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

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

Specific remarks:

- To avoid misunderstandings, I would use “classifier” throughout the manuscript instead of “systems”
- 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.
- 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.
- L.34 : “such signal processing”, what are you referring to?
- 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”.
- 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.

- 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.
- L.139 : Define “pre-eruptive processes ». Do you mean everything that happens in between eruptions, or events that can be interpreted as eruption precursors?
- 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? Same questions for the LAPALMA2021 database.
- L.247 : You do not explain how you choose the probability threshold.
- L.252 : You do not explain what the “desired result” is.
- 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.
- 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.
- 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.
- L.295 : “a subset”, how do you construct it? What portion of the dataset does it represent?
- L.326 – 341: “All results … during training” : this should be in the Methodology section.
- L.339 - : “the model”, which one?
- L.340 : What is “early stopping” ?
- 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 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.
- 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.
- 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.
- L.475-476 ; “previously hidden information (…) can be obtained”

- L.537 : It is strange to have a paragraph "Summary of findings", and a paragraph "Conclusion". The conclusion is precisely about summarizing the findings.
- 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.
- L.572 : You do not investigate unsupervised learning techniques in your work, so your work does not "demonstrate" that these approaches are more successful.
- 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.

Minor remark :

- Title : "catalogus" -> catalogues
- L.34 "frequency", it is not clear whether you speak of the signal frequency content or of the occurrence frequency.
- L.49-52 : there are too many references. you should develop on a few of them to explain their main results / methods.
- L.54 and following : why is this part in italic?
- L.83 : Deceptio -> Deception
- L.118 : "our hypothesis", what are your referring to?
- L.124 : How can you have 8 channels on a three-components seismic sensor? Chat are these channels?
- L.144 : "UMAP", give a reference, how did you compute it?
- L.212 : "domain information", what do you mean?
- L.213 : You do not explain what Ys and Yt are.
- 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.
- 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.
- L.291 : "our initial hypothesis", at this point, the reader may not remember what your initial hypothesis is.
- L.295 : "Each" -> each
- L.496 : "because the" -> because of the
- L.512 – 516 "The first row ... respectively". This should be in the legend, not in the main text.
- L.532 : There a missing number after "Figure"