March 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 **all four 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 Dr. Gordon Woo 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.

The limitations of the paper should be more clearly presented.

Answer to comments of Reviewer#2

Dear reviewer#2, We are very thankful for your thoughtful suggestions. Below, we present how we have addressed them.

In this revised version of the manuscript, the authors have clarified their work and the purpose of their research. In my opinion their work is worth publishing however I do think some points still need to be clarified / improved.

- Overall, the manuscript is well written but is sometimes very verbose (e.g. "to dive into these results" l.498, "Considering this information, we now proceed to discuss the results", …). It sometimes makes the reading difficult, I would simplify the text to highlight the conclusions and observations.

In the new version of the manuscript, We have made some sentences simpler to enhance readability.

- You still do not describe the features used to classify the signals. You do not have to describe them in details, especially if this is done in another study (otherwise put it in Supplementary). But you still need to describe them broadly in the manuscript.

The objective of this work is the application of the weakly supervised approach to create reliable seismic catalogs with less human effort. This approach can be used both with parameterized signals and with the raw waveform itself (if a sufficiently large dataset is available). Therefore, we understand that the description of the parameterization paradigm is not a goal of this work. This is why, in the experiments section, we briefly include a small description of this parameterization and reference. Our prior work provides a detailed description of the applied raw signal parameterization procedure.

In the actual version of the manuscript, we had introduced this paragraph: 'The data stream illustrates continuous or streaming analysis (allowing near real-time processing). To carry out the recognition step using the network seed (trained with the MASTER-DEC dataset), streaming or continuous signals are filtered between 1 and 20 Hz and split into frames or windows; the same feature extraction algorithm used in MASTER-DEC is applied. For each window, a feature engineering pipeline based on a logarithmic frequency scale filter bank is applied. This pipeline reduces the dimensionality of the input vector associated with each analysis window (compared to raw signals), which facilitates the training and convergence of the systems, as it increases the separability of the data based on well-studied features in the literature (see Titos et al., 2024 for a detailed understanding of the parameterization pipeline).' '

- The methodology is now more clearly explained, but it still needs to be improved :

o I understand the aim of the authors when they first present the overall method in Section 3.1 before explaining the application to the different cases, but it is hard to follow as we need some elements of section 3.2 to clearly understand section 3.1. Thus I suggest the following structure for the methodology section, that follows the overall structure of the article : 1) Description of the pre-trained systems (including all the technical details given at the beginning of section 4 and some insights on the accuracy/scores of the classifier on the MASTER Dataset), 2) Application of the pre-trained systems to event detection and classification, 3) Direct transfer learning, 4)Weakly supervised approach, 5)Outline of experiments on the POPO2002 and LAPALMA2021 datasets.

Regarding the structure of the article, we have followed the suggestions of all the reviewers from the previous version. After reviewing your recommendations and those of the others, we found that the simplest and clearest structure is the one we have outlined in the manuscript.

First, we introduce the proposed methodology. Second we describe the experiments: 1. a classical knowledge transfer experiment; 2. an experiment with a weakly supervised approach where we use a seismic catalog to compare results with our approach; 3. and finally, we conducted an experiment with a set of seismic signals for which no catalog is available. Through these experiments, we demonstrate the capability of the proposed model to automatically improve existing catalogs or build them from scratch.

We believe this structure most effectively aligns with the previous suggestions and enhances the clarity of the methods and results presented.

o It is still not clear to me what are the implications of the assumptions made l.210 and following. You assume that conditional distribution are the same l.213, but then acknowledge that they could be different (l.216). As said in my first review, I don't understand the logical link "Therefore" l.220. Are you suggesting that using weakly-supervised approaches allows to overcome the problem that conditional distributions are not the same? If so, why?

The wording of our hypothesis in the manuscript has been rewritten to improve its clarity.

The conditional distributions may differ between the source and target domains, which is a common challenge in domain adaptation tasks. Weakly-supervised approaches, such as pseudo-labelling, do not completely overcome this problem, but they provide a practical way to mitigate its effects under certain assumptions:

1. **Leveraging High-Confidence Predictions:**
   Weakly-supervised methods rely on the model trained on the source domain (D_S) to generate probabilistic predictions for the target domain (D_T). By selecting only those instances in D_T with high per-class probability (i.e., high confidence), we assume that these predictions are more likely to be correct. This approach implicitly assumes that, for high-confidence predictions, the conditional distributions $P(Y|X_s)$ and $P(Y|X_t)$ are approximately similar, at least for the shared classes between domains.
2. **Reducing the Impact of Distribution Mismatch:**
   While the conditional distributions may differ globally, weakly-supervised methods focus on the subset of target data where the model's predictions are most reliable. This subset is likely to have a smaller discrepancy between $P(Y|X_s)$ and $P(Y|X_t)$, as the model's confidence reflects a degree of similarity in the feature-label relationships. By iteratively refining the pseudo-labels and retraining the model, we can gradually adapt the model to the target domain's conditional distribution.
3. **Handling Shared and Novel Classes:**
   In the context of open set domain adaptation, where the target domain may contain classes not present in the source domain, weakly-supervised methods help identify and separate shared classes from novel ones. High-confidence pseudo-labels are typically assigned to shared classes, while low-confidence predictions may indicate novel classes or domain-specific variations. This selective approach reduces the risk of negative transfer caused by mismatched conditional distributions.

4. **Justification:**
   While weakly-supervised methods do not guarantee that $P(Y|X_s)$= $P(Y|X_t)$, they provide a computationally efficient and scalable way to adapt models to new domains when labelled target data is scarce. The key assumption is that the model's high-confidence predictions in the target domain are sufficiently accurate to bootstrap the adaptation process, even if the conditional distributions are not identical.

In summary, weakly-supervised approaches do not entirely overcome the problem of differing conditional distributions, but they offer a practical framework to mitigate its effects by focusing on high-confidence predictions and iteratively refining the model's understanding of the target domain. This makes them a valuable tool in scenarios where obtaining labelled target data is expensive or impractical.

o You must provide more details on how you carry out the direct transfer learning approach. I'm not an expert but I understand there are different approaches.

Section 3.1 Methodology has been rewritten with the intention of including those aspects that help clarify the transfer learning and domain adaptation approach followed in this work. Above in this letter, we have also described in detail how we carried out the knowledge transfer.

o You must explain more clearly how classified events are compared the database events. From my understanding, the scores are computed on the labels associated to consecutive time windows of fixed length. If this is the case you must state it explicitly in the Methodology. You must also explain how you transform the datasets into labels associated to time windows.

In Section 4, where we describe the results of the experiments, we have added this paragraph indicating how we map the information from the seismic catalog to labels: To perform a robust analysis of system performance based on the accuracy metric (%) and build confusion matrices, it is necessary to transform the information contained in the catalog into labels from which the study can be conducted. Since in experiments 1 and 2 we start with a seismic catalog that contains annotations for the start and end of each event present in each seismic signal, once the signals are preprocessed and windowed, we can associate a label with each window. In this way, each window can be analyzed based on its classification according to its label.

- Although integrating the LAPALMA2021 dataset is interesting, I do not really see a clear link with the main subject of this paper, that is transfer learning. Indeed, you apply directly the Master dataset classifier to the dataset and explore how it allows you to detect events. So there is no added value on the "transfer learning" subject. In my view, to remain in the scope of the paper, you would need for example to compare the results of the Master dataset classifier, to results of the classifier re-trained on thePOPO2002 dataset (by direct transfer learning and/or weakly supervised transfer learning). That would show how data from different volcanoes can be combined to classify events on a new volcano.

As we have argued previously, the experiments and results presented in this article address the suggestions of the different reviewers. To this end, three distinct experiments have been conducted. In the first experiment, classical transfer learning is carried out, where a model trained with Deception Island data is retrained with data from Popocatépetl. The goal of this experiment is to assess how well the system can adapt to the labeled data in the catalog and achieve highly effective results, with a performance of around 90%.

In the second experiment, the goal is to introduce our weakly supervised learning methodology and demonstrate how the catalog obtained using this methodology greatly differs from the preliminary catalog available for Popocatépetl. To do this, similarity results are shown, where it is observed that only 50% of the events initially annotated in the catalog are recognized. Meanwhile, many other events that are now recognized were never considered previously.

Finally, in the third experiment, included in the first round of revisions, the aim is to demonstrate how effective our methodology is for building catalogs from scratch, where no prior information exists. For this, we use data from the 2021 La Palma eruption and compare our approach with a widely used tool like Phasenet, which is also based on AI.

We believe that these three experiments cover the full spectrum of the use of the proposal introduced here, with different use cases. And we do this in response to the demands of previous reviewers.

- In my view the Results section must be expanded a little bit to highlight the main results. Instead of just stating that results are given in Table XX and Figure XX, comment the objectively (e.g. the best accuracy score are obtained with XX, the event with the highest confusion rate is XX, …). Then you can discuss and interpret these Results in the Discussion section.

Given the difficulty raised by several reviewers in following the workflow and following the template of some articles published in this journal, we believe that to facilitate this, it is necessary to include a separate section for describing the results and another for discussing them. Therefore, in Section 4 (Results), we simply describe the obtained results and their meaning. In Section 5 (Discussion), we analyze these results in detail for each experiment.

- It is interesting to see the influence of the probability detection threshold on the Results, why not do it for the POPO2002 experiment as well? How would the confusion matrices of Tables 5 and 7 change with a different probability threshold? Besides, I don't think you mention the probability threshold you use to derive the Results presented in Section 4.

As the reviewer points out, the detection probability threshold is a crucial parameter in the weakly supervised algorithm proposed here, as it controls the system's sensitivity. A very high threshold will only allow the inclusion of events highly similar to those learned in the source domain. A very low threshold will include more diverse events, ultimately enabling domain adaptation.

However, in the context of the classical transfer learning experiment, specifically regarding Table 5, the results remain unchanged because the probability threshold does not exist. In this case, events are classified by assigning the seismic category with the highest probability in the output layer, meaning they are always classified into the most probable category.

Finally, we have added to the experiment description using the weakly supervised approach that the selected detection probability threshold was 50%, aiming to include as many events as possible, even if they were less rigorous.

- You do not clearly explain why you test three different classifiers (RNN-LSTM, Dilated- LSTM and TCN). Is it to determine the best method? To study the variability of results depending on the classification methodology? Although interesting, this is beyond the main scope of this paper which deals with the pros and cons of direct / weakly supervised machine learning techniques. So you should investigate this point in a dedicated Discussion paragraph, rather than throughout the Results section. You can say in the methodology that you tested different methods and retain only one for the main results presentation, but investigate the influence of the classifier in the Disucssion. The same remarks stands for the size of the training dataset : In Section 4.1 you test 20% and 40%, but you do not carry out the same sensitivity analysis for the other applications. I would use the same percentage for all tests (e.g. 40%), and if you deem it important discuss the influence of the training test size in the discussion.

The reason for including these three methodologies and not others in the paper was primarily to test the robustness of the method. We agree with the reviewer that any other methodology capable of analyzing temporal signals could have been used, from Hidden Markov Models to Transformers. However, since this study builds upon previous work using pre-trained and already published systems, we chose these three so that readers can easily find extensive information about these systems and their characteristics, facilitating and streamlining the reading of this paper. Otherwise, we would

have had to describe both the proposed models and their training before addressing classical transfer learning and the weakly supervised approach.

Regarding the percentage of the dataset used for training, we would like to clarify that we included it as an illustrative example to show that when performing classical transfer learning between related domains, it is not necessary to use a very large training dataset to achieve good results. This allows most of the data to be used for testing while still obtaining a high performance, close to 89%.

In the case of the weakly supervised approach, the size of the training dataset depends on the complexity of the signals, and it is up to the user to determine the appropriate size. In this study, we decided to set it at 40% to better analyze the number of detected events, even in scenarios where the training set is relatively small.

- Although the objective of the paper is not to point out that weakly supervised TL approaches can detect more events that direct TL approach or direct application of pre- trained classifiers, I would still expect a quantified comparison on this point. In this respect, I would include the results of the pre-trained classifier, and of the direct transfer learning approach. Besides, you do not clearly show that weakly supervised approaches allow to build less biased catalogues in comparison to other approaches. You do show that events that are not detected in the manually constructed catalogues are identified by weakly supervised classifiers, but you do not show clearly that direct transfer learning are less efficient in building less biased catalogues. In this perspective, the advantage of using weakly supervised approaches in comparison to direct transfer learning approaches is not clearly shown in your manuscript. For instance, how would direct transfer learning approaches for the seismic signal presented in Figure 4?

The results of the pre-trained classifier and the direct transfer learning approach are included in the manuscript. Once again, we would like to clarify that classical transfer learning uses a pre-trained model as a starting point to train a new system with data from a new seismic catalog, in our case, using labels under a supervised learning paradigm. The results are presented in Tables 4 and 5 in Section 4 and discussed in Section 5.

Regarding the results of the pre-trained classifier, these can be found in Table 6. This table consists of two result columns: blind test and weakly supervised. The blind test column corresponds to the results obtained by the pre-trained system when compared with the POPO2002 catalog without re-training. The weakly supervised column presents the results obtained after re-training with the data included in the new dataset, compared to the annotations in the same POPO2002 catalog.

As seen from the results in both tables, along with Table 8, classical transfer learning techniques before re-training are responsible for creating the training dataset for domain adaptation and, as such, contribute to the creation of less biased catalogs. Therefore, the weakly supervised algorithm simply uses the events recognized and labeled by the pre-trained system as labels and training events, adjusting the system to the characteristics of the new events. Applying this use case to the example in Figure 4, the pre-trained system will recognize the inserted LP events and include them in the new database if they meet the detection probability threshold criterion. In this way, once retrained, the model will be able to detect these types of events if they are present in the traces.

As previously mentioned, the results in Table 8 show that pre-trained models detect many events that are not annotated in the catalog, since the weakly supervised approach originates from these systems.

- You must improve the legends of all Figures. The reader must be able to understand their content without referring to the manuscript.

All the legends have been improved for the sake of clarity.

Specific remarks:

- To avoid misunderstandings, I would use “classifier” throughout the manuscript instead of “systems”

The proposed system is not simply an implementation of a machine learning-based classification algorithm. These systems are built around events that are precisely delineated in time, commonly known in the literature as isolated event classification systems. The final output of our system is one that, given the signal's waveform, detects an event and assigns the appropriate label to a specific class. In other words, the system performs both event detection and classification tasks.

- I would mention the data used in the manuscript in the abstract, in its present form it is rather general and the reader does not know how the authors reached, in practice, their conclusions.

The abstract has been modified including the general idea behind the work and the dataset used: ‘When a system trained on a master dataset and catalog from Deception Island Volcano (Antarctica) is used as a pseudo-labeller in other volcanic contexts, such as Popocatépetl (Mexico) and Tajogaite (Canary Islands) volcanoes, within the framework of weakly supervised learning, it can uncover and update valuable information related to volcanic dynamics’

- Table 1 : Add the acronyms used for the events in the first column. Make it clear in the Table / the legend what classification/names you use in your work.

The acronyms have been added to the table.

- L.34 : “such signal processing”, what are you referring to?

Signal processing in this context refers to the spatio-temporal analysis of the seismic signals for comprehending the underlying physics behind the eruptions, and thus understanding why they occur. In the previous sentence, we discussed signal processing to analyze volcanic dynamics, which is why we refer to this signal processing.

- L.94 : You must expand and explain chat Transfer Learning consists in, with references and examples in the literature. Otherwise, a reader that is not familiar with this concepts will not understand what you mean by “re-train”.

In the introduction, we have added a reference to one of the most widely cited works on the transfer learning paradigm. In the methodology section, the previous version already included a brief description of this concept. We have cited the same work again to provide clearer guidance for the reader.

- L.96-99 (“The outcomes … volcanic dynamics”) and l.102 – 104 (“The outcomes … dynamics”) : This is a conclusion of your work, it should not be in the introduction.

These sentences appear in the introduction as they provide a general overview of our methods and findings from the experiments. If the reviewer thinks they should be removed, we are open to doing so. However, we feel these sentences help the reader better understand the context of the paper.

- L.130 “over various time periods or at different volcanoes): I agree that you processing can minimize the difference in signals due to the sensor type, but you do not eliminate the variations associated to temporal evolution of the volcanic system, nor the variations associated to differences in volcanic processes or associated to different paths properties between the source and the sensor.

The sentence has been revised to incorporate the reviewer’s suggestion.: ‘This filtering minimizes the influence of the sensorization used for signal recording and ensuring the comparability of the data recorded by different sensors over various time periods or at different volcanoes (it does not fully eliminate variations related to the temporal evolution of the volcanic system, nor those stemming from differences in volcanic processes or path properties between the source and the sensor)’

- L.139 : Define “pre-eruptive processes ». Do you mean everything that happens in between eruptions, or events that can be interpreted as eruption precursors?

With pre-eruptive processes we refer to a set of phenomena occurring within a volcanic system before an eruption. We added this explicative sentence in the

manuscript: ‘set of geological, geophysical, and geochemical phenomena occurring within a volcanic system before an eruption.’

- L.156 : Although I understand you may not have all the information on the sensors (but do check it, if you use mseed files you should have access to metadata), you must at least say how you got the data. Is it on a public repository? Is there a paper describing the acquisition and data? Where were the stations positioned on the volcano slopes?

In Section 2: Seismic Data and Catalogs, we provide a detailed explanation of all information related to the sensors and databases available. Most data files are in binary format, containing only waveform information. The data were provided by three different observatories, as noted in the acknowledgments. Therefore, in the Data Availability section, we recommend contacting the corresponding author.

Same questions for the LAPALMA2021 database.

- L.247 : You do not explain how you choose the probability threshold.

The following sentence has been included for the sake of the clarity: ‘The system's sensitivity is directly influenced by the chosen threshold: a lower value increases sensitivity, allowing more events to be included but potentially reducing specificity. Conversely, a higher threshold enhances specificity by selecting only the most confident detections, though at the risk of lowering sensitivity The threshold value will be determined by the user based on their needs when addressing the problem. In our case, we have set it at 60%, allowing the inclusion of a greater number of events and better adaptation to the new domain.’

- L.252 : You do not explain what the “desired result” is.

We thank the reviewer for this observation. This was a drafting error. The sentence has been corrected to: "Repeat steps 2 to 4 iteratively until the results converge and no further improvements are observed in the catalog creation, or until the user deems it appropriate."

- L.257 : “some of these methods may not be as effective …” : Be more specific, give examples. Besides, this part should be in the introduction when you explain why (weakly supervised) transfer learning approaches are needed.

This sentence aims to highlight that some of the most widely used techniques in the recognition of continuous seismo-volcanic signals, both offline and in real-time, where a signal may contain multiple seismic signals and the goal is to detect and classify all of them, are not as effective as they should be. Since this is the experimental framework section, this paragraph aims to explain to the reader that, given the nature of the signals and the goal of the problem, many of the classification systems used are not suitable. It also introduces or

justifies the use of the systems proposed in this study (LSTM, Dilated-LSTM, and TCN). We believe that including this in the introduction could be confusing for the reader, as the introduction only addresses the problem of catalog construction, and the type of architecture used to carry out this task is secondary. As we have already mentioned in this response letter, we used these three architectures as a baseline because we started with systems trained on MASTERDEC, and there are publications that support their results. We believe this paragraph is well-placed in the experimental framework section because it motivates the reader to understand the choice and use of these architectures.

- L.231 : What difference do you make between “continuous” or “streaming”? besides you should make it clear at some point that all transfer learning approaches can’t be used in real time.

The main difference between continuous and streaming lies in the type of signal analysis. Continuous analysis involves the examination of signals with the goal of detecting and classifying different types of events. Therefore, streaming refers to the real-time or near real-time analysis of continuous signals, where data is processed as it is received, meanwhile continuous (offline) analysis involves the examination of pre-recorded signals, where the data is analyzed retrospectively.

The sentence has been modified: ‘Some of these methods may not be as effective for the specific challenges posed by continuous or near real-time data processing’

- Figure 3 : Shouldn’t the lines in C) sum to 1? I.e. a frame is necessarily classified as one of the 5 categories? If there’s no detected event, then the window should be classified as noise.

Since these are classifiers with a softmax layer in the output layer, the sum of each output unit corresponds to the probability of the input belonging to each seismic category. Therefore, each input will always be classified into one seismic category, and the number of seismic categories will depend on the catalog from which the classifier was built. In our case, the seismic categories are 5 (Noise, TRE, HYB, LPE, and VTE). In this regard, any input window will always be classified with one of the 5 labels. Once the signal has been analyzed and all the obtained labels are generated, a post-processing step is applied, and consecutive windows with the same label are grouped together, which we interpret as part of the same event (see Titos et al. 2018). If labels of very short duration (e.g., a single frame) of different seismic categories appear consecutively, that part of the signal is detected and classified as an ‘unknown event’, as there is no pattern indicating its association with a specific event. On the other hand, if multiple consecutive labels from the same event are obtained, their average probability of belonging is analyzed, and if it exceeds

the established probability threshold, they are added to the new training dataset.

- L.295 : “a subset”, how do you construct it? What portion of the dataset does it represent?

As described in the manuscript, the aim of this work is to construct a robust seismic catalog with minimal human effort. To achieve this, we start with a system that we consider a "master" system, as it was built from a database that we also consider "master." As outlined in the methodology and Figure 3, the idea is to analyze a subset of data from a new volcano to include some of the detected events from that new subset into a new database, and train the new system with this new data to build a classifier adapted to the new volcanic environment. When we refer to the subset, we are specifically talking about that subset of data from the new volcanic environment that will be used to create the database that drives the domain adaptation of the classifier between volcanoes. Therefore, if this subset is too large, the remaining data to test the robustness of the new catalog will be limited. This is why, as shown in Section 4.2, which describes the obtained results, we used 40% of the POPO2002 data, reserving 60% to test the robustness of the method.

- L.326 – 341: “All results … during training” : this should be in the Methodology section.

Since we are discussing a characteristic specific to the training and setting up of the systems, we believe this feature should be placed in the experimentation section.

- L.339 - : “the model”, which one?

Corrected. It was a drafting error.

- L.340 : What is “early stopping” ?

As the text indicates, early stopping is a widely used regularization technique in the deep learning field to prevent overfitting of the systems.

- Table 4 and 6 : As mentioned in my main remarks, it is not clear why you test 20% and 40% for the training dataset. Besides as you focus in the following on 5 categories only, I would keep the results of the 7 categories for the discussion (if relevant). Thus, I would only keep Table 5 and add a column for the accuracy of the direct transfer learning approach.

Table 4 refers to the results obtained from the POPO2002 dataset when applying a classical transfer learning approach. The inclusion of 20% and 40% of the database in the training set is simply to show the reader that this approach is capable of learning the information contained in the catalog and achieving highly effective results, close to 90%. Including more training data could slightly improve recognition results, but our goal is not this. Instead, we

aim to demonstrate that the systems are learning the information contained in the catalog without disregarding valuable information, meaning they are learning in a biased way.

The inclusion of 5 and 7 categories has the same objective: to inform the reader that when training a system with a predefined catalog, the results are very good regardless of the number of categories included in the training. However, since the objective of the work is to apply a weakly supervised approach, using a pre-trained system on a master database with 5 seismic categories, we are limited to working with and extracting results from only 5 categories. It is impossible to extract results with 7 categories as our system only recognizes 5.

- Table 5 and 7 : Why did you not include the 7 categories of the POPO2002 dataset? Even if you can't predict all categories, it's interesting to see how they are classified. Otherwise, explain in the text why you do not display the 5 categories in the tables.

The objective of the work is to apply a weakly supervised approach, using a pre-trained system on a master database with 5 seismic categories, we are limited to working with and extracting results from only 5 categories. It is impossible to extract results with 7 categories as our system only recognizes 5.

We have tried to make this clear in the text, line 387: 'As previously stated, since MASTER-DEC consists of five seismic categories and the weakly supervised approach builds on pre-trained models, the results presented here include only these 5 seismic categories.'

- Table 8 : As stated in the main comments, I would add the number of detected events for the original classifier and for the classifier obtained with the direct transfer learning approach. You should also add a line for the HYB events.

The results in Table 8 are the same as those in Table 6; however, one shows the number of original events and the number of recognized events, while the other shows the percentage of matching events between the approaches and the original catalog. Regarding the hybrid events, they are not included in Table 8 because none are detected, and in the original catalog, as shown in Table 3, there is only one.

- L.419 "The vast majority", l.422 "many times" : You must quantify these statements.

Besides, your remarks questions indeed the validity of the accuracy score computation. Couldn’t you compute differently using events rather than time windows? E.g. for an event with start time t1 and end time t2, you label it with the label most represented in the successive tie windows. It would be a more robust accuracy estimation, eliminating “artifacts” associated to SNR or nested subevents, and prove that (i) you do detect rather correctly the events of the catalogue, and (ii) are able to refine the events duration and detect sub-events.

When we refer to the "vast majority," it is difficult to quantify the exact number, as what we are describing is that, in the original catalog, what was initially known as "garbage" or "tremor" in our system is associated with valid event labels. Analyzing the information in Table 7, the confusion ratio of each seismic category can be observed. Regarding the validation of the system performance, applying the approach suggested by the reviewer is complicated, as the start and end labels of each event depend on the subjectivity of the human operator who created the catalog. Therefore, to compare at the event level, we would need to define when a detected event is considered correct, even if the start or end does not exactly match the annotations in the catalog. This is why we opted for window-based recognition, which ultimately indicates what percentage of the events are being recognized. Additionally, once the recognition at the window level is obtained, a grammar is applied, from which the number of recognized events is derived, as noted in Table 8.

- L.475-476 ; “previously hidden information (…) can be obtained”

Hidden information has been changed to unannotated information for the sake of clarity.

- L.537 : It is strange to have a paragraph “Summary of findings”, and a paragraph

“Conclusion”. The conclusion is precisely about summarizing the findings.

The summary of findings and conclusions, although similar, address different aspects of the previous review processes. In the summary of findings, the general results and conclusions of the experiments are described. This section was suggested by a reviewer to clarify the experimental framework of the work, which seems to be quite confusing for readers not familiar with these techniques. On the other hand, the conclusions address the overall and final points of the work.

- L.552 : If you mention the issue of membership threshold in the conclusion, I would expect you to investigate this issue not only for the construction of a new catalogue from scratch, but also for the weakly supervised methodology to train the classifier.

This is an interesting suggestion. The choice of threshold is a very important parameter, as it determines which future events will be included or excluded

from the new training database, from which the adaptation to the new domain will be carried out. In this study, we have decided to set a very low threshold, around 60%, to include as many events as possible in the new database. Although studying the effect of the threshold would be interesting, we believe it is highly dependent on the specific objectives of the observatory or the problem being addressed. That is why we have only considered this analysis with LAPALMA2021, because Including this analysis for POPO2002 would significantly extend the work.

- L.572 : You do not investigate unsupervised learning techniques in your work, so your work does not “demonstrate” that these approaches are more successful.

We agree with this comment. In this study, we did not evaluate unsupervised learning techniques. The reason for not including them is that our focus was on studying semi-supervised learning techniques. However, we are currently working on unsupervised approaches based on constructive learning to analyze their capabilities. It is also important to highlight that using unsupervised learning techniques inherently requires a posteriori analysis by experts to interpret the clusters identified by the system. Since our goal is to minimize human review efforts, we believe that utilizing a master database could assist in constructing less biased catalogs.

- L.575 : I agree that using data from several catalogues could help develop “universal” monitoring tools, and you have the opportunity to investigate this in your work: use classifier trained on the master dataset, transferred with weakly supervised approaches to the Popo2 catalogue, and tested on the LAPALMA dataset. You could then compare the result with the ones obtained with the original classifier from the Master dataset, transferred directly to the POPO2 catalogue, and tested on the LAPALMA dataset. This would be very interesting.

We agree with the reviewer that this experiment is very interesting. In fact, the authors of this study previously evaluated it. However, including this experiment in the manuscript presents a challenge. As we have argued in the text, a complete catalog of the seismic-volcanic data from La Palma is not available; there is only a catalog that includes some of the earthquakes detected by human operators during the seismic crisis. Therefore, it would be impossible to conduct a comparative performance analysis of the systems without a reliable reference catalog.

In this regard, we are collaborating with INVOLCAN technicians to create a more comprehensive seismic catalog that will allow us to carry out this experiment. We take this opportunity to invite the reviewer to collaborate with us on this new study. We encourage him/her to contact us, and if she/he has any other databases with reliable annotations, we would be happy to include

tit in future work. This study would analyze both the LAPALMA dataset and his/her database, using MASTER-DEC and POPO2002 as master datasets within a weakly supervised learning framework.

Minor remark :

- Title : “catalogus” -> catalogues

Corrected

- L.34 “frequency”, it is not clear whether you speak of the signal frequency content or of the occurrence frequency.

Corrected. We were talking about occurrence frequency.

- L.49-52 : there are too many references. you should develop on a few of them to explain their main results / methods.

All the references were included to highlight the variety of models developed for recognizing volcano-seismic signals using machine learning approaches. Additionally, we felt it was important not to omit any, as excluding some could create gaps in the discussion. To keep the text concise while maintaining a rigorous state of the art, we decided to include all of them

- L.54 and following : why is this part in italic?

This is the most significant challenge when constructing such systems, given the unique nature of the volcanoes and the data. As a result, we aimed to emphasize this challenge.

- L.83 : Deceptio -> Deception

Corrected

- L.118 : “our hypothesis”, what are your referring to?

We introduced our hypothesis earlier, around line 86: ‘We hypothesize that, often, automatic recognition systems are not capable of modeling the spatial-temporal evolution of seismic events. Instead, they learn to recognize the probabilistic pattern-matching observed in their training data. In other words, rather than simply learning to characterize volcanic dynamics by describing the latent physical model, catalog-induced learning biases the system’s performance as it learns the description of the data annotated in the catalog, potentially discarding useful data that describes volcanic dynamics. Therefore, we conclude that using systems trained with a master database (complete and large) as pseudo-labeler, could help create less biased catalogs from which the systems can be retrained and adapted to different volcanic environments.’ This sentence refers to our hypothesis.

- L.124 : How can you have 8 channels on a three-components seismic sensor? Chat are these channels?

We believe that the reviewer has not properly understood the sentence describing the sensorization used in the data collection for Deception Island. The text literally states: The Deception Island dataset (hereafter referred to as MASTER-DEC) was created using seismic data collected during the 1994-1995 campaign organized by the Andalusian Institute of Geophysics (IAG) with a short-period array of 8 channels. The array consisted of a three-component Mark L4C seismometer with a lower frequency band of 1 Hz and five Mark L25 sensors with a vertical component frequency of 4.5 Hz, electronically extended to 1 Hz. As can be seen, the array consists of one three-component seismometer and five single-component vertical sensors, adding up to a total of 8 channels.

- L.144 : “UMAP”, give a reference, how did you compute it?

Included. The application of UMAP approaches are explained in Supplementary material.

- L.212 : “domain information”, what do you mean?

This section has been rewritten almost in its entirety to enhance understanding and incorporate the mathematical foundations suggested by the reviewer.

- L.213 : You do not explain what Ys and Yt are.

This section has been rewritten almost in its entirety to enhance understanding and incorporate the mathematical foundations suggested by the reviewer.

- L.230 : You have not yet explained what are RNN-LSTM, Dilated-RNN and TCN, you do it only l.264. As stated in my main comments, the Methodology section can be re-organize to avoid this kind of problem.

Modifying the entire methodology section would conflict with the comments made by the previous reviewers. This sentence has been included in the paragraph to address the lack of information: ‘For our experimental framework, we will base our approach on the pre-trained systems previously published in Titos et al. (2018, 2022, 2024). These systems include Recurrent Neural Networks (RNN), Dilated Recurrent Neural Networks (Dilated-RNN), both utilizing LSTM cells, and Temporal Convolutional Networks (TCN). These models, referred to as RNN-LSTM, Dilated-LSTM, and TCN, generate a probabilistic event detection matrix with per-class membership outputs’.

- L.279 : “three systems”, I understand that you refer to RNN-LSTM, Dilated-RNN and TCN trained on the Master datasets, but when first reading the sentence it is not obvious.

Corrected

- L.291 : “our initial hypothesis”, at this point, the reader may not remember what your initial hypothesis is.

We have completed the sentence by including a brief reference to our hypothesis: ‘To test our initial hypothesis—that automatic recognition systems often fail to model the spatial-temporal evolution of seismic events, relying on probabilistic pattern-matching from training data, which can introduce biases and overlook valuable information about volcanic dynamics—and following…’

- L.295 : “Each” -> each

Corrected

- L.496 : “because the” -> because of the

Corrected

- L.512 – 516 “The first row … respectively”. This should be in the legend, not in the main text.

Corrected

- L.532 : There a missing number after “Figure”

Corrected

Answer to comments of Reviewer#3

Dear reviewer#3, We are very thankful for your thoughtful suggestions. Below, we present how we have addressed them.

The Authors describe the automatic labelling of seismic activity in volcano environment. The manuscript is interesting because it deals with a very current topic, that is, the automatic management of large amounts of data that would require manual work that is very expensive in terms of time and human resources. A solution that has been widely used in recent years uses machine learning techniques, that is, training an algorithm to make decisions by replacing us. The issue at this point becomes that of minimizing the errors made by the algorithm so that the results are reliable. To this end, using a database as a reference point to train a machine learning algorithm that can then be applied to other databases is essential. The Authors declare great knowledge of Deception Island volcano and use the seismic database of this volcano as benchmark for other 2 volcanoes database to build an automatic machine learning based procedure aimed to recognizing the seismic event type among 5 possible event sources in order to detect pre-eruttive signals. The topic is of great general interest, because the labelling of seismic events concerns modern seismology in general because of the proliferation of increasingly dense seismic networks that collect an ever-increasing amount of data but the paper needs a major revision before publication because Author should do an effort to explain their work in a more concise way.

All sections are too long and repetitive and fail to stress the most important parts of the method and data processing. The manuscript fails to indicate in clearly and concise way the necessary and important parameters used by the (proposed/used it is not clear) methodology and fails to describe the dataset used.

We agree that some parts of the text are repetitive and lengthy. This is a result of the revision process that was carried out. In the first round of revisions, we had to adapt the manuscript to the suggestions of eight reviewers. Each reviewer proposed improvements in different sections, which extended the text and sometimes made it repetitive. We have attempted to address this by conducting a detailed review and removing sections that do not contribute to the manuscript. However, it is difficult to carry out such an extensive revision with so many reviewers without the text being affected. In the methodology section, we have attempted to clarify which parameters are necessary to implement the algorithm. In this case, there is only one key parameter: the probabilistic membership threshold. Everything else corresponds to training parameters, which are extensively described in the referenced articles.

Since the Authors state that the code and data are available upon request, it is necessary to show an example of data, to explain how the data is acquired, treated, processed and used, using explanatory figures. The manuscript then fails to describe the software and the dataset to reproduce their results.

This article incorporates a weakly supervised methodology to the results obtained from systems that have been previously published. The recognition systems used in this work, namely LSTM, Dilated-LSTM, and TCN, are not only referenced in the paper (having been published and analyzed with the same MASTER-DEC dataset), but are also widely known within the scientific community (Hochreiter, S., Schmidhuber, J., 1997; Schmidhuber, J., 2015; Chang et al. 2017; Lea et al., 2017).Therefore, we believe that they do not require a detailed description in the text, but rather just a citation. A similar approach applies to the seismic signal preprocessing pipeline (Titos et al. 2024). We make use of a widely used and well-known pipeline in the scientific community, based on a log frequency scale filter bank.

The citations are not in the correct position in the text and do not help the reader understand what they refer to. So, it is not clear if authors used a particular software for their automatic labelling or they propose a new software that they wrote themselves. The data used are not well described together with the construction of the sub-dataset for training. Maybe a schematic sketch of the seismograms' processing can help. The use of acronymous should help the reading but are explained after their use. The language is too qualitative (what do Authors mean for reliability, for acceptance? Did the Author set any thresholds? Which are the parameters used to build the initial sub-datasets? How are results affected by these decisions?) and the reading is in some parts frustrating since it is hard to get to the heart of the problem that is: can we leave these algorithms work alone? If so, which is the uncertainty of the results?

Citations: we sometimes place them at the end of the text when referring to a general idea, or within a sentence when we are referring to a specific concept within that sentence, as is the case described by the reviewer. We have tried to italicize the cases where the concept within the sentence itself is important, in order to avoid misinterpretations.

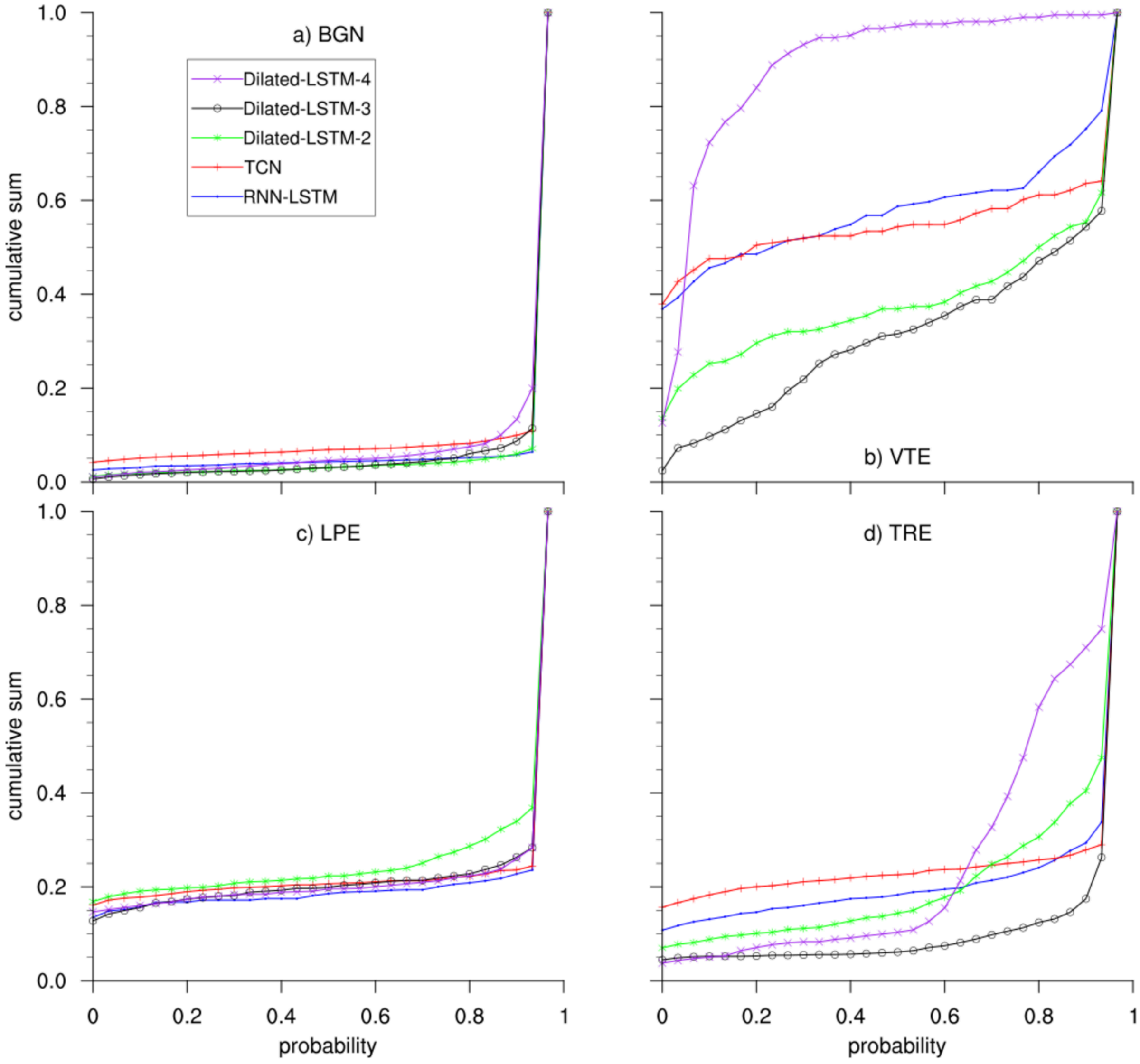
Automatic labelling: This work proposes a methodology to create robust seismic catalogs while reducing the cost or human effort required for their development. To achieve this, systems trained on a master database are used as labelers to generate a new database, which is then used to retrain these systems and adapt them to a new volcanic environment. In this study, this software (developed by us and previously published) is utilized, and only the code necessary for selecting events—using a weakly supervised approach—is developed in this work to determine which events will be used for system retraining.

Data sets: We have used three different databases, all of which have been thoroughly documented. We have provided details on both the databases themselves and the sensorization used to record these data (when such information is available). Regarding preprocessing, we have outlined the general pipeline and referenced articles that describe this process in depth, as it is a widely used and well-known concept in the field. We have attempted to adjust the language to improve readability and comprehension, as well as to organize acronyms and references more systematically. Finally, by "reliability," we refer to the system's robustness in detecting and classifying seismo-volcanic events. Specifically, we use this concept to highlight that the reliability of an automatic recognition system is traditionally assessed by its accuracy across all events in a catalog. A system with an average performance below 75% is generally considered unreliable. However, this often happens not from the system's ability to differentiate between events but from how the catalog is constructed. If seismic categories are not homogeneous and events of different natures are grouped under the same type, system performance declines. Inconsistent categorization undermines learning, leading to recognition accuracy below 70%.

Threshold and hyperparameter: The choice of the threshold is a very important and interesting issue. We are considering writing a new paper analyzing this parameter, as including it in this work would make it excessively long. In this study, we opted for a low threshold to allow the inclusion of a larger number of events in the retraining database in order to demonstrate the usefulness of the method. Choosing a higher threshold would bias the dataset toward events that are almost identical to those in the master database, thereby reducing the system's ability to adapt to the new domain. Therefore, when creating a retraining database, only the threshold needs to be determined. However, the user can also adjust other typical training parameters of neural network-based models, though this is not strictly necessary—rather, it is a user decision to improve convergence or adaptation to the new domain. In this work, none of these parameters are modified. We simply choose the threshold and keep the hyperparameters inherited from the previously trained models fixed.

Uncertainty: The uncertainty of the results can be analyzed from two perspectives.

The first is through the detection matrices. If the detection results in probabilistic terms are very high, we can assume that the detected events are reliable. As shown in the attached image, which describes the cumulative distribution function for each event in the master database, in the case of LPEs, for example, 90% of the events are detected with probabilities higher than 85%. A similar pattern is observed for Noise and TRE events. A slightly less robust recognition is only noticeable in the case of earthquakes; however, given their characteristics, the systems still detect them quite effectively.



The second approach would involve a case-by-case analysis by an expert on the volcano, where each detection and classification would be validated or discarded. However, this is a very costly task, so we rely on the class membership probabilities obtained by the systems, based on the previous argument.

I think that the manuscript should be rewritten because some parts must be moved in other sections (some examples are reported in the comments' list below). The sections are introduced with phrases that can be eliminated without changing the meaning of the speech. The manuscript should assume a consequential structure aimed at making it clear exactly what the data are, how data are treated and processed, how data are organised, how data are used in what sequence, how the software works and how each choice made previously can influence the subsequent choices and the final results. When a new approach is proposed, it is essential that the comparison between the decision of the algorithm and that of the human seismologist be clear in order to assign a reliable uncertainty to the results.

As we mentioned earlier, this article was reviewed in its first iteration by 8 reviewers, each with a different background and expertise. Each of these reviewers conducted an exhaustive review, and the authors addressed almost all the changes, which is why the article has changed significantly since its first version. Some reviewers suggested including experiments and results comparing different artificial intelligence approaches or architectures for creating catalogs with different databases, while others suggested expanding the introduction and structuring the discussion and results section. Therefore, we structured the article into 6 sections. The introduction, where we motivate the problem of catalog construction and why monitoring systems might be biased. The second section describes the data and the volcanoes under study. In this section, we describe both the database and the available catalogs. In section 3, we describe the proposed methodology and the experimental framework of the work, where the three experiments conducted to test the robustness of the methodology are outlined. In section 4, we present the results for each experiment. In section 5, we describe and discuss the results of each experiment, and finally, in section 6, we conclude the work. We believe this structure is appropriate for describing a work that is complex to understand for those who are not familiar with these types of techniques. Although the structure suggested by the reviewer could also be suitable, we must maintain a balance between the suggestions from different reviewers, which is why we cannot address such a profound structural change, especially when, in the second review, 2 out of the 4 reviewers suggested only minor changes. What we have done is a thorough reading of the article, removing redundant and repetitive phrases, lightening the text, and making it easier to understand.

In the following, some non-exhaustive comments:

Line 83 Deceptio instead of Deception.

Corrected

Line 140. “For the current study, we extracted a subset of reliable data, consisting of 2,193 seismic events”. Can the Authors quantify the meaning of reliable? Do they refer to any quality indicator of location as residuals, hypocentral errors,…

When we refer to reliable data, we mean that we have selected seismic events that meet prototype standards and that the geophysical experts who have been monitoring the volcano since 1986 agree that these events correspond to the type they have been categorized as.

Line 144. What are the event parameters that make up the UMAP projection and that are represented graphically? It is not explained so the Figure 2 does not make any sense.

The text states: *Figure 2 illustrates the UMAP (Uniform Manifold Approximation and Projection) projection (McInnes et al. 20218), showing the distribution of the five MASTER-DEC event types within the feature representation space.* The feature representation space refers to how each event analysis window is represented with UMAP, after performing parameterization using the log frequency scale filter bank, which is the chosen preprocessing pipeline. We have added a sentence in the text that describes how the events have been parameterized using a logarithmic scale filter bank, which represents how the energy of the different events is distributed across the various frequency bands, in order to facilitate understanding.

Sentence included: 'The representation space aligns with a log frequency scale filter bank, which captures the energy distribution of each event across various frequency bands. For a more detailed explanation of how the workflow constructs the feature vectors, please review Titos et al.,2024.'

Table 2. Did the Author classify the seismic events for this paper, or the classification refers to another paper? If they did the classification in this paper, they should explain how they did it. If not, they should refer to the exact citation in the table caption.

None of the databases used in this work were created here. All the databases are inherited and constructed by experts from the different volcanoes. In the case of Table 2, which presents the events in the MASTER-DEC database, we simply describe some of the event characteristics. The construction of this database was carried out by a team of expert geophysicists with extensive experience in monitoring the volcano. As referenced in the text, they have published a wide range of works providing a comprehensive understanding of the structure and dynamics of Deception Island volcano through numerous campaigns conducted since 1986. The database corresponds to the data collected during the 1994-1995 seismic campaign. This information is already described in the text. In the case of Table 3, which describes the POPO2002 database, the data was analyzed and labeled by experts from that volcano and provided by Dr. Raúl Aránbula for conducting the study and experiments in this work. Finally, the LAPALMA2021 database was provided by Dr. Luca D'Auria, along with the seismic catalog describing the recorded earthquakes obtained by INVOLCAN staff.

Line 163. "the signals were first filtered to match" do they authors mean 1-20 Hz filter? Please specify in the text.

This sentence explains that during the data preprocessing, before applying the filter bank, a filter is applied to adjust the sampling frequency of both databases to 50 Hz, ensuring that the obtained feature vectors are comparable. This step is necessary because the Deception Island data corresponds to the

1994-1995 campaigns, while the data from POPO2002 and La Palma come from more recent campaigns, where the seismic sensors recorded signals at higher frequencies, such as 100 Hz.

Line 177. “the inclusion of this use case could be of interest” rephrase.

Corrected

Table 3. See the comment to Table 2.

The same response applies to Table 2. The POPO2002 database was constructed by experts from that volcano during the 2002 seismic campaign. All available information regarding the construction of the database is included in the text.

Line 179. Same as line 163.

Corrected

Line 183-189. The introduction to Section 3 is quite confused. The first sentence is too long and the meaning is lost. Please rephrase. Regarding “once its functioning is understood” do you mean understood by you or by the reader? It is a rude language that insinuates that the reader might not understand. The manuscript needs a thorough rereading and rewriting to be better understood.

To streamline the reading of the article and avoid redundancy, we have removed this sentence. In the previous version, it was simply a reminder of the initial hypothesis we aimed to test. Regarding "once its functioning is understood," we just wanted to explain that, once the methodology is described, we would proceed with the experiments. This part of the paragraph has also been revised to prevent any misinterpretations.

New paragraph: ‘This section outlines the methodology and experiments conducted in this work. The proposed algorithm will be described, followed by a detailed explanation of the three experiments conducted. The results of these experiments will be presented in the results section’

Line 202-206. After 9 pages of introduction, very long and confused, finally the authors start to describe their work. It is not clear to me whether the methodology is proposed or applied because often the references are not in the correct position in the text, that is, at the end of the paragraph to which they refer, but in the middle of the speech as if the sentences were paraphrases. Example: “The goal is to address a domain adaptation task (Kouw and Loog, 2019; Farahani et al., 2021) to reduce the cost of developing a reliable seismic catalog and database for a new given dataset with minimal initial human supervision.” Do the citations refer to the entire methodology used, or do they refer only to the execution of that part of the task?

We agree with the reviewer that the introduction is quite long and sometimes unclear. However, as mentioned previously, this is a result of the suggestions

from the 8 previous reviewers, each of whom had a different perspective on the work.

Regarding the references, we sometimes place them at the end of the text when referring to a general idea, or within a sentence when we are referring to a specific concept within that sentence, as is the case described by the reviewer. We have tried to italicize the cases where the concept within the sentence itself is important, in order to avoid misinterpretations.

Line 215. The Authors declare “Such assumptions have important implications” but they spend few word to explain the difference between marginal distribution and conditional distribution. How did the Author choose the events that must belong to the two distributions? Did they compare the results with other choices?

In accordance with another reviewer's suggestion, this section has been completely rewritten. The meanings of marginal and conditional distribution are now better explained.

Line 223. “events showing characteristics similar” how the authors identify similarity? Can they quantify this choice? Similarity refers to some characteristics of the seismograms, of the frequency content, of amplitude, of magnitude, of location, of signal length, of coda waves, of body waves,….?

In the context of this work, when we refer to similar characteristics, we mean events that have similar waveforms and spectral content. Translating this similarity into the feature space, we refer to points in the feature representation space located in nearby regions (Figure 2 in the manuscript). Since, as we have described throughout this letter and the manuscript itself, our analysis windows or frames are parameterized using a filter bank, what we are representing is the energy distribution in each of the bands covered by each filter. Therefore, similar events will occupy similar regions in the feature representation space.

Line 230. The acronymous are used here for the first time and are not explained. Maybe a citation is needed here?

According to the comments from other reviewers, this paragraph has been modified, and the references describing the baseline models have been included.

Line 233. How long is the frame or the window? Does this choice affect the results? And how much?

In this work, each analysis window or frame has a duration of 4 seconds, with an overlap of 3.5 seconds with the previous frame. The duration and overlap are parameters that can affect system performance. However, this study does not focus on these aspects, as we are using previously published models whose best results were achieved with these encoding characteristics. The

adjustment of the window size and the overlapping between them was extensively studied in Titos et al. (2018) and Titos, 2018 (Doctoral Thesis Dissertation).

Line 260-265. This sentence should come earlier in the text. Section 3.1 is too general and do not help to understand the method. I propose to eliminate it or to include it in the subsequent sections where the methodology is explained.

Section 3.1 has been partially rewritten. First, we have expanded the formal description of the domain adaptation problem. Second, we have improved the wording of the proposed methodology. Finally, regarding lines 260-265, we have simply included this information earlier in the text, specifically in the paragraph describing the recognition models used in the proposed methodology.

Section 3.2.1 it is another introduction. Did the Authors mean that they used the approach of Weiss et al.? They should be more concise.

In Section 3.2, we simply define the first experiment within our experimental framework. The reference to Weiss et al. corresponds to the concept of Transfer Learning itself. Essentially, what we aim to convey in this section is that instead of building a system from scratch, we will retrain existing models using the available data and labels from POPO2002. This approach is known as classical transfer learning.

Line 320. “This section presents the results supporting the experiments outlined in the previous section” it is obvious. Please avoid these explanations in the text.

Corrected

Line 323. Did the Authors compare the automatic results with a manual inspection of the data automatically labelled in order to evaluate the accuracy between the automatic choice and your best accurate human one?

The results and the accompanying images in the results section, where a detailed analysis is conducted, correspond to the manual inspection referred to by the reviewer.

Line 361. “The y-axis corresponds to the real label or ground-truth and the x-axis corresponds to predicted labels.” Is this the correct labelling of the master dataset to which the results must be compared to?

The results in Table 7 correspond to the confusion matrix obtained by the different systems, using the manual annotations from the POPO2002 catalog as a reference.

Line 411. “Once the construction of catalogs through transfer learning has been discussed, we are now ready to discuss the use of weakly supervised pseudo-labeling approaches.” As Line 320.

Removed

Line 413. I beg to differ with this statement. Results are not clear and comparison between the automatic labelling of the master dataset with the manual labelling is missing or not well explained.

The sentence the reviewer refers to in this comment is: "Thus, although system performances range between 85% and 90%, this does not always reflect a complete or unbiased seismic catalog. Rather than solely learning to characterize volcano dynamics based on an underlying physical model, the systems may be learning the information contained within the catalog itself. Consequently, catalog-induced learning could limit a system’s ability to generalize, potentially obscuring information relevant to advancing our understanding of volcanic behavior." This sentence aims to convey that when a system is trained with a predefined seismic catalog (constructed under specific circumstances and for a particular purpose), the training process itself adapts the way the systems detect and classify different seismic events to minimize errors compared to the catalog annotations. In contrast, when using a pseudo-labeler built from a master database, the system detects and classifies the different events without the implicit human bias. We believe that this conclusion is concise and clear, and does not require a detailed discussion.

Discussion section. This section is too long and include figures that belong to the results and that can be useful in earlier part of the manuscript. I suggest a deep reorganization of the paper.

This article follows the structure of others published in this journal and the suggestions of several reviewers. In the results section, we present the outcomes obtained within the experimental framework. In the discussion section, the results are discussed in detail, and both the pros and cons of the methodology are argued. This is why most of the figures are found in this section, as the summarized results in tables are presented in the results section.

In the manuscript the reference to the figure is sometimes written as Fig. and other times as Figure. Please check.

Corrected. According to the journal's writing template(as far as we know), when the word "Figure" begins a sentence, it must be written in full. However, when "Figure" is part of a sentence, it can be referenced as "Fig." In any case, since this is a drafting error, we will consult with the journal to correct these issues before the final publication.

Answer to comments of Reviewer#4

Dear reviewer#4, We are very thankful for your thoughtful suggestions. Below, we present how we have addressed them.

General Comments

In this second iteration of the manuscript, the authors have clearly devoted significant effort to refining and restructuring their work. The result is a substantially improved document that showcases clearer objectives, methods, and outcomes. Across all sections, the organization and writing style have been noticeably enhanced, making the overall manuscript much more coherent and accessible.

The Introduction is particularly strong, providing both a concise background and a clear statement of the research motivation. In addition, Section 2, which focuses on seismic signals and data catalogs, has been reconstructed in a way that captures the essential details of catalog construction and usage. This section now offers a thorough explanation of how seismological data is collected, cataloged, and analyzed, setting a solid foundation for the subsequent methodological discussion.

The Methodology section has also undergone a marked improvement compared to the previous version. The authors' decision to outline each step more systematically—especially how the three experiments are structured—makes it much easier for readers to follow the logic and replicate the work. Notably, the emphasis on **pseudo-labeling** as part of their weakly supervised learning strategy deserves commendation. By using a pre-trained model as a pseudo-labeler and then re-training with the newly labeled data, they demonstrate an innovative approach to semi-supervised or weakly supervised classification in seismo-volcanic signals.

Regarding the Discussion, one of the central points the authors address, which is particularly interesting for the field, is the relatively low recognition rate compared to existing reference catalogs. They offer a plausible explanation that these catalogs, while established, may be incomplete or biased toward particular classes of events. Consequently, a strict comparison against them can underestimate the efficacy of the new system.

Along the same line, the authors highlight the **quality vs. quantity** dilemma. While the weakly supervised methodology might introduce some degree of noise or misclassification, it also increases the overall number of detected events, thus expanding the catalog. According to their description, it would be ideal for future users of this methodology to strike a balance by conducting manual checks on a fair portion of newly labeled events to verify their authenticity. These checks not only help mitigate the risk of accumulating errors from pseudo-labels but also lend credence to the claim that genuinely overlooked events are being discovered. Nevertheless, **we recommend** that the authors (and future users) **explore**

additional statistical consistency checks and cross-comparison with alternative detection methods to further strengthen the reliability of these expanded catalogs in subsequent research projects. By systematically verifying or filtering pseudo-labeled events—through model agreement, confidence thresholds, statistical checks, and domain-expert reviews—one can reduce the risk of error accumulation and improve the quality of the final training data.

From a contextual usefulness standpoint, the authors argue that any additional events— correctly identified or carefully verified—enrich our understanding of volcanic processes, potentially offering earlier or more nuanced insights into volcanic unrest. They stress that while it is important to measure success against established reference catalogs, it is equally crucial to recognize the value in uncovering smaller or subtler events that might have gone undetected.

As a result, even if the system does not perfectly align with existing catalogs, it may enhance real-time monitoring, inform hazard assessments, and ultimately lead to more comprehensive research in volcano seismology.

Nonetheless, further elaboration on the potential pitfalls of pseudo-labeling, along with additional quantitative or expert-driven validations, would strengthen the overall argument.

Despite these minor weaknesses, this manuscript now provides a valuable contribution to the application of machine learning within volcano seismology. The authors' demonstration of how to construct and refine catalogs, leverage pre-trained models, and evaluate performance across multiple experiments will be extremely useful in guiding future research. Overall, the revision is a notable success, and the text should serve as a new reference for continued advances in the automated recognition and analysis of seismic-volcanic signals.

About very Minor writing issues:

• Introduction.

- line 52: A period "." is missing after "etc".

Corrected.

- line 54: Maybe lose instead of loss?

Corrected.

- line 54: an interesting topic: "monitoring systems loss effectiveness when recognizing events over time.." it would be ideal to include some references to support this point.

Corrected. We have included references to two of our articles where we tested how the systems perform when using data from the same volcano, obtained from different seismic campaigns.

- line 83: Deception misspelled.

Corrected.

- line 108: there is an extra period ".".

Corrected.

- line 110: "volcano" misspelled.

Corrected.

- Seismic data and catalogs.

- line 123: as stated in fig.1., the data was also collected in 1996 and 2001-2002?

Corrected. It was a drafting error. Dato was collected in 1994-1995.

- lines 150 and 151: are we using "Popocatépetl" with or without an accent?

Corrected.