Jun 12, 2025

Editor-in-Chief

Natural Hazards and Earth System Sciences

Dear Editor,

I am pleased to submit this cover letter regarding the original research article (nhess-2024-102) entitled “Could seismo-volcanic catalogues be improved or created using weakly supervised approaches with pre-trained systems?” by Titos M., et al., for consideration in NHESS. We have carefully reviewed the feedback from reviewers and greatly appreciate the time and effort they have invested in evaluating our work.

We hope that these revisions, alongside the provided documentation of changes, meet the reviewers’ expectations and adequately address their feedback.

Thank you for the opportunity to improve our work. We look forward to your advice.

Yours sincerely

Manuel Marcelino Titos Luzón
Postdoctoral Researcher, University of Granada, Spain
mmtitos@ugr.es.

ANSWER TO THE REVIEWER'S COMMENTS

In the following, we have provided detailed answers to the comments of the reviewers. The original texts from the reviewers are in normal font. Our answers are in bold font. We would like to take this opportunity to thank the reviewers for their valuable comments and for their time and resources.

Answer to comments of Reviewer#1

We would like to thank reviewer#1 for the careful reading of this manuscript. Furthermore, below are those comments that need more clarification.

Insufficient data have been used for any useful scientific inference.

The major revision primarily focuses on the inclusion of more data. We would like to respectfully clarify that the primary objective of our manuscript is not to introduce a new model architecture. Rather, our work focuses on leveraging previously published and operational models as pseudo-labelers within a weakly supervised learning framework. This approach is intended to support the generation of seismic event catalogs with reduced manual effort—an aspect we consider central to the contribution of the study.

We fully acknowledge the value of the reviewer's suggestion to evaluate the approach on additional datasets. However, this is only feasible when reliable labeled data is available for comparison. In practice, the creation of high-quality seismic catalogs is a time-consuming and resource-intensive task, and such labeled datasets remain scarce. This challenge further motivates the goal of our work.

In this work, we have used the two labeled datasets that are currently available to our research group. Additionally, we contacted one of our collaborators, who kindly provided a dataset from a volcanic crisis along with its earthquake catalog. At this point, we can only express to the reviewer that we are open and willing to apply our method to any reliably labeled dataset they can recommend or share with us in order to fulfill their suggestion.

This is the third time we have been assigned a "major revision," and unfortunately, we do not have access to additional data. Seismic observatories are often reluctant to share their datasets, and when they do, the data is typically not labeled. It is simply not feasible for us to continue addressing this request without access to the necessary data.

Therefore, we kindly ask the reviewer to guide us on where and how we can obtain such datasets in order to carry out the suggested analysis. Without this, it will not be possible for us to further address this concern.

Answer to comments of Reviewer#2

We would like to thank reviewer#2 for the careful reading of this manuscript and the thoughtful comments that have improved the quality of this manuscript. Furthermore, below are those comments that need more clarification.

Detailed Review Report

General Impressions

The authors clearly propose that traditional machine learning models for seismic event detection often carry biases due to training on specific, limited catalogs. Their approach, utilizing pseudo-labeling based on pre-trained systems to enhance model generalization, is compelling and well-motivated. The manuscript presents a highly relevant and interesting investigation into improving seismic-volcanic catalogs through weakly supervised machine learning techniques.

After multiple rounds of review, significant improvements have been made. However, readability and methodological clarity remain core concerns. Below, I outline detailed recommendations to address these issues constructively.

- *About Methodology Section.*

• Section 3.1 (Methodology Clarity):

The manuscript currently states that the proposed method aligns with the open-set domain adaptation paradigm, explicitly designed to handle novel event categories. However, the authors subsequently note a significant limitation: the method only labels events within categories already present in the master database. This limitation appears to directly contradict the previously stated open-set capability. I recommend clarifying this contradiction explicitly.

The authors should specify whether the approach is truly open-set (capable of detecting and handling unseen seismic categories) or acknowledge clearly that it is currently limited to closed-set scenarios.

Although the assumptions clearly indicate that label spaces may only partially overlap, and thus novel categories could be present, the authors later explicitly state their methodology can only label categories that exist in the master database. Therefore, the authors should clarify how their approach practically handles (or does not handle) the novel categories mentioned in their assumptions. If the method currently doesn't handle these novel categories, explicitly stating this limitation and distinguishing clearly between the theoretical scenario and the actual method implementation would strengthen the manuscript.

We appreciate the reviewer's observation. We agree that there is a tension between the theoretical framing of our work within the *open-set domain adaptation* paradigm and the actual capabilities of the implemented method. Since we base the creation of new catalogs on a weakly supervised technique using models previously trained on master catalogs, what we are implicitly performing is a knowledge transfer between domains. It is true that across domains (volcanic environments), classes not present in the training catalogs may appear. This is why we acknowledge that the technique has limitations and to develop a more universal pseudo-labeler, a master database containing a broader range of seismic categories would need to be constructed. However, by examining the probability matrices, it is possible to identify when and

where such events appear. To conduct a comprehensive analysis of all events, regardless of whether they are included in the training catalogs, an unsupervised approach would be required—something that falls outside the scope of this article. Nevertheless, we have added a sentence in the manuscript clarifying this limitation: "Although our method only labels categories present in the master catalog, potential novel classes in the target domain may still be revealed through analysis of the probabilistic detection matrices, especially when combined with unsupervised techniques for event discovery"

• Main issues in the Methodology (Experiment 3.2.1):
1) Insufficient methodological details to reproduce the experiment
o Problem: Currently, the authors only briefly mention:
- Three model architectures (RNN-LSTM, Dilated-LSTM, TCN),2)
- Pre-training on MASTER-DEC, then re-training with POPO2002

However, readers might ask:
- How were the models pre-trained initially (hyperparameters, training set sizes, epochs, loss functions, etc.)?
- What specific transfer learning strategies were applied (e.g., layer freezing, fine-tuning, learning rate adjustments)?
- How exactly were data split (train-validation-test)?
- What were the evaluation metrics or validation methods?
- Did authors address class imbalance or category distribution?

Without these details, readers cannot reproduce the experiments. It seems that some important information about this is in section 4 (results). We suggest to the authors that change the text to the methodology section.

This has been one of the major challenges the manuscript has faced. Given the large number of reviewers, maintaining a clear structure has proven impossible. Some reviewers requested that the methodology remain free of implementation details and suggested including those details in the results section. Others, noting that the manuscript does not present the development of a new model, recommended simply referencing the articles where each model's technical details are described. Regarding the transfer learning techniques, what we perform is a fine-tuning of all the parameters of the architectures in order to adapt each pre-trained model to the specific volcanic environment.

We have included the following sentence in Section 3.1 (Methodology – Re-training): "This approach applies a transfer learning strategy in which all model parameters were fine-tuned, experimenting with different learning rates and regularization techniques, and employing early stopping to prevent overfitting.

2)Confusion caused by two alternatives, stating:
o Problem: The authors propose two alternatives, stating:
- Option A: Only use the 5 categories in common with the MASTER-DEC catalog.
- Option B: Adapt the model output to accommodate all 7 categories present in POPO2002 by updating the output layer only.

But, critically:

- Authors state vaguely: “these two approaches have no major implications from a ML perspective”.
- This statement is confusing because, practically, these two options have very different implications:

Option A completely excludes new categories, simplifying the task significantly. Option B involves at least minor model changes (output layer modification), and crucially, implies retraining with novel data categories (a clearly significant ML implication). This confusion significantly weakens methodological clarity.

We appreciate the reviewer highlighting this point, as it is indeed important. When we mention that switching between 5 and 7 classes has no major implications, we are referring specifically to the fact that, in terms of performance, model complexity, and number of parameters, both configurations are practically equivalent. In both cases, the models retain the parameters of all layers except for the output layer, which increases from 5 to 7 units. The retraining process remains the same, with the only difference being that the model is now trained to recognize 7 classes instead of 5.

We have changed the text and included this sentence for the sake of the clarity: From a ML perspective, both approaches follow standard procedures, although they differ slightly in implementation. In the first case, where only five seismic categories are considered, the models are fully retrained using the new catalog. In the second case, which includes seven categories, the output layer is modified to accommodate the two additional classes, while the pre-trained parameters from the original model are retained. The model is then fine-tuned on the new data, allowing efficient adaptation without retraining from scratch.

• Second Experiment (3.2.2):
Clearly distinguish the novelty of Experiment 2 by explicitly stating upfront that the primary difference from Experiment 1 is the source of labels (pseudo-labels rather than true annotations). Emphasizing this difference early in the description would enhance readability.

We have included this sentence at the beginning of the paragraph to enhance readability: “The second experiment differs from the previous one primarily in the source of the labels used for training: instead of relying on true annotations, it leverages pseudo-labels generated by the pre-trained models themselves”.

- About Results Section.
• Lines 334-357: This methodological detail is beneficial and should be moved explicitly into the methodology section for clarity and improved reproducibility.

As we previously mentioned, this information is placed in this section following the suggestion of earlier reviewers. Given the large number of reviewers who have revised the manuscript, it has been very challenging to establish a clear structure, and we have had to continuously relocate parts of the text between different sections. We kindly ask the reviewer that, if they are not comfortable with keeping this content here, to suggest again where it should be placed, and we will do our best to

accommodate the change. However, we believe that implementation details are better presented alongside the results rather than in the methodology section.

• Line 362: The statement regarding "two experiments" conducted at this stage is confusing and should be simplified in the methodology section.

What we mean by this statement is that we have conducted two separate experiments, one with 5 classes and another with 7. This is entirely independent of the methodology. Our intention is simply to demonstrate that, when retraining the models using the labels from a given database, the performance results remain high regardless of the number of classes.

• Should the reader be beneficed with more transfer learning details? (e.g., fine-tuning strategies, freezing layers explicitly, loss functions and training epochs?).

We have included new sentences along the manuscript to enhance de readability of the proposed technique.

• First Experiment Results:
While the overall self-consistency result (e.g., 77.38% accuracy for the RNN-LSTM model) provides a general sense of model performance, the confusion matrix reveals important class-specific differences—most notably, the relatively low recall for VT events (0.51) compared to much higher values for noise (0.97) and other event types. This suggests that the model may be biased toward the dominant class (likely noise), potentially inflating the global performance metric. I recommend that the authors include additional evaluation metrics, such as precision, recall, and F1-score for each class, as well as macro-averaged or balanced accuracy scores. These would provide a more nuanced understanding of how well the model generalizes across all event types, especially the underrepresented or more challenging classes like VT. Including this information would strengthen the assessment of the model's real-world applicability in diverse seismic scenarios.

Table 5 refers to the best results obtained in Table 4, which were achieved using a training split of 40% of the total data and considering 5 seismic classes. That said, the reviewer's analysis makes a lot of sense; however, we did not incorporate it into the manuscript because, in this case, weighted results or metrics such as the F1-score do not provide much additional information.

If we analyze Table 3 in detail, which is the basis for the retraining process, it shows that the number of VTE events — which are later the worst recognized — is 371, while BGN, one of the best-recognized classes, has 340 events. As can be seen, there is no strong imbalance. However, considering the nature of the events themselves and the spectral description given by the parametrization scheme used (based on filter banks on a logarithmic scale), noise-type events are much easier to discriminate than VTE events. As can also be seen in Figure 2, VTE events share regions of the

representation space with many other event types, which complicates their recognition.

Finally, Figure 6 is another clear example of the complexity involved in detecting VTE events. Some of the VTE events labeled in POPO2002 do not consistently share spectral characteristics with VTE in MASTER-DEC. This may be because catalogs are often built using data from multiple seismic stations, with strong attenuation and source effects, as well as rules or conditions imposed for signal identification.

Therefore, the original labeling of an event does not always match the spectral content and waveform of the analyzed signal, since it may vary depending on the station being analyzed. As a result, if the analyzed signal does not align with the characteristics of the prototype event used to build the system, it will be labeled or associated with the prototype that probabilistically most resembles it. This behavior reduces the recognition rate for this seismic category.

Precision, Recall, and F1 Score for RNN-LSTM architecture

Class	Precision	Recall	F1 Score
BGN	0.836	0.97	0.898
TRE	0.780	0.78	0.780
HYB	N/A	0	0
VTE	0.864	0.51	0.640
LPE	0.680	0.85	0.750

Precision, Recall, and F1 Score for Dilated-LSTM architecture

Class	Precision	Recall	F1 Score
BGN	0.768	0.96	0.855
TRE	0.650	0.69	0.669
HYB	N/A	0	0
VTE	1.000	0.31	0.473
LPE	0.565	0.78	0.654

Precision, Recall, and F1 Score for TCN architecture

Class	Precision	Recall	F1 Score
BGN	0.766	0.98	0.863
TRE	0.819	0.68	0.742
HYB	N/A	0	0
VTE	0.819	0.59	0.688
LPE	0.735	0.86	0.791

The weighted precision, recall, and F1-score are calculated by taking into account the number of events in each class as weights. This method adjusts for class imbalance by assigning more importance to classes with more samples. The weighted metric is computed as the sum of each class's metric multiplied by its number of events,

divided by the total number of events across all classes. This gives a more representative overall performance metric that reflects the actual distribution of the data.

$$\text{Weighted Metric} = \frac{\sum_{i=1}^{C} N_i \cdot M_i}{\sum_{i=1}^{C} N_i}$$

C: total number of classes

Ni: number of events in class ii

Mi: metric value (e.g., precision, recall, F1-score) for class ii

Weighted precision, recall and F1-score for each class using the TCN architecture.

Class	Precision	Recall	F1-score
BGN	0.98	0.98	0.98
TRE	0.68	0.68	0.68
HYB	0.00	0.00	0.00
VTE	0.59	0.59	0.59
LPE	0.86	0.86	0.86

Weighted precision, recall and F1-score for each class using the Dilated-LSTM architecture.

Class	Precision	Recall	F1-score
BGN	0.96	0.96	0.96
TRE	0.69	0.69	0.69
HYB	0.00	0.00	0.00
VTE	0.31	0.31	0.31
LPE	0.78	0.78	0.78

Weighted precision, recall and F1-score for each class using the RNN-LSTM architecture.

Class	Precision	Recall	F1-score
BGN	0.97	0.97	0.97
TRE	0.78	0.78	0.78
HYB	0.00	0.00	0.00
VTE	0.51	0.51	0.51
LPE	0.85	0.85	0.85

As a conclusion, we can say that the recognition of VTE is not so much influenced by the class imbalance of the dataset, but rather by the complexity of an event that shares spectral features with others throughout its temporal evolution.

• Second Experiment Results:
While the weakly supervised fine-tuning improved global accuracy, the model's ability to detect meaningful seismic events—especially VT and LP types—remains limited, with VT nearly absent in the confusion matrix. The dominance of the noise class likely inflates the global metric. Additionally, the model's detection rate far exceeds the label count, which may reflect over-sensitivity rather than true discovery. More rigorous evaluation, including precision-recall analysis, event-level validation, or expert review of excess detections, would strengthen confidence in the weak supervision pipeline.

We agree with the reviewer that validating the detected events or conducting an exhaustive expert review would increase confidence in the detections. However, this task is unfeasible from a human standpoint due to the large number of detections generated by the different architectures.

We would like to emphasize that Table 7 does not indicate that the models are incapable of detecting VTE and LPE, for instance. In fact, Table 8 shows the number of detections each architecture makes for each event type, confirming that there are many more detections than those reflected in the original catalog. What Table 7 indicates is that the detections of these events do not match with the catalog labels.This is why we performed an error analysis and provided graphical examples to support the reported recognition percentages.

Furthermore, to test the reliability of the results, we also conducted experiments on La Palma database where only earthquake events are annotated. We compared both matching and non-matching annotations in order to evaluate sensitivity and the false positive rate for this type of event. By analyzing Figure 9, we conclude that the vast majority of detected events match those in the catalog, considering them well recognized.

• Third Experiment Results:
The use case of applying weakly supervised models during a pre-eruptive crisis is compelling and highlights the practical value of such approaches. However, the presentation of results—particularly the so-called "recognition results" table—is unclear. It is not evident whether the numbers reflect validated detections, raw counts, or comparisons to any ground truth. The sudden introduction of PhaseNet, while relevant, is also only partially integrated, with no evaluation metrics provided to contextualize its outputs or compare them to the proposed models. A more transparent and consistent presentation of results, including quantitative comparisons, ground truth validation, and clearer labeling of what each table or number represents, would greatly improve the interpretability and impact of this section.
While the discussion highlights VT confusion rates exceeding 60% in some cases, this appears to reference only the worst-performing model (Dilated-LSTM). The other models achieve higher recall (e.g., 59% for TCN), and the average across all three models is closer to 47%, not 40%. A more balanced summary would acknowledge this range to accurately reflect performance variability across architectures.

We have modified the caption of Table 9 to indicate that it shows the number of earthquakes recognized at each station. As we argue in the text, the seismic catalog for this dataset contains 247 events. This study can be divided into two experiments. On the one hand, we analyze how many events are detected by PhaseNet, an AI-based model for seismic phase detection at each station, and on the other hand, how many are detected by our approach. For the PhaseNet experiment, we include the result tables and figures, and we also provide an analysis highlighting the cases where catalog events and detected earthquakes match (Figure 8b).

In the case of our models, we conduct a visual analysis of the results and present a discussion in the text to avoid including additional images and to keep the discussion more concise:
"Regarding the detection of events identified by the systems but not annotated in the catalog, on average, RNN-LSTM and Dilated-LSTM detected approximately 60 earthquake-type events, while TCN identified over 150. Figure 11 presents a couple of examples of such earthquakes. The PSDs reveal that they share characteristics consistent with those of earthquakes. However, as indicated by the probabilities shown at the top of the figure, their partial similarity in spectral content prevented them from being classified with higher confidence."

As we previously mentioned, conducting a thorough analysis of the results and maintaining a readable structure in the article has been challenging due to the large number of outputs. We believe that the way the results are currently organized is the simplest and most efficient way to analyze, compare, and discuss them.

Summary about results:
Throughout the results and discussion sections, the manuscript refers to “confusion matrices” and reports numerical values (e.g., 0.51, 0.31, 0.59 for VT events across models) without clearly stating whether these represent recall or confusion rates. However, the structure of the matrices—particularly the fact that each row sums to one—strongly suggests that the values correspond to per-class recall, i.e., the proportion of correctly classified instances for each true class. This is the standard interpretation for row-normalized confusion matrices in the machine learning literature. The ambiguity around this point makes the discussion difficult to follow and may contribute to the impression of poor presentation. For instance, the statement that “confusion rates exceed 60%” appears to refer to only the worst-performing model and does not align with the higher recall values seen in other models unless the reader assumes a confusion rate = 1 - recall. For the sake of clarity and consistency, it is essential that the manuscript explicitly define how these matrices are computed and what the reported values represent. This will not only improve readability but also help readers interpret the results accurately.

We have added the following sentence in Section 4 (Results), describing how the confusion matrices were constructed and what they represent: “For each experiment, tables describing the system performances in terms of accuracy, along with detailed confusion matrices are presented. These confusion matrices were constructed by comparing the model predictions against the labeled events in the catalog. This approach allows for a granular analysis of the classification behavior, revealing not

only the global accuracy but also class-specific performance, misclassification patterns, and possible confusion between seismic event types".

While the qualitative example shown in Figure 4 is compelling and suggests the model is capable of discovering events missed during the initial labeling, these anecdotal demonstrations are not sufficient to validate the effectiveness of the weakly supervised system. To move beyond suggestive visuals and convincingly argue for the scientific value of these new detections, the study would benefit from a more rigorous validation approach—such as expert review, waveform similarity analysis, or cross-comparison with independent models like PhaseNet. Without such steps, the claim that these new detections are not false positives remains speculative and limits the broader impact of the proposed method.

We fully agree with the reviewer. As we have previously argued, validating the detected events or conducting an exhaustive expert review would indeed increase confidence in the detections. However, this task is unfeasible from a human standpoint due to the large number of detections generated by the different architectures. This is precisely why we included the experiment and comparison with PhaseNet. It should be noted that PhaseNet cannot be used for comparison with other types of events beyond earthquakes, as it would not be meaningful. Moreover, we cannot compare segmentation performance since PhaseNet only detects P and S phases. Therefore, our analyses are necessarily limited to comparisons with the available labeled data.

We kindly ask the reviewer, if they are aware of any master and reliable annotated database to share it with us. We would be happy to run the experiments and provide a full analysis of the results accordingly.

- About the Discussion Section:

• Line 416: Verify if percentages presented in the discussion exactly match the results section; discrepancies would confuse readers.

Corrected

While the qualitative example shown in Figure 4 is compelling and suggests the model is capable of discovering events missed during the initial labeling, these anecdotal demonstrations are not sufficient to validate the effectiveness of the weakly supervised system. To move beyond suggestive visuals and convincingly argue for the scientific value of these new detections, the study would benefit from a more rigorous validation approach—such as: expert review, waveform similarity analysis, or cross-comparison with independent models like PhaseNet (we'll talk about this later). Without such steps, the claim that these new detections are not false positives remains speculative and limits the broader impact of the proposed method.

We completely share the reviewer's view. As discussed earlier, verifying the detected events or carrying out a thorough expert validation would certainly enhance the reliability of the results. However, given the sheer volume of detections produced by the various architectures, such an undertaking is not practically feasible. For this reason, we incorporated the experiment involving PhaseNet as a point of comparison. It is important to clarify that PhaseNet is specifically designed for detecting seismic phases in earthquake signals and is not suitable for evaluating other event types. Additionally, a direct comparison in terms of event segmentation is not applicable, as PhaseNet only identifies P and S phase arrivals. Consequently, our evaluations are necessarily confined to comparisons against the labeled data available in the catalog.

The discussion attributes the weak performance of the model on volcano-tectonic events (VTEs) to discrepancies in labeling criteria, subjective annotation boundaries, and prototype mismatches. While labeling inconsistency is a known challenge in volcano seismology, VTEs are typically among the most well-defined and reliably detectable seismic signals due to their impulsive, high-frequency nature. Numerous existing models (e.g., PhaseNet) have shown robust detection of such events across different volcanoes. The fact that the system recovers only 5% of annotated VTEs suggests that the problem may lie more in the modeling strategy or prototype selection than in catalog inconsistency alone. A more balanced discussion should consider whether the weak supervision framework fails to generalize to realistic variability within VTEs and whether model or prototype refinement could improve performance.

We also share the reviewer's conclusion and agree that many of the labels in the catalog were likely assigned based on information derived from multiple seismic stations. However, when analyzing a single vertical-component signal from a distant station—where attenuation and propagation effects are present—the waveform may not exhibit the typical spectral or shape characteristics associated with that event type. As a result, the system may fail to recognize it. This is supported by the La Palma experiment, where the system successfully detects earthquakes when the signals are clear.

The comparison with PhaseNet in the third experiment raises concerns regarding methodology.
The authors assess PhaseNet's performance by comparing the number of detected phases across different score thresholds, arguing that only detections above 0.8 correspond well with the labeled dataset. However, this approach overlooks the fact that many valid seismic picks— especially low-amplitude or emergent phases—often have lower phase scores (e.g., 0.3–0.6), yet still align with cataloged arrivals. Furthermore, raw pick counts do not constitute a meaningful evaluation metric unless aligned with ground truth picks using a timing tolerance. To make a valid comparison, the authors should report precision, recall, and pick timing accuracy against the labeled dataset across multiple thresholds. Without this, the argument that PhaseNet underperforms is not well supported and may misrepresent the model's actual capabilities.

We believe we may not have conveyed our point clearly enough. What we intended to express is that setting the probability thresholds above 80% greatly reduces the

number of detected phases, not that the detected events correspond only to phases with probabilities above 80%. The text literally states:

"For example, for values close to 80%, only approximately 722 P-phases and 503 S-phases at PLPI; and 282 P-phases and 216 S-phases at PPMA are detected. This significantly reduces the number of potential events that could be included in the catalog. Figure 8b shows the match between detections and the cataloged events. Of these 247 annotated events, PhaseNet detects 206 P-phases and 199 S-phases at PLPI; and 157 P-phases and 28 S-phases at PPMA, all without applying any probability threshold. Again, when setting the phase score threshold greater than or equal to 80%, the detections decrease to 163 P-phases and 164 S-phases at PLPI, and 116 P-phases and 21 S-phases at PPMA."

This perfectly aligns with the reviewer's conclusion that many phases have low scores. We think there may have been a misunderstanding regarding this point. Finally, Figure 8b presents the PhaseNet results, taking into account the timing of the P- and S-wave picks, exactly as the reviewer suggests. This comparison allows us to evaluate how well the detections align with the cataloged arrivals based on their temporal correspondence.

- About Summary of Findings Section:
• Figure 10: Clearly label differences between rows 3 and 4.
Corrected

• Ensure consistent PSD plotting style across Figures 10, 11, and 12 for clarity.

Figures 11 and 12 display a different PSD style because the objective of Figure 10 is specifically to allow the reader to easily identify visual differences between the smoothed PSD shape of the average and that of the analyzed event. This stylistic choice enhances the clarity of the comparison and supports the interpretation of the spectral content.

Other (Very) Minor Remarks:
- Abstract:
• Lines 2 & 4: avoid unnecessary repetition of word "however" within the same paragraph; consider synonyms or rephrasing to improve readability.
Corrected
- Introduction:
• Line 37: a space is missing, "..crises.However"
Corrected
• Line 51: there is an extra space; "Canario et al., 2020 ;"
Corrected
• Lines 57-63: suggestion: another challenge is that upgrades and updates to seismic instrumentation over decades complicate the review of historical seismicity, as the digital signals may not share a consistent framework.
Corrected
• Lines 56, 66, 71, 95, etc.: There are inconsistencies in citation formatting throughout

the manuscript. Please ensure that references within parentheses follow the standard format, e.g., “(Weiss et al., 2016)2, rather than “(Weiss et al. (2016))”.

We are following the template guidelines, and if the manuscript is accepted for publication, we will coordinate with the editorial team to ensure the correct formatting. We are unsure why, despite adhering to the guidelines, the references are not being cited correctly.

• Lines 168: space missing at “MASTER-DEC(1-50HZ)”

Corrected

- Methodology and experimental framework:

• Line 247: repetitive vocabulary again (stream).

Corrected

- Discussion.

• Lines 445-446: review grammar of “According to such table, on average, only 5% of the analysis windows labeled as VTE in the original catalog were recognized by the retrained systems.”

Corrected

• Line 473: please check the text “(Fig. 7)a.”

We did not find any error in the text.