Reviewer#7:

Dear reviewer#7, We are very thankful for your thoughtful suggestions. Below, we present how we have addressed them.

Summary

I have now had the opportunity to read and review the manuscript “How can seismo-volcanic catalogues be improved or created using robust neural networks through weakly supervised approaches?”. Where the authors use machine learning techniques and a dataset from Deception Island as the master catalog to create and compare a new catalog for Popocatepetl in Mexico. While there are a lot of caveats and author interpretations in this research, the science, information and methods are interesting. The manuscript shows a small progress in ML techniques that can be used as the basis for future research. Below I list a few major comments for the review along with some line-by-line comments. Additionally, I would like to make a note about the subject matter. I feel this research would be more suited for a different journal. I was a bit surprised when I saw this manuscript was submitted to Natural Hazards and Earth System Sciences.

Major comments:

-What about other signals when building the model? There is a lot of source noise in volcanic terrains, how do these methods work when you introduce for example mass flows, edifice collapse, rock falls, ballistics, etc. In the same train of thought, how did leaving these out affect the outcomes. Furthermore, how about teleseismic earthquakes, how does the classification work on these?

This question is very interesting, and we thank the reviewer for bringing it up. As we have mentioned in our responses to other reviewers and in the text itself, our proposal serves as a use case and an example of operation applied to the databases we have available. If, during the training of the master system, events that may appear in other volcanic environments are not considered, those events will not be recognized, as the master system will only recognize those events it has been trained on. This highlights the importance of sharing databases and catalogs from which master models can be built that can be universally used. In this work, we aim to illustrate how the system functions and the implications of catalog bias, as well as how it could be improved.

Our work is not a universal system and therefore cannot be applied to volcanoes with very different dynamics. This is why we have compared it to a database of similar events. However, the proposed methodology could indeed be used among volcanoes of similar nature, provided there are reliably labeled databases available. We believe this is the important aspect of the work; we are not offering a universal tool, but rather a universal methodology to address the significant issue of incompleteness in seismic catalogs.

-“Early warning” is capitalized on line 24, but is not anywhere else, stay consistent throughout.

-Have you looked at the source depth of the signal, differing characteristics can occur depending on depth, you may have a problem similar to the attenuation issue.

In this work, we do not focus on the analysis of the events; instead, we rely on the information contained in the catalogs, assuming that this work has been carried out previously. It is from these catalogs, which we consider reliable, that we conduct our experiments to validate our initial hypothesis.

-I think the length of the training dataset is too short, how can we get a sense of what goes on at a volcano in just two months of data. Similarly, please explain why you are using a pre-eruptive model on a volcano that is in a phase of unrest. The difference in signal characteristics are going to be different, also the types of signals as I mentioned before.

-Some acronyms do not match, I tried to correct some in my Line-by-line comments, but it got too out of hand. A good example is the constant change between VT and VTE.

-While the frequency band of 1-20 Hz is fine, I am wondering about the difference between sensors. This paper does not mention any details about the sensors. What is the sampling rate of each sensor, are they all the same, are they different at different volcanoes. The details about each sensor are very important in knowing which frequency range can be used. Furthermore, how about is the sensor broadband or not. Is every signal from the vertical competent? If so how about using horizontal components?

In the new version of the manuscript, we will attempt to include more detailed information about the instrumentation. However, we would like to emphasize that both databases have been sampled at 50 Hz and that the response of the systems has been removed, making it possible to analyze the signals effectively from the outset.

-How did you choose which time window to use? What if there is a signal longer than 4 seconds, e.g. tremor, mass flow?

The choice of the analysis window is inherited from previous studies. In those works, tests were conducted with different window sizes, and the best results were obtained using 4-second windows. While the results might vary with different window sizes, in this work, this parameter does not hold significant importance. We simply used the window size that performed best for Deception Island.

Addressing the reviewer's second question, it appears that the methodology has not been fully understood. The reviewer asks what happens when a signal longer than 4 seconds (the proposed analysis window) is received. The answer is:

nothing changes. That signal will be segmented into 4-second windows with a 3.5-second overlap, and each window will be analyzed separately. For each window, we will obtain a label, and the combination of those labels will give us the trace analysis and event recognition.

The potential issue could arise in the opposite case, when an event lasts less than 4 seconds. In that case, we could have two events within the same window, and we must choose which label to assign. In our approach, during training we assign the label of the event with the longest duration within the window.

-You only train on one volcanic environment or master. I would like to see what the results would be if you used multiple environments from different volcanoes to make the master.

We agree with the reviewer that this test would be a great contribution both to the article and to the field. However, we only have access to the databases and catalogs analyzed here. In this regard, we are fully open to collaborating and testing our proposal on as many volcanoes as necessary, provided that we are given access to reliable data and catalogs from which conclusions can be drawn. This is an important issue in the field, as data is rarely made public, and access is often restricted.

-Most of the text in the methods section should be in the introduction. I suggest making a section in the introduction describing different kinds of methods people used in the past and then in the methods, explain the techniques you used for this research. Most of everything before section 3.1 should be in the introduction.

We will aim to implement these suggestions in the revised version of the manuscript if it is accepted for publication.

-I would like to see some comments about computing power and time. Some ML models and processes take lots of commuting resources as well as extended processing lengths. I would like to see a paragraph discussing these stats in the manuscript. What would I need to reproduce or do a similar computation at my observatory?

This is also an interesting point, and we appreciate the reviewer's input. From a computational perspective, since the base of the weakly supervised algorithm is already trained, its recognition process would be immediate, providing results within seconds. The only computationally intensive aspect could be the retraining using the pseudo-labeled database. However, considering that the models described here consist of several hundred thousand parameters, the training time, as observed in other referenced works along the manuscript conducting similar analyses, would range from several minutes to an hour. Therefore, it could be applied daily, weekly, or monthly, keeping the system updated in just a matter of minutes.

-I would like to know how much human work or time goes into creating this new catalog. Since it is a supervised learning technique, you still need human input and review, so how much time/effort are we gaining?

Creating a catalog from scratch is a very time-consuming task. It requires reviewing and analyzing signals using various types of analysis. By applying an algorithm like the one we propose, the human operator can obtain a tentative catalog that includes the event type, start, and end times. Therefore, the operator's task becomes validating whether what the system produces is valid, which is much faster and simpler than manually isolating, cutting, and analyzing the signals.

-In Lines 400-408: The training missed labeled tremor events, and you say this error was not actually an error, how can this be? The algorithm mislabeled, which means it did not work. Furthermore, reading your explanation further signals that this technique cannot be completed universally across different volcanoes. The attenuation affects you mention, points to the fact this would be difficult to do universally. A human had to go back in and review every event to make sure the event was labeled correctly, so how does this save time or is a better option?

Lines 400-408 detail the reason behind the observed performance drop of all tested systems, ranging from 20% to 33%, due to discrepancies between manual and automatic labeling concerning tremor and LPE-type events. In Figure 4A, two high-energy, low-frequency events are labeled as LPE in the POPO2002 catalog. However, these events resemble TRE-type events from Deception rather than LPEs from Deception. Consequently, during recognition and pseudo-labeling, the system assigns TRE labels to both events. Statistically, this would be considered an error when compared to the catalog, but after consulting with experts, these events can indeed be labeled as TRE, since both their duration and waveform match a TRE-type event. Therefore, what is computed as an error (detection) is not truly an error upon evaluation.

Regarding the system's universality, as previously mentioned, the aim is not to provide a universal system but rather a universal methodology applicable to any volcano. In terms of review efficiency, a human operator would typically need to detect, isolate, analyze, and classify events. With our approach, the operator simply needs to validate the classification offered by the system, as it already provides classification, isolation, and event segmentation.

-I would like to see more one-on-one comparison statistics in reference to Table 6. It is great the algorithm found more events but how many of the catalog events did it find and how many of the “human” events did it miss? Also, how many of these “new” events are real? Do the humans perform better for certain signal types and vice versa? How does each signal classification compare to one another.

This analysis is highly interesting and could greatly enrich the work. However, such validation should be performed by one or more experts, and given the large

number of recognized events, it would take some time to obtain reliable statistics. We are considering the idea of randomly sampling different types of recognized events, analyze them and gather statistics that could be extrapolated to the entire set of recognized events.

-There is a lot of repetitive nature of some paragraphs, try to go over the manuscript and cut some of this out.

We will make an effort to review the text and reduce the existing redundancies.

-A point on universality, every volcano is different even in the pre-eruption context of this manuscript. Some volcanoes do not even display signs of activity before erupting, so how can these ML techniques be considered universal at this point?

Authors did not intend to convey that our system is universal. In fact, throughout the text, we propose the premise that this methodology could become universal if multiple catalogs and volcanoes of different nature are considered in the training process with a master database: "The use of more sophisticated pseudo-labelling techniques involving data from several catalogues could help to develop universal monitoring tools able to work accurately across different volcanic systems, even when faced with unforeseen temporal changes in monitored signals." This sentence was included precisely because we understand that every volcano is different, even in the pre-eruption context. Therefore, by incorporating information from different volcanoes, the system could have more universal applicability and could be used in observatories where reliable information or catalogs are lacking. The only challenge, as we have already mentioned, is the availability and access to such data. Once again, we are open to collaborating on this effort, considering different volcanoes and catalogs. We invite the reviewers to join this initiative and work together on a project that leads to a universally applicable system.