Dear Editor,

Thank you for the insightful comments and reviews, which have allowed us to improve the content of the paper and the result analysis.
Our main revision deals with what both reviewers questioned: the confidence in the annotations. Regarding the presence of uncertain events, arguments have been added to highlight the goal of the proposed benchmark. Most works in the geophony acoustic field followed the same methodology: picking events of interest and, then, using the time and location of the sources, classifying them. This, therefore, justifies our tolerant approach in the sense that we look for a model able to detect all signals that “could” be of geophonic nature, letting their exact classification to later stages not covered by this study. We however decided to cope with the uncertainty issue by differentiating the evaluation dataset in two categories: a “complete” one, containing all geophony annotated events, and a “conservative” one, containing geophony events annotated by at least three annotators. The ROC curves and AuC tables now provide values for two more datasets. We hope the argument and these complements fill in the gaps pointed out in the methodology in terms of relevance and completeness.
We also added Aude Lavayssière as a co-author, given her contribution in establishing the methodology and participating in the annotation campaign.
Finally, we found and fixed a minor bug in the code which altered the model scores. As a consequence, the AcousticPhaseNet score has improved, leading to further discussion.

We hope these changes answer the reviews.

With our best regards,

Pierre-Yves Raumer & co-authors

Please find below our replies to the reviewers’ queries/comments.

Reviewer A:

Thanks for your many minor proposals, which helped us to improve the overall quality of the paper.

R: I just felt a little short the discussion of the paper and I had difficulties to understand some parts of methodology. I did not see any citation or work compared with in discussion.
A: The part on methodology has been modified to be more explicit. For example, some explanations have been given about ROC curves and AuC. A comparison with two other works (the SGBT original paper and PhaseNet) has been added, however, the limited amount of papers focusing on geophony detectors prevents us from doing extensive comparisons. A more in-depth explanation for the proposed benchmarking framework has been added to make it clear to the reader that the data set is intended to enable the evaluation of potential new models. This should make our overall scope clearer.

R: If the signal is coming from ice, you should to use the ice or iceberg term.

A: After a brief review of the literature, “cryogenic” or ice-generated events look more frequently used than “cryophony”. We thus replaced the latter with the former. As examples, we can cite Pinto and Chandrayadula, 2021 JASA (DOI 10.1121/10.0003444) or Dziak et al., 2015 PLoS ONE (DOI 10.1371/journal.pone.0123425).

Reviewer B:

Thanks for your suggestions and questions.

R: The “geophony” labelled group includes the “uncertain” labels (L169-170) which includes also non-seismic events according to L137-138 (“..,and uncertain when the level of ambiguity between T-, P- or H-waves, or with any other acoustic event, was deemed too high for the annotator.”). The models are evaluated against this group and “uncertain” events make up 6784 of 6897 (98%) of them. As the study and related framework are aimed at targeting geophonic signals, any other acoustic signals not related to geophony should be excluded from a ground-truth dataset. It could also be argued for that it is in line with the aims of the study to evaluate the models against any kind of acoustic signal, but then this should be discussed more, as of now the study clearly states that it targets geophonic events.

A: The philosophy behind the proposed benchmark is to develop methods to detect all events that “could” be of geophonic origin (i.e. seismic or volcanic in our case). We then expect, in later steps not covered by this paper, an a-posteriori classification (mostly based on the spatio-temporal context of the signals). This is justified by similar works in hydroacoustics, that, indeed, confirm the signal nature a-posteriori. Eexperts even tell us that it may be impossible to be perfectly sure of the nature of certain signals without first contextualizing them. We added a discussion to justify this choice and its relevance. The idea of letting the exact classification to a-posteriori contextualization implies a tolerant approach in the detection part. This is how the annotators were instructed to annotate every signal that may be geophonic but to mark them as “uncertain” if they looked so. However, a non-geophonic signal may be picked by an annotator, but then it is unlikely that it will be picked as such by several annotators. The answer to the next point addresses this issue

R: Around 50% of all events from the geophony group are only seen by one annotator (Figure 5). What kind of events are those only seen by a single annotator and how useful are these events in the context of geophony? Possibly the true positive rate is influenced by the models triggering on weak signals not necessarily related to geophony? It would be interesting to see how the models perform using more “obvious” geophony events as ground-truth, e.g., events that were annotated by at least 3 annotators. Is the true positive rate significantly higher for those and do the observed performance differences between the models remain? This would give a better understanding of the datasets and models performance.

A: We also think that the agreement among annotators is an important issue. To provide a more complete study and answer this point, we made two branches from the two evaluation datasets : a “complete” one, similar as it was before, and a “conservative” one when 3 or more annotators agree, following your suggestion. The models have then been tested against the two datasets, leading to 4 ROC curves and 4 lines in the AuC table. As expected, the models all perform better with the conservative set than with the complete

one. The best models remain the same in both cases. Examples of signals from and out of the conservative set have been added in the supplementary materials.

R: For the evaluation of the models performance, it would also be important to understand better the nature of geophonic signals that were missed. For example, it would be interesting to see if there are geophonic events that were missed by all models but seen by ≥ 3 annotators. Are there any of the few T and H labeled events in the geophony group that were missed? Are most of the missed events only seen by 1 annotator and not very strong? Also, are there any patterns in the false positives in between the models? Did they all trigger on a similar signal? A few examples of representative events missed or falsely triggered for the cases would be good (could go in the supplement and be briefly discussed in the evaluation).
A: Examples of false positives shared by all models have been added in the supplementary materials. In addition, such examples were given for signals both from and out of the new "conservative" set.

R: To compare and analyze the annotations between the annotators it would be good to also have absolute numbers and not only proportions for the results shown in Figure 4. For example, annotator 1 only annotated very few "uncertain" events compared to annotator 4 and, thus, the proportions of H phases are very different between them. However, it could be that the absolute numbers of T and H phases are actually not that different between the two. The absolute numbers could be added to the figure or also as a table in the supplement for example.
A: Indeed, some interesting observations can be done from this. Both absolute and relative histograms are now available, with the corresponding values written above each bar. The difference among annotators is still important, even in the absolute case.

R: Acoustic data section: The information about the choices for the datasets could be provided here already. What was the reason to choose exactly those three time periods as datasets for the study when there is 14 years of data available also for the training dataset? Why are they representative for the aims of the study? This is later addressed partially in the annotations section (L127) but could be mentioned here when introducing the data sets. While the information about the OHASISBIO network is given in detail (deployment period 2009-2023, station distances), this information is missing about the HYDROMOMAR network. Is it still deployed at the moment? What are the station distances? One could also think of leaving the station distances from the text as it is shown in Figure 1 and Figure 2. However, it should be consistent between both networks.
A: Justifications have been added and the inequality of descriptions has been fixed. Regarding the station distances, an order of magnitude has been given in the text in addition to the two maps.

R: L293-294: The percentages for TP + FP do add up too 100,4 % for the HYDROMOMAR-2013 data set (100,1% OHASISBIO-2020). Rounding error?
A: The relevance to give these values has been reconsidered and they were removed. However, the values given were accuracy and FP. Accuracy is (TP+TN)/(P+N), so the addition with the FP rate which is FP / N should not give 100%.