Reviewer A:

General

The work addresses the compilation of a machine-learning ready dataset of recorded waveforms and associated metadata for the Pacific NW (PNW) of the United States. To this end, it is very appreciable that the authors follow the guidelines and the standardization proposed in the literature making the resulting dataset fully compatible with the SeisBench platform which has been designed to support seismologist to test and benchmark (interchangeably) machine learning models and datasets.

The manuscript describes concisely the various stages required to assemble the dataset while providing a general description of its main features and how these evolved through time because of instrumentation and software upgrading. It is also quite appreciable the inclusion of different types of seismic (exotic) sources. On the other hand, I found that the Earthquake Transformer (EQT) P- and S-wave onset picking comparison with the other EQT trained datasets available in SeisBench should be performed differently using a common test dataset (see below). One other aspect that can be important for the users of the dataset would be the inclusion of additional trace quality metadata to ease the identification of waveform data have been affected by digitization or sensor problems. This last can be quite useful when selecting the training data subset.

Finally, I believe that the article can benefit from a dedicated "discussion" section in which to focus on the features of the presented dataset. For example, it would be relevant a comparison with the results obtained by Münchmeyer et al. (2022) that also performed a similar analysis. Some discussion and recommendations on future work that have been presented in the conclusions could be moved to the discussion section and extended possibly. Overall, I recommend the publication of the paper after *minor* revision.

Questions, comments and suggestions

L10-L17: I found the abstract somewhat too concise. It would be important for the reader to understand immediately what is the size of the dataset in terms of number of earthquakes (and other exotic sources) and of independent 3C streams. Similarly, it would be important to provide the magnitude range of the earthquakes included.

L71: While I understand the difficulties in including seismic source like non-volcanic tremors or low amplitude low-frequency earthquakes (LFEs), it would be appreciable that the authors comment on the reasons why these sources are difficult to include in a dataset like this.

L181: Does the download of ~70 TB miniSEED refer to continuous data streams ? If correct please make it clear that the download involved continuous waveforms. Also, it would be important to know how long this big download took so that the reader can have an appreciation of the difficulties inherent with the analysis of big data.

L196: board → broad

L208: board → broad

L208-L209: regarding the offset (see Fig. S4), there does not seem to be a way to identify this kind of problems from the metadata alone. The addition of *trace* metadata like *median_counts* and *mean_counts* as it was done in INSTANCE through the application of the *MSEEDMetadata* module of Obspy can certainly help to identify this kind of behavior and help the user to remove them form the dataset before performing the analysis.

L254-L255: The sentence *"The model trained with the STEAD data set has the best picking accuracy"* should be slightly reworded. If I understand it, the authors refer to the more peaked distribution of the residuals (or the *phase-pick quality* referred in the previous sentence). This because the term "accuracy" has a precise meaning in the ML classification jargon.

L258-L259: this trade-off between detection and picking accuracy seems quite peculiar. Münchmeyer et al. (2022) did a quite extensive study across several dataset and it could be quite valuable to compare the results obtained with those presented in the supplement tables S2, S3, S4 of that work. Also, I observe that all the mean residuals are negative implying that the manual onset picking is always somewhat late. This issue should be addressed in the discussion.

L260-L267: I have some major concern for the results of the comparison presented in Figure 4. It seems that the entire ComCat dataset is used for training the model while exploiting the STEAD dataset through transfer learning. If so, the comparison made in Fig. 4 is not really appropriate because the statistics (MAE, RMSE, Mean) of the results obtained with the ComCat dataset follow from the training with the same dataset whereas all the others do not. For proper comparison, the analysis should be repeated after splitting the ComCat dataset into "training" and "test" data. After the model has been trained, the appraisal should be done using the test data subset on both the newly trained EQT model of ComCat and for all the other models presented in Figure 4. The discussion of this paragraph should then reflect the results of the analysis just outlined.

L324-L325: I am not clear why one event (Mw 6.4 Northern California, 20 December 2022) has been included in the magnitude bin between 6 and 7 because no data streams have been included in the dataset for this event (cf Table 1). If this the case, it would be better to reword the sentence and say that the dataset includes earthquake waveforms in the range 0≤M<6.

L401-L404: In addition to the switch in the ratio between positive and negative polarities, have the authors detected any geographic pattern of this ratio? If this occurs, it would be important to the reader to know.

L431: I have cloned the GitHub repository and, when trying to use the notebooks, I experienced some difficulties because of lack of documentation. I also tried the google colab access but I did not have the privileges to open it. In any event, it is very appreciable that the authors have made available their entire work but it would need more documentation. My recommendation is that the authors prepare a small test dataset which can be promptly downloaded and prepare some notebooks that show how to access and select the data, and plot them.

Figures

Figure 3: please change 50 to 60 km in the sentence *"Some events are color-coded white because they are deeper than 50 km"* the color scale ends at 60 and not 50 km. Also, it would be a good idea to insert the depth km value labels (not all) for the depth isolines of the slab.

Figure 4: This figure (and also the S1) should be redone using the results of the "test" dataset applied to all the Earthquake Transformer training models. Are the RMSE values the uncertainties to the MAE ? Also, in this figure the picking completeness attained by the INSTANCE and SCEDC trained datasets for P appears comparable if not better than that obtained from the trained PNW dataset. Similarly, ETHZ, SCEDC and INSTANCE perform better than PNW for S.

Figure 6: the Mh magnitudes (green) are barely visible (if not invisible). May be that the adoption of a log scale would help the visualization although it would make less visible the changes in the number of data available through time.

Figure 7: it would be good to specify that the units of the amplitude are in counts in this figure and all those showing seismograms in the supplement.

Recommendation: Revisions Required

Reviewer C:

Dear authors,

I find your dataset and the accompanying paper a good and important addition to the collection of ML-ready datasets. It's nice that you follow the SeisBench standard, making the dataset easily usable with existing codes. Another important thing that your dataset brings is the presence of a nicely organized collection of 'exotic' events. There are also some minor issues that I think should be improved before publishing the paper. The minor issues are:

1. It would be good to expand the sentence in the introduction on Seisbench data format explaining the idea behind it.
2. P2-L95: Please reference the Figs. S10 and S11 here.
3. P4-L148: Please reference Table 1 in the last sentence to allow the reader to see the number of the kept events immediately.
4. P5-L188: Do you have an easy way to add information about clipping into the metadata? It would be one of the ways to allow the users of the dataset to select high-quality data only. For this purpose, you could also label the traces which had missing channels, offsets, etc.
5. It would be good to have the order of the supplement figures follow the order in which they are referenced (currently e.g. Figs. S20, S21, S23, S24, S19 are referenced before Figs. S10, S4, S11).
6. P6-L260: Please explain better your retraining procedure, i.e. what data did you use for retraining, and what data for validation/testing?
7. It should be noted that STEAD probably contains some of the data that you are testing on, i.e. it was trained on them. This means that there is some information leakage and that the comparison between the model trained on STEAD and the models trained on other datasets is not fair.
8. Do you plan to make the dataset available through SeisBench (when the paper is published)? I think this would be a good way to make your dataset easier to find. If you plan to do it, you could add this to the Code and Data Availability section.

Recommendation: Revisions Required

Reviews of Ni et al., “ Curated Pacific Northwest AI-ready Seismic Dataset”

Reviewer-A comments

The work addresses the compilation of a machine-learning ready dataset of recorded waveforms and associated metadata for the Pacific NW (PNW) of the United States. To this end, it is very appreciable that the authors follow the guidelines and the standardization proposed in the literature making the resulting dataset fully compatible with the SeisBench platform which has been designed to support seismologist to test and benchmark (interchangeably) machine learning models and datasets.

The manuscript describes concisely the various stages required to assemble the dataset while providing a general description of its main features and how these evolved through time because of instrumentation and software upgrading. It is also quite appreciable the inclusion of different types of seismic (exotic) sources. On the other hand, I found that the Earthquake Transformer (EQT) P- and S-wave onset picking comparison with the other EQT trained datasets available in SeisBench should be performed differently using a common test dataset (see below). One other aspect that can be important for the users of the dataset would be the inclusion of additional trace quality metadata to ease the identification of waveform data have been affected by digitization or sensor problems. This last can be quite useful when selecting the training data subset.

Finally, I believe that the article can benefit from a dedicated “discussion” section in which to focus on the features of the presented dataset. For example, it would be relevant a comparison with the results obtained by Münchmeyer et al. (2022) that also performed a similar analysis. Some discussion and recommendations on future work that have been presented in the conclusions could be moved to the discussion section and extended possibly. Overall, I recommend the publication of the paper after minor revision.

L10-L17: I found the abstract somewhat too concise. It would be important for the reader to understand immediately what is the size of the dataset in terms of number of earthquakes (and other exotic sources) and of independent 3C streams. Similarly, it would be important to provide the magnitude range of the earthquakes included.

We have added more statistics about the data set in the abstract section as suggested.

L71: While I understand the difficulties in including seismic source like non-volcanic tremors or low amplitude low-frequency earthquakes (LFEs), it would be appreciable that the authors comment on the reasons why these sources are difficult to include in a dataset like this.

We add the discussion of this in the conclusion section.

L181: Does the download of ~70 TB miniSEED refer to continuous data streams ? If correct please make it clear that the download involved continuous waveforms. Also, it would be

important to know how long this big download took so that the reader can have an appreciation of the difficulties inherent with the analysis of big data.

We now mention that these mSEED are continuous data, and the download took 2 months to finish. But the curated data set of trimmed waveforms is smaller.

L196/L208: board → broad

Thanks for pointing out this typo.

L208-L209: regarding the offset (see Fig. S4), there does not seem to be a way to identify this kind of problem from the metadata alone. The addition of *trace* metadata like *median_counts* and *mean_counts* as it was done in INSTANCE through the application of the *MSEEDMetadata* module of Obspy can certainly help to identify this kind of behavior and help the user to remove them from the dataset before performing the analysis.

Thanks for pointing out this. As this is also pointed out by reviewer A, we checked waveform from 2002 when potential offsets are possible. Then, we manually labeled if the traces have any visible offsets (trace_has_offset attribute).

L254-L255: The sentence *"The model trained with the STEAD data set has the best picking accuracy"* should be slightly reworded. If I understand it correctly, the authors refer to the more peaked distribution of the residuals (or the *phase-pick quality* referred to in the previous sentence). This is because the term "accuracy" has a precise meaning in the ML classification jargon.

We reworded it to be ”quality in phase picks relative to the (ground truth) analyst's picks", which has been used in the previous sentence.

L258-L259: this trade-off between detection and picking accuracy seems quite peculiar. Münchmeyer et al. (2022) did a quite extensive study across several dataset and it could be quite valuable to compare the results obtained with those presented in the supplement tables S2, S3, S4 of that work. Also, I observe that all the mean residuals are negative implying that the manual onset picking is always somewhat late. This issue should be addressed in the discussion.

Thanks for pointing this out. Yes, all of the mean residuals are negative, indicating that the ML-picks tend to be earlier than the manual pick. Referring to Table S3 from Münchmeyer et al. (2022), we also notice a similar pattern that EQTransformer tends to have negative mean residuals (e.g., trained with STEAD, SCEDC, and ETHZ). This also may be due to the fact that we set a lower threshold for the phase picking (0.1). The retraining reduces such bias, although does not eliminate it.

L260-L267: I have some major concern for the results of the comparison presented in Figure 4. It seems that the entire ComCat dataset is used for training the model while exploiting the STEAD dataset through transfer learning. If so, the comparison made in Fig. 4 is not really appropriate because the statistics (MAE, RMSE, Mean) of the results obtained with the ComCat dataset follow from the training with the same dataset whereas all the others do not.

For proper comparison, the analysis should be repeated after splitting the ComCat dataset into "training" and "test" data. After the model has been trained, the appraisal should be done using the test data subset on both the newly trained EQT model of ComCat and for all the other models presented in Figure 4. The discussion of this paragraph should then reflect the results of the analysis just outlined.

> We agree and apologize for this oversight. The training was done using 70% of the data set for training and 30% for testing (we did not split the data set into a validating set). We replot Figure 4 using only the results from the testing set only. Figure S18, S30 and S31 have also been corrected accordingly. Figure S32 and S33 don't need to be changed since none of the strong motion streams are shown in the training data.

L324-L325: I am not clear why one event (Mw 6.4 Northern California, 20 December 2022) has been included in the magnitude bin between 6 and 7 because no data streams have been included in the dataset for this event (cf Table 1). If this is the case, it would be better to reword the sentence and say that the dataset includes earthquake waveforms in the range 0≤M<6.

> Thank you for catching this typo. We have added other parentheses to show the number of streams from these events.

L401-L404: In addition to the switch in the ratio between positive and negative polarities, have the authors detected any geographic pattern of this ratio? If this occurs, it would be important to the reader to know.

> We have not seen an obvious geographical pattern of this change in polarity ratio. It rather correlates in time with the adoption to AQMS in 2012. The network seismologists on the team were not aware of such systematic change. Therefore, this work generates new future research directions.

L431: I have cloned the GitHub repository and, when trying to use the notebooks, I experienced some difficulties because of lack of documentation. I also tried google colab access but I did not have the privileges to open it. In any event, it is very appreciable that the authors have made available their entire work but it would need more documentation. My recommendation is that the authors prepare a small test dataset which can be promptly downloaded and prepare some notebooks that show how to access and select the data, and plot them.

> Sorry for the inconvenience you have experienced using the notebook and colab provided. We have updated the notebook and the access so that anyone could access it through colab. We also prepared a micro version of the dataset (50 3C stream total, 9 MB) that is uploaded to GitHub. The new notebook on colab would pull this small dataset.

Figure 3: please change 50 to 60 km in the sentence "Some events are color-coded white because they are deeper than 50 km" the color scale ends at 60 and not 50 km. Also, it would be a good idea to insert the depth km value labels (not all) for the depth isolines of the slab.

Thanks for pointing out this. The color scale for events is actually on the lower left, which ends at 50 km. Since this potentially makes some confusion, we correct this scale to the same as the plate depth down to 60 km. We also add the depth label as suggested.

Figure 4: This figure (and also the S1) should be redone using the results of the "test" dataset applied to all the Earthquake Transformer training models. Are the RMSE values the uncertainties to the MAE? Also, in this figure the picking completeness attained by the INSTANCE and SCEDC trained datasets for P appears comparable if not better than that obtained from the trained PNW dataset. Similarly, ETHZ, SCEDC and INSTANCE perform better than PNW for S.

The RMSE values are the uncertainties to the MAE, which has been added to the captions. We have corrected the figures as mentioned in the previous discussion. We also agree that the P-wave picking completeness from INSTANCE and SCEDC (and ETHZ, SCEDC and INSTANCE for S-wave) are indeed better than PNW, as they all picked more arrivals than PNW on all ground truth picks. However, all of the models mentioned, while having more picks, have slightly worse picking quality in terms of larger MAE and RMSE than PNW. They also show a very large S-wave picking bias compared to the PNW.

Figure 6: the Mh magnitudes (green) are barely visible (if not invisible). Maybe that the adoption of a log scale would help the visualization although it would make less visible the changes in the number of data available through time.

We agree that the Mh events are too few to be visible on the plot. However, having a log scale plot would make the number of events less visible through time given that they are almost at the same order of magnitude. Thus, we would want to keep this figure as it is.

Figure 7: it would be good to specify that the units of the amplitude are in counts in this figure and all those showing seismograms in the supplement.

We now add this in the caption of each figure.

Reviewer-C comments

Dear authors,
I find your dataset and the accompanying paper a good and important addition to the collection of ML-ready datasets. It's nice that you follow the SeisBench standard, making the dataset easily usable with existing codes. Another important thing that your dataset brings is the presence of a nicely organized collection of 'exotic' events. There are also some minor issues that I think should be improved before publishing the paper. The minor issues are:

1. It would be good to expand the sentence in the introduction on Seisbench data format explaining the idea behind it.

We expand the discussion of SeisBench in the introduction section.

2. P2-L95: Please reference the Figs. S10 and S11 here.

This has been added to the new submission.

3. P4-L148: Please reference Table 1 in the last sentence to allow the reader to see the number of the kept events immediately.

We edit the text as suggested.

4. P5-L188: Do you have an easy way to add information about clipping into the metadata? It would be one of the ways to allow the users of the dataset to select high-quality data only. For this purpose, you could also label the traces which had missing channels, offsets, etc.

Thanks for suggesting this. We now save the number of missing channels in the metadata (trace_missing_channel attribute). We also visually check waveforms from 2002 when potential offsets are possible and manually labeled if the traces have any visible offsets (trace_has_offset attribute).

5. It would be good to have the order of the supplement figures follow the order in which they are referenced (currently e.g. Figs. S20, S21, S23, S24, S19 are referenced before Figs. S10, S4, S11).

We reorder the references of figures and tables in the supplementary material.

6. P6-L260: Please explain better your retraining procedure, i.e. what data did you use for retraining, and what data for validation/testing?

We add information about dataset splitting (70% for training and 30% for testing). The labels are triangular with 10-sample half-width. We use the same loss function that Mousavi et al. (2020) used to train the original EqT (weighted sum of loss from P-, S- and detection branches). All information has been added to the main test.

7. It should be noted that STEAD probably contains some of the data that you are testing on, i.e. it was trained on them. This means that there is some information leakage and that the comparison between the model trained on STEAD and the models trained on other datasets is not fair.

We acknowledge the existence of some streams contributed by PNSN in the STEAD dataset. But this only makes at most ~3% of STEAD. What's more, PNW does not apply filters to the data, while STEAD is band-passed from 1 to 45 Hz, which makes our testing data different from STEAD. We added such a comment to the main text. To make a better comparison, one should only compare PNW-retrained to INSTANCE, SCEDC and ETHZ.

8. Do you plan to make the dataset available through SeisBench (when the paper is published)? I think this would be a good way to make your dataset easier to find. If you plan to do it, you could add this to the Code and Data Availability section.

Yes. We will make data available through SeisBench. We now mention this in the Code and Data Availability section and will create a pull request on SeisBench's github repository.