Authors' Response to Reviewers

May 15, 2024

Dear Editor Prof. Christian Huber,

We are grateful for the reviewers' helpful comments. Please find the attached item-by-item reply to 3 reviewers of the manuscript titled "*Deep-learning-based phase picking for volcano-tectonic and long-period earthquakes*" [#2024GL108438].

In the revised manuscript, we used blue color to highlight the changes suggested by reviewer #1, teal color to highlight the changes suggested by reviewer #2, orange color to highlight changes suggested by reviewer #3 and purple color for other changes.

Best wishes,

Yiyuan Zhong
Yen Joe Tan

To reviewer 1 (changes marked by blue color):

Overview

In this paper, "Deep-learning-based phase picking for volcano seismicity" by Zhong and Tan, the authors present the results of a retrained phase picker designed to improve automated picks on volcanic earthquakes, especially long-period earthquakes. Overall, I found the paper to be well-written, and most of my comments only highlight places where I needed additional clarification. I'll mention that I am not equipped to fully review the details of the methodology; I have some basic knowledge of deep learning but have never trained a model myself and must trust that the methods presented here are standard practice.

My only major concern is that there is a lot of information contained in the supplement. Part of me appreciates the exhaustive and transparent assessments contained therein, but the other worries that important information may be overlooked because it's in the supplement. I recognize that a limitation of GRL is the length (of text and number of figures), but I suspect there is a more efficient way to utilize the 3 figures to accompany the text. I've tried to point out some possibilities in the discussion of the figures below, but I include them as suggestions rather than required changes.

Find below my comments by section/line/figure number:

Comments

Comment 1: Title Line 1: Would it be worthwhile to better highlight the focus on long-period seismicity explicitly somewhere in the title?

Reply 1: Thank you for the suggestion. Please see the updated title: "Deep-learning-based phase picking for volcano-tectonic and long-period earthquakes".

Comment 2: Key Points Line 7: Replace "the first" with "a"; I find it hard to believe that this is the first time an international dataset of volcanic waveforms has ever been assembled.

Reply 2: Thank you for the comment. We have changed "the first" to "a" in line 7 in the revised manuscript, even though we are unaware of any other global dataset of volcano seismicity waveform.

Comment 3: Key Points Line 10: I think this line is more accurate without the phrase "volcanic earthquake" given that the pickers also don't perform well on the tectonic low-frequency earthquakes

Reply 3: Thank you for the comment. We have deleted "volcanic". Please see line 10 in the revised manuscript.

Comment 4: Abstract Line 16: I suggest this line to read as, "...has yet to be fully evaluated. ..."

Reply 4: Thank you for your suggestion. We have changed "...has yet to be evaluated. ..." to "...has yet to be fully evaluated. ...". Please see line 16 in the revised manuscript.

Comment 5: Abstract Lines 21-23: I suggest this sentence to be reworded, perhaps as "...significantly better, including when tested on two data sets where no training data were used: volcanic earthquakes from Northern California, and tectonic low-frequency earthquakes along the Nankai Trough."

Reply 5: Thank you for your suggestion. We have reworded the sentence. Since we have included volcanic earthquakes from the Cascade Volcanoes in the revised manuscript, we use "along the Cascadia subduction zone" instead of "from Northern California".

Please see lines 21-23.

Comment 6: Plain Language Summary Lines 28-39: The first part of this summary (Lines 28-33) is a nice explanation of the problem statement. After this point, the summary reads very similar to the abstract. Perhaps some additional work is merited in summarizing the rest for a general audience?

Reply 6: Thank you for your suggestion. Please see lines 35-43 in the revised manuscript.

Comment 7: Introduction Lines 75-77: The difficulty in detecting (and therefore cataloging) long period earthquakes also makes studying their source mechanisms difficult, which might be worth some mentioning here as a hurdle your work aims to help overcome.

Reply 7: Thank you for the suggestion. We have mentioned this in lines 82-84 in the revised manuscript.

Comment 8: Dataset Lines 97-120: This section is an important one, and I didn't feel it was quite complete. I have many general questions:

- How far back do the datasets go in time? I presume it's quite heterogeneous.
- How many pre-eruptive and eruptive sequences are included as part of the dataset?
- Is the dataset dominated by a few highly prolific volcanoes (and if so, which ones and of what style of volcanism)?
- Why are the Cascades volcanoes excluded? Catalog and waveform data should be available with numerous LPs for the 2004 eruption of Mount St. Helens.
- How are LPs defined? I assume they are flagged explicitly by an analyst as such, and you are relying on that as the label?
- Are waveforms from stations only available if a pick is made?
- How were VT waveforms culled to have similar numbers as LPs?

- How were the noise waveforms chosen? Why 20,000?

- How do these waveform and event numbers compare to the STEAD and INSTANCE datasets?

Reply 8: Thank you for these questions and comments.

1. How far back do the datasets go in time?

We have introduced the time spans of the datasets. Please see lines 109-114 in the revised manuscript and Figure S4 in the revised supplement.

2. How many pre-eruptive and eruptive sequences are included as part of the dataset?

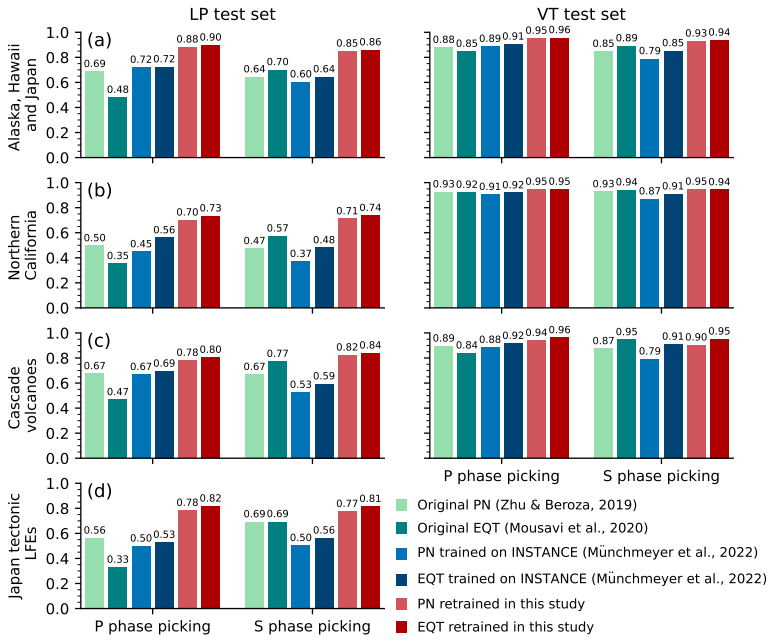
We have not made an analysis of the pre-eruptive and eruptive sequences, which are not directly available from the catalogs we downloaded. It would take some time to identify and count the eruptive sequences from the available catalogs. Since we have included as many waveforms as possible, we think this analysis may not be necessary in terms of improving the model performance.

3. Is the dataset dominated by a few highly prolific volcanoes (and if so, which ones and of what style of volcanism)?

We have pointed out the prolific volcanoes. Please see lines 116-124 and Figure 1 in the revised manuscript.

4. Why are the Cascades volcanoes excluded? Catalog and waveform data should be available with numerous LPs for the 2004 eruption of Mount St. Helens.

We have included 810 LP waveforms and 810 VT waveforms from events near Cascades volcanoes cataloged by the Pacific Northwest Seismic Network. The test result is shown in the following figure (the third row).



Since the Cascades volcanoes and the volcanoes in the Northern California are in adjacent regions, we merged the two test results in Figure 3b in the revised manuscript.

Please see lines 113, 202-203, 261-262, Figure 1d and Figure 3b in the revised manuscript.

5. Are waveforms from stations only available if a pick is made?

Continuous waveforms are available for downloading but we only use the time windows around available manual picks.

6. How were VT waveforms culled to have similar numbers as LPs?

We first select as many LPs with both P and S manual picks as possible from the available catalogs and then randomly choose a similar number of VTs with both P and S manual picks. Please see Text S3 in the revised supplement.

7. How are LPs defined? I assume they are flagged explicitly by an analyst as such, and you are relying on that as the label?

Yes, they are already flagged by analysts. For example, in the catalogs downloaded from JMA, there is a marker indicating whether the event type is a natural earthquake, artificial event or low-frequency earthquakes. Please see line 131 in the revised manuscript.

As for labelling, we do not use event type labels during training since our model only focuses on picking the phase arrival and does not perform classification. The LP flags are only used to divide the test set into an LP test set and a VT test set to distinguish the model performances for LP and VT waveforms. Please see lines 220-222 in the revised manuscript.

8. How were the noise waveforms were chosen? Why 20,000?

We choose the noise waveforms at the same stations as the event waveforms by visual inspection. Please see lines 131-132 in the revised manuscript.

Manual inspection of pure noise waveform is time consuming, and we found including the noise waveforms did not improve our model performance a lot which might be because event waveforms also contain information of noise. There are also models that are not trained on pure noise waveforms, such as the original PhaseNet (Zhu and Beroza, 2019), PickBlue (Bornstein 2024). Therefore, we stopped collecting more noise waveforms when we had collected 20,000 noise waveforms.

Bornstein, T., Lange, D., Münchmeyer, J., Woollam, J., Rietbrock, A., Barcheck, G., ... & Tilmann, F. (2024). PickBlue: Seismic phase picking for ocean bottom seismometers with deep learning. Earth and Space Science, 11(1), e2023EA003332.

Zhu, W., & Beroza, G. C. (2019). PhaseNet: a deep-neural-network-based seismic arrival-time picking method. Geophysical Journal International, 216(1), 261-273.

9. How do these waveform and event numbers compare to the STEAD and INSTANCE datasets?

The waveform number of our dataset (334,390) is only one fourth of that of STEAD (1,265,657) or INSTANCE (1,291,537). For the event numbers, our dataset has more events (73,516) than INSTANCE (54,008) but less events than STEAD (441,705).

Please see lines 145-146 in the revised manuscript for the comparison of waveform numbers. We do not compare the event number because it could be distracting.

Comment 9: Dataset Line 116: Would it be accurate to say that your dataset has a wider distribution of frequency index by design?

Reply 9: Thank you for the comment. We have modified the sentence to: "... our data set has a wider distribution of frequency index by design ...". Please see line 143.

Comment 10: Dataset Lines 118-120: I agree that most approaches tend to focus on a single or at most a few volcanoes to lean on expert knowledge of that system's behavior. Your goal is different in scope and relies on a different type of knowledge (i.e., analyst picks) which requires a larger dataset to get the number of events required to train on. While I can't refute the statement in these lines, I might recommend that you focus instead on the design aims of the dataset and that these can only be fulfilled by a large and necessarily global dataset.

Reply 10: Thank you for the comment. Please see lines 147-149.

Comment 11: Evaluation Lines 144-149: There is a lot of information in Text S2 on how these are defined that might be useful to bring back into the main text. The section on the definitions of positive and negative were especially illuminating.

Reply 11: Thank you for the comment. We decide to keep Text S2 in the supplement because it would make the main text too long if we moved the Text S2 to the evaluation section.

Comment 12: Evaluation Line 156: You mention that there is not a significant systematic change in signal-to-noise ratio with frequency index and reference Figure S17. I wondered around here how much the performance varied as a function of SNR, both overall and separated out between VTs and LPs. You have a lot of plots with statistics in the supplement, and I was surprised that looking at performance explicitly as a function of SNR was not among them.

Reply 12: Thank you for the suggestion. We have included tests of performance as a function of SNR. Please lines 187-191 in the revised manuscript and Figures S15,S17 in the revised supplement.

Comment 13: Training Line 174: Perhaps it's worth mentioning explicitly that these waveforms also have lower frequency indices here?

Reply 13: Thank you for the suggestion. We have added this statement. Please see lines 206-207 in the revised manuscript.

Comment 14: Training Line 217: I was encouraged by the performance of your model in its residuals. I think stating a {plus minus}2σ range in words here could be useful.

Reply 14: Thank you for the suggestion. Please see lines 256-257 in the revised manuscript.

Comment 15: Training Line 221: Northern should be capitalized (here and throughout the paper/supplement).

Reply 15: Thank you for the suggestion. We have replace "northern California" with "Northern California" throughout the paper and the supplement.

Comment 16: Training Lines 231-232: Again, another reference to SNR. I note looking at the supplemental figures that the SNR for the LPs in Northern California is low compared to the VTs. Part of this is because the dedicated analyst for many, many years (Mitch Pitt) was meticulous about trying to make picks on any LP he could find, including low amplitude and noisy ones. At some point it might be worthwhile to include his picks as part of a training dataset, but that's future work for someone else.

Reply 16: Agreed.

Comment 17: Training Lines 233-239: I appreciate the tests on non-volcanic low frequency index earthquakes, as this convinces me of its transportability. Perhaps this means that including tectonic LFEs as part of a training dataset might also be beneficial?

Reply 17: Thank you for the comment. We will consider including the tectonic LFEs and the waveforms in Northern California for training and update the model as version 2 in the future.

Comment 18: Discussion Lines 249-251: Indeed, though perhaps they can contribute to creating more templates?

Reply 18: Thank you for the suggestion. We have added a statement in line 292 in the manuscript.

Comment 19: Discussion Line 252: Perhaps "are a step toward improving" instead of "can help to improve" here?

Reply 19: Thank you for the suggestion. We have change "can help to improve" to "are a step toward improving". Please see line 294 in the revised manuscript

Comment 20: Discussion Lines 255-259: I suppose this highlights a need in the community to cultivate a good group dataset that covers the wide variety of volcanic signals observed.

Reply 20: Thank you for the comment. We agree and we have included your comment in the revised manuscript. Please see lines 301-303.

Comment 21: Discussion Lines 259-261: I don't know if this is a fair statement given the different goals of each study. Rather, I think you could highlight that your study explicitly balances the dataset with the distribution of frequency content in mind.

Reply 21: Thank you for the comment. We have highlighted the balanced distribution of frequency content in lines 304-305 in the revised manuscript.

Comment 22: Figure 1: The maps here are so busy and difficult to read they are nearly useless. The basemap is pretty but distracts from the data. The earthquakes are so numerous they cover most of the volcanoes, and at this scale the location structure of the seismicity is hard to see. What I'd suggest is to remove the locations of the VT and LP seismicity entirely, and color the volcanoes by the number of both within 50 km, perhaps with some way of signifying the proportion of each. The tectonic LFEs can remain as dots since they aren't associated with a center, and rather the Nankai Trench as a whole. This suggestion is partially to address some of my questions from the Dataset section regarding how much seismicity from a single center potentially dominates the dataset. I recognize this is basically asking for a complete rework of the figure, but as it is, the figure does not contribute as much as it should for the space it occupies.

Reply 22: Thank you for the suggestion. In the revised Figure 1, we have replaced the colored basemap with a grayscale basemap and we only show volcanoes colored by the number of LPs and VTs within 50km. We have moved the locations of VT and LP seismicity to the supplement (Figure S1).

Please see Figure 1 in the revised manuscript and Figure S1 in the supplement.

Comment 23: Figure 2: I think this is a very illustrative figure as is. If you decide to pull some more information out of the supplement into the main figures, I'd appreciate some information showing the distribution of waveform numbers as a function of frequency index. Perhaps something like Figures S7d and/or S10?

Reply 23: Thank you for the suggestion. We have included Figure S10 in the Figure 2b. The SNR distributions are also included in the revised Figure 2c.

Please see Figure 2 in the revised manuscript.

Comment 24: Figure 3: This figure really highlights how much of an improvement your model is for events with lower frequency contents, while also being transparent that it still

(understandably) struggles a bit. I noted that there is an extra space in the legend in "Münchmeyer" that should be removed. Also, you usually refer to the tectonic events as "LFE" rather than "LP" and should change the axis label in part c.

Reply 24: Thank you for the comment. We have removed the extra space and changed "LP" to "LFE". Please see Figure 3 in the revised manuscript.

Comment 25: Figure S1/S2: Similar issue as Figure 1 with the busy basemap.

Reply 25: Thank you for the comment. We have replaced the colored basemap with grayscale basemap. Please see Figure S1-S3 in the supplement.

Comment 26: Figure S4: Typo "neareat" at the end of the caption.

Reply 26: Thank you for pointing out the typo. Please see Figure S11 in the revised supplement.

Comment 27: Figures S5-S9: I rather like these distribution histograms. Figure S6 really highlights the difference in SNR between the rest of the dataset and Northern California for its LPs.

Reply 27: Thank you for the comment.

Comment 28: Figure S10: As mentioned earlier, I think this figure belongs in the main text somehow. If you do include it, I might recommend rethinking the color choices as blue and purple are difficult to discriminate.

Reply 28: Thank you for your comment. We have changed the colors of the original Figure S10 and moved it to Figure 2b in the revised manuscript.

Comment 29: Figures S11-S12: Can you include the total number of traces/events for each panel?

Reply 29: Thank you for your suggestion. We have included the number of traces/events. Please see Figures S8-S9 in the revised supplement.

Comment 30: Figure S13-S14: You don't discuss magnitudes in the main text and for LPs it's somewhat questionable if the magnitudes are meaningful. I think SNR is probably fine as a proxy for size, so I'm not sure if these figures are necessary.

Reply 30: Thank you for the comment. We have removed these two figures in the revised supplement.

Comment 31: Figure S18: Again, I feel like I see something resembling a tradeoff between SNR and performance with FI and would appreciate seeing how performance looks as a function of SNR directly.

Reply 31: Thank you for your suggestion. We have addded the tests of performance variation as a function of SNR in Figure S15 and Figure S17.

Comment 32: Figures S19-S20, S22-S23: I appreciated these figures a lot. {plus minus}0.5 s for a pick on an LP is encouraging to me. Perhaps part of that is that I see this as a tool for assisting an analyst or helping make detections in the first place, rather than replacing one entirely.

Reply 32: Thank you for the comment.

Comment 33: Figures S24-S27, S29: Again, I appreciate the transparency in these figures. It's just a lot of figures!

Reply 33: Thank you the comment. Indeed, there are a lot of figures. We have also combined some figures in the original supplement. For example, Figures S5-S6 in the original supplement are combined to Figure S7 in the revised supplement. Figure S9ab and Figure S7 in the original supplement are combined into Figure S8 in the revised manuscript. Figure S9cd and Figure S8 in the original supplement are combined into Figure S9 in the revised supplement.

I hope you find these comments useful. I think the paper is generally in good shape as is, despite my lengthy commentary. I look forward to the release of the model and perhaps getting some time to properly play with it.

To reviewer 2 (changes marked by teal color):

Overview

This paper presents an analysis of how well popular deep learning based phase pickers, PhaseNet, and EqTransformer, perform at the task of volcano seismicity monitoring (e.g., volcano tectonic (VT) and long-period (LP) earthquakes). They show that as the frequency content becomes predominately lower, which is common for these types of waveforms, the performance of the existing algorithms deteriorates significantly. They then assemble a comprehensive dataset of waveforms from many volcanoes around the world (containing many VT and LP events) and re-train these models with this dataset, and show performance comparisons of the newly trained models with the previous models. They show a marked improvement in precision and recall, among other metrics, with the re-trained models, which is significant. They show some specific test cases of good performance on volcanic events from

geographic areas not included in training (e.g., northern California and the Nankai Trough in Japan), which demonstrates good generalization performance.

This is a very impressive and interesting paper. The authors have raised an important question about a largely ignored topic: the "effectiveness" of current deep learning pickers at monitoring volcano seismicity - which is notable, since these methods are currently being used in many studies to monitor volcano seismicity (with varying degrees of success), but to my knowledge, the tradeoffs shown here about performance deteriorating with frequency content has not previously been appreciated, nor has an effective solution (the re-trained models) been proposed. So, the authors have clearly highlighted and addressed a very important problem, and the writing is clear and succinct. I have only minor comments to improve the clarity of presentation in some places. I look forward to using this picker in future studies myself.

Comments

Comment 34: Line 88, Page 4, note that "Automatic detection for a comprehensive view of Mayotte seismicity", Retailleau et al., 2022, uses PhaseNet, and detects LP events.

Reply 34: Thank you for the comment. To avoid misunderstanding, we have changed "... in their analyses" to "... in their training and tests". In addition, "Automatic detection for a comprehensive view of Mayotte seismicity, Retailleau et al., 2022" has been added in the review of recent studies.

Please see line 87 and lines 96-99 in the revised manuscript.

Comment 35: Line 94, Page 5, briefly say "how" you train new models, and why it's successful.

Reply 35: Thank you for your comment. Considering the limitation on text length, we decide to leave details of training to Section 4 "Training deep-learning phase pickers for volcano seismicity".

Comment 36: Line 105, Page 5, expand a bit on "Figures S3-S14 for other properties of the data"

Reply 36: Thank you for the suggestion. We have added more description about the supplementary figures. Please see lines 125-128 in the revised manuscript.

Comment 37: Line 108, Page 5, I am not sure it's wise to purposefully exclude VT events just to balance them with LP events; why not be comprehensive, and supply all (of a high enough quality) VT events? It is easy to balance the ratio of either dataset during training by just sampling proportionally.

Reply 37: Thank you for the comment. Based on our test results, our model can work well under the current strategy of data selection. Since there are huge number of VTs, downloading all of them would require more time and larger computer storage. Maybe we can explore this training strategy in the future.

Comment 38: Fig. 2. Is it possible to plot the x-axis as only a single scalar value (not an interval)? This would make it easier to read and the fact that it's an interval can be explained in the caption.

Reply 38: Thank you for the comment. We have change the x-axis to single scalar values. Please see Figure 2 in the revised manuscript.

Comment 39: Line 137, Page 7, when explaining that "Considering that...", you use both the Instance version of PhaseNet, the original PhaseNet, and your own PhaseNet; it doesn't come across completely clearly. I was under the impression only the Instance version of PhaseNet and your own are trained based on this statement, but I see later that is incorrect (and you actually use three versions).

Reply 39: Thank you for your comment. We have added a sentence before introducing the INSTANCE version of PhaseNet/EQT: "The original versions of PhaseNet and EQTransformer are tested here since they are most widely used."

Please see lines 168-169 in the revised manuscript.

Comment 40: Line 182, Page 10; Interesting, and I think promising data augmentation strategy, but one aspect I am wondering: if you normalize the waveforms before merging two different samples, it will mostly only see cases where two adjacent arrivals are of similar amplitude? In the real applications, will it then mostly miss a low energy arrival occurring nearby a high energy arrival? (Since it is not being trained on these cases).

Reply 40: Thank you. We used the word "rescaled" to mean the waveform is normalized and then multiplied by a factor. We have rephrase this sentence in the revised manuscript to make it more clear. Please see lines 216-217 in the revised manuscript.

Comment 41: Line 196, Page 10, word "initialization" repeated

Reply 41: Thank you for the comment. We have rephrased this sentence. Please see lines 233-234 in the revised manscript.

Comment 42: From Fig 3; Do the authors have any insight why the PhaseNet trained on Instance performs better on LP events than the original PhaseNet? If so, can they comment somewhere in the text on this?

Reply 42: Thank you for the comment. We have updated the test results of the original PhaseNet, because the component order of input waveforms for the original PhaseNet is different from any other pretrained model in seisbench (see https://github.com/seisbench/seisbench/issues/260#issuecomment-1888817044) but we originally assumed the same component order. Now we have corrected the test results for the original PhaseNet, and double-checked the other models involved in this study.

In addition, we used the conservative version of EQTransformer in the original Manuscript. Now we choose to use the non-conservative version because it was the original EQTransformer in the paper of Mousavi et al. 2020 (see https://github.com/seisbench/seisbench/pull/102#issuecomment-1161970617).

After updating the test results of the original PhaseNet and original EQTransformer, we find that neither the original PhaseNet/EQTransformer nor the models trained on INSTANCE show advantage. Please see Figures 2-3 and lines 246-253. Figures S12, S15-S28 and Tables S5-S6 are also updated accordingly.

Comment 43: Line 217, Page 12, by how much are the picking residuals reduced with the re-trained models? (this text can be added)

Reply 43: Thank you for the suggestion. We have included 2σ (two standard deviation) of picking residuals of the retrained EQTransformer in the main text for reference. However, to avoid making the text too lengthy, we decided not to make a detailed comparison of picking residuals of our retrained models and the original models. Interested reader can refer to the supplement.

Please see lines 256-257 in the revised manuscript.

Comment 44: Line 261, Page 14; I may have missed it, but how did the authors end up with so many labeled LP events? Was there just a sufficient number of them in previously developed routine catalogs, or was some special processing employed to obtain the dataset?

Reply 44: Thank you for your comment. We didn't employ any special processing. The LPs events are flagged by analysts in JMA and USGS. For example, in the catalogs downloaded from JMA, there is marker indicating whether the event type is a natural earthquake, artificial event or low-frequency earthquakes.

Comment 45: Lines 262 - 265, Page 14; Is this statement somewhat misplaced? It doesn't seem to make sense with what came before it.

Reply 45: Thank you for your comment. This statement is to explain why we didn't test the other models mentioned in this paragraph on our data set. We have removed this sentence in the revised manuscript to avoid confusion. Please see lines 307-309 in the revised manuscript.

Comment 46: Line 308, Page 15, This sentences ending is worded somewhat strangely "... suboptimal with biases".

Reply 46: Thank you for your comment. We have changed it to "... suboptimal due to potential biases". Please see line 352 in the revised manuscript.

To reviewer 3 (changes marked by orange color):

Overview

The authors compiled a dataset of volcanic low-frequency events to be used in machine learning applications and retrained proven ML based phase picking algorithms (EQT and PhaseNet) using this dataset. They demonstrate the improvements in phase picking using the PhaseNet and EQT algorithms on volcanic signals from events unseen during training.

The compiled dataset might be of interest to the larger seismological community, not only for volcanology but also for the research of very low frequency earthquakes. It is not as extensive as the benchmark datasets such as STEAD, but it is definitely an important start towards an global benchmark dataset. I commend the authors and encourage them to continue building upon on this dataset.

I am pleased to recommend this manuscript for publication in GRL after revisions. Below are my suggestions/comments the authors need to address in the revision.

Comments

Comment 47: The dataset description needs to be more detailed. I am missing the number of stations used in training (and validation). Do certain stations have an higher representation in the dataset? So how many percent of the waveforms you compiled are from what number of stations. I could imagine that there is also a distance dependence in the dataset. I also assume that the vast majority of waveforms are from the Japanese stations but without more detailed information this is just a guess. If it is the case however, there might be an issue with the transferability of the trained model.

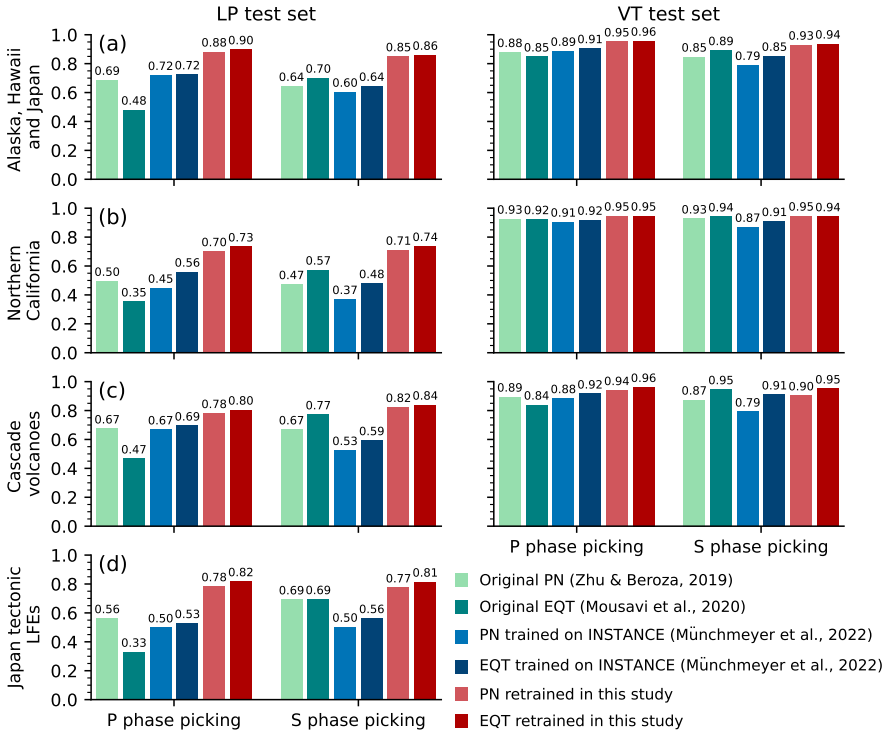
Also what is the depth distribution of the events you have chosen. Some simple histograms might already be enough and be of great value to present your compiled dataset.

Reply 47: Thank you for the comment. We have included the number of stations in lines 109-114 in the revised manuscript. In addition, we have colored the stations by waveform numbers in Figure S2 in the revised supplement. In addition, the distributions of event depths, epicentral distances and back azimuths are shown in Figures S8-S9. We have also added more description about the supplementary figures in lines 125-128 in the main text.

Comment 48: I would also suggest to add another test, maybe just a demonstration which does not have to be exhaustive on the performance of your algorithm to another test area. Maybe you could demonstrate the Phase picking on one or two selected VLF events from e.g. Mexico?

Reply 48: Thank you for the suggestion. We have added a test in the Cascades volcanoes, as is also suggested by reviewer #1.

The test result is shown in the following figure.



Since the Cascades volcanoes and the volcanoes in the Northern California are in adjacent regions, we merged the two test results in Figure 3b in the revised manuscript. Please see line 113, Figure 1d and Figure 3b in the revised manuscript.

Comment 49: Please also state explicitly if you retrained (transfer learned) EQT and PhaseNet or if you (as I assume) trained them from scratch.

Reply 49: Thank you for the comment. We have compared training from scratch and transfer learning, and then choose the one with the highest F1 score on the validation set. The comparison is shown in Table S4. For EQTransformer, the model of transfer learning shows marginally higher F1 score (averaged for P picking and S picking), while for PhaseNet the model trained from scratch shows marginally higher F1 score.

Please see lines 233-237 in the revised manuscript and Table S4 in the revised supplement.

Comment 50: I tested the jupyter notebook you kindly send along. It works fine so far, but I would recommend to get in contact with the seisbench team (Jannes Münchmeyer, jannes.muenchmeyer@gfz-potsdam.de) to discuss full implementation of your pickers as a variant to choose from. Should be easy enough from a technical standpoint.

Reply 50: Thank you for the suggestion. We will contact the seisbench team to make our model as an option in the seisbench.

Comment 51: L22: Suggestion to change the line to: "-quake waveforms from northern California, from which no training data is used and tectonic"

Reply 51: Thank you for the comment. We have rephrased this sentence by changing "... California where no training data are used" to "two data sets where no training data were used: volcanic earthquakes along the Cascadia subduction zone and tectonic low-frequency earthquakes along the Nankai Trough."

Please see lines 21-22 in the revised manuscript.

Comment 52: L32: Please strike the "a type of artificial intelligence".

Reply 52: Thank you for the suggestion. We have removed "a type of artificial intelligence" in the revised manuscript.

Please see line 34 in the revised manuscript.

Comment 53: L52: I suggest also citing: "Kriegerowski, Marius, et al. "A deep convolutional neural network for localization of clustered earthquakes based on multistation full waveforms." Seismological Research Letters 90.2A (2019): 510-516."

Reply 53: Thank you for the suggestion. Please see line 57 in the revised manuscript.

Comment 54: L86 Maybe also cite Manley et al, 2022

Reply 54: Thank you for the suggestion. Please see lines 93-99 in the revised manuscript.