Dear Reviewer,

We are pleased to submit this answering letter regarding the original research article by Carthy J., et al., for consideration in NHESS. We have carefully reviewed the feedback and greatly appreciate the time and effort they have invested in evaluating our work.

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

The manuscript presents an interesting and potentially high-impact study on the application of Self-Supervised Contrastive Learning to volcano-seismic datasets. The work addresses a relevant issue—the cost and time-intensive nature of labeling volcanic datasets—and I believe it is valid and merits publication in Natural Hazards and Earth System Sciences (NHESS).

However, before the manuscript can be accepted, it requires some revisions to improve its overall completeness, the fluency of the writing, and the factual contextualization of the work. Please address the following main points:

1. Introduction, Problem Framing, and the "Lack of Labels" Claim

The introduction currently feels a bit sparse, especially regarding the description of the seismic signals and the presentation of the methods. Furthermore, the authors state that there is a lack of available labels for training supervised machine learning models on volcano-seismic tasks. Is this absolute statement entirely accurate? While creating specialized ML datasets is indeed time-consuming, volcanic observatories worldwide have maintained extensive, labeled seismic catalogs for decades. The authors should nuance this statement, perhaps clarifying the distinction between the availability of raw observatory catalogs and curated ML-ready datasets. Additionally, it lacks clarity regarding the "practical monitoring objectives" this specific method aims to achieve. The authors should better clarify how their approach stands out from the state of the art and how it practically aids monitoring compared to existing methods.

We fully agree with the reviewer's comment and appreciate the opportunity to clarify this important point.

Regarding the description of the seismic signals and the presentation of the methods, we have thoroughly revised this section to improve its clarity, readability, and to better explain the motivation for the proposed methodology. The reviewer could check these changes in the Diff_file of the new version.

With respect to the lack of available labels for developing robust automatic recognition systems, our intention was not to imply that volcano observatories lack seismic catalogs. On the contrary, many observatories around the world have compiled extensive seismic databases and catalogs over several decades. However, the key limitation we intended to highlight is the scarcity of publicly available, high-quality, machine learning–ready datasets that can be reliably used

either to develop automatic monitoring systems or to serve as benchmarks for the objective comparison of different methodologies. This limitation arises for two main reasons (between others). First, most seismic datasets and catalogs belong to individual observatories and are generally not publicly accessible. In our own experience, despite contacting several observatories, access to these data has rarely been granted, and when it was, the shared catalogs were often incomplete or lacked sufficient metadata for systematic machine learning studies. Second, as discussed in Titos et al. (2025), many historical volcano-seismic catalogs were manually compiled by experts based on time-frequency characteristics and wavefield properties. Although expert interpretation is indispensable, manual labeling may introduce unavoidable qualitative and quantitative biases. These biases can originate from the scientific understanding available at the time of annotation, operational constraints during periods of intense volcanic unrest (where only the most significant events can be cataloged) or situations in which overlapping seismic signals are grouped into a single event despite corresponding to different physical processes. A representative example is the 2011 El Hierro eruption, where the occurrence of continuous VTE events generated a high-frequency signal resembling volcanic tremor because hundreds of VTEs overlapped each hour (Ibáñez et al., 2012; Díaz-Moreno et al., 2015).

Therefore, when referring in the Introduction to the lack of reliable labeled datasets, we specifically mean the limited availability of curated, consistently annotated, and publicly accessible catalogs with a well-defined level of uncertainty, suitable for the robust development, training, and objective evaluation of machine learning models. We will revise the manuscript to make this distinction explicit and avoid any ambiguity. However, in the new version of the manuscript, we have revised the text to better convey the underlying idea of our approach and to avoid potential misunderstandings.

2. Dataset Justification and Ground Truth (Mt. Etna)

Regarding the Mount Etna dataset, the authors claim that there is no ground truth public catalog available and that the dataset was analyzed in an exploratory manner for this reason. This assertion is highly questionable. Mount Etna is one of the most heavily instrumented and closely monitored volcanoes in the world, managed by INGV, which undoubtedly maintains detailed and robust operational catalogs. Is it truly possible that no certain labels exist for Etna? If the authors mean that a specific, pre-formatted public machine-learning dataset was not readily accessible for their study period, they must state this explicitly. Claiming a general "lack of ground truth" for Etna undermines the credibility of the premise and overlooks the extensive monitoring efforts present at this volcano.

We thank the reviewer for this important comment and agree that the original wording could be interpreted as questioning the monitoring efforts carried out at

Mount Etna. This was not our intention. Mount Etna is indeed one of the best-monitored volcanoes in the world, and the operational work performed by the INGV is exemplary and has been fundamental to the advancement of volcano monitoring around the world.

Our intention was instead to emphasize that, for the specific time period selected in this study, no publicly available and validated catalog was available that could be used as ground truth for a supervised evaluation. We deliberately selected such a period to perform a blind assessment of the proposed methodology under realistic monitoring conditions, where expert annotations are not available beforehand. This allows us to evaluate the robustness and transferability of the method in a scenario that closely resembles its intended operational use.

To avoid any misunderstanding, we have revised the manuscript to explicitly state that our claim refers exclusively to the lack of a publicly available reference catalog for the selected study period, rather than to the absence of operational catalogs or monitoring activities at Mount Etna. We sincerely apologize if the previous wording suggested otherwise.

New sentence included in section 4.4 Qualitative Analysis Aids: "For the Mount Etna dataset, we deliberately selected a period for which no publicly available validated reference catalog existed yet, enabling a blind qualitative assessment of the proposed methodology under realistic operational conditions."

3. Fluency and Completeness of the Methods

The methodology section requires more thorough explanations and a more fluid narrative. Currently, the presentation of the techniques (such as SimCLR and TS-TCC) feels unbalanced. I recommend expanding this section to provide a broader, more balanced overview, properly justifying the choice of all the adopted strategies (rather than focusing heavily on TS-TCC at the expense of SimCLR) and ensuring the text flows more naturally for the reader.

In the new version of the manuscript we delve into the different techniques in order to avoid unbalanced explanations. We will expand this section providing a broader and balanced overview, justifying our choice.

4. Discussion and Conclusions

The Discussion and Conclusion sections need to be better argued and characterized. Currently, they appear slightly weak compared to the amount of work that was conducted. It is crucial that the conclusions refer more specifically and explicitly to the obtained results and the figures presented in the text, explaining in detail how the metrics and visible clusters support the final claims regarding the utility of SSL for the volcanic observatory community.

We agree that the Discussion and Conclusion sections should better reflect the extent of the experimental work presented in the manuscript. Following the reviewer's recommendation, we will substantially revise these sections to provide a more detailed interpretation of the obtained results, explicitly relating the reported metrics and the patterns observed in the figures. We will also strengthen the discussion of how these findings support the potential utility of self-supervised learning for volcanic seismic monitoring and for the volcanic observatory community.

5. Improvement of Figures (Waveforms)

The images presented in the manuscript need to be revised and replaced. In particular, the figures displaying the waveforms (e.g., the time and frequency representations of the clusters) require significant graphical improvement to make them clearer and more readable. These figures are fundamental for visually supporting the effectiveness of the cluster separation proposed in the text.

I am confident that by addressing these points, the manuscript will reach the level of clarity and maturity required for publication in NHESS.

I have provided a few suggested references below. However, I highly recommend expanding the reference list further, particularly in the introduction, to better frame the study.

We thank the reviewer for these helpful suggestions. We agree that the graphical quality and readability of the figures should be improved, particularly for the waveform and time–frequency representations. We will revise and replace these figures to provide clearer and more informative visualizations, with particular attention to the aspects that support the interpretation of cluster separation.

We also appreciate the suggested references. We will carefully consider them and expand the reference list, particularly in the Introduction, in order to provide a broader and more appropriate context for the study and better position our contribution within the existing literature.

Answer to comments of Reviewer#2

This is a useful paper on an important problem. The comparison of SimCLR and TS-TCC on the Llaima, Colima and Etna datasets is relevant. The Llaima result is encouraging, and the Etna example is probably the most useful part of the paper because it is closer to a real monitoring catalogue.

I do not think the paper is ready in its present form. The main issue is that the study is presented as unsupervised, but labels appear to be used at several stages when the pipeline is chosen. The evaluation is also too limited, the baseline is weak, and the data and code statement needs to be replaced. These points need substantial revision.

Main comments

1. **The current pipeline is not fully unsupervised**

This is the main issue for me. The encoder is selected using supervised accuracy on Llaima. The manuscript also says that contrastive parameters are selected partly from supervised classification, UMAP is hand-tuned to obtain a discriminative representation, and the number of clusters is fixed from the known number of classes. It is also not clear whether the best clustering algorithm was chosen after looking at the test labels.

Please show clearly where labels are used: architecture choice, SSL tuning, UMAP settings, cluster number, clustering method and final evaluation. Ideally, rerun the main experiment with those choices fixed without labels. Otherwise, describe the work as label-informed or semi-supervised rather than fully unsupervised.

We agree with the reviewer that the manuscript should more clearly distinguish between the unsupervised and supervised components of the proposed methodology. We acknowledge that the current version may not make sufficiently clear to the reader where class labels are used and how they influence the different stages of the workflow. In the next revision of the manuscript, we will explicitly clarify the role of labels in model selection, SSL parameter tuning, UMAP configuration, clustering, and final evaluation. This will allow the reader to clearly distinguish the representation learning and clustering stages from the supervised steps used for model assessment and comparison.

1. **The evaluation needs more information**

Please add Adjusted Rand Index and Adjusted Mutual Information. Colima is quite imbalanced, so V-measure alone is not enough. A class-by-cluster contingency table would also show whether the improvement is shared across event types or is mainly coming from the large classes.

Tables 2 and 3 should include all combinations of representation and clustering method, not only the best one. Repeat the runs with several random seeds and report the mean and spread. Give the UMAP seed, initialisation, metric and software version. It would

also help to cluster in the encoder space, or after PCA, so the reader can see how much of the result depends on UMAP.

For Etna, add at least internal cluster scores and a stability check. The visual result is interesting, but the current interpretation is entirely qualitative.

We agree that a more comprehensive quantitative evaluation is needed to better assess the clustering performance and its robustness, particularly for the imbalanced datasets. In the revised manuscript, we will complement the V-measure with the Adjusted Rand Index (ARI) and Adjusted Mutual Information (AMI), and we will include class-by-cluster contingency tables to provide a more detailed assessment of the performance across the different event types. We will also expand Tables 2 and 3 to report the different combinations of representations and clustering methods rather than focusing only on the best performance configuration. To further assess the robustness and reproducibility of the results, we will be repeated using multiple random seeds, reporting the corresponding mean and variability..

Following the reviewer's suggestion, we will additionally compare clustering performed directly in the encoder space. This comparison will help quantify the extent to which the observed clustering performance depends on the dimensionality-reduction step.

Finally, for the Etna experiments, we agree that the current analysis is primarily qualitative. Therefore, in the revised manuscript we will complement the visual analysis with internal clustering metrics and a stability analysis across different random seeds. These additions will provide a more quantitative and reproducible assessment of the clustering structure observed in the Etna data.

- **The augmentation choice needs a test**

Window warping changes the local time scale and therefore changes frequency content. Frequency is also used to separate and interpret the event groups in this paper. I am not convinced that the representation should be invariant to that change without further evidence.

A small ablation would be enough: jitter only, jitter plus amplitude scaling, and jitter plus window warping, using the same training budget and more than one seed. Please also explain why each transformation is a plausible nuisance change for these signals and not a possible change of event class.

We thank the reviewer for raising this important point. We agree that window warping modifies the local temporal scale and can therefore alter the frequency content of the seismic signals. Since frequency-related characteristics are also relevant for distinguishing and interpreting the event groups considered in this study, the assumption of invariance to this transformation requires further justification.

In the revised manuscript, we will therefore include an ablation study comparing three augmentation strategies proposed by the reviewer: (i) jitter only, (ii) jitter combined with amplitude scaling, and (iii) jitter combined with window warping. All configurations will be trained using the same training set and evaluated across multiple random seeds to assess both performance and variability.

We will also clarify the motivation for each transformation. In particular, we will explicitly discuss the extent to which each transformation preserves the characteristics that are relevant for event discrimination, and we will clarify the magnitude of the applied transformations.

- **The baseline is too weak**

UMAP applied to the prepared waveform is a useful lower bound, but it is not a strong baseline for learned feature extraction. Please add at least one conventional feature set, such as spectral or cepstral descriptors, and one learned non-contrastive baseline such as a convolutional autoencoder. Scattering features would also be relevant.

If the paper keeps a broad claim about TS-TCC, a more recent time-series SSL method would be useful as well. If that is outside the scope, narrow the claim to a comparison of the two selected methods.

Applying UMAP directly to the prepared waveforms provides a useful simple reference, but does not constitute a sufficiently strong baseline for evaluating learned representations, we agree with the reviewer. In the revised manuscript, we will therefore include additional baselines covering both conventional and learned feature extraction. Specifically, we will evaluate a conventional set of spectral/cepstral features and a convolutional autoencoder as a learned non-contrastive representation. These additional experiments will allow us to distinguish the contribution of conventional signal-processing features, non-contrastive representation learning, and contrastive representation learning with TS-TCC. Although we acknowledge that scattering features would also provide a relevant comparison, we have decided not to include them in the revised experimental analysis in order to keep the scope of the additional experiments focused and manageable. Therefore, in the revised manuscript, we will prioritize the two complementary baselines described above: conventional spectral/cepstral features and a convolutional autoencoder as a learned non-contrastive representation.

Regarding the comparison with more recent time-series self-supervised learning methods, we agree that such a comparison would be valuable for a broader assessment of TS-TCC. However, rather than extending the scope of the study to a general benchmark of time-series SSL methods, we will narrow the corresponding claims in the revised manuscript and explicitly frame the experiments as a comparison between the selected representation-learning approaches. This will avoid making broader claims about the state of the art that are not directly supported by the experimental design.

- **The novelty claim should be toned down**

The paper should not say that it introduces contrastive learning to volcano-seismic analysis. The contribution is a comparative application of SimCLR and TS-TCC to three volcano-seismic datasets, including an automatically picked Etna catalogue. That is still a useful contribution.

The related-work section also needs to be expanded. Please discuss the close seismic SSL work, including Yang et al., SeisCLIP and SeisLM, and the more recent time-series methods TS2Vec, CA-TCC and TF-C. Unsupervised seismic clustering and cross-volcano transfer work, including the studies by Ardid and colleagues, should also be placed in context.

We will therefore revise the manuscript and clarify that the main contribution is a comparative evaluation of SimCLR and TS-TCC for volcano-seismic representation learning across three datasets, including the automatically picked Etna catalogue. We appreciate the reviewer's observation that this comparative analysis still represents a useful contribution.

We also agree that the Related Work section should be expanded to better position the study within the existing literature. We will discuss closely related seismic self-supervised learning studies, including the work by Yang et al., SeisCLIP, and SeisLM. We will also include more recent self-supervised learning approaches for time series, such as TS2Vec, CA-TCC, and TF-C, and clarify their relationship to the methods evaluated in this study.

We will accordingly revise both the Introduction and Related Work sections to ensure that the novelty claims are appropriately scoped and supported by the existing literature.

- **The work must be reproducible**

The current data and code statements amount to asking readers to contact the corresponding author. That is not enough for a benchmark paper. Please release the code, configuration files, splits, seeds and package versions in a public versioned repository.

The Llaima dataset DOI should be given directly. For Colima and Etna, release the processed event sets if possible. If the waveforms cannot be redistributed, provide the station and channel information, exact time windows, pick lists, filtering steps and scripts needed to recreate the datasets.

We agree that reproducibility is essential for a benchmark paper and that requesting readers to contact the corresponding author is not sufficient.

We will improve the Data and Code Availability statement by providing a public, versioned repository containing the source code, configuration files, dataset split

definitions, random seeds, and software dependency information, including package versions. We will also provide clear instructions to reproduce the experiments and evaluation procedures reported in the manuscript.

Regarding the datasets, we will provide the DOI for the Llaima dataset directly in the revised manuscript. For Colima and Etna, we will assess the availability and redistribution conditions of the processed event sets and release them whenever permitted. If the original waveforms cannot be redistributed due to data access or licensing restrictions, we will provide the available metadata and the necessary information to facilitate dataset reconstruction, including station and channel identifiers, event time windows, pick lists, preprocessing and filtering parameters, and the corresponding scripts.

- **The Etna section needs a few basic details**

Please report the number of detections before filtering, the number removed by each filter, and the final number of events in each split. Explain how the spectra and envelopes in Figure 11 were calculated. Cluster 2 contains only 99 samples, so it needs some discussion: is it a stable rare family, an outlier group, or an artefact of the chosen settings?

The paper also uses the word noise for both waveform noise and mistakes from automatic picking. Please keep those two things separate.

The event-selection procedure, the calculation of the signal representations, and the interpretation of small clusters need to be described more explicitly.

First, we will revise the manuscript to report the number of detections before filtering, the number of events removed at each filtering step, and the final number of events retained in each data split. We will clarify the sequence of filtering operations and provide the corresponding counts to make the selection procedure fully traceable.

Second, we will expand the description of Figure 11 to explain how the spectra and signal envelopes were calculated, including the relevant preprocessing steps and parameter settings.

Third, we agree that Cluster 2, which contains only 99 samples, requires further discussion. We will examine its characteristic waveforms and spectral properties and assess its sensitivity to the clustering configuration. We will then discuss whether the cluster appears to represent a potentially meaningful rare event family, a group of outliers, or a structure that depends on the selected clustering parameters.

Finally, we will revise the terminology throughout the manuscript to distinguish between background waveform noise, referring to moise components present in

the recorded signals, and automatic picking errors, referring to incorrect or spurious detections produced by the detection or picking algorithm.

Minor comments

1. The raw citation keys "smith1987iris, suarez2008international" are still in the introduction and need to be fixed.
2. The manuscript needs a careful language check. Examples include "volcano-seimsic", "eachother", "disciminating", "reproducable", "expectd", "window wawrping" and "generealising".
3. Use one form consistently: SimCLR, K-means, HDBSCAN and self-supervised when used as an adjective.
4. Spell out the LAV station name and provide network, station, location and channel codes for all datasets.
5. State the value of beta used in the V-measure equation.
6. Give the train, validation and test sizes for Llaima and Colima, and all final split sizes for Etna.
7. Event-wise maximum-amplitude normalisation removes absolute amplitude information. Please discuss what that means for monitoring applications.
8. Check the UMAP figures for print and colour-vision accessibility. Marker shapes would help.
9. The conclusion should be more measured. The Colima score and the qualitative Etna result are promising, but they do not yet show generally high performance across noisy volcano-seismic datasets.
10. The author-contribution statement is too broad. Please state who prepared the data, implemented the methods, designed the experiments and interpreted the results.

All minor comments will also be carefully addressed, and the corresponding changes will be highlighted in the revised manuscript to facilitate the reviewer’s assessment.