Round 1

Comments by Handling Editor Hongyu Sun:

1. Publishing a dataset to support AI applications in seismology is always welcome, especially when it fills a gap in the community. However, it would be helpful to provide more detailed information about the dataset so that readers can easily understand and use it. For example:

- Where do the event catalogs come from (e.g., which agencies and geographic regions)?
- How were these catalogs built?
- What is the magnitude of completeness of the catalogs used to construct the training dataset?
- The maximum magnitude in the dataset is 2.0; did you remove all events above this threshold?

A summary table of this information, along with a map showing the spatial distribution of the datasets, would greatly benefit the readers.

2. The procedure for incorporating noise samples is somewhat unclear. In line 134, you mention that, to obtain noise samples from continuous data recorded at Rittershoffen, we first building a catalog using SeisComp and a preliminary PhaseNet model (Details are given in Building an induced seismicity catalogue.). What is the difference between these two in the context of obtaining noise samples?

In the section "Building an Induced Seismicity Catalogue," you state: "Once we found models that perform well on our test data, we applied these models to one month of continuous data from Rittershoffen." This raises a few questions:

(1) Was the PhaseNet model used here the preliminary version or a fully trained one?

(2) How do you define the “preliminary” PhaseNet model?

(3) Later in the same section, you mention: "We trained three different models from scratch for induced seismicity using different numbers of noise samples." Did you already have a pre-existing collection of noise samples when building the catalog?

Overall, the description of the procedure is somewhat confusing. Including a clear workflow diagram that outlines the steps, number of models compared, and how each was trained and fine-tuned would be very helpful.

3. The manuscript emphasizes the importance of noise from Rittershoffen. However, what is the role of the STEAD dataset in this context? Was the noise data from Rittershoffen insufficient for training? Should future users combine STEAD with local noise data when training their own models?

Minor comments:

Line 11: “Applying current published models to induced seismicity data leads to only a few events being detected and deep-learning pickers are not able to outperform well-established workflow in seismology.”

This is a strong claim and should be stated with more caution. Have you tested all available published models to support this conclusion? For example, you mention only three single-station deep-learning models up to 2020, while the field has progressed rapidly since then. More recent multi-station approaches, such as PhaseNO (Sun et al., 2023), have demonstrated improved performance in detecting low-SNR signals compared to PhaseNet and EQTransformer.

Line 33: “Instead of calculating explicit features from three component seismic waveforms…”

Please note that single-component waveforms can also be used in such workflows by repeating the component three times.

Line 69: “the final data set consists of 171 175 waveforms from 40 228 different seismic events.”

The abstract states that PhaseNet was trained with 170,000 three-component waveforms from 40,000 events. Please use the exact numbers consistently throughout the manuscript.

Line 152: “A pick is counted as a true positive if the probability distribution exceeds a certain decision threshold (pick threshold in Fig. 6) within ±0.25 s of the true phase arrival. A false positive is a predicted phase arrival that does not match any true arrival within a larger window (typically ±2 s)”

Please clarify how the ±0.25 s and ±2 s thresholds were determined. Why are different time windows used? For example, there are works using 0.5 s for both time windows.

Table 2. I suggest adding brief descriptions of the parameters listed, especially for readers who may not be familiar with PyOcto.
Figure 5. The left-y axis seems to be incorrect. Do you have tens of thousands picks in six hours of data?

Line 216. You mention that the threshold for PN-STEAD is 0, meaning every time point is considered a pick. This seems problematic and not meaningful for phase picking. Please verify and revise accordingly.

Reviewer A: Cindy Lim Shin Yee

In this study, the authors investigate whether using differently trained PhaseNet models (an extensively trained model) could outperform the original PhaseNet model when applied to induced seismicity from a geothermal site. Specifically, the authors compared re-trained PhaseNet from scratch using a local induced seismicity dataset (piSDL), transfer learning with the piSDL dataset (TF original), the original model (PhaseNet), the model trained with STEAD (PN stead) and transfer learning with the piSDL data on the STEAD-trained PhaseNet (TF stead). They also present the piSDL dataset as a training dataset that consists of low magnitude, low SNR induced events. The authors also suggest that including noise samples when re-training PhaseNet can decrease the number of false picks, however, the authors have yet to show that adding noise samples decreases the number of false picks (i.e., increases precision). Overall, I found the article and its results interesting. The study presents useful insights into re-training deep learning models for induced seismicity detection, and the results could contribute meaningfully to the field. However, some key aspects—such as quantifying performance improvements, clarifying methodology, and restructuring certain sections for better readability—need to be addressed. I have outlined specific points below to improve the clarity and robustness of the manuscript.

All the best,

Cindy Lim Shin Yee

Main points:

Abstract

- Line 18-19: It is unclear in the abstract what the authors mean by which "new model" outperforms PhaseNet's original published model- training from scratch with piSDL or transfer learning with piSDL or all the newly trained models that included piSDL yielded better results: please clarify this here.
- Line 18-19: It is also important to quantify what "outperforms" means in this case (e.g. detecting x% more events than the original PhaseNet or reducing the false positive rate by x%).
- Line 20-21: The authors need to quantify what they mean by "more events" (i.e. x% more events than those published by an agency).

Introduction

- Pg 2, Lines 38-40: This is a good place to separate the paragraphs so paragraph 1 is about the background of automated phase picking and the deep learning phase picking models and paragraph 2 can be about the different datasets (where the authors introduce the induced seismicity dataset)

- Pg. 2, Lines 43-45: There have been a few studies that prove the contrary (i.e. that existing, original models like PhaseNet and EQTransformer have performed well in different geographical regions): e.g., Scotto di Uccio et al (2023) showed success with EQTransformer in Southern Italy. Pitta-Slim (2023) with EQTransformer in New Zealand and Lim et al (2025) in using the both PhaseNet and EQTransformer to detect induced seismicity (even down to low magnitudes of -3). It would be useful to discuss the mixed success of these original models to provide context for the necessity of re-training.
- Pg.2, Lines 45-47: If the authors argue that models fail for small source-receiver distances and low SNR, a reference would strengthen this claim. Previous studies (e.g., Lim et al., 2025) have shown that while original models (GPD, PhaseNet, EQT) can detect induced seismicity, their detection rates tend to decline for smaller induced events. Discussing this would help contextualise the performance of the newly trained models.
- Pg. 2, Lines 48-49: This might be a good place to separate into paragraphs 2 and 3, where paragraph 3 covers the work the paper has done and where the authors present the new dataset for training these original models

Data Set:

- Pg.2, Lines 62-63: It would be interesting to know what the % split of the data is from location-wise to ensure there is minimal bias during testing (e.g. if most of the data is from where your test is)
- Pg. 3, Lines 69-70: Perhaps it is more important/interesting to include % split of the locations of where the seismic events and a % split of the different instruments/stations data (e.g., broadband, etc) from the final training dataset (either a list of proportions of the data in the supplementary e.g. like in the PhaseNet paper)
- Pg.3, Lines 72: Did the authors set the correct sampling rates for the detection/data reading for PhaseNet during the testing? i.e,. set the sampling_rate variable to 100 Hz for 100 Hz data and 300 Hz for 300 Hz data? This is an important technical detail as it can affect the way PhaseNet reads the data.

Methods:

- Pg. 7, Lines 126-127: Could you show the split of the data with regards to its location and instrument type for the training, validation and testing sets?
- Pg. 7, Lines 133-134: The order of the figures need to be arranged as “Figure 5” is called before “Figure 4” (later referenced in the Results section). The authors need to correct the order so the reference figure here is referenced as Figure 4.
- Pg. 8, Lines 151-152: The referenced figure here should be labelled as Figure 5 to maintain reference order.
- Where is the reference to Tab. 3 in the methods section (where the authors show the different models trained with piSDL dataset)? These models need to be specified earlier in the Methods section.

Results:

- Pg 10, Lines 181-182: Figure 4b has to be relabelled to Figure 6 (as there were two referenced Figures (originally Figure 5 and 6) before this Figure).
- Pg. 10, Line 187: "Probably, many of these picked phase onsets are false picks." These results need to be quantified with respect to a ground-truth or with manual selection. The quantification of the number and proportion of false picks is important to justify that the "piSDL without noise" model produces too many false picks and that "piSDL with noise" does not have a low recall value (i.e., piSDL with noise is missing phases).
- Pg.10. Lines 187-189: "Adding noise samples… enabled PhaseNet to learn how noise samples differ from seismic phase onsets…"- this line is an interpretation of the results and needs to be separated to be in the Discussions section. Or in the results section, the authors could write that "As we observed that training PhaseNet with our piSDL with noise samples yielded 33 P and 35 S-arrival detections. This could imply that adding noise samples from the analysed stations… enabled PhaseNet to learn…, improving its ability to distinguish between both phase and noise."
- Pg. 10, Lines 189-190: "After training PhaseNet with our induced… detected only 33 P and 36 S arrivals." This result needs to be compared with the groundtruth of the continuous half an hour (e.g., manually picked data) so that we know how each model is performing with respect to its recall rate (number of missed events) and precision (number of false picks).
- Pg. 10, Lines 190-191: I'm unsure what the authors mean by this line. Does it mean that when training PhaseNet with noise samples, the rate of picking is reduced AND there are more noise samples added to the dataset?
- Pg.10, Lines 192-195: Instead of posing the question with "Perhaps", the authors can restructure these lines by stating: "To investigate if there is an optimum size for the noise dataset, we gathered up to 60,000 noise samples…" - also, this might be a good place to break into a new paragraph for improved readability.
- Pg. 10, Line 195-196: "Since each model works slightly differently, we trained ten models for each bin of noise samples." - how and why did the authors bin the noise samples? Please justify/clarify this.
- Pg. 10, Lines 204-296: In Figure 5 alone, it looks like the correct picks wrt the catalogue (%) is decreasing when the noise samples (%) are increased… it is hard to tell whether "adding 5-15% of noise samples reduces the number of false picks". In the figure, there might be merit to adding between 0 and 5% of noise samples as it increases the % of correct picks for the P picks in catalogue (black line) but it is unclear for the other lines as the % of correct picks wrt catalogue does not increase after 0% of noise samples. It would be more useful to estimate the precision (low precision for high number of FPs) for the different % noise samples to show whether precision increases with % of noise samples in the training dataset.

Results: application to continuous data

- Pg. 12, Lines 235-236: "To test our new models…" - which new models are the authors referring to? Do they mean the models in Tab. 3? Tab. 3 is yet to be referenced in the paper (it is actually referenced too late in this results section Lines 246-247. This needs to be in the Methods section!!!!!

- Pg. 12, Lines 247-248 “Both transfer learned models have been trained with 20,000 noise samples from Rittershoffen and …” which models are the authors referring to? Is it any of the PN piSDL1-3 models? Or are the authors referring to the TF original and TF STEAD? Please clarify this line for ease of readability
- Pg. 15, Line 259-260: “After manual analysis of **all** events…” Do the authors mean all events detected by all the models or the events both detected by the PN piSDL3 and SeisComp?

Discussion:

- Pg. 17-18, Lines 283: “Adding noise samples from both STEAD and Rittershoffen was pivotal in reducing false picks, as illustrated in Figure 5.” - again, Figure 5 does show that the number of picks for both phases do decrease as the % of noise samples increase however, the % of correct picks with respect to the catalogue also decreases past 0% noise samples (with the exception of an increase for the black line, indicating P picks in catalogue (%))...
- Pg. 18, Lines 286-287:”Conversely, adding 10-15% noise samples optimised the model’s ability to differentiate between noise and seismic phases” - need to show the precision and recall values to strengthen this claim
- Pg.18, Lines 291: “... we selected randomly 30s time windows from continuous data where no seismic event was known in this 30 s. This selected window was then added to the noise dataset.” - Do the authors mean that their noise dataset consists of thousands of these random 30s time windows? If that is the case, this line should have been clarified/stated in the Methods section as this part of how the authors construct the noise dataset.
- Pg. 18, Lines 292-294: “Since training of PhaseNet requires a large number of waveform samples… manual inspection of the noise samples is not possible…” - arguably, it is very important for a robust training dataset to ensure that the noise samples should not contain events. I commend the authors for stating this fact and understand that it is difficult to manually inspect all of them, however, it is then difficult to robustly state that using all the unchecked “noise samples” reduces the number of false picks.
- Pg. 18-19, Lines 319-321: “Further improvements… our workflow did not **sort out** events with source locations outside of the analysed network…” - what do the authors mean by “sort out”? Do you mean that the events with source locations outside of the analysed network were excluded or not excluded?

References:

- Lim, C.S., Lapins, S., Segou, M. and Werner, M.J., 2025. Deep learning phase pickers: how well can existing models detect hydraulic-fracturing induced microseismicity from a borehole array?. *Geophysical Journal International*, *240*(1), pp.535-549.
- Scotto di Uccio, F., Scala, A., Festa, G., Picozzi, M. and Beroza, G.C., 2023. Comparing and integrating artificial intelligence and similarity search detection techniques: application to seismic sequences in Southern Italy. *Geophysical Journal International*, *233*(2), pp.861-874.

- Pita-Sllim, O., Chamberlain, C.J., Townend, J. and Warren-Smith, E., 2023. Parametric testing of EQTransformer's performance against a high-quality, manually picked catalog for reliable and accurate seismic phase picking. *The Seismic Record*, *3*(4), pp.332-341.
- Wong, W.C.J., Zi, J., Yang, H. and Su, J., 2021. Spatial-temporal evolution of injection-induced earthquakes in the Weiyuan Area determined by machine-learning phase picker and waveform cross-correlation. *Earth and Planetary Physics*, *5*(6), pp.520-531.

Minor points:

Abstract

- Pg. 1, Line 8: "Training deep-learning picking models… can be **easily** done through…"
- Pg. 1, Line 10: "... in the low magnitude **(comma)** close distance region" or "... in the low magnitude**(dash)**close distance region"
- Pg.1 Line 15-16: "Noise samples were added **in the training data** to reduce the number of false picks"
- Pg. 1, Lines 16-18: "**In this study**, we noticed that a good earthquake training data set and noise samples from the analysed area are **both** important to detect more seismic events with a newly trained PhaseNet model"
- Pg. 1, Lines 20:22: "The newly created seismicity catalogue contains **more (increase of x% events)**… , and also, **successfully recalled most (x% events)** that have already been detected.

Introduction

- Pg. 1, Lines 24: restructure sentence to "**For example**, precise onset times of different seismic phases are essential for accurate source locations and travel time tomography."
- Pg. 2, Lines 33-34: "Instead of calculating explicit features… deep-learning algorithms **learned** from large training datasets to determine…"

Data Set:

- Pg. 2, Lines 63-64: "However, after **initial** training of the PhaseNet model and testing it with our dataset…"
- Pg. 2-3, Lines 66-69: "Additionally, all samples **are** likely mislabelled by analysts, extremely weak…"

Methods: model training

- Pg. 7: Lines 120-123: The authors could also justify using the original pre-trained model for transfer learning as there are more studies that show that the original PhaseNet can pick induced seismicity (Wong et al, 2021; Lim et al, 2025).
- Pg. 9, Line 166-167: "...reduce the number of false picks, when **training/optimising** (?) our trained models."

Results:

- Pg. 10, Lines 192-195: “with different numbers of noise samples, **n_noise = (model_number -1) * 5000,** i.e., first model trained with zero noise samples, a second with 5000 and the 13th model with all 60,000 noise samples.”
- Pg. 10, Line 211: “**However,** there is a trade-off between precision and recall.”

Discussion:

- Pg. 18, Line 297-298: “Notably, the transfer-learned STEAD model and the PN piSDL models (**PhaseNet model trained from scratch**)…” - Please clarify if you mean the PhaseNet model trained from scratch using the piSDL dataset?
- Pg.18, Lines 319-321: “Further improvements… and still(**comma**) challenges in associating overlapping seismic events remain.”
- Pg. 19, Lines 323-325: “Applying a model trained from scratch without earthquake waveforms from Rittershoffen results in a similar number of detections as **for** the original model, even though the model was trained with noise samples from the site.”

Reviewer B

Overall, the paper presents a clear, structured, and important advancement in seismic event detection for induced seismicity applications using deep learning. The methodology, which involves training neural networks with various strategies and then applying these to continuous seismic data, is well-executed and relevant to current seismological challenges.

However, several important aspects require clarification and further refinement.

Major Comments:

#1 The results, especially those summarized in Table 4 and Figure 9, raise some questions. For instance, there is a substantial discrepancy between the high number of picks detected (over 100,000) and the relatively low number of associated events (max 39 events). The authors should re-evaluate the correctness of their velocity model or recheck their association strategy and parameters to ensure the reliability of the results. Currently, this discrepancy appears unusual and undermines the confidence in the final event detections.

Additionally, including a map showing station locations and the distribution of detected seismic events would enhance the clarity and credibility of the results. The authors should also clearly define criteria used to determine "common events" between catalogues (e.g., specific thresholds for temporal and spatial coincidence).

Minor suggestions:

#1 Data Description: The authors mention that 321,946 waveform samples were collected from various sources. A more explicit description of these sources, including geographic distribution and agencies, is necessary for completeness and reproducibility.

#2 Data Sampling Rate: The manuscript mentions sampling rates of 100 Hz and 300 Hz. Clarifying the distribution or proportion of data recorded at each sampling rate would provide important context for understanding any potential influence of sampling rates on model performance.

#3 Waveform Representation and Analysis: Given that the Rittershoffen dataset cannot be directly shared, the authors should include more waveform figures from their test results. Additional waveform examples, along with more comprehensive discussions on specific detections, especially challenging or unique cases, would enhance the manuscript's depth and relevance.

In summary, addressing these suggestions will substantially improve the manuscript’s clarity, reliability, and impact.

Summary of resulting changes

Thanks for your time to review our work and to improve the manuscript. The main changes in the revised manuscript are, that we added a Figure (Fig. 2) to shows how we optimised our dataset for induced seismicity by removing waveforms, training preliminary PhaseNet models to build earthquake catalogues, removing false event detections after manually inspections and cutting out noise samples to train our final PhaseNet models. These models are compared at the end of the work. Further, we added a table that shows more details each single dataset (Table 1). However, as you noticed, we had thousands of picks when we tested our models with respect to the number of noise samples. The left y-axis is correct (thousands of picks) and that was the main reason why we included noise samples to the training. Figure 10 (comparison of catalogues from different models) is a little bit easier now, since we removed the doubling when comparing catalogues against each other. We also added a second example the main manuscript to show the differences between our new trained PhaseNet model and the original model. The supplementary material includes even more examples, showing all three components of each station. Further, we added to the supplementary a Figure showing the event and station distribution of each dataset, pie diagrams of the channels, details about the number of picks and a map where we compare the locations of the detected events.

In the following is the detailed response to the suggestions of the handling editor, reviewer 1 & 2. Please note that changes in the abstract have not been marked in different colours by Latexdiff.

Comments by Handling Editor Hongyu Sun:

1. Publishing a dataset to support AI applications in seismology is always welcome, especially when it fills a gap in the community. However, it would be helpful to provide more detailed information about the dataset so that readers can easily understand and use it. For example:

- Where do the event catalogs come from (e.g., which agencies and geographic regions)?
- How were these catalogs built?
- What is the magnitude of completeness of the catalogs used to construct the training dataset?
  I added a table (Table 1) that summarises the agencies, magnitude of completeness and which catalogues have been pick manually, automatic or both. Further a map for each dataset, showing all events and most of the station locations is supplemented.
- The maximum magnitude in the dataset is 2.0; did you remove all events above this threshold?

  No, the maxmimum magnitude in our dataset is 4.5 (can be seen in Fig. 4a). We only took all available events from the Swiss Seismological Service (SED) with M_L <= 2 to increase the number of events in our dataset. Additionaly, this dataset as a high quality since it only contains manual picks. I corrected this point in the first sentence in chapter 2 (Data set), by adding the term natural events. So our dataset contains induced seismic events with no limit in magnitude and natural (tectonic) events with a maximum magnitude of 2.

A summary table of this information, along with a map showing the spatial distribution of the datasets, would greatly benefit the readers.

2. The procedure for incorporating noise samples is somewhat unclear. In line 134, you mention that, to obtain noise samples from continuous data recorded at Rittershoffen, we first building a catalog using SeisComp and a preliminary PhaseNet model (Details are given in Building an induced seismicity catalogue.). What is the difference between these two in the context of obtaining noise samples?

In the section "Building an Induced Seismicity Catalogue," you state: "Once we found models that perform well on our test data, we applied these models to one month of continuous data from Rittershoffen." This raises a few questions:

(1) Was the PhaseNet model used here the preliminary version or a fully trained one?

(2) How do you define the “preliminary” PhaseNet model?

(3) Later in the same section, you mention: "We trained three different models from scratch for induced seismicity using different numbers of noise samples." Did you already have a pre-existing collection of noise samples when building the catalog?

Overall, the description of the procedure is somewhat confusing. Including a clear workflow diagram that outlines the steps, number of models compared, and how each was trained and fine-tuned would be very helpful.

The preliminary PhaseNet model was only trained with our induced seismicity dataset and no noise samples. Therefore, the first derived catalogue contains many false detections and after a manually inspection and sorting out of these false detections, we were able to exclude noise sample from that catalogue and to train PhaseNet with our new dataset and noise samples. I added a Figure (Fig. 2) that shows the worklfow, i.e. optimising the training dataset, training a preliminary PhaseNet model, cutting out noise samples, using a preliminary seismicity catalogue and train the full models. However, now the manuscript has 11 Figures.

3. The manuscript emphasizes the importance of noise from Rittershoffen. However, what is the role of the STEAD dataset in this context? Was the noise data from Rittershoffen insufficient for training? Should future users combine STEAD with local noise data when training their own models?

As mentioned in the discussion, we did not check if our noise samples are purley noise since we only used our derived catalogue to cut out noise samples. This means, that seismic events that were only registered at single stations could be extracted as noise samples, since these events are not part of the fully associated seismicity catalogue. Therefore, we were a little careful not only taking noise samples from Rittershoffen. Since STEAD has a well derived noise database, we also inlcuded noise samples from STEAD. I think it would be great to see what happens if a PhaseNet model is only trained with noise samples from STEAD or from the site of interest (here Rittershoffen). I add this point to the discussion to keep this is mind for future studies.

Minor comments:

Line 11: “Applying current published models to induced seismicity data leads to only a few events being detected and deep-learning pickers are not able to outperform well-established workflow in seismology.”

This is a strong claim and should be stated with more caution. Have you tested all available published models to support this conclusion? For example, you mention only three single-station deep-learning models up to 2020, while the field has progressed rapidly since then. More recent multi-station approaches, such as PhaseNO (Sun et al., 2023), have demonstrated improved performance in detecting low-SNR signals compared to PhaseNet and EQTransformer.

You are totally right. I corrected this in the abstract to emphasize that we only trained PhaseNet and also added PhaseNo for further improvements to the discussion and conclusion.

Line 33: “Instead of calculating explicit features from three component seismic waveforms…”

Please note that single-component waveforms can also be used in such workflows by repeating the component three times.

Yes this point is clear to me that PhaseNet can also predict picks from single component waveforms. However, in this study we only use three component seismic waveforms to train PhaseNet and also most of the published datasets include three component seismic waveforms.

Line 69: “the final data set consists of 171 175 waveforms from 40 228 different seismic events.”

The abstract states that PhaseNet was trained with 170,000 three-component waveforms from 40,000 events. Please use the exact numbers consistently throughout the manuscript.

This is corrected throughout the whole manuscript.

Line 152: “A pick is counted as a true positive if the probability distribution exceeds a certain decision threshold (pick threshold in Fig. 6) within ±0.25 s of the true phase arrival. A false positive is a predicted phase arrival that does not match any true arrival within a larger window (typically ±2 s)”

Please clarify how the ±0.25 s and ±2 s thresholds were determined. Why are different time windows used? For example, there are works using 0.5 s for both time windows.

To have a strict evaluation criterium when testing the models, we selected ±0.25 s for the pick uncertainty. Otherwise the pick is not counted as a true positive. However, if a pick occurs in a window with ±2 s wrt the true pick and has a pick uncertainty > 0.25 s, then the predicted pick is counted as a false positive pick.
I think taking only 0.5 s for the window size to search for predicted picks and also using 0.5 s for the uncertainty are not good enough for a robust model evaluation, especially finding false positives. However, I have added a few sentences to this section to emphasize that point.

Table 2. I suggest adding brief descriptions of the parameters listed, especially for readers who may not be familiar with PyOcto.

I added a column with a short description of these parameters.

Figure 5. The left-y axis seems to be incorrect. Do you have tens of thousands picks in six hours of data?

Unfortunately, we have such a high number of picks when training PhaseNet without noise samples. This effect is due to the low SNR events in our dataset. I think Figure 6b (in new manuscript) and Figure S4 clearly show that we predict many false picks when no noise samples are included during training and that still many false picks are detected also if PhaseNet was trained with noise samples. I think this is one critical part. Either we miss many events, for example when applying PhaseNet's original model or we have to deal with many false picks, but in our case, we can remove these picks after phase association. However, as you have already mentioned, approaches like PhaseNo might be able to improve the problem by using spatial information from neighboring stations.

Line 216. You mention that the threshold for PN-STEAD is 0, meaning every time point is considered a pick. This seems problematic and not meaningful for phase picking. Please verify and revise accordingly.

The models have been tested for pick thresholds in the range [1e-3, 1]. The optimal pick threshold was derived from the closest value on the precision-recall curve to the point [1, 1]. Therefore, the optimal threshold would be 1e-3, but we rounded only up to two decimal points. However, also 1e-3 is an extremely bad optimum for a pick threshold.
I added the test range to the text and also added, that the PN STEAD model does not work on our test dataset.

Reviewer A:

In this study, the authors investigate whether using differently trained PhaseNet models (an extensively trained model) could outperform the original PhaseNet model when applied to induced seismicity from a geothermal site. Specifically, the authors compared re-trained PhaseNet from scratch using a local induced seismicity dataset (piSDL), transfer learning with the piSDL dataset (TF original), the original model (PhaseNet), the model trained with STEAD (PN stead) and transfer learning with the piSDL data on the STEAD-trained PhaseNet (TF stead). They also present the piSDL dataset as a training dataset that consists of low magnitude, low SNR induced events. The authors also suggest that including noise samples when re-training PhaseNet can decrease the number of false picks, however, the authors have yet to show that adding noise samples decreases the number of false picks (i.e., increases precision). Overall, I found the article and its results interesting. The study presents useful insights into re-training deep learning models for induced seismicity detection, and the results could contribute meaningfully to the field. However, some key aspects—such as quantifying performance improvements, clarifying methodology, and restructuring certain sections for better readability—need to be addressed. I have outlined specific points below to improve the clarity and robustness of the manuscript.

All the best,

Cindy Lim Shin Yee

Main points:

Abstract

- Line 18-19: It is unclear in the abstract what the authors mean by which “new model” outperforms PhaseNet’s original published model- training from scratch with piSDL or transfer learning with piSDL or all the newly trained models that included piSDL yielded better results: please clarify this here.
  Thanks for this comment, you are totally right and I corrected this sentence by writing that the models trained with our dataset and noise samples outperform PhaseNet’s original model and traditional methods in seismology.
- Line 18-19: It is also important to quantify what “outperforms” means in this case (e.g. detecting x% more events than the original PhaseNet or reducing the false positive rate by x%).
  We were able to detect up to 62% more events in comparison to the catalogue from the agency. I corrected this point in the abstract
- Line 20-21: The authors need to quantify what they mean by “more events” (i.e. x% more events than those published by an agency).

  See bullet point above.

Introduction

- Pg 2, Lines 38-40: This is a good place to separate the paragraphs so paragraph 1 is about the background of automated phase picking and the deep learning phase picking models and paragraph 2 can be about the different datasets (where the authors introduce the induced seismicity dataset)
  Thanks a lot, I added a paragraph.
- Pg. 2, Lines 43-45: There have been a few studies that prove the contrary (i.e. that existing, original models like PhaseNet and EQTransformer have performed well in different geographical regions): e.g., Scotto di Uccio et al (2023) showed success with EQTransformer in Southern Italy. Pitta-Slim (2023) with EQTransformer in New Zealand and Lim et al (2025) in using the both PhaseNet and EQTransformer to detect induced seismicity (even down to low magnitudes of -3). It would be useful to discuss the mixed success of these original models to provide context for the necessity of re-training.
  I think in our study, we clearly show the PhaseNet’s original published model does not perform well at our test site in Rittershoffen. Therefore, retraining is in our case necessary. Of course, other studies show different behaviors but I think, this are different cases and every PhaseNet user has to decide whether retraining is necessary or not for his/her dataset.
- Pg.2, Lines 45-47: If the authors argue that models fail for small source-receiver distances and low SNR, a reference would strengthen this claim. Previous studies (e.g., Lim et al., 2025) have shown that while original models (GPD, PhaseNet, EQT) can detect induced seismicity, their detection rates tend to decline for smaller induced events. Discussing this would help contextualise the performance of the newly trained models.
  As mentioned above, our results also show that the original PhaseNet model does not perform well at our test site. However, I also added literature how also noticed that the recall

rate of the original model when picking P and S phases is relatively low.
Dai, Z., Zhou, L., Hu, X., Qu, J., & Li, X. (2023). Generalization of PhaseNet in Shandong and its application to the Changqing M4. 1 earthquake sequence. *Earthquake Science*, *36*(3), 212-227.

- Pg. 2, Lines 48-49: This might be a good place to separate into paragraphs 2 and 3, where paragraph 3 covers the work the paper has done and where the authors present the new dataset for training these original models

  There was already a paragraph, however, this was/is not visible because of the perfect length of the sentence to fill up a line.

Data Set:

- Pg.2, Lines 62-63: It would be interesting to know what the % split of the data is from location-wise to ensure there is minimal bias during testing (e.g. if most of the data is from where your test is)
  In the end of section 3.2, where the split of the dataset is introduced, I added the information that each single dataset is split by 70% for training, 20% for validation and 10% for testing.
- Pg. 3, Lines 69-70: Perhaps it is more important/interesting to include % split of the locations of where the seismic events and a % split of the different instruments/stations data (e.g., broadband, etc) from the final training dataset (either a list of proportions of the data in the supplementary e.g. like in the PhaseNet paper)
  I added a Figure to the supplementary to show the distribution of channels for the whole dataset. However, the split will be available in the published dataset in SeisBench but only for the datasets we can make available to the public.
- Pg.3, Lines 72: Did the authors set the correct sampling rates for the detection/data reading for PhaseNet during the testing? i.e,. set the sampling_rate variable to 100 Hz for 100 Hz data and 300 Hz for 300 Hz data? This is an important technical detail as it can affect the way PhaseNet reads the data.

  Data with different sampling rates are automatically resampled in SeisBench. Each SeisBench model that works with seismological data has the sampling rate as an argument. Also the sampling rate is necessary in the metadata file. When reading the waveform data, SeisBench uses the sampling rate information from the metadata file and automatically resamples the waveform data to the required sampling rate. However, in section 3.1 (Model training), I added that data with different sampling rates are automatically resampled in SeisBench.

Methods:

- Pg. 7, Lines 126-127: Could you show the split of the data with regards to its location and instrument type for the training, validation and testing sets?
  I have added a figure with the channel naming for the whole dataset and for the splits. Additionaly, I added a figure only with the events from Rittershoffen which were used fro training, validation and testing. I think in general this is good point where our work shows some problems, since our model is trained with events from the same region where we also tested the model. Indeed, the training dataset does not include events from the test time, however, probably some events might be very similar. Therefore, we trained a model without data from Rittershoffen and found that this model only performs poorly. We

conclude that in our case waveform data from the site of interest are extremely important to guarantee that most of the events will be found when applying the fully trained PhaseNet model in the future.

- Pg. 7, Lines 133-134: The order of the figures need to be arranged as “Figure 5” is called before “Figure 4” (later referenced in the Results section). The authors need to correct the order so the reference figure here is referenced as Figure 4.
- Pg. 8, Lines 151-152: The referenced figure here should be labelled as Figure 5 to maintain reference order.
  Thanks, I corrected both figures.
- Where is the reference to Tab. 3 in the methods section (where the authors show the different models trained with piSDL dataset)? These models need to be specified earlier in the Methods section.

  I moved the table to section 3.4 and added a sentence that we trained three different models with our dataset for induced seismicity and three different noise datasets.

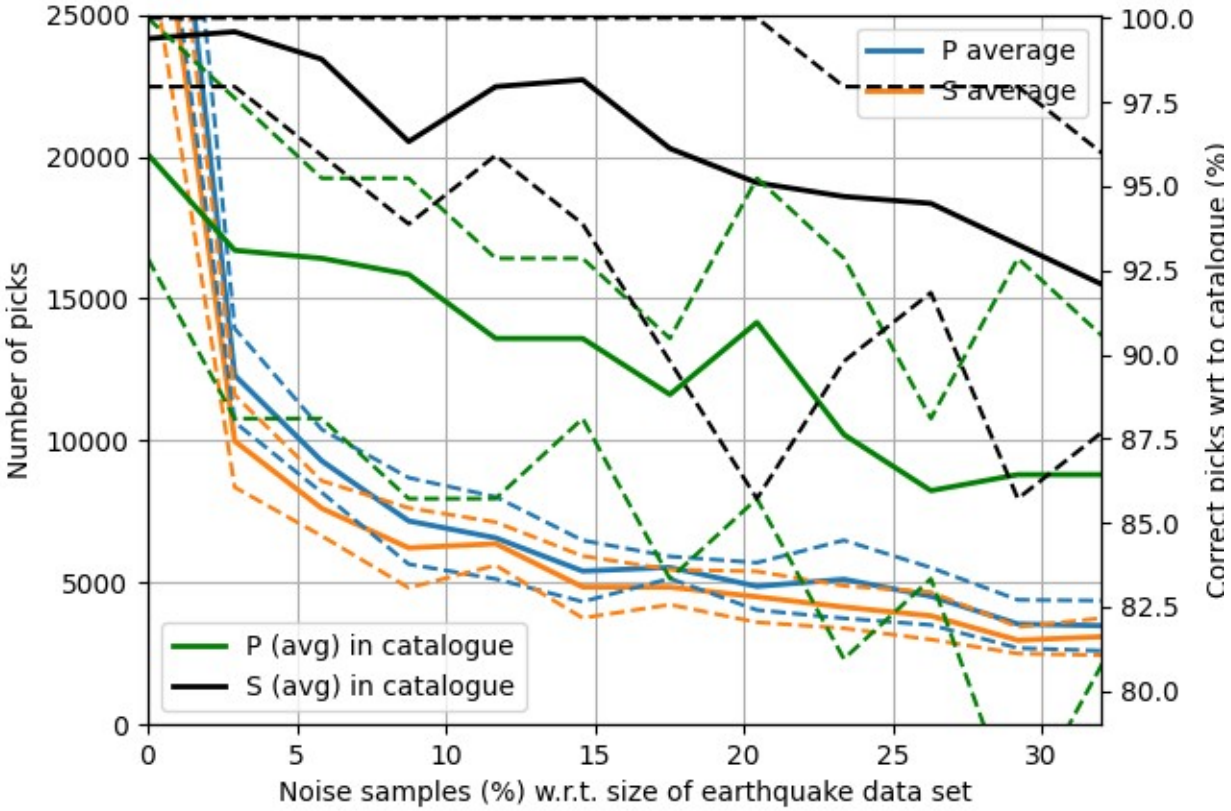
Results:

- Pg 10, Lines 181-182: Figure 4b has to be relabelled to Figure 6 (as there were two referenced Figures (originally Figure 5 and 6) before this Figure).
  I don’t think this is necessary. Otherwise I have to restructure the whole text and in Figure 5 (before Figure 4), I would loose the direct comparison of the original model, piSDL without noise and with noise.
- Pg. 10, Line 187: “Probably, many of these picked phase onsets are false picks.” These results need to be quantified with respect to a ground-truth or with manual selection. The quantification of the number and proportion of false picks is important to justify that the “piSDL without noise” model produces too many false picks and that “piSDL with noise” does not have a low recall value (i.e., piSDL with noise is missing phases).
  After manually inspection of many picks, we conclude that most of the picks are false picks. Furthermore, only two events in the given time window are know. In addition, we show with our noise sample tests that we were able to reduce the number of picks by adding noise samples to the training and still have a high recall in detection correct P and S phases. However, you can never know what is the ground truth, since there can also be very noisy events which are also only visible at single stations.
  After manually checking of the waveforms, I found 13 P and S picks in that time window. I added these information to the sentence.
- Pg.10. Lines 187-189: “Adding noise samples… enabled PhaseNet to learn how noise samples differ from seismic phase onsets…”- this line is an interpretation of the results and needs to be separated to be in the Discussions section. Or in the results section, the authors could write that “As we observed that training PhaseNet with our piSDL with noise samples yielded 33 P and 35 S-arrival detections. This could imply that adding noise samples from the analysed stations… enabled PhaseNet to learn…, improving its ability to distinguish between both phase and noise.”
  Thanks for that point. I correct the sentence and restructured this part a little.
- Pg. 10, Lines 189-190: “After training PhaseNet with our induced… detected only 33 P and 36 S arrivals.” This result needs to be compared with the groundtruth of the continuous half

an hour (e.g., manually picked data) so that we know how each model is performing with respect to its recall rate (number of missed events) and precision (number of false picks).
14 P and 14 S arrivals are the ground truth. Both piSDL models were able to pick all phases, however the original model missed seven P phases but also picked all S phases. I added these information to the section.

- Pg. 10, Lines 190-191: I'm unsure what the authors mean by this line. Does it mean that when training PhaseNet with noise samples, the rate of picking is reduced AND there are more noise samples added to the dataset?
  Yes, these sentence was weird. Now it should be clearer: When training PhaseNet with an earthquake data set and noise samples, the number of picks is reduced the more noise samples are added to the data set.
- Pg.10, Lines 192-195: Instead of posing the question with "Perhaps", the authors can restructure these lines by stating: "To investigate if there is an optimum size for the noise dataset, we gathered up to 60,000 noise samples…" - also, this might be a good place to break into a new paragraph for improved readability.
  Thanks for this great suggestion.
- Pg. 10, Line 195-196: "Since each model works slightly differently, we trained ten models for each bin of noise samples." - how and why did the authors bin the noise samples? Please justify/clarify this.
  I think this was a misunderstanding by the reviewer, however, I corrected this sentence.
- Pg. 10, Lines 204-296: In Figure 5 alone, it looks like the correct picks wrt the catalogue (%) is decreasing when the noise samples (%) are increased… it is hard to tell whether "adding 5-15% of noise samples reduces the number of false picks". In the figure, there might be merit to adding between 0 and 5% of noise samples as it increases the % of correct picks for the P picks in catalogue (black line) but it is unclear for the other lines as the % of correct picks wrt catalogue does not increase after 0% of noise samples. It would be more useful to estimate the precision (low precision for high number of FPs) for the different % noise samples to show whether precision increases with % of noise samples in the training dataset.

  I clearly see the point to have smaller gradations by adding noise samples during the training. However this would mean that have to train once again 60 models (or even more). Here, I only give rough numbers how many noise sample can be added to reduce the number of false picks for our case. I think further studies have to show if the approach of adding 5 – 15% of noise samples to the training dataset could be an improvement for everybody. So far, we are the first study that used this approach. I also calculated the upper and lower limits of correct picked P and S phases (see Fig. below, dashed lines) but this Figure is too confusing even for the supplementary. Also I think using precision could improve the figure, however, for people how are not familiar with metrics from machine learning, this kind of figure is much easier to understand.



Results: application to continuous data

- Pg. 12, Lines 235-236: “To test our new models…” - which new models are the authors referring to? Do they mean the models in Tab. 3? Tab. 3 is yet to be referenced in the paper (it is actually referenced too late in this results section Lines 246-247. This needs to be in the Methods section!!!!!
  All models that are listed in table 5. I added in parenthesis what is meant by models.
- Pg. 12, Lines 247-248 “Both transfer learned models have been trained with 20,000 noise samples from Rittershoffen and …” which models are the authors referring to? Is it any of the PN piSDL1-3 models? Or are the authors referring to the TF original and TF STEAD? Please clarify this line for ease of readability
  I added in parenthesis that TF original and TF STEAD are meant by transfer learned models.
- Pg. 15, Line 259-260: “After manual analysis of **all** events…” Do the authors mean all events detected by all the models or the events both detected by the PN piSDL3 and SeisComp?

  Events detected by PN piSDL3 and SeisComp. This is now clearer for the reader.

Discussion:

- Pg. 17-18, Lines 283: “Adding noise samples from both STEAD and Rittershoffen was pivotal in reducing false picks, as illustrated in Figure 5.” - again, Figure 5 does show that the number of picks for both phases do decrease as the % of noise samples increase

however, the % of correct picks with respect to the catalogue also decreases past 0% noise samples (with the exception of an increase for the black line, indicating P picks in catalogue (%))…

I don't understand this point. The overall number of picks is given in the text and caption of Fig. 5 (42 P- and 49 S-picks). The main point I want to emphasize with our approach is that we are reducing the number of false picks. Of course there is a trade off between reducing the number of false picks and also missing true picks. However, in many cases missing a few picks does not hinder for correct phase association. Furthermore, many earthquake catalogues built width PhaseNet (or similar approaches) detect many events. But till today there is not method how to filter out false detections and thus reducing the number of false picks is one way to reduce the number of false detections in seismic catalogues. And by adding 5-10% noise sample we still detect more than 90% of the S picks and more than 95% of the P-picks.

- Pg. 18, Lines 286-287:"Conversely, adding 10-15% noise samples optimised the model's ability to differentiate between noise and seismic phases" - need to show the precision and recall values to strengthen this claim

  Please see my points above why I don't use precision and recall here.
- Pg.18, Lines 291: "... we selected randomly 30s time windows from continuous data where no seismic event was known in this 30 s. This selected window was then added to the noise dataset." - Do the authors mean that their noise dataset consists of thousands of these random 30s time windows? If that is the case, this line should have been clarified/stated in the Methods section as this part of how the authors construct the noise dataset.

  We describe in section 3.2 how we gathered our noise dataset. But yes, our noise dataset consists of thousands of 30 s time windows.

  "After manually inspecting the catalogue and removing false detections, 30 s three-component noise waveforms were randomly extracted from all stations at the Rittershoffen site."
- Pg. 18, Lines 292-294: "Since training of PhaseNet requires a large number of waveform samples… manual inspection of the noise samples is not possible…" - arguably, it is very important for a robust training dataset to ensure that the noise samples should not contain events. I commend the authors for stating this fact and understand that it is difficult to manually inspect all of them, however, it is then difficult to robustly state that using all the unchecked "noise samples" reduces the number of false picks.

  Thanks for that comment, but then I have to ask why we picked so many false picks when we trained without noise samples? Perhaps taking a cleaner noise dataset (i.e. all noise waveforms were checked previously for events), the number of correct picked picks is higher than for our case. However, also taking our naive approach for gathering noise sample reduces the number of false picks. In the end the term "ground truth" is always a bit difficult since we never know what is the ground truth. Of course some events might be very noisy and denoising method might help to find these events. I think future studies can use this approach very well.
- Pg. 18-19, Lines 319-321: "Further improvements… our workflow did not **sort out** events with source locations outside of the analysed network…" - what do the authors mean by "sort out"? Do you mean that the events with source locations outside of the analysed network were excluded or not excluded?

We did not exclude events which are outside of the network. I corrected this sentence.

References:

- Lim, C.S., Lapins, S., Segou, M. and Werner, M.J., 2025. Deep learning phase pickers: how well can existing models detect hydraulic-fracturing induced microseismicity from a borehole array?. *Geophysical Journal International*, *240*(1), pp.535-549.
- Scotto di Uccio, F., Scala, A., Festa, G., Picozzi, M. and Beroza, G.C., 2023. Comparing and integrating artificial intelligence and similarity search detection techniques: application to seismic sequences in Southern Italy. *Geophysical Journal International*, *233*(2), pp.861-874.
- Pita-Sllim, O., Chamberlain, C.J., Townend, J. and Warren-Smith, E., 2023. Parametric testing of EQTransformer's performance against a high-quality, manually picked catalog for reliable and accurate seismic phase picking. *The Seismic Record*, *3*(4), pp.332-341.
- Wong, W.C.J., Zi, J., Yang, H. and Su, J., 2021. Spatial-temporal evolution of injection-induced earthquakes in the Weiyuan Area determined by machine-learning phase picker and waveform cross-correlation. *Earth and Planetary Physics*, *5*(6), pp.520-531.

Minor points:

Abstract

- Pg. 1, Line 8: "Training deep-learning picking models… can be **easily** done through…"
- Pg. 1, Line 10: "... in the low magnitude **(comma)** close distance region" or "... in the low magnitude**(dash)**close distance region"
- Pg.1 Line 15-16: "Noise samples were added **in the training data** to reduce the number of false picks"
- Pg. 1, Lines 16-18: "**In this study**, we noticed that a good earthquake training data set and noise samples from the analysed area are **both** important to detect more seismic events with a newly trained PhaseNet model"
- Pg. 1, Lines 20:22: "The newly created seismicity catalogue contains **more (increase of x% events)**… , and also, **successfully recalled most (x% events)** that have already been detected.

All points have been corrected

Introduction

- Pg. 1, Lines 24: restructure sentence to "**For example**, precise onset times of different seismic phases are essential for accurate source locations and travel time tomography."
- Pg. 2, Lines 33-34: "Instead of calculating explicit features… deep-learning algorithms **learned** from large training datasets to determine…"

learnt (BE) and learned (AE) ;)

I corrected all points

Data Set:

- Pg. 2, Lines 63-64: "However, after **initial** training of the PhaseNet model and testing it with our dataset…"
- Pg. 2-3, Lines 66-69: "Additionally, all samples **are** likely mislabelled by analysts, extremely weak…"

I corrected all points

Methods: model training

- Pg. 7: Lines 120-123: The authors could also justify using the original pre-trained model for transfer learning as there are more studies that show that the original PhaseNet can pick induced seismicity (Wong et al, 2021; Lim et al, 2025).
  I added your suggestion.
- Pg. 9, Line 166-167: "...reduce the number of false picks, when **training/optimising** (?) our trained models."

  None of your suggestions. Instead of taking I used applying

Results:

- Pg. 10, Lines 192-195: "with different numbers of noise samples, **n_noise** = **(model_number -1) * 5000,** i.e., first model trained with zero noise samples, a second with 5000 and the 13th model with all 60,000 noise samples."
  I think this is clear without using an equation.
- Pg. 10, Line 211: "**However,** there is a trade-off between precision and recall."

  Is corrected

Discussion:

- Pg. 18, Line 297-298: "Notably, the transfer-learned STEAD model and the PN piSDL models (**PhaseNet model trained from scratch**)…" - Please clarify if you mean the PhaseNet model trained from scratch using the piSDL dataset?
- Pg.18, Lines 319-321: "Further improvements… and still(**comma**) challenges in associating overlapping seismic events remain."
- Pg. 19, Lines 323-325: "Applying a model trained from scratch without earthquake waveforms from Rittershoffen results in a similar number of detections as **for** the original model, even though the model was trained with noise samples from the site."

  All points are corrected

Many thanks for your suggestions. I think the most problematic part you saw is about precision and recall for the correct number of noise samples. Here, we have shown that adding noise samples to the training dataset is a good approach to reduce the number of false picks. In our case it was sometimes hard to say if we really know all picks or some picks are not in the catalogue since they were not associated (i.e. events that were only recorded at a single station). I hope, readers and also other studies can use these results to better pick induced seismicity events (especially in noisy environments) and to reduce the number of false picks. For me, this is at the moment one of the main challenges when building seismicity catalogues with deep-learning. We are able to associate thousand of events but there is no method that double checks whether the associated events are all real events.

Recommendation: Revisions Required

Reviewer B:

Overall, the paper presents a clear, structured, and important advancement in seismic event detection for induced seismicity applications using deep learning. The methodology, which involves training neural networks with various strategies and then applying these to continuous seismic data, is well-executed and relevant to current seismological challenges.

However, several important aspects require clarification and further refinement.

Major Comments:

#1 The results, especially those summarized in Table 4 and Figure 9, raise some questions. For instance, there is a substantial discrepancy between the high number of picks detected (over 100,000) and the relatively low number of associated events (max 39 events). The authors should re-evaluate the correctness of their velocity model or recheck their association strategy and parameters to ensure the reliability of the results. Currently, this discrepancy appears unusual and undermines the confidence in the final event detections.

You are totally right and I did not find any other study who used PhaseNet (or similar approaches) to build seismicity catalogues during phase with low seismicity rates. Most studies building catalogues on earthquake sequences and then they have also many picks but are able to associate many events. At the moment, we are analysing such an earthquake sequence in Rittershoffen, where we have more than 300 events in 3 h and we also have a high number of PhaseNet picks. However, in this study, analysed one month of continuous data and twelve stations, which results in a high number of picks but a low number of associated events. Before associating the full month, we selected the parameters for PyOcto by testing it on a few hours of continuous with well known events. After finding the best parameters, we applied this setting to the full month of data. We also tested GaMMA for phase association. GaMMA has less parameters than PyOcto but it associates even less events.
I added to the supplement a figure that shows six hours of continuous data from Rittershoffen and only a single event was associated during this time period. Additionally, I added a table that shows the details how many picks have been predicted at each station.

Additionally, including a map showing station locations and the distribution of detected seismic events would enhance the clarity and credibility of the results. The authors should also clearly define criteria used to determine "common events" between catalogues (e.g., specific thresholds for temporal and spatial coincidence).

I added a map showing the detected events to the supplement. Further I added the following sentence, that we manually checked for common events by checking the origin times and predicted picks: “To find common events between different catalogues, we first compared the origin times. Since we only have a very limited number of events, we then compared the predicted picks and waveforms for each event and station manually. If both the origin time and the picks at the stations match, the events from different catalogues were counted as common events.”

Minor suggestions:

#1 Data Description: The authors mention that 321,946 waveform samples were collected from various sources. A more explicit description of these sources, including geographic distribution and agencies, is necessary for completeness and reproducibility.

I added a table that summarises these results and additionally a map showing all the events, most of the stations and regions is in the supplement.

#2 Data Sampling Rate: The manuscript mentions sampling rates of 100 Hz and 300 Hz. Clarifying the distribution or proportion of data recorded at each sampling rate would provide important context for understanding any potential influence of sampling rates on model performance.

SeisBench automatically resamples waveform data to the required frequency, here 100 Hz. To emphasize this point, I added this point in section 3.1 (Model training).

#3 Waveform Representation and Analysis: Given that the Rittershoffen dataset cannot be directly shared, the authors should include more waveform figures from their test results. Additional waveform examples, along with more comprehensive discussions on specific detections, especially challenging or unique cases, would enhance the manuscript's depth and relevance.

I added an example to Fig. 10 that shows how our PhaseNet model trained on the induced seismicity dataset picks very noisy waveforms. More examples of this kind are added to the supplementary material. A paragraph, discussing these Figures, focusing on these noisy examples, is added to the discussion. Predicting these kind of noisy waveforms is the most challenging aspect for our new trained PhaseNet model.

In summary, addressing these suggestions will substantially improve the manuscript’s clarity, reliability, and impact.

Recommendation: Revisions Required

Round 2

Reviewer A: Cindy Lim Shin Yee

The authors presented a revised version of the manuscript “Picking Induced Seismicity with Deep Learning (piSDL)”.

In the previous version of the manuscript, I raised concerns about the need for clarification on a few instances within the manuscript and the quantification of performance metrics, particularly with respect to how adding noise samples affected precision and false pick rates.

In the revision and response, the authors have thoughtfully addressed these concerns. While the broader challenge of validating all picks in a noisy induced seismicity context remains, the manuscript now more clearly explains the rationale and the observed effects of including noise samples in re-training PhaseNet. All minor points raised in the previous review have also been fully addressed.

With these improvements, I find the manuscript to be a meaningful and well-structured contribution- particularly to the discussion around training datasets for deep learning models aimed at induced seismicity detection. I believe that the manuscript is suitable for publication in Seismica, and I thank the authors for addressing all the comments in reviewing the manuscript.

All the best,

Cindy Lim Shin Yee

Reviewer B

The authors have made substantial improvements in response to the editor's and reviewers' comments, and have nearly addressed all of my original concerns. I only have several additional comments on the revised manuscript:

1. I find the updated Figure 2 helpful in illustrating the overall workflow. However, the manuscript provides only minimal explanation (Lines 145–147), which is insufficient—especially considering that even machine learning experts may find the figure complex. For example, why is the first step necessary, and how exactly is it implemented? You mention training PhaseNet on a preliminary dataset—does that mean you used the entire dataset for training and then applied it to itself to remove single-phase waveforms? Furthermore, since you eventually remove events from Rittershoffen, why include them in the initial step at all? Perhaps your intention is simply to show how the induced dataset is constructed via this workflow. Either way, the workflow should be described in much more detail in the manuscript if it is included as a figure.
2. You mention low signal-to-noise ratio (SNR) multiple times in both the response and manuscript. However, I am curious why you did not apply any filtering to improve the SNR. I recall that both PhaseNet and EQT implementations in seisbench perform certain preprocessing steps, including filtering. For example, in Figure 11, filtering could make the P- and S-wave arrivals much more visible. It's also possible that applying filtering to continuous data might help detect more events.
3. As I mentioned in my previous review, I remain concerned about how you define whether two detections correspond to the "same event." In your response, you mentioned that manual checking was used, but the process remains unclear in the manuscript. For example, what specific criteria or time/location tolerances were used during manual verification? Did you check arrival times, epicentral distance, or waveform similarity? Please provide a clearer explanation of the procedure in the text to ensure transparency and reproducibility.
4. Following up on the editor's comment regarding the 0.25s time window, I believe one important factor is the epicentral distance. Based on Figure 4c, your dataset seems to contain more events with shorter epicentral distances, which would naturally result in smaller picking errors. This rationale would support your choice of a tighter threshold and should be stated clearly in the manuscript. Additionally, you may also consider the following strategy. I recall that the authors of the referenced study explored using multiple thresholds and averaging the results, which could serve as a useful approach to consider. Si, X., Wu, X., Li, Z. et al. An all-in-one seismic phase picking, location, and association network for multi-task multi-station earthquake monitoring. Commun Earth Environ 5, 22 (2024). https://doi.org/10.1038/s43247-023-01188-4
5. I am still concerned about your use of noise waveforms from the STEAD dataset. Since many of them come from global stations, the characteristics of the noise may differ significantly from those in Rittershoffen. Considering that noise waveforms are relatively easy to obtain, I suggest selecting ones from sensors similar to yours, or from other induced seismicity regions such as Oklahoma. At the very least, if you choose to continue

using STEAD, you should restrict the selection to stations located in Europe or other induced earthquake zones for better consistency.

The revised manuscript demonstrates clear progress and reflects thoughtful effort in addressing the previous comments. I appreciate the improvements made to the figures and structure, which enhance the overall clarity. That said, I believe a few areas would still benefit from further elaboration. Clarifying these points will help strengthen the manuscript and improve its reproducibility and impact.

Summary of changes

Dear Editor,
Dear Reviewers,

many thanks for your time to review our work a second time and to improve our manuscript. In the revised manuscript we added two phrases to the caption of Figure 2, added one sentence how we found common events when comparing different seismicity catalogues and also followed the suggestion of the reviewer why we used 0.25 s as an uncertainty for the prediction of a pick. Other open points raised by the reviewer are commented below.

Changes made in the manuscript are marked in blue.

Comments by anonymous reviewer:

- I find the updated Figure 2 helpful in illustrating the overall workflow. However, the manuscript provides only minimal explanation (Lines 145–147), which is insufficient—especially considering that even machine learning experts may find the figure complex. For example, why is the first step necessary, and how exactly is it implemented? You mention training PhaseNet on a preliminary dataset—does that mean you used the entire dataset for training and then applied it to itself to remove single-phase waveforms? Furthermore, since you eventually remove events from Rittershoffen, why include them in the initial step at all? Perhaps your intention is simply to show how the induced dataset is constructed via this workflow. Either way, the workflow should be described in much more detail in the manuscript if it is included as a figure.

  Thanks for that comment. You are right, we used the presented workflow to construct the final dataset, since we noticed after a first training of PhaseNet with all waveforms that the application on continuous data leads to many false picks. In the Figure description we already write, that this workflow (especially the first and second step) is used to create the final induced seismicity dataset. Moreover, in lines 65 – 73 we describe very detailed how we obtained the final dataset and we also refer to Figure 2 in this part. I think it is not necessary to add more explanation to lines 145 – 147 because of doubling and we describe in this section how the set up the noise dataset.
  It is true that we removed events from Rittershoffen for our final dataset. In lines 66 – 70 we describe that extremely weak earthquakes were not successfully detected by our preliminary PhaseNet model. Figure 1 shows some examples of these waveforms.

- You mention low signal-to-noise ratio (SNR) multiple times in both the response and manuscript. However, I am curious why you did not apply any filtering to improve the SNR. I recall that both PhaseNet and EQT implementations in seisbench perform certain preprocessing steps, including filtering. For example, in Figure 11, filtering could make the P- and S-wave arrivals much more visible. It's also possible that applying filtering to continuous data might help detect more events.

  EQT filters the data in the range 1 – 45 Hz and PhaseNet takes unfiltered data. Yes, you are right filtering the data could make the waveforms in Figure 11 more visible. We had similar ideas and applied different approaches, i.e. filtering during and training and prediction,

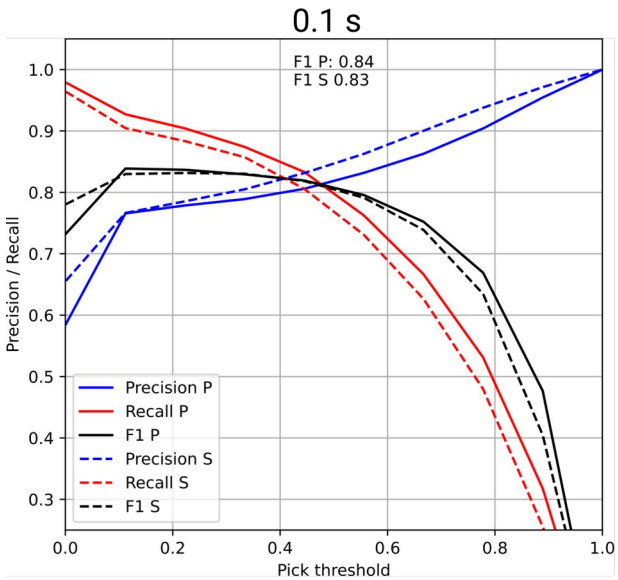
especially for very weak earthquake signals. We did not notice any improvement when we applied this approach. Furthermore, applying a pre-processing like filtering is already feature engineering and deep-neural networks have the power to learn these features on their own. Additionally, using filtered waveforms is one further step for practitioners when they easily apply trained PhaseNet models on their data.

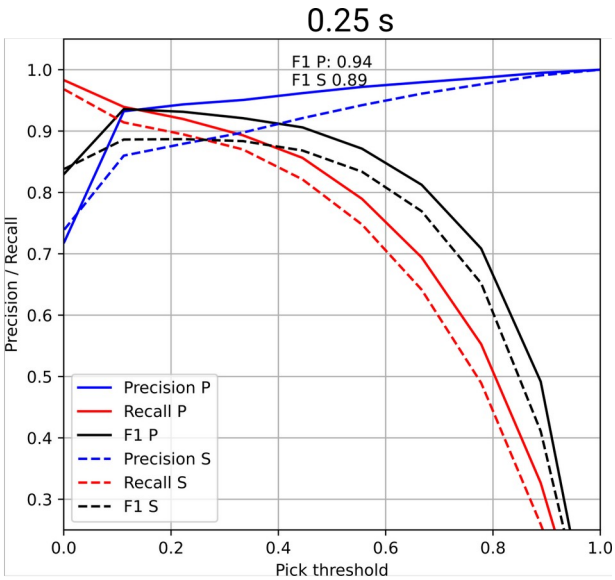
- As I mentioned in my previous review, I remain concerned about how you define whether two detections correspond to the "same event." In your response, you mentioned that manual checking was used, but the process remains unclear in the manuscript. For example, what specific criteria or time/location tolerances were used during manual verification? Did you check arrival times, epicentral distance, or waveform similarity? Please provide a clearer explanation of the procedure in the text to ensure transparency and reproducibility.

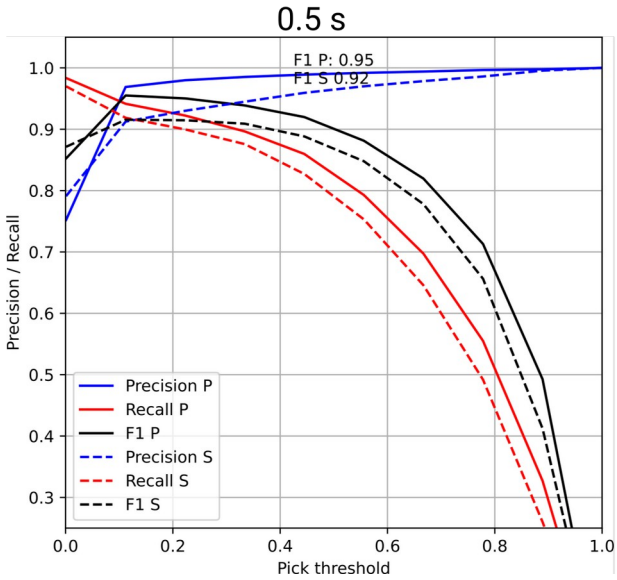
  A common event of two catalogues was found if the origin time of both events lies within +- 0.2 s and we manually checked the picked arrival times of common events against each other (lines 254 - 257). We added this criterion to the text. However, for a more precise comparison further criteria are needed to compare seismicity catalogues against each other. In our case this was not necessary since we only have a very small number of events.

- Following up on the editor's comment regarding the 0.25s time window, I believe one important factor is the epicentral distance. Based on Figure 4c, your dataset seems to contain more events with shorter epicentral distances, which would naturally result in smaller picking errors. This rationale would support your choice of a tighter threshold and should be stated clearly in the manuscript. Additionally, you may also consider the following strategy. I recall that the authors of the referenced study explored using multiple thresholds and averaging the results, which could serve as a useful approach to consider. Si, X., Wu, X., Li, Z. et al. An all-in-one seismic phase picking, location, and association network for multi-task multi-station earthquake monitoring. Commun Earth Environ 5, 22 (2024). https://doi.org/10.1038/s43247-023-01188-4

  Thanks for the argument for the 0.25 s criterion. I added this to line 161.
  Below I attached a Figure that shows the precision-recall curve (PRC) for three different uncertainties (0.1 s, 0.25 s and 0.5 s). Using a residual of 0.1 s for model tests does result in low values for precision, recall and f1-score over the tested pick thresholds, since we are very close to the manual error and the standard deviation of our Gaussian window is 0.1 s. On the other hand, using 0.5 s as an error for a positive pick is very similar to 0.25 s. To conclude, using multiple thresholds and averaging over the results does not improve the precision, recall and f1-score. Further, using larger uncertainties works better for induced seismicity due to weak events. In our case we show that the uncertainty of 0.25 s leads to valid results when testing our models.







- I am still concerned about your use of noise waveforms from the STEAD dataset. Since many of them come from global stations, the characteristics of the noise may differ significantly from those in Rittershoffen. Considering that noise waveforms are relatively easy to obtain, I suggest selecting ones from sensors similar to yours, or from other induced seismicity regions such as Oklahoma. At the very least, if you choose to continue using STEAD, you should restrict the selection to stations located in Europe or other induced earthquake zones for better consistency.

  As mentioned, obtaining the noise dataset from Rittershoffen was done automatically from the derived catalogue. This means we only used time windows when no known event from our previously derived catalogue was in the selected time window. Our proposed method to select noise windows does not ensure that randomly selected time windows contain low magnitude earthquakes that were only visible at a single station (lines 321 – 324). For other regions one has to be very careful when selecting noise windows. Therefore, we decided to work with the noise samples from STEAD since this dataset contains carefully selected noise samples. We don’t think it will make a big difference when only working with noise samples from Europe and also from similar waveforms. Instead PhaseNet learns from a large variety of different noise samples, including those from the site in Rittershoffen. We

also raise the question (lines 316 – 318) what would happen if we only use noise samples from Ritterhoffen or STEAD and do not combine them. Follow up studies might give an answer to that question. We show that using noise samples in addition to an earthquake training dataset leads to less false picks, which is one of the key message of our manuscript.

Round 3

Reviewer B

Thank you for the authors' detailed responses to my comments. I am satisfied with the revisions and have no further questions. The manuscript now meets the standards for publication.