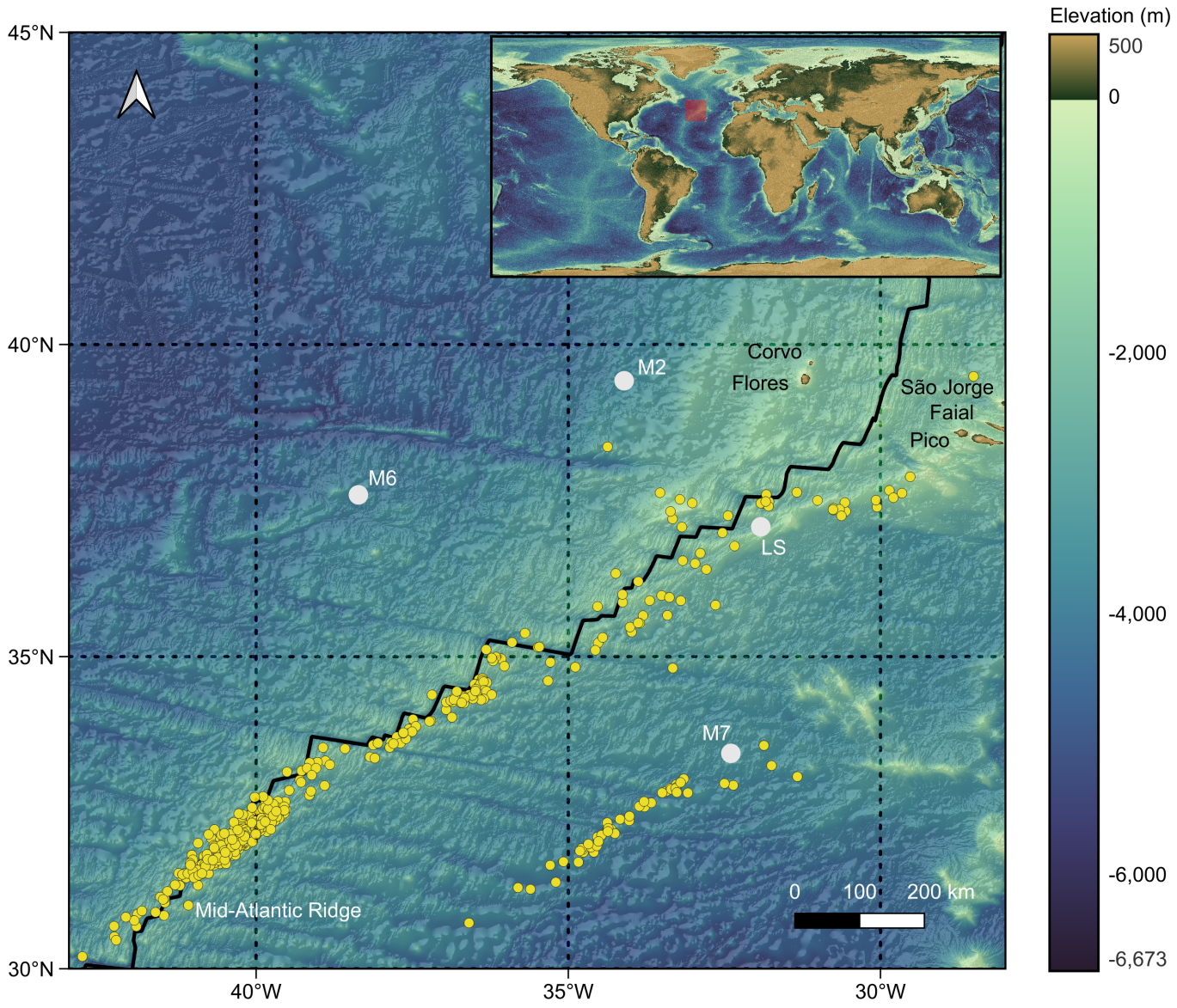

Figure 2 Location map of the HYDROMOMAR hydrophone network, south of the Azores Archipelago in the Central Atlantic Ocean. The HYDROMOMAR-2013 dataset includes recordings from the 4 sites shown on the map (LS, M2, M6, M7). The black line marks the mid-Atlantic ridge, which is responsible for most of the geophonic signals recorded. The yellow dots correspond to seismic events that occured in 2013 (Giusti, 2019), used to select the optimal period of annotation for the HYDROMOMAR-2013 dataset. Bathymetry were taken from GEBCO grids (Kapoor, 1981).

revealed by previous annotations. Its relevance in the context of geodynamic studies made it a relevant candidate for a representative benchmarking use-case.

The HYDROMOMAR network, operated in the Atlantic Ocean between 2010 and 2020 (Perrot, 2010), consisted of up to 5 stations, spaced hundreds of kilometers apart south of the Azores Archipelago (Figure 2). It had a similar objective of monitoring the seismic activity of the mid-Atlantic ridge. The third dataset, hereinafter referred to as HYDROMOMAR-2013, spans 11 to 17 March 2013 and includes recordings from 4 stations LS, M2, M6 and M7. This period has been chosen because of its dense seismic activity revealed by Giusti (2019).

3.2 Annotations

The OHASISBIO-2018 dataset was annotated in an analysis of events along the Southwest Indian Ridge, resulting in a dataset containing 6,767 manually picked events associated with either underwater earthquakes or volcanic eruptions (Ingale et al., 2021). However, it is crucial to note that this annotation, focused on three large seismic swarms bounded temporally and spatially, is incomplete and cannot serve as a comprehensive theoretical ground truth annotation. Consequently, in a binary classification task, all picked events have been considered as positive identifications (i.e., of geophonic nature), while the remainder of the dataset cannot be designated as negative (i.e., void of any events). To address this challenge and obtain negative samples, we performed a random sampling of 200 s-long windows across the entire recording period for all stations. A manual inspection of these windows allowed identification of 2,462 additional positive signals and 24,690 segments considered free of any geophonic signal. This process resulted in a dataset comprising 9,229 double-

Dataset	Duration	Annotation type	Complete positive events	Conservative positive events	Negative 200 s-windows	Total annotations (including negatives)
OHASISBIO-2018	~50,000 h	Partial (from literature)	9,229	N/A	24,690	33,919
OHASISBIO-2020	864 h	Full coverage (APLOSE)	3,838	1,073	28,053	31,891
HYDROMOMAR-2013	576 h	Full coverage (APLOSE)	3,059	1,003	18,517	21,576

Table 1 Datasets summary, providing the duration, coverage of manual annotation and number of annotations for each of the published datasets.

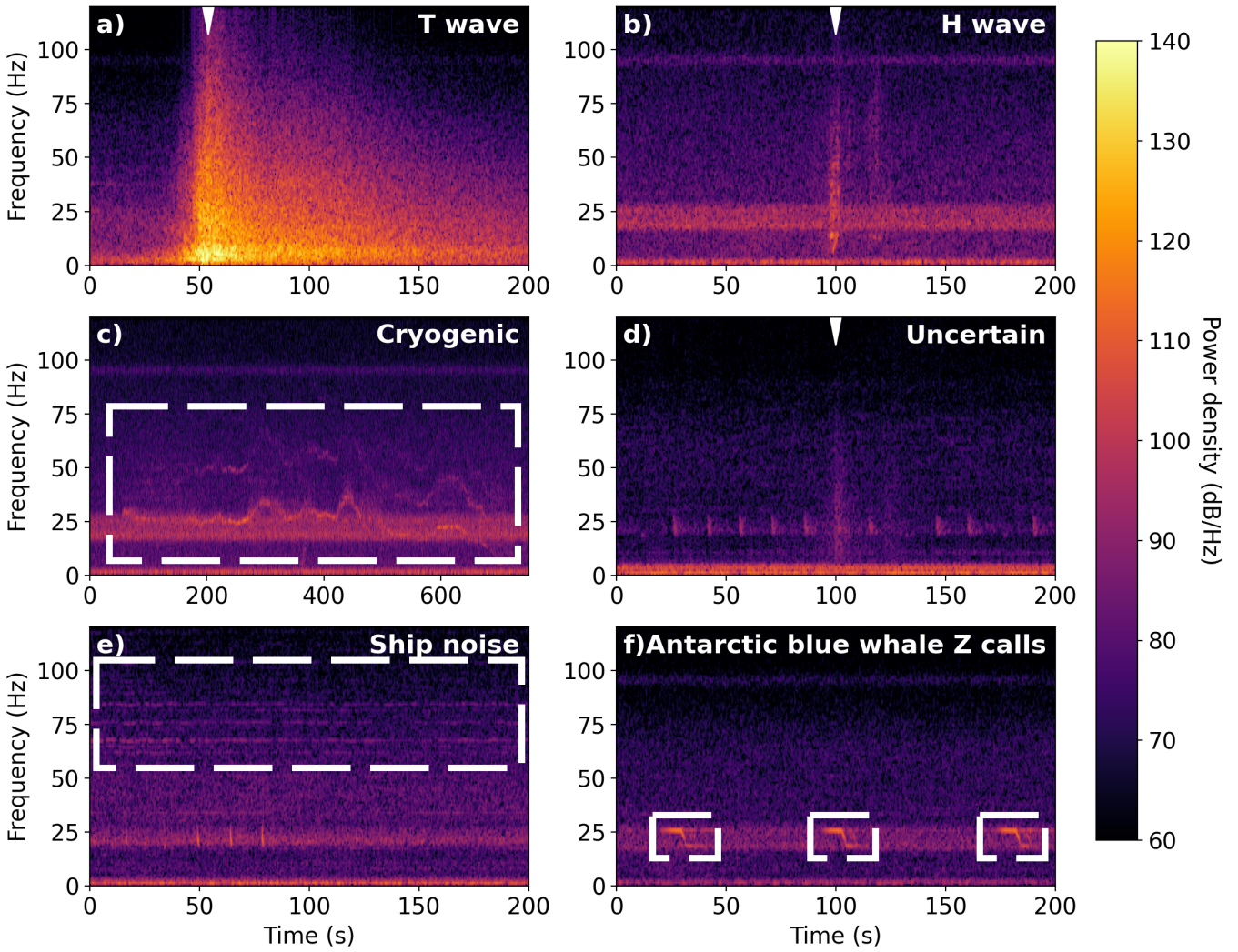

Figure 3 Spectrograms of typical signals observed in either OHASISBIO-2018, OHASISBIO-2020, or HYDROMOMAR-2013 dataset. a) T-wave recorded by RTJ in 2020; b) H-wave recorded by WKER2 in 2020; c) Cryogenic signal recorded by ELAN in 2020 (time axis different from others); d) event designated as "uncertain" by the annotators recorded by LS in 2010, surrounded by several whale calls; e) ship noise showing as horizontal lines recorded by M6 in 2010; f) Antarctic blue whale Z-calls recorded by SSEIR in 2018. White arrows and white rectangles show the signal of interest ToA and time-frequency bounds. Spectrograms are normalized between 60 and 140 dB.

checked positive events and 24,690 known negative or void segments.

OHASISBIO-2020 and HYDROMOMAR-2013 were annotated with the goal of a comprehensive detection of seismic events. Given this property, and the fact that these datasets represent temporal and geographical contexts different from OHASISBIO-2018, we consider them particularly suitable to serve as evaluation datasets. Table 1 summarizes the datasets used in this work, along with the annotations they contain.

For this study, manual annotations were performed by 5 different people using the open source annotation platform APLOSE (Annotation PLatform for Ocean Sound Explorers, Keribin et al. (2024), supplementary

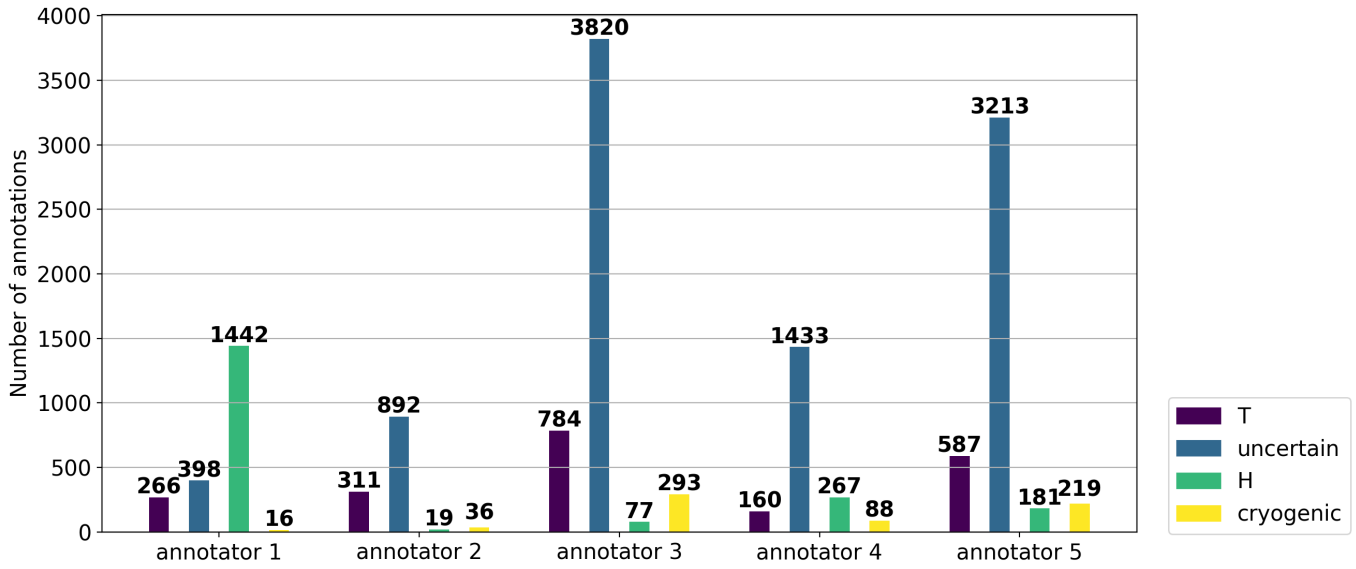

Figure 4 Number of annotations of each label in the annotation catalog produced by different people. T label corresponds to T-waves and H label corresponds to impulsive, magmatic events.

information S1 and Figure S1). This group included 4 expert annotators, who already participated in manual picking sessions which led to publications. We thus consider them as experienced. The last person had only seen a few geophonic signals, and is thus qualified as novice. Each annotator was presented with 2,000-second window spectrograms, generated using a STFT with a Hamming window and segments of 256 points overlapped by 50%. The 2,592 spectrograms were sequentially displayed one by one. The annotation labels were "T" for T-waves or hydroacoustic P-waves if any, "H" for H-waves, and "uncertain" when the level of ambiguity between T-, hydroacoustic P- or H-waves was deemed too high by the annotator. In addition, although it was not the main focus of this work, numerous cryogenic events associated to long-duration ice tremors (as in Royer et al., 2015; Leroy et al., 2018a) were recognized and a generic label "cryogenic" has therefore also been made available to the annotators. All signals, except cryogenic signals, were annotated with a unique sampling time corresponding to the estimated maximum of energy. Cryogenic signals have been surrounded by time-frequency boxes due to the difficulty of choosing a precise time in long signals. Figure 3 shows some spectrogram examples of signals belonging to the datasets. 14,502 annotations have been produced by the 5 annotators, with an average annotation time of 1 hour per station, resulting in a total working time of around 10 hours per annotator. The annotations made by the novice annotator turned out to be very conservative, after a random inspection of some of them. Indeed, only the "obvious" events were selected. This led us to consider this set of annotations the same as the others, as we were confident that it would not add many false positives.

The variability of the 14,502 resulting annotations was then analysed. Figure 4 shows the absolute number of annotations among the available labels. The corresponding relative values are given in supplementary information S2 and Figure S2. Figure 4 shows that most annotators often chose the "uncertain" label (~70% of the annotations), except for one annotator. This suggests that visually distinguishing between T-waves, hydroacoustic P-waves and H-waves can be challenging, leading conservative annotators to select this label. Examining the distribution of T and H labels among events not labeled as "uncertain", it appears that T is the most frequent label for three annotators, while H is the most frequent for the remaining two. Cryogenic signals amounted to less than 5% of the total annotations, underscoring the prevalence of volcano-tectonic events during the studied periods. This predominance in OHASISBIO may be attributed to the fact that the OHASISBIO-2020 data were recorded in June, outside the austral summer period, when cryogenic noise is known to be particularly significant in the southern Indian ocean (*e.g.,* Royer et al., 2015; Leroy et al., 2018a). In the HYDROMOMAR dataset, which was not in the summer period, cryogenic signals have not been particularly studied or noticed. For further analysis, a figure showing the distribution of labels depending on the stations is available as Figure S3. We simply observe that MAD-W at 26°S provided fewer cryogenic annotations than ELAN at 56°S, which is closer to cryogenic event sources.

We grouped same-label annotations that were temporally close to each other using a time-window of $\pm$ 10 s around the annotations. For cryogenic events, due to the time-frequency boxes annotation method, a group was formed when two boxes had a non-zero intersection. Figure 5 illustrates the distribution of group sizes for each label. The results show that there is more consensus in identifying cryogenic signals than other signals, with as many groups containing two or more annotations as singleton groups. In contrast, T and H signals are primarily identified by one annotator at a time. A visual inspection of these singletons often indicated a scenario where one annotator selected it as T-wave, whereas one or several others labeled it as "un-

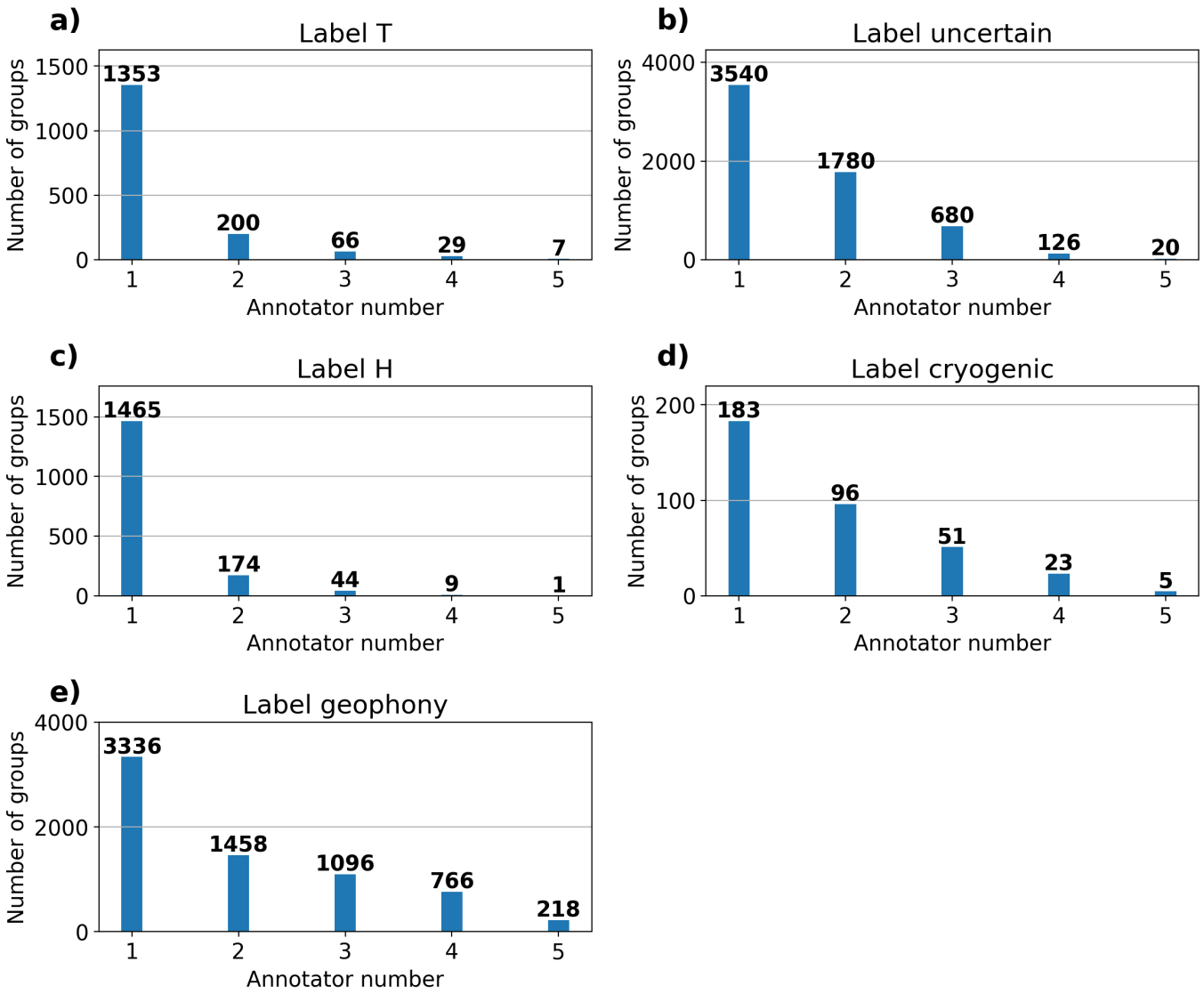

Figure 5 Distribution of groups as a function of size. The annotations from the different annotators were grouped together according to their temporal proximity. The resulting histogram shows the number of groups of each size, with the size varying from 1, for singletons (events seen by a single annotator), to 5, for groups with complete agreement between all 5 annotators. The "geophony" label merges all the others except cryogenic.

certain". Consequently, a new group-finding algorithm was employed, allowing annotations with different labels (among "T", "H" and "uncertain") to be grouped together if they were temporally close (still within 10 s), resulting in heterogeneous groups. The outcome, labeled "geophony" according to the definition given in the introduction, shows much more consensus among annotators.

To streamline the analysis and obtain a final catalog, these groups have been recorded in a catalog. Labels "T" and "H" have been assigned for groups of size two or more, when all annotators agreed on the label. All other groups except cryogenic were assigned to "uncertain". For all these labels, the event time of each group was determined by averaging the picked time of each annotation within the group. For cryogenic annotations, overlapping boxes were averaged, and their starting and ending times were retained in the resulting catalog. The process resulted in a catalog containing: 81 events labeled as T, 32 events labeled as H, 6,784 events labeled as uncertain and 359 time segments labeled as cryogenic.

Figures 4 and 5 highlight the great difficulty for annotators to distinguish labels "H" and "T". This difficulty and the methodology applied to obtain groups led to the high prevalence of "uncertain" events. Two annotators were convinced of seeing a substantial number of "H" events, while the others often disagreed, preferring to label the same events as "uncertain" or even "T" in some cases. Previous analyses of geophonic hydroacoustic events often overcame potential ambiguities with an a posteriori classification. For example, Bazin et al. (2022) compared the positions of hydroacoustic events with that of successive lava flows mapped by bathymetric surveys as evidence for a volcanic origin. Ingale et al. (2021) used the position and the spatio-temporal distribution of hydroacoustic events as criteria to determine the nature of their sources. This apparent difficulty to decipher "H" from "T" label led us to include them in a common label "geophony" for the detection part of this work. All events from "T", "H" or "uncertain" class were

considered as belonging to this class, leaving their exact classification to later stages of contextualization not covered in this work. Thus, the classification step mentioned thereafter is of a binary nature.

The resulting catalog also recorded the number of annotations merged to obtain each event, the number of annotations of the same class, and the number of different annotators participating in this group. This last number accounts for the agreement among the annotators. A qualitative visual inspection showed that all events recognized by at least three annotators had a high signal-to-noise ratio, while events only seen by one annotator were often difficult to recognize, even though they may still be of geophonic nature. As the objective of this work is to develop methods to detect any signal that may be of geophonic nature, even these "non-obvious" events have been kept in the catalog, leaving their exact classification to further steps. Some examples of signals seen by one and three annotators are available in supplementary figures S4 and S5.

The primary focus of this work was the detection of possible T-, hydroacoustic P- and H-waves, collectively referred to as geophony. For the three datasets, two tasks were formulated:

- Task A: given a window of a fixed duration of 100s, determine whether or not it contains at least one geophonic signal.
- Task B: given a window of a fixed duration of 100s, containing at least one geophonic signal, estimate the time instant of the(se) signal(s).

In practice, with the aim of establishing the most complete seismic-event catalog possible from unlabeled data, a technique capable of solving Task A can be used to determine the windows of interest and then, a technique capable of solving Task B can be applied to the windows of interest to obtain a comprehensive catalog. For both techniques, we analysed the datasets OHASISBIO-2020 and HYDROMOMAR-2013 in two ways: a "complete" dataset where all geophony events are considered, and a "conservative" dataset where only geophony events annotated by at least three annotators are considered. This enables us to evaluate both the ability of the models to detect uncertain events and their reliability at detecting "obvious" ones, that are unlikely to be missed by humans. Moreover, this also enables to assess the impact of the annotation aggregation method on the model evaluations.

4 Automatic detection methods

To address tasks A and B, four models were evaluated, three of which were considered for benchmarking, with the last one proposed as an original model. We considered time-windows of 100 s of the two classes, except for a particular model. In training mode, positive windows were uniformly and randomly sampled between -45 and +45 seconds around known geophonic events. This random sampling was implemented to prevent models from learning to focus exclusively on the central portion of the signal. Negative windows were extracted in the middle of negative ranges given by the annotations.

4.1 Benchmarking existing models

4.1.1 Stochastic Gradient Boosted Trees (SGBT)

SGBT is a boosting technique that involves growing multiple trees of limited length and aggregating them to form a complex model. Trees are stacked one at a time, and designed to reduce the overall loss. Sukhovich et al. (2014) applied this technique on features obtained through a DWT to classify hydroacoustic signals into P-waves, T-waves, iceberg-generated or ship-generated. The objective was to provide a low-cost detection solution, in terms of resources and time. The DWT and the SGBT training or execution times are minimal, often less than one second on a laptop for the use case detailed in this work. This aspect is relevant for embedded systems in recording instruments such as MERMAID (Simons et al., 2009).

The process begins with the application of DWT which is performed on the raw acoustic time series, retaining eight scales and using biorthogonal wavelets, which are two wavelets with vanishing moments of respectively two and four. The energy of each scale is averaged over time, and the relative importance of each scale is used as a feature for training the SGBT. The trees in the SGBT model are constrained to a depth of 3.

The training process of SGBT models can be tuned by modifying two hyperparameters, the learning rate and the maximal number of trees. The Tree Parzen Estimator algorithm has been used to perform Bayesian optimization, an optimization paradigm leveraging a priori knowledge about parameters and enabling refinement of the guessed distribution of the optimal ones (Bergstra et al., 2011). The learning rate was selected from an a priori logarithmically uniform distribution in the range [0.01, 1] and the number of trees was sampled from an a priori uniform distribution in the range [1, 1000]. 10,000 iterations were performed to estimate the optimal hyperparameters using the hyperopt library (Bergstra et al., 2013).

It is worth noting that, unlike other models, SGBT does not directly leverage the time-series nature of the data but rather uses averaged features. This approach dilutes small signals when dealing with 100 s time-windows. To address this issue, this work also considered 20 s time-windows as input data, using random shifts of $\pm$ 9 seconds instead of 45. The model was implemented with the scikit-learn library version 1.3.1 (Pedregosa et al., 2011).

4.1.2 ResNet-50

ResNet-50 is a CNN architecture that incorporates skip-connections, obtaining several properties such as the ability to mitigate the vanishing gradient problem during back-propagation through the network. It demonstrates high accuracy on the ImageNet classification task, achieving a Top-1 accuracy of approximately 75% (He et al., 2016). Notably, the model has proven effective for transfer learning on diverse tasks using images,

making it a popular choice for such applications. In this work, we use the weights obtained by the network after training on the ImageNet database. Log-spectrograms are generated from raw acoustic time-series using 256 points per segment, a 50% overlap, and a Hamming window. The spectrograms undergo normalization, with minimum and maximum values empirically set to -35 dB and 140 dB, respectively. The spectrograms are saved as grayscale images with a size of 186x129x1. The ResNet pipeline used in this work initially reshapes the spectrograms to a size of 224x224x3 with linear interpolation and duplication for the channels and applies the standard preprocessing steps of ResNet-50. All layers of the model, except the last convolutional layer, are frozen. Subsequently, following a MaxPooling operation, five layers of fully connected neurons with 50% dropout are applied. The last layer uses a sigmoid activation function to perform a binary classification task. The network comprises 2,793,473 trainable parameters, primarily concentrated in the fully connected layers. A stochastic gradient descent algorithm with batches of size 64 was used for training, with a binary cross-entropy loss. The binary cross-entropy is a notion borrowed from information theory, used in this context to quantify the divergence between two probability distributions: the ground truth values and the network outputs. The binary cross-entropy $H(p, q)$, with p and q being two distributions on X, is expressed as follows:

$$H(p,q) = \sum_{x \in X} -p(x)\, log(q(x)) - (1 - p(x))\, log(1 - q(x)) \tag{1}$$

This model, like other neural networks in this work, was implemented with the Tensorflow library (Abadi et al., 2015).

4.1.3 AcousticPhaseNet

PhaseNet is a U-Net-like CNN, designed with one-dimensional convolutions, specifically for seismic data analysis. Its original purpose is to estimate the probability of the presence of P- and S-waves at each time step (Zhu and Beroza, 2019), a mechanism that we call time segmentation. Thus, for both P- and S-waves, the neural network learns a function $f : [0, 1]^T \rightarrow [0, 1]^T$ where T is the time dimension of the input signal. To do so, the network compresses the data using 1D convolutions, a process called downsampling, before expanding such that the input and output have the same shape, a process called upsampling. Residual connections are added between the downsampling part and the upsampling part, which causes PhaseNet to be a U-Net-like model. For our work, PhaseNet potentially addresses task B. Additionally, by extracting the maximum output of the network for a given time-window, the model results can be simplified from time segmentation to binary classification results, allowing for evaluation on task A.

The input consists of raw waveform values from the provided time-windows. The data is normalized such that each sample corresponds to an instantaneous sound pressure level between -35 dB and 140 dB. This leads to 24,000 values, stored on four bytes, which are then linearly resampled to 32,768 values, aligning with a power of two.

Similarly to the PhaseNet approach, the ground truth is composed of zeroes at each time step, except around known events where an absolute-value function is employed, reaching a width of 10 s. The width is chosen to cover twice the approximate time resolution of the tool used for annotations of the OHASISBIO-2018 dataset, preventing a high penalization of the model in case the annotation pick time is not exact.

For our purpose, modifications were made to adapt the model to hydroacoustic data. Firstly, the number of input channels was reduced from three (N–S, E–W, vertical movement) to one (pressure). Secondly, two more upsampling and downsampling blocks were added to account for the higher sampling rate and longer duration of the time-windows, resulting in many data points. The resulting network, called AcousticPhaseNet, has 118,242 trainable parameters, distributed among 1D convolutions.

A stochastic gradient descent algorithm with batches of size 64 was used for training, with a binary cross-entropy loss. To face the difficulty posed by the 32,768 output values, the loss was simply taken as the sum of individual cross-entropies of each time step. Overall, the formula used can be summed up as follows, with X the sample space and $t_i \in x$ the i_{th} time step of the sample $x \in X$:

$$H(p,q) = \sum_{x} \sum_{t_i \in x} -p(t_i)\, log(q(t_i)) - (1 - p(t_i))\, log(1 - q(t_i)) \tag{2}$$

4.2 A custom model: Time Spectrogram Segmentation Network (TiSSNet)

The rationale behind PhaseNet and the usage of spectrograms in the acoustic community led us to create a new Fully Convolutional Network (FCN). This FCN takes spectrograms as inputs and provides an estimation of the probability of an event presence at each time step, thus performing time segmentation as defined for PhaseNet and learning a function $f : [0, 1]^{T \times F} \rightarrow [0, 1]^T$ where T and F are the time and frequency dimensions of the spectrograms. The network incorporates successive two-dimensional convolutions with an asymmetric stride, dividing the number of bins along the frequency axis by two or four at each step while keeping the number of bins along the time axis unchanged. Following several downsampling blocks, when the data only has one bin along the former frequency axis left, a sigmoid function is applied to ensure that the output for each time bin lies in the interval [0,1]. This output can be interpreted as the probability of the presence of a signal of interest as a time series, represented in one dimension.

Similar to ResNet, the input data is generated from raw acoustic time-series and transformed into log-spectrograms. For this model, the input spectrograms are linearly reshaped to contain 128 frequency bins, enabling to optimally use powers of two as strides, where strides refer to the number of pixels separating neighboring kernel applications in the convolution.

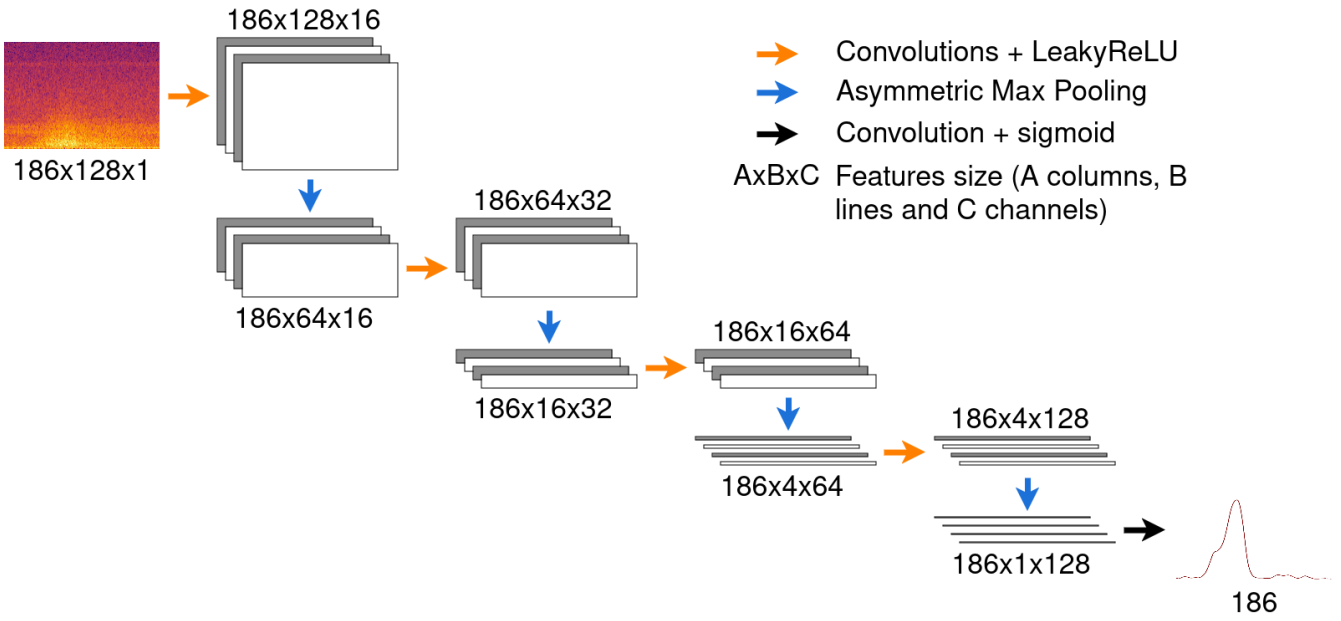

Figure 6 The proposed TiSSNet network architecture is designed to process a 100 s spectrogram with 128 bins of frequency. The orange arrows represent series of three stacked convolutions, using LeakyReLU activation function and a kernel varying with the features size, keeping a width of 8 but a height of 8, 5, 3 and 2 from left to right. Blue arrows represent a Max Pooling layer with asymmetric kernels, keeping a width of 1 and a height of 2, 4, 4 and 4 from left to right. The last convolutional layer, in black, applies a sigmoid on the data to output a result in the range [0,1].

	SGBT (20s)	SGBT (100s)	ResNet-50	AcousticPhaseNet	TiSSNet
AuC (OHASISBIO-2018)	0.9357	0.9058	0.9534	0.9700	**0.9812**
AuC (OHASISBIO-2020 "complete")	0.8495	0.7744	0.8090	**0.9357**	0.9264
AuC (OHASISBIO-2020 "conservative")	0.8903	0.8184	0.9092	**0.9775**	0.9684
AuC (HYDROMOMAR-2013 "complete")	0.9012	0.8435	0.8557	0.9018	**0.9463**
AuC (HYDROMOMAR-2013 "conservative")	0.9348	0.8843	0.9254	0.9605	**0.9725**
Trainable parameters	/	/	2,793,473	118,242	1,038,161
Training time	**321 ms**	**321 ms**	1h22	1h10	2h55

Table 2 Summary results, giving Task A scores (expressed in term of AuC of ROC), number of parameters, measured training time and the resources used for each of the models. Task A refers to the binary classification of time-windows, which can either contain a geophonic event or not. The trainable parameters consist of convolutional kernel and fully connected weights and biases, thus being non applicable for the SGBT model. For each line, the best score, when applicable, is written in bold.

The ground truth values are computed as in the AcousticPhaseNet case. The architecture of the network is described in Figure 6. The network has 1,038,161 trainable parameters, distributed among 2D convolutions.

A stochastic gradient descent algorithm with batches of size 64 was used for training, with a binary cross-entropy loss as defined by equation 2.

5 Evaluation of models for automatic detection

At inference time, the models are applied on consecutive time-windows, covering all periods of this study. Receiver Operating Characteristic (ROC) curves, within the context of binary classification, illustrate the performance of models by plotting the True Positive (TP) rate as a function of the False Positive (FP) rate. Given that binary classifiers generally output a value within the interval $[0, 1]$, a threshold has to be chosen to take a binary decision. By varying this threshold, different TP and FP rates can be determined and thus the ROC curve can be plotted. The Area under the Curve (AuC) of ROC curves serves as a performance metric for classifiers, less arbitrary than common metrics such as accuracy or recall measured with a fixed threshold. The AuC score ranges between 0 and 1 with a higher value indicating better classification performance. Table 2 presents the AuC scores of each classifier on each of the three datasets (two of which being also considered in conservative configuration). The number of trainable parameters for each model and the associated training time on the OHASISBIO-2018 dataset are also given. Training times were estimated based on a single training of

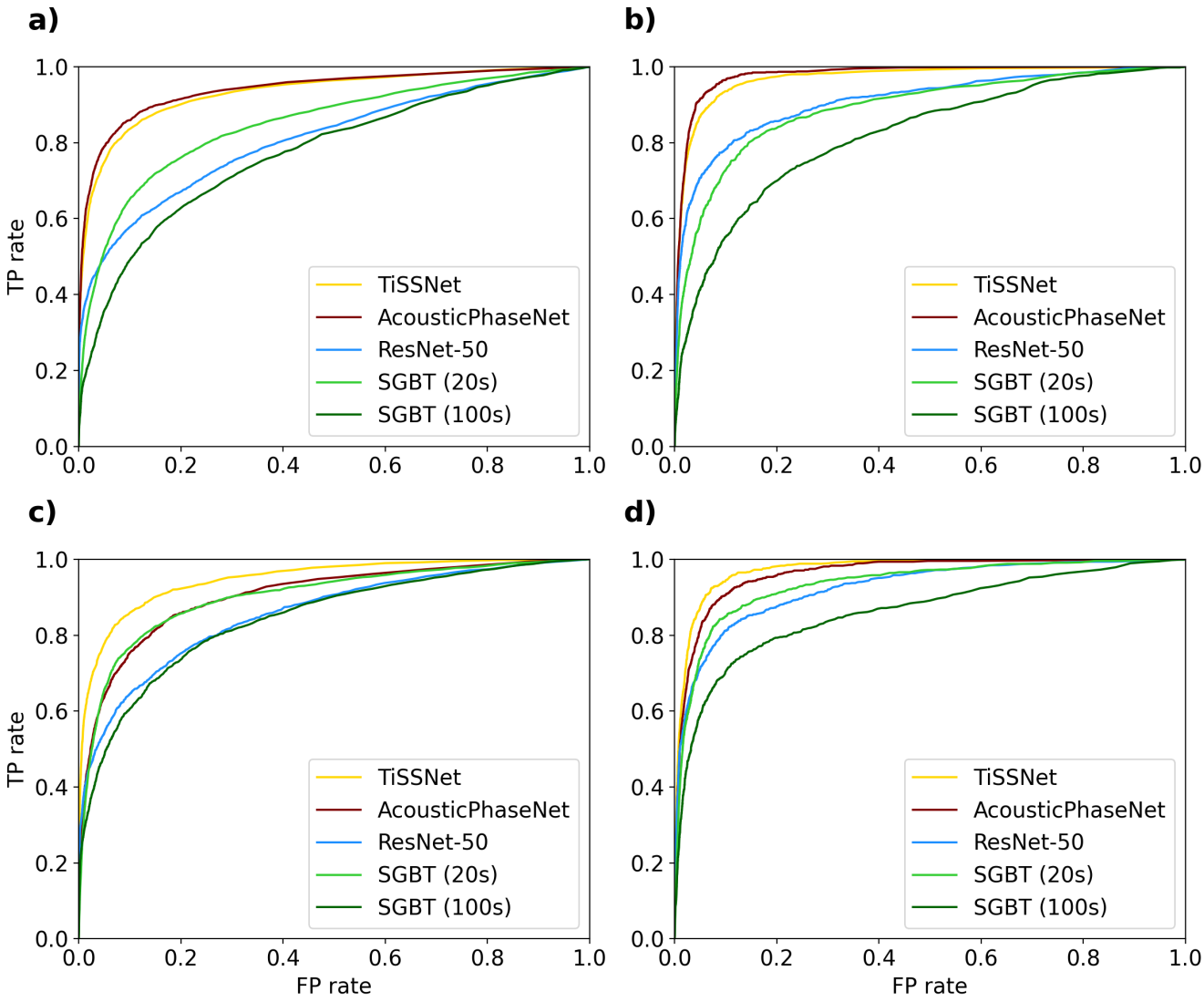

Figure 7 ROC curves of the models, showing True Positive (TP) rate as a function of False Positive (FP) rate, applied to the a) complete and b) conservative OHASISBIO-2020 datasets; c) complete and d) conservative HYDROMOMAR-2013 datasets.

150 epochs for the neural networks, while it was averaged on 1,000 consecutive trainings for the SGBT model. For OHASISBIO-2018, serving as a training set, a 5-FOLD cross-validation was employed to evaluate AuC before training the models on the entire dataset for evaluation on OHASISBIO-2020 and HYDROMOMAR-2013.

Figure 7 shows the ROC curves for the different models, including TiSSNet, applied to the OHASISBIO-2020 and HYDROMOMAR-2013 dataset, in both complete and conservative cases.

To evaluate the results of task B, both time-segmenter models are applied to the complete evaluation datasets. A peak-finding algorithm implemented in scikit-learn is employed to convert time-wise probability estimations into distinct events. A maximum time difference of 10 s between two peaks is chosen arbitrarily. This decision aligns closely with the selected time resolution of the time-axis labels in the APLOSE annotation tool, and to the width of the absolute value function used for training those models as previously mentioned. Then, a selection criterion retains only the peaks closest to ground truth annotations, provided they are less than 10 s apart. Finally, the time differences between these peaks and their corresponding ground truth annotations are calculated for the time-segmentation models (Figure 8).

6 Discussion and conclusion

The main goal of this work was not only to provide models to deal with hydroacoustic data for seismology, but most importantly to propose a common benchmarking framework for the community. In addition to being made public, the annotated data can still be further improved and will be kept up to date. The resulting benchmarking framework, composed of datasets, a methodology and models, should help to evaluate the performance of new geophony detection tools.

Inconsistent human annotations have been discussed in section 3.2. Interestingly, our analysis of their annotations shows they had a great difficulty in distinguishing T- from H-waves. This may be attributed to the recent interest on H-waves, causing annotators to poorly know their characteristics. Further works focusing on these signals may lead to a better understanding

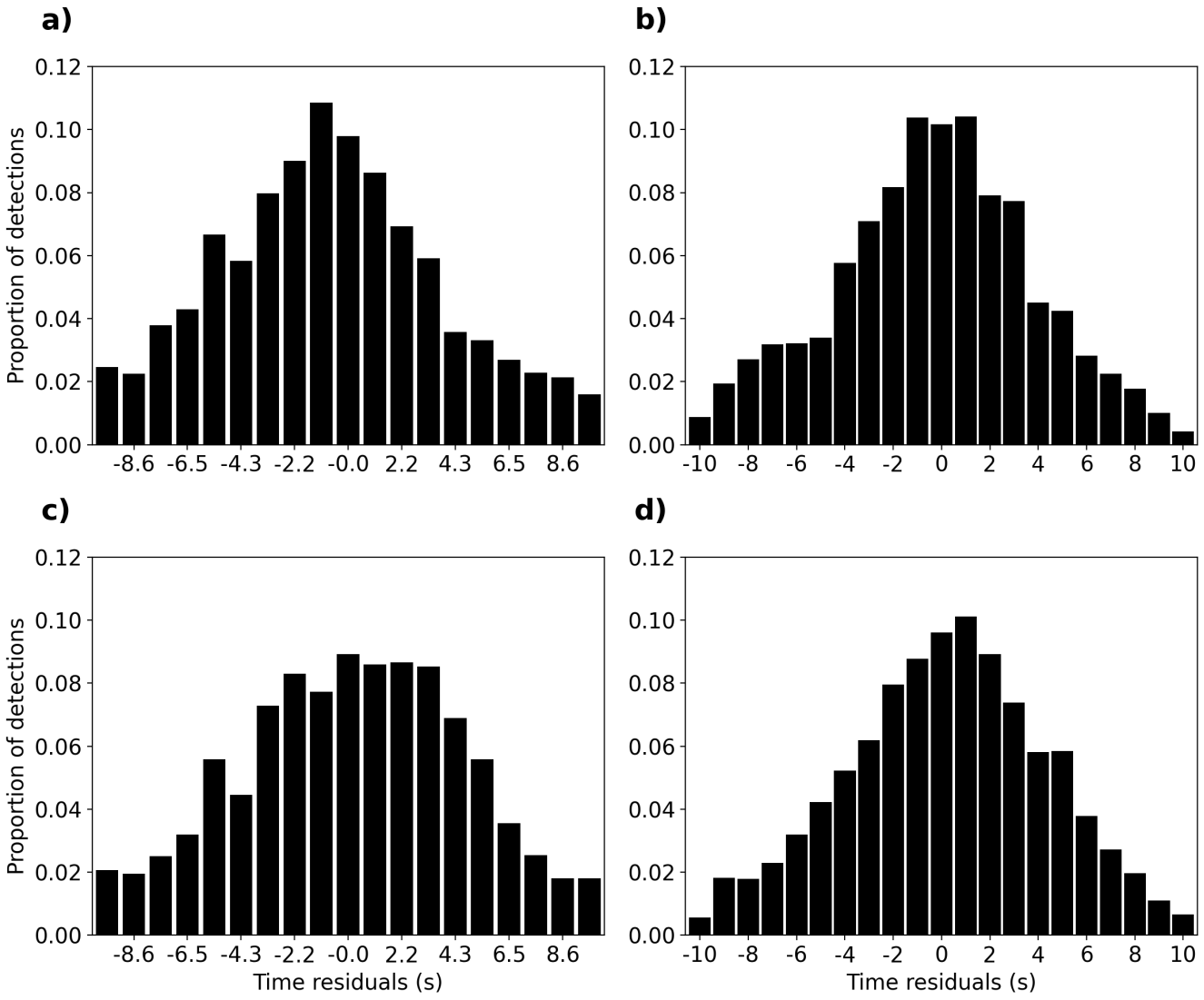

Figure 8 Distribution of pick errors of TiSSNet and AcousticPhaseNet on the two evaluation datasets: a) TiSSNet and b) AcousticPhaseNet on OHASISBIO-2020 dataset; c) TiSSNet and d) AcousticPhaseNet on HYDROMOMAR-2013. For a) and c), the x-axis increments are multiples of the spectrogram time resolution, which is close to 0.55 s.

of their generation mechanism and to the emergence of more examples to compare to. This could lead the way to the emergence of a multiclass catalog, instead of the current binary classification catalog. For these reasons, the annotation campaign will remain open and editable for any researcher interested in further verifying, adding and/or correcting current annotations upon request. This was actually a key feature of our annotation campaign, and part of the intended design of the APLOSE platform. Creating an annotation campaign kept open in the long term will favor the appropriation of our dataset by the community, in particular because the quality of annotations can thus be directly assessed by anyone. This could improve the coverage of the data and most importantly the reliability of the annotations.

Table 2 shows that the two time segmentation models, AcousticPhaseNet and TiSSNet, outperform the classification-based models (SGBT and ResNet-50). This superiority may be because these models implicitly encounter more negative samples during training, as positive windows selected for training are predominantly composed of negative time segments. Classification models, on the other hand, only see those windows as positive ones, without knowing which segment of the input is responsible for this label. SGBT exhibits lower AuC values than other models when using 100 s windows, particularly on the evaluation datasets. Given the much higher score of the model on 20 s windows, we attribute most of this difference to the dilution of the signal of interest in the long windows. Solving this problem would require more features in the SGBT inputs. For example, the 100 s inputs could be divided into five segments and the obtained features could be aggregated to a two-dimensional time-scale matrix. Another point is the low dimensionality of the SGBT input features, which possibly fail to capture the relevant discriminant information of the base signal. Yet, SGBT is more computationally efficient, with a training and evaluation time four orders of magnitude smaller than the other models. The ResNet-50 based model quickly overfits, likely due to the small size of the dataset, and a lighter architecture could lead to a lower training variance and

thus a better score. TiSSNet demonstrates a slightly better AuC than AcousticPhaseNet on the training set, although AcousticPhaseNet performs slightly better on the OHASISBIO-2020 dataset. AcousticPhaseNet, however, greatly falls behind the score of TiSSNet for HYDROMOMAR-2013 and thus fails to demonstrate a great generalization capability for this environment, geographically different from the training dataset environment. TiSSNet has the advantage of using a meaningful time-frequency representation, which offers coherent 2D support that is easier to work with convolutions than raw waveforms. However, this requires 2D kernels, resulting in a higher number of parameters than the 1D convolutions of AcousticPhaseNet. Moreover, working with 1D convolutions enabled AcousticPhaseNet to work in a U-Net-like architecture. In the end, this work both shows the applicability of PhaseNet to the hydroacoustic field, and the efficiency of a similar spectrogram-based model. Sukhovich et al. (2014), in their data analysis, obtained a true positive rate of about 98.7% with a false positive rate of about 1.3%, which is much more than that measured in OHASISBIO-2020 and HYDROMOMAR-2013. TiSSNet, for example, has a false positive rate of 23.5% when reaching a similar true positive rate. However, this difference is because Sukhovich et al. (2014) used the same set of samples to build the training and evaluation datasets. The different sensors were, moreover, very close to one another, limiting the diversity of soundscape induced by geographical variability. The score they obtained is thus comparable to the cross-validation performed in our work on the OHASISBIO-2018 dataset. Moreover, the classification model described in our work has been used in the same conditions as the other tested models.

Considering task B, TiSSNet and AcousticPhaseNet have a time residuals distribution roughly centered on zero for both evaluation datasets, with a smaller kurtosis in the case of HYDROMOMAR-2013 for TiSSNet (Figure 8). The residuals variance for TiSSNet can be attributed to model imprecision and the coarse-grained time-resolution of the spectrograms which cause the model to approximate the detection peak location to a neighboring time bin. Zhu and Beroza (2019) obtained, with a similar detection capability, a much more precise picking, with a deviation typically smaller than 0.5 s, thus 20 times more precise than that shown in Figure 8. The main reason we advance is that seismic phases often last a few seconds, while T-phases easily reach tens of seconds. Moreover, the uncertainty in the annotator pickings, especially given the high duration of spectrograms shown in APLOSE, may contribute to model incompressible time residuals. In the end, the overall precision of TiSSNet and AcousticPhaseNet meet the requirements of the field, but still leaves room for improvements.

The performances of the models, considering both tasks A and B, are as expected slightly better on events in the "conservative" dataset than those in the "complete" one. However, the ranking of models are even, showing a similar impact on them. Some examples of the ground truth events missed by all models in task A have also been visually inspected. They did not show any notable pattern that could explain the common error. Examples on this topic are given in the supplementary materials, in figures S7 and S8.

Considering both tasks, it is important to note that improvements are still needed. One may argue that task B could be the target of non-parametric algorithms, that, for instance, could search for the maximum of energy in the considered time-window. The literature provides limited studies focused on this task, which awaits a more extensive investigation of this specific problem. Task A seems more challenging and leaves room for improvement as shown by the ROC curves in this work and the abundant literature focusing on acoustic classification. In the end, we propose a benchmarking framework consisting of a training dataset, two evaluation datasets, two tasks to solve and some models to compare with, including an original one. We are convinced that the proposed framework may ease the development of new models, because it was tailored to enable their evaluation, and because we proposed a state-of-the art framework open to improvements.

Acknowledgements

The authors wish to thank the captains and crew of RV Marion Dufresne, Thalassa and Le Suroit for the successful deployments and recoveries of the hydrophones of the OHASISBIO (Royer (2009), doi: 10.18142/229) and HYDROMOMAR (Perrot (2010), doi: 10.18142/263) experiments. The authors also thank the persons who participated in the annotation campaigns on APLOSE, and the APLOSE team itself for the helpful support. Pierre-Yves Raumer was supported by a fellowship from the University of Brest and from the Regional Council of Brittany (ARED), and through the Interdisciplinary Graduate School for the Blue Planet (Isblue), co-funded by ANR (ANR-17-EURE-0015). The authors also acknowledge Ifremer for providing some computer infrastucture (Datarmor, Dataref) used in this work, and, for the elevation grid used for the maps, the GEBCO Compilation Group (2023) GEBCO 2023 Grid (doi: 10.5285/f98b053b-0cbc-6c23-e053-6c86abc0af7b). Finally, the authors are grateful to the reviewers Guilherme de Melo and Matthias Pilot, and Seismica Editor Stephen Hicks for their insightful remarks and relevant suggestions, which contributed to improve this paper.

Data and code availability

The modules and notebooks used to generate the results may be found on Github at https://github.com/PYLRR/OHASISBIO_dataset and Zenodo at https://zenodo.org/records/10458857 (Raumer), where a README file explains the installation and usage procedure. A custom GUI enabling the exploration of the data is also shared in the same repository. The data used in this paper is available on Dataref at https://sextant.ifremer.fr/record/b618b24e-82f9-4b3b-9753-048e1f043ca6 (Raumer et al., 2024), hosted by Ifremer. They can be accessed through Datarmor, a computing infrastructure from Ifremer, or with a web browser.

Competing interests

Authors declare no competing interests of any type regarding this work.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, 2015. https://www.tensorflow.org/. Software available from tensorflow.org.

Bazin, S., Royer, J.-Y., Dubost, F., Paquet, F., Loubrieu, B., Lavayssière, A., Deplus, C., Feuillet, N., Jacques, É., Rinnert, E., et al. Initial results from a hydroacoustic network to monitor submarine lava flows near Mayotte Island. *Comptes Rendus. Géoscience*, 354(S2):1–17, 2022. doi: 10.5802/crgeos.119.

Bergstra, J., Bardenet, R., Bengio, Y., and Kégl, B. Algorithms for Hyper-Parameter Optimization. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. https://proceedings.neurips.cc/paper_files/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf.

Bergstra, J., Yamins, D., and Cox, D. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 115–123, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. https://proceedings.mlr.press/v28/bergstra13.html.

Bermant, P. C., Brickson, L., and Titus, A. J. Bioacoustic Event Detection with Self-Supervised Contrastive Learning. *bioRxiv*, pages 2022–10, 2022. doi: 10.1101/2022.10.12.511740.

Bouffaut, L., Dréo, R., Labat, V., Boudraa, A.-O., and Barruol, G. Passive stochastic matched filter for Antarctic blue whale call detection. *The Journal of the Acoustical Society of America*, 144(2):955–965, 2018. doi: 10.1121/1.5050520.

de Melo, G. W., Parnell-Turner, R., Dziak, R. P., Smith, D. K., Maia, M., Do Nascimento, A. F., and Royer, J.-Y. Uppermost mantle velocity beneath the Mid-Atlantic Ridge and transform faults in the Equatorial Atlantic Ocean. *Bulletin of the Seismological Society of America*, 111(2):1067–1079, 2021. doi: 10.1785/0120200248.

Dubus, G., Adam, O., Duc, P. N. H., Torterotot, M., and Cazau, D. Leveraging citizen science in manual annotation for deep learning in underwater passive acoustic studies. In *2023 Signal Processing Symposium (SPSympo)*, pages 1–5. IEEE, 2023a. doi: 10.23919/SPSympo57300.2023.10302716.

Dubus, G., Torterotot, M., Duc, P. N. H., Beesau, J., Cazau, D., and Adam, O. Better quantifying inter-annotator variability: A step towards citizen science in underwater passive acoustics. In *OCEANS 2023-Limerick*, pages 1–8. IEEE, 2023b. doi: 10.1109/OCEANSLimerick52467.2023.10244502.

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

Dziak, R., Bohnenstiehl, D., Matsumoto, H., Fox, C., Smith, D., Tolstoy, M., Lau, T., Haxel, J., and Fowler, M. P-and T-wave detection thresholds, Pn velocity estimate, and detection of lower mantle and core P-waves on ocean sound-channel hydrophones at the Mid-Atlantic Ridge. *Bulletin of the Seismological Society of America*, 94(2):665–677, 2004. doi: 10.1785/0120030156.

Fox, C. G., Radford, W. E., Dziak, R. P., Lau, T.-K., Matsumoto, H., and Schreiner, A. E. Acoustic detection of a seafloor spreading episode on the Juan de Fuca Ridge using military hydrophone arrays. *Geophysical Research Letters*, 22(2):131–134, 1995. doi: 10.1029/94GL02059.

Fox, C. G., Matsumoto, H., and Lau, T.-K. A. Monitoring Pacific Ocean seismicity from an autonomous hydrophone array. *Journal of Geophysical Research: Solid Earth*, 106(B3):4183–4206, 2001. doi: 10.1029/2000JB900404.

Giusti, M. *Apport des données hydroacoustiques pour l'étude de la sismicité de la dorsale médio-Atlantique nord*. Theses, Université de Bretagne occidentale - Brest, Mar. 2019. https://theses.hal.science/tel-02292753.

Giusti, M., Perrot, J., Dziak, R. P., Sukhovich, A., and Maia, M. The August 2010 earthquake swarm at North FAMOUS–FAMOUS segments, Mid-Atlantic Ridge: geophysical evidence of dike intrusion. *Geophysical Journal International*, 215(1):181–195, 2018. doi: 10.1093/gji/ggy239.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Ibrahim, A. K., Zhuang, H., Chérubin, L. M., Schärer-Umpierre, M. T., and Erdol, N. Automatic classification of grouper species by their sounds using deep neural networks. *The Journal of the Acoustical Society of America*, 144(3):EL196–EL202, 2018. doi: 10.1121/1.5054911.

Ingale, V. V., Bazin, S., and Royer, J.-Y. Hydroacoustic observations of two contrasted seismic swarms along the Southwest Indian ridge in 2018. *Geosciences*, 11(6):225, 2021. doi: 10.3390/geosciences11060225.

Ingale, V. V., Bazin, S., Olive, J.-A., Briais, A., and Royer, J.-Y. Hydroacoustic Study of a Seismic Swarm in 2016–2017 near the Melville Transform Fault on the Southwest Indian Ridge. *Bulletin of the Seismological Society of America*, 113(4):1523–1541, 2023. doi: 10.1785/0120220213.

Kapoor, D. C. General bathymetric chart of the oceans (GEBCO). *Marine Geodesy*, 5(1):73–80, 1981. doi: 10.1080/15210608109379408.

Keribin, E., Morin, E., and Vovard, R. APLOSE: a scalable web-based annotation tool for marine bioacoustics - public repository. doi: 10.5281/zenodo.10468000.

Kong, Q., Sobieraj, I., Wang, W., and Plumbley, M. Deep Neural Network Baseline for DCASE Challenge 2016, 2016.

Leroy, E. C., Royer, J.-Y., Bonnel, J., and Samaran, F. Long-term and seasonal changes of large whale call frequency in the southern Indian Ocean. *Journal of Geophysical Research: Oceans*, 123(11):8568–8580, 2018a. doi: 10.1029/2018JC014352.

Leroy, E. C., Thomisch, K., Royer, J.-Y., Boebel, O., and Van Opzeeland, I. On the reliability of acoustic annotations and automatic detections of Antarctic blue whale calls under different acoustic conditions. *The Journal of the Acoustical Society of America*, 144(2):740–754, 2018b. doi: 10.1121/1.5049803.

Luo, W., Yang, W., and Zhang, Y. Convolutional neural network for detecting odontocete echolocation clicks. *The Journal of the Acoustical Society of America*, 145(1):EL7–EL12, 2019. doi: 10.1121/1.5085647.

Matsumoto, H., Dziak, R., Mellinger, D., Fowler, M., Haxel, J., Lau, A., Meinig, C., Bumgardner, J., and Hannah, W. Autonomous hydrophones at NOAA/OSU and a new seafloor sentry system for real-time detection of acoustic events. In *OCEANS 2006*, pages 1–4. IEEE, 2006. doi: 10.1109/OCEANS.2006.307041.

Mesaros, A., Heittola, T., and Virtanen, T. A multi-device dataset for urban acoustic scene classification. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, pages 9–13, November 2018. https://dcase.community/documents/workshop2018/proceedings/DCASE2018Workshop_Mesaros_8.pdf.

Mousavi, S. M., Ellsworth, W. L., Zhu, W., Chuang, L. Y., and Beroza, G. C. Earthquake transformer - an attentive deep-learning model for simultaneous earthquake detection and phase picking. *Nature communications*, 11(1):3952, 2020. doi: 10.1038/s41467-020-17591-w.

Okal, E. A. The generation of T waves by earthquakes. *Advances in Geophysics*, 49:1–65, 2008. doi: 10.1016/S0065-2687(07)49001-X.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Perrot, J. HYDROMOMAR, Hydroacoustic observatory of the Mid-Atlantic Ridge (MOMAR area). Technical report, University of Brest, Brest, France, 2010. doi: 10.18142/263.

Rasmussen, J. H. and Širović, A. Automatic detection and classification of baleen whale social calls using convolutional neural networks. *The Journal of the Acoustical Society of America*, 149 (5):3635–3644, 2021. doi: 10.1121/10.0005047.

Raumer, P.-Y. A public benchmarking hydroacoustic dataset for geophonic signals detection task: code. doi: 10.5281/zenodo.10458857.

Raumer, P.-Y., Bazin, S., Cazau, D., Ingale, V. V., and Royer, J.-Y. Donnees hydro-acoustiques passives 240 Hz annotees en ocean Atlantique 2013 et ocean Indien Sud 2018 et 2020, 2024. doi: 10.12770/b618b24e-82f9-4b3b-9753-048e1f043ca6.

Retailleau, L., Saurel, J.-M., Zhu, W., Satriano, C., Beroza, G. C., Issartel, S., Boissier, P., Team, O., Team, O., et al. A Wrapper to Use a Machine-Learning-Based Algorithm for Earthquake Monitoring. *Seismological Research Letters*, 93(3):1673–1682, 2022. doi: 10.1785/0220210279.

Royer, J. OHASISBIO-Hydroacoustic observatory for the seismicity and biodiversity in the Indian Ocean. Technical report, University of Brest, Brest, France, 2009. doi: 10.18142/229.

Royer, J.-Y., Chateau, R., Dziak, R., and Bohnenstiehl, D. Seafloor seismicity, Antarctic ice-sounds, cetacean vocalizations and long-term ambient sound in the Indian Ocean basin. *Geophysical Journal International*, 202(2):748–762, 2015. doi: 10.1093/gji/ggv178.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Saurel, J.-M., Retailleau, L., Deplus, C., Loubrieu, B., Pierre, D., Frangieh, M., Khelifi, N., Bonnet, R., Ferrazzini, V., Bazin, S., et al. Combining hydro-acoustic sources and bathymetric differences to track the vent evolution of the Mayotte eruption, Mozambique Channel. *Frontiers in Earth Science*, 10:983051, 2022. doi: 10.3389/feart.2022.983051.

Simons, F. J., Nolet, G., Georgief, P., Babcock, J. M., Regier, L. A., and Davis, R. E. On the potential of recording earthquakes for global seismic tomography by low-cost autonomous instruments in the oceans. *Journal of Geophysical Research: Solid Earth*, 114 (B5), 2009. doi: 10.1029/2008JB006088.

Sukhovich, A., Irisson, J.-O., Simons, F. J., Ogé, A., Hello, Y., Deschamps, A., and Nolet, G. Automatic discrimination of underwater acoustic signals generated by teleseismic P-waves: A probabilistic approach. *Geophysical research letters*, 38(18), 2011. doi: 10.1029/2011GL048474.

Sukhovich, A., Irisson, J.-O., Perrot, J., and Nolet, G. Automatic recognition of T and teleseismic P waves by statistical analysis of their spectra: An application to continuous records of moored hydrophones. *Journal of Geophysical Research: Solid Earth*, 119(8):6469–6485, 2014. doi: 10.1002/2013JB010936.

Tepp, G. and Dziak, R. P. The Seismo-Acoustics of Submarine Volcanic Eruptions. *J Geophys Res Solid Earth*, 126(4), Apr. 2021. doi: 10.1029/2020JB020912.

Tolstoy, I. and Ewing, M. The T phase of shallow-focus earthquakes. *Bulletin of the Seismological Society of America*, 40(1): 25–51, 1950. doi: 10.1785/BSSA0400010025.

Williams, C. M., Stephen, R. A., and Smith, D. K. Hydroacoustic events located at the intersection of the Atlantis (30°N) and Kane (23°40'N) Transform Faults with the Mid-Atlantic Ridge. *Geochemistry, Geophysics, Geosystems*, 7(6), 2006. doi: 10.1029/2005GC001127.

Zhong, M., Castellote, M., Dodhia, R., Lavista Ferres, J., Keogh, M., and Brewer, A. Beluga whale acoustic signal classification using deep learning neural network models. *The Journal of the Acoustical Society of America*, 147(3):1834–1841, 2020. doi: 10.1121/10.0000921.

Zhong, M., Torterotot, M., Branch, T. A., Stafford, K. M., Royer, J.-Y., Dodhia, R., and Lavista Ferres, J. Detecting, classifying, and counting blue whale calls with Siamese neural networks. *The Journal of the Acoustical Society of America*, 149(5):3086–3094, 2021. doi: 10.1121/10.0004828.

Zhu, W. and Beroza, G. C. PhaseNet: A deep-neural-network-based seismic arrival-time picking method. *Geophysical Journal International*, 216(1):261–273, 2019. doi: 10.1093/gji/ggy423.

Zhu, W., Biondi, E., Li, J., Yin, J., Ross, Z. E., and Zhan, Z. Seismic arrival-time picking on distributed acoustic sensing data using semi-supervised learning. *Nature Communications*, 14(1):8192, 2023. doi: 10.1038/s41467-023-43355-3.