Reviewer#2:

Dear reviewer#2, thank you for your valuable revision. In the following paragraphs, we respond to each of your suggestions.

This manuscript can be better written, and its science better executed. As it stands, the manuscript appears too ambitious in scope. The results being presented are insufficient to deliver the intended scientific message, and the methods described lack sufficient details for reproducibility and related discourse.

In this work, the authors put a strong emphasis on using weakly supervised frameworks to improve seismo-volcanic catalogs for eruption forecasting or early warning. While there is novelty in the application of weakly supervised approaches in the context of volcano seismology, far too many details are left out on the seismology front, and the discussion in relating (improved) seismic catalogs to eruption onset or characterization is clearly absent.

Consider the following issues:

1. Throughout the manuscript, the authors utilize catalogues derived from Deception Island and Popocatepetl. Strong words are used to assert their robustness and quality, yet readers lack information on the related monitoring network, duration of observation (Deception Island), and contemporaneous volcanic activity for which each catalog was constructed. There is some passing discussion on "seismic attenuation processes" and "source radiation patterns", as well as the proposed underlying mechanisms behind each signal type from literature, but how do we know for sure if there is no information on the seismic network geometry, source-receiver distance, or eruption style being recorded?

This work comes from a database that has been widely used in previously published studies. Throughout the paper, references describing the monitoring periods, the network's geometry, and the instrumentation itself have been provided. Additionally, these studies already contain clear references and figures depicting the waveforms and spectrograms of the signals used. We believe that including this information again in this work would be redundant. However, if the reviewer thinks it would be helpful, we will provide a detailed description of this information in the next version of the manuscript.

2. Even if we were to assume that the catalogs are 100% accurate in their labels, this does not mean that they are necessarily suited for machine learning applications. When building a classification model, at least some care must be taken to balance the labeled dataset, especially if accuracies are being used as a metric. A perfectly labeled catalog could still be deficient in certain classes which the model hopes to classify. In such cases, the resultant biases need to be more thoroughly discussed.

We agree with the reviewer's comment; however, the collection of events on a volcano is subject to its activity and the observation period during which the

database was created. In this regard, we have attached Cumulative Distribution Functions (CDF) for the reviewer's reference, which represent the recognition level achieved by the classifier, once trained, on this database within the test set. **These probabilities can be used as an indicator of the confidence level in the recognition**. As can be observed, noise, tremor, and long-period events each achieve recognition rates above 90% probability of belonging in 90% of the evaluated events during the test period. This demonstrates that both the database design and the training process are adequate. It is important to highlight that the recognition of earthquakes shows a decline, but this is due to the inherent nature of the database, its imbalance typical of a real-world scenario, and the specific characteristics of the event. Given the results obtained here, we believe that transferring this knowledge to other volcanoes or to different observation periods of this same volcano is justified. We could say that our model is capable of robustly recognizing different types of events.

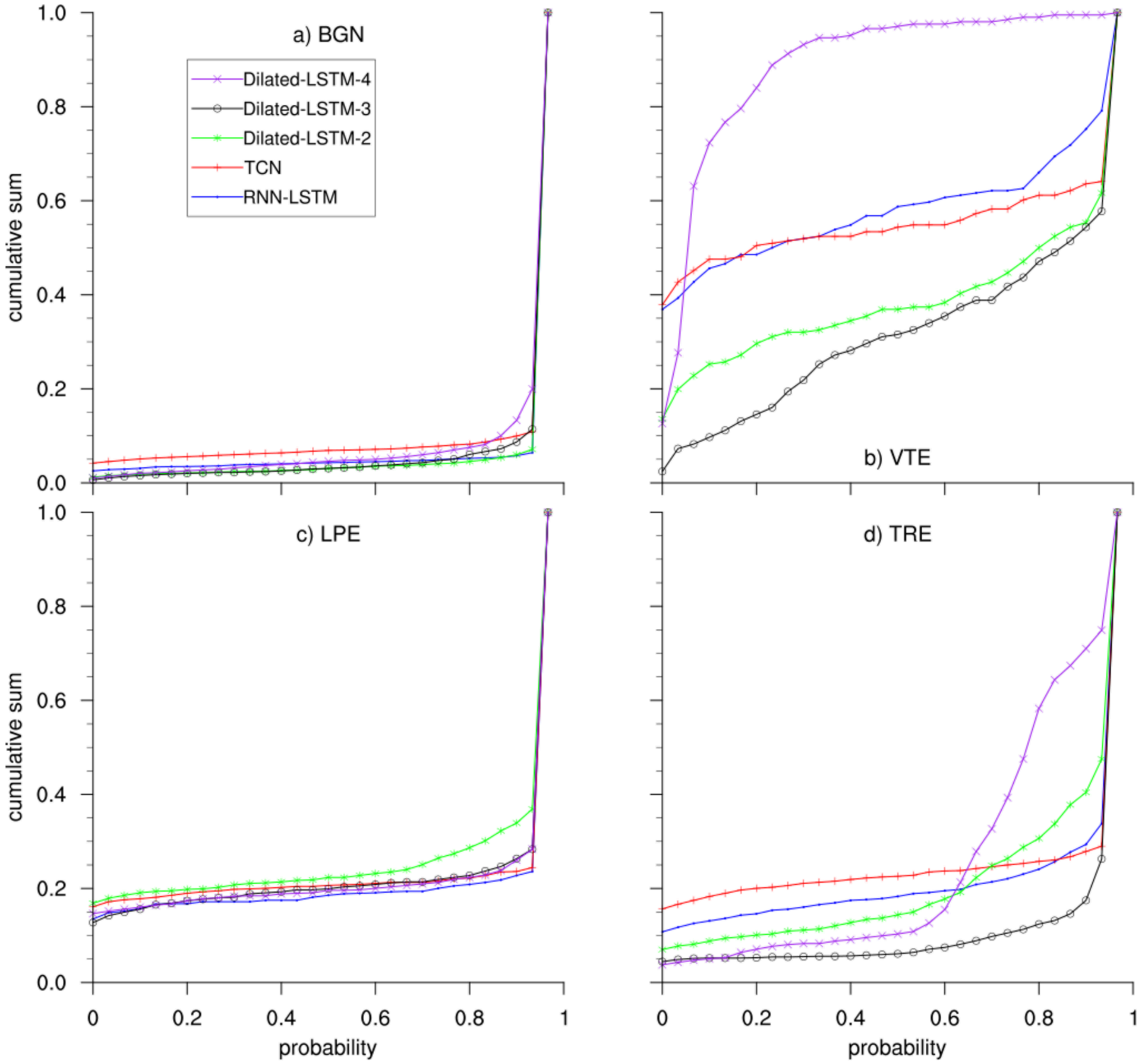

Figure 1: Cumulative distribution functions (CDFs) of probabilities for the predicted seismic categories. It is important to note that these CDFs are the classification

probabilities for the four types of volcano-seismic events—long period event (LPE), volcano-tectonic earthquake (VTE), background noise (BGN), and tremor (TRE)—obtained with different Recurrent Neural Networks using Long Short Term Memory cells (RNN–LSTM), Temporal Convolutional Networks (TCN), and Dilated Recurrent Neural Network using Long Short Term Memory cells (Dilated–LSTM) architectures with 2,3 and 4 hidden layers.

3. The authors introduce a set of (typical) labels used in volcano seismology, but fail to show clear examples from each dataset until late in the paper. An early figure showing the different classes from each volcanic setting (waveforms and spectrograms) would have been really informative on how the human experts had distinguished the different signal types, and what classifications they are hoping to achieve with their models.

As we previously mentioned, we did not include the waveforms and spectrograms of the events because this database has been widely cited and referenced in previous works, which form the foundation of this study. Therefore, we considered that information redundant. However, if the reviewer deems it necessary, we will include and thoroughly describe this information in the next version.

4. Although the algorithm framework is shown in Figure 1, there is data pre-processing and feature engineering step is too opaque. What does the "stream data" entail exactly? Why was a bandpass filter of 1-20 Hz chosen? How many stations are being used to constrain each label? Are we looking at the vertical component only? Is the instrumentation the same at Deception Island and Popocatepetl? What is the feature space here? If the features were indeed learned in the "deep learning" sense, it was not entirely clear to me how they were computed.

The pipeline presented in Figure 1 has been widely used for the automatic recognition of continuous seismo-volcanic signals and has served as the baseline for several previously published works. In this context, stream data refers to the continuous flow of data that is generated and transmitted in real time. This data is produced constantly and is processed or analyzed as it is received. Consequently, it is connected to the recognition block (detailed in Figure B). Essentially, this illustrates that the data can be processed in quasi-real-time and continuously, meaning it is not composed of isolated and segmented events.

Regarding the bandpass filter, we chose the range of 1-20 Hz because the frequency content associated with the source models (proposed by Ibáñez, J.M. et al. (2000)) describing seismo-volcanic events is primarily located within this frequency band. Furthermore, since we are using a filter bank on a logarithmic scale, we enhance the resolution of the analysis in the lower frequency ranges.

Given that we are working with established databases, multiple stations are not being used in this work to establish or constrain the label of each recognized event at test time. This is something that could easily be incorporated and has been done in other studies. In summary, this work utilizes information from a

pre-established catalog (composed by labels obtained for different seismic stations), which serves as the basis for conducting comparative statistics.

With respect to instrumentation and data components, this study only considers the vertical components of the recorded seismic data. The instrumentation details related to the data from Deception Island, which form the foundation of this work, are referenced throughout the manuscript. As for the instrumentation details of the POPO database, we did not have access to them. We only have access to the catalog and preprocessed seismic data, where the instrumental response has been corrected. Both databases are recorded at a sampling frequency of 50 Hz.

Lastly, we address the feature space. As previously mentioned, each signal is windowed into 4-second segments with a 3.5-second overlap. For each window, a bank of 16 filters on a logarithmic scale is applied, and the first and second derivatives of each component of the 16-feature vector are calculated to enrich the contextual information. This results in a 48-feature vector per window. Thus, the feature space referred to in the text corresponds to the space of the parameterized features. The models use these features to extract hidden information in the form of nonlinear relationships and to recognize the events. In conclusion, this work applies a feature engineering process to help the models more effectively learn to characterize each event type and its temporal evolution. Therefore, while the model learns autonomously, the features have been previously extracted based on expert knowledge of the problem at hand.

5. The authors mention that volcano monitoring and eruption forecasting involves a multidisciplinary approach. However, much of the manuscript is aimed at improving a catalog using machine learning techniques, which only involves the discipline of volcano seismology. The translation of this information into understanding unrest and hazards is absent. If the authors were to show that rapid catalog improvement could result in near-real-time characterization of real volcanic unrest, it would have made a far more convincing case. Unfortunately, this was not done or shown.

This work focuses on improving seismic catalogs, which can assist in the multiparametric monitoring process supported by volcanological observatories. At no point has this work aimed to eliminate or downplay the importance of the multiparametric approach. On the contrary, this work is centered on the improvement or real-time construction of seismic catalogs based on the knowledge acquired from a master database. Even so, our work does not aim to be a universal tool for creating catalogs for any volcano. Instead, it seeks to highlight a significant issue in volcanic monitoring from a seismic perspective and provide a methodology through which, using techniques purely based on seismic observations, we can develop a robust tool that can easily adapt to different volcanic environments, creating effective and reliable catalogs that enhance our understanding of volcanic dynamics. To achieve this, each observatory will need to set up its system based on the available data or the similarities of its volcano with others that have public data and catalogs that

support the development of the tool. This was the intended goal of the work, but in light of the feedback received, we realize that we may not have been clear enough for our idea to be interpreted correctly.

6. A key issue in volcano seismology machine learning literature is that volcanoes do not behave uniformly over time. Unrest signatures can vary from eruption to eruption, and from volcano to volcano. One way to make the applicability of this work more convincing could be to show its "temporal transferability" for one volcano in between different eruptive periods, before showing its applicability at a completely different volcano (i.e. "volcano transferability") as the authors have attempted in this work.

As we mentioned earlier, our work does not aim to be a universal tool for creating catalogs for any volcano. Instead, it seeks to highlight a significant issue in volcanic monitoring from a seismic perspective and provide a methodology through which, using techniques purely based on seismic observations, we can develop a robust tool that can easily adapt to different volcanic environments, creating effective and reliable catalogs that enhance our understanding of volcanic dynamics. To achieve this, each observatory will need to set up its system based on the available data or the similarities of its volcano with others that have public data and catalogs that support the development of the tool.

As the manuscript stands, it seems more suited for a journal like IEEE, where novel applications of machine learning techniques are discussed. In the context of NHESS or any other Earth Science journal, I would hope to see a more rigorous discussion of (1) the labeled dataset, (2) the different ML architectures, (3) the contextual volcanic unrest for which the seismic signals are observed, and (4) the relation between seismic catalogs and eruption forecasting.

We believe that the novelty of this work lies not in improving an AI algorithm, but in highlighting the problem of training and creating automatic recognition models with biased catalogs, and how this issue can potentially be addressed using weakly supervised techniques. This is why we chose to submit this work to this journal. We are not attempting to create an AI model that universally generates catalogs. Our goal is to provide the reader with the perspective that a model with a high recognition rate in a given database may be biased by the information it has learned from that database. What might initially seem like a very robust model may not be as strong as we think. To illustrate this, we present a weakly supervised approach based on models trained with a master database, demonstrating that a model trained from scratch can learn to recognize the labeled information in the catalog while leaving useful information hidden. As we've mentioned, this approach is not intended to be universal, as there are many forms of semi supervised learning that can be employed. The idea is to show the volcanological community that the development of automatic seismic-volcanic signal recognizers can be biased, even when they have robust recognition rates. Consequently, the use of semi-supervised techniques could be a potential solution.

To conclude, I would like to inform the reviewer that this work is a continuation of all our previous research works. Most of the questions you may have can be addressed by consulting these references. Additionally, many of our publications have appeared in IEEE journals, which tend to attract a more technically-focused audience, particularly in machine learning. This is why we decided to submit this work to NHESS, aiming to increase its visibility within the volcanology community.