Review of Seismica manuscript "Picking Regional Seismic Phase Arrival Times with Deep Learning" by Aguilar and Beroza.

Thank you for sharing this submission with me. In the manuscript, Aguilar and Beroza developed a Deep Learning model, with a focus of performing earthquake phase picking at distances up to 20 degrees. The model features longer window length input and can pick and classify secondary phases (Pn, Pg, Sn, Sg). The manuscript is likely to be a valuable contribution to the earthquake monitoring community, a good start points of secondary phase picker, and a great showcase of CREW dataset's usefulness. Below are several comments and questions that the authors would need to address, which would improve the manuscript.

L76: For a direct comparison, it is suggested to report the number of parameters of PhaseNet (hereafter referred as PN).

Five extra convolution layers effectively produces a longer receptive field. A similar strategy was also used in Shi and Denolle (2023) to extend the time window. Can authors provide more details on how the choice was quantitatively made? It will be beneficial for future studies on other type of phases, where the length of time windows may vary, and window adaptation is necessary.

Given that SKYNET has some similar structures with PN in the deeper layers, a common question to ask is, would the transfer learning of PN be applied here, that takes the advantage of pre-trained PN model and improve the performance.

L93: consider providing more details about the noise generator used here, and the statistical features of the noise. Why did the authors introduce this specific type of noise data, as in Figure S1. Furthermore, many datasets contain noise samples. Why are these noise samples are not used here?

L145: as a convolution model, the input size of PN is not fixed. One can train PN with CREW dataset but with 300 sec inputs. I am curious how the author would justify that the PN's receptive field is narrower than SKYNET, and that the broadening of the field makes SKYNET superior over PN.

It seems to me that Section 6 lacks some details and need more justification. Please see comments below.

- The training process of the multi-phase picker is a bit confusing for me. Did the author train the model with only 6k samples? Is any data augmentation implemented here?

- And again, did the author implement transfer learning here? The SKYNET trained on the whole CREW dataset can be used and fine-tuned with 6k samples.

- The authors should provide more statistics or figures on this small dataset, which are used to train the multi-phase picker. It may be helpful identifying potential overfitting.

- L169: “we trained during 12000 stages of samples of 200 randomly chosen examples” is a bit confusing to me. Please explain what “stage” indicates here, and where these 200 samples are chosen from.

- L177: Only 3.1k out of 22k examples are kept. I am curious of any sign of overfitting here, as the false positive rate is very high. The authors should better justify the effectiveness of this multi-phase picker, probably experiment with another major events with clear primary and secondary phases and illustrate their results like Figure 3/6/7.

I tried to run the code in Listing 1 but failed. The functions here is not synced with the script in your github repository. Please update your scripts (or this manuscript), with appropriate documentations.

Miscellaneous

Abstract: please be accurate on the use of plural, i.e., models.

Figure 2: consider add some examples from short-period stations.

Figure 4: consider only show the one event that the waveform corresponds to, not all from the swarm, on the map.

Two missing figure labels: Figure 4 caption and L170.

Figure 6: there are several stations shown on the map but has no waveform or picks on the left. The authors may improve this figure by removing stations that are not used or including these waveforms if they can be picked.

The manuscript may benefit from a subplot showing phase arrival times versus source-station distance of the training and testing dataset, probably in the supplementary material.

Afterall, please improve the usages of the inline citation throughout the manuscript, and especially in the introduction section.

Reference

Shi, Q., & Denolle, M. A. (2023). Improved observations of deep earthquake ruptures using machine learning. Journal of Geophysical Research: Solid Earth, 128, e2023JB027334.

Round 1

Reviewer D

For author and editor

Dear Authors

First and foremost I want to thank the authors for this interesting work, the curation of the CREW dataset and training of a new ML model beyond regional distances. I have read the manuscript "*P Picking Regional Seismic Phase Arrival Times with Deep Learning*" with great interest. The work will improve our observational capabilities, which will lead to more complete earthquake catalogs and ultimately contribute to the understanding of the Earth's dynamics. After carefully reviewing the work I think, that the work would benefit from moderate revisions.

I hope that these comments are helpful, and I look forward to seeing a revised version of the paper.

Please see my comments below.

Sincere regards

Review Seismica 1431

Picking Regional Seismic Phase Arrival Times with Deep Learning

November 12, 2024

Dear Authors

First and foremost I want to thank the authors for this interesting work, the curation of the CREW dataset and training of a new ML model beyond regional distances. I have read the manuscript "*P Picking Regional Seismic Phase Arrival Times with Deep Learning*" with great interest. The work will improve our observational capabilities, which will lead to more complete earthquake catalogs and ultimately contribute to the understanding of the Earth's dynamics. After carefully reviewing the work I think, that the work would benefit from moderate revisions.

I hope that these comments are helpful, and I look forward to seeing a revised version of the paper.

Please see my commented PDF and comments below.

Sincere regards

Comments

General

The synthetic noise is a good approach. However the generation of synthetic noise is not specified. The generation of meaningful noise and training data augmentation is key for a good generalization of the model. Maybe also real noise samples extracted from the most noisy station can be extracted and added to augment the input/training waveforms.

Why were triangular labels chosen? PhaseNet uses Gaussian labels. Please clarify.

Please compare SKYNET against stretched (rescaled) PhaseNet input. It would be interesting to see how this performs. Essentially put another figure

4 into the supplement with 2x stretched PhaseNet input. See Shi et al., 2024 (https://essopenarchive.org/users/551624/articles/740608-from-labquakes-to-megathrusts-scaling-deep-learning-based-pickers-over-15-orders-of-magnitude). Rescaling in code here: https://github.com/pyrocko/qseek/blob/dev/src/qseek/images/seisbench.py#L296

Now the repository at https://github.com/albertleonardo/skynet is a bit hidden. To increase the visibility and the benefit for the community it would be great to include the model into SeisBench. Which is the established library for seismic phase pickers. I am sure the maintainer would welcome the contribution to the open-source framework.

Code formatting standards are important for clear communication. Please format the listing in 8.1 according to PEP8 (https://peps.python.org/pep-0008/). Commonly *ruff* (https://github.com/astral-sh/ruff) is used for this automatic task. Please also format the code in the GitHub repository for better readability and review. This point may seem pedantic, is however crucial for sustainable software development and maintenance.

Please perform computational benchmark of SKYNET/PhaseNet to compare the throughput of the deeper and shallow model, and how the performance could be improved (bloat16 / quantization?). This is an important aspect for scalable analysis of large continuous waveform dataset (e.g. https://github.com/pyrocko/qseek)

Make it crystal clear in the beginning of the manuscript that two models are presented and how they differ.

Introduction

L28 small → local and regional distances

Fig 1: Add information about what type of instruments are included in the dataset (broadband / short-period?).

Fig 4 Change color or PhaseNet/SKYNET S arrival. In general the plot is too busy. Replace station names with meaningful distances, add vertical grid to guide the eye. Remove top and right spines. Move map. Add generic axis label.

Fig 7 Here picks are shown as lines. For consistency please show the raw annotations as in Figure 4 and 6. Mention the depth of the earthquake in the caption. Remove the figure title.

Fig 8 Choose better colors for the different arrivals. Remove redundant axis labels. Focus on the key information you want to communicate.

L50 *… earthquakes recorded from … → earthquakes recorded at distances*

L93 Be more specific what kind of synthentic noise was generate and added.

L115 Remove *clearest use case*, this is a judgement.

L136 What magnitudes? ML, MW? How were they estimated for the additionally detected events?

Minor Comments

L6 *Sparse instrumental coverage for much of the Earth requires working with regional* ***seismic phase arrivals*** *for effective seismic monitoring.*

L7 *Machine learning* ***seismic*** *phase pickers …*

L12 Wording, repeated *model*

L14 Remove ML abbreviation

L181 4 → four

Review of

Picking Regional Seismic Phase Arrival Times with Deep Learning

by *A. L. Aguilar* and *G. C. Beroza*

submitted to *Seismica*

Summary and General Comments

The authors present a new deep learning seismic phase picker to identify P- and S-wave arrivals at regional distances, that is, distances up to 20° for which widely used deep learning phase pickers (*e.g.*, Zhu and Beroza, 2019; Mousavi et al., 2020) do not perform well. This work will greatly help improve the quality of seismic monitoring in sparsely instrumented areas and thereby contribute to the better understanding of, for example, subduction zones and stable continental regions. Their application to the picking of secondary phases is also very promising. The manuscript is well organized and written, the figures are clear, I therefore only have minor comments and suggestions.

Comments

Time window probed by each feature of the latent space

At the beginning of section 3 (lines 70-75), the authors mention the dimensions of the feature space in the deepest layer: 32x30. My understanding is that "30" corresponds to the direction of the "transformed" time axis. So, does that mean that the kernel in the deepest layer probes 300/30=10 s of the original time series? And, ultimately, what is the time duration upon which every sample of the output channels is based? I don't have a good understanding of how the kernel size and stride value can answer my question (although I know they are the key to my question). The reason I'm asking is that this kind of information would be useful for people interested in using your model to analyze time series longer than 300 s without a windowing approach but, instead, taking advantage of the fact that the convolution architecture can slide through any duration.

Line by line comments

- Line 14: "ML" was not defined.
- Line 26: It looks like the Park and Schultz references should be in the same parenthesis.
- Line 54: Max or standard deviation normalization?
- Figure 2: Which earthquake does the label correspond to when several earthquakes are mixed together? The biggest one?
- 99: "the height of the peak of the predictions" doesn't read very well, and it's used again at line 125. Why not talk about probability values instead?
- Line 108: This is an important observation. If the residuals are not gaussian, then using the SKYNET picks to locate earthquakes with least-squares optimization wouldn't make much sense (although the SKYNET picks will most likely be used like that). Could you comment on that? L1-norm optimization assumes errors are distributed according to the Laplace distribution.
- Figure 4 caption and line 170: Figure references are broken.
- Line 159: There's an extra "arrival".
- Line 164: Should be "This type" or "These types".

References

S. M. Mousavi, W. L. Ellsworth, W. Zhu, L. Y. Chuang, and G. C. Beroza. Earthquake transformer—an attentive deep-learning model for simultaneous earthquake detection and phase picking. *Nature communications*, 11(1):3952, 2020.

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

C-Review

Review of Seismica manuscript “Picking Regional Seismic Phase Arrival Times with Deep Learning” by Aguilar and Beroza.

Thank you for sharing this submission with me. In the manuscript, Aguilar and Beroza developed a Deep Learning model, with a focus of performing earthquake phase picking at distances up to 20 degrees. The model features longer window length input and can pick and classify secondary phases (Pn, Pg, Sn, Sg). The manuscript is likely to be a valuable contribution to the earthquake monitoring community, a good start points of secondary phase picker, and a great showcase of CREW dataset’s usefulness. Below are several comments and quesHons that the authors would need to address, which would improve the manuscript.
Thank you for taking the time to review our work

L76: For a direct comparison, it is suggested to report the number of parameters of PhaseNet (hereafter referred as PN).
Noted and done, lines 79-80 now contain the number of parameters in both models, 38,734 for PN and 79,442 for our SKYNET model.

Five extra convolution layers effectively produces a longer receptive field. A similar strategy was also used in Shi and Denolle (2023) to extend the time window. Can authors provide more details on how the choice was quantitatively made? It will be beneficial for future studies on other type of phases, where the length of time windows may vary, and window adaptation is necessary.
This point goes back to the development of the Curated Regional Earthquake Waveforms (CREW) dataset, which we used for training our models. For earthquakes recorded at 20 degrees of source to receiver distance, we found the S minus P time to be around 200 seconds. Thus, we chose a 300 second window that would capture enough context before the P arrival and after the S arrival for effective processing.

Given that SKYNET has some similar structures with PN in the deeper layers, a common question to ask is, would the transfer learning of PN be applied here, that

takes the advantage of pre- trained PN model and improve the performance.
The models presented are very similar to PhaseNet indeed. It is possible to do transfer learning, but due to the relatively small number of trainable parameters in our new models, we chose to train from scratch.

L93: consider providing more details about the noise generator used here, and the statistical features of the noise. Why did the authors introduce this specific type of noise data, as in Figure S1. Furthermore, many datasets contain noise samples. Why are these noise samples are not used here?

The datasets that contain noise samples are made of waveforms shorter than 300 seconds. Procuring 5 minute long waveforms free of uncataloged earthquakes is a difficult task, for simplicity we opted to generate noise examples instead. This is not optimal, but the results of the model in finding new earthquakes are testament to its performance regardless of the choice of the noise used.

Section 2 of the supplementary material provides more information on noise generation, which is at its core Gaussian noise with added complexities. We also expanded Figure S2 (previously S1), which now displays 6 noise examples instead of 2 as in the previous version.

L145: as a convolution model, the input size of PN is not fixed. One can train PN with CREW dataset but with 300 sec inputs. I am curious how the author would justify that the PN's receptive field is narrower than SKYNET, and that the broadening of the field makes SKYNET superior over PN.

We estimated the receptive field size of PN and SKYNET to be ~10K and ~40K respectively (see response to E-review for details). This confirms your statement above that "Five extra convolution layers effectively produces a longer receptive field".

In theory the input size is not fixed, but in practice all code implementations create computation graphs that need to declare the sizes of the tensors. Our deeper architecture has a wider receptive field (as mentioned above, 40K vs 10K sample points) that allows information to be shared over long times in the waveforms, aiming to capture long S minus P times, expected from earthquakes recorded at several hundred and over one thousand kilometers away.

We do not claim that SKYNET is superior to PhaseNet, they are different models trained on different data with different objectives, and the difference in performance is shown in supplemental Figure 4, that shows the realms in which each model has been trained to work on.

It seems to me that Section 6 lacks some details and need more justification. Please see comments below.

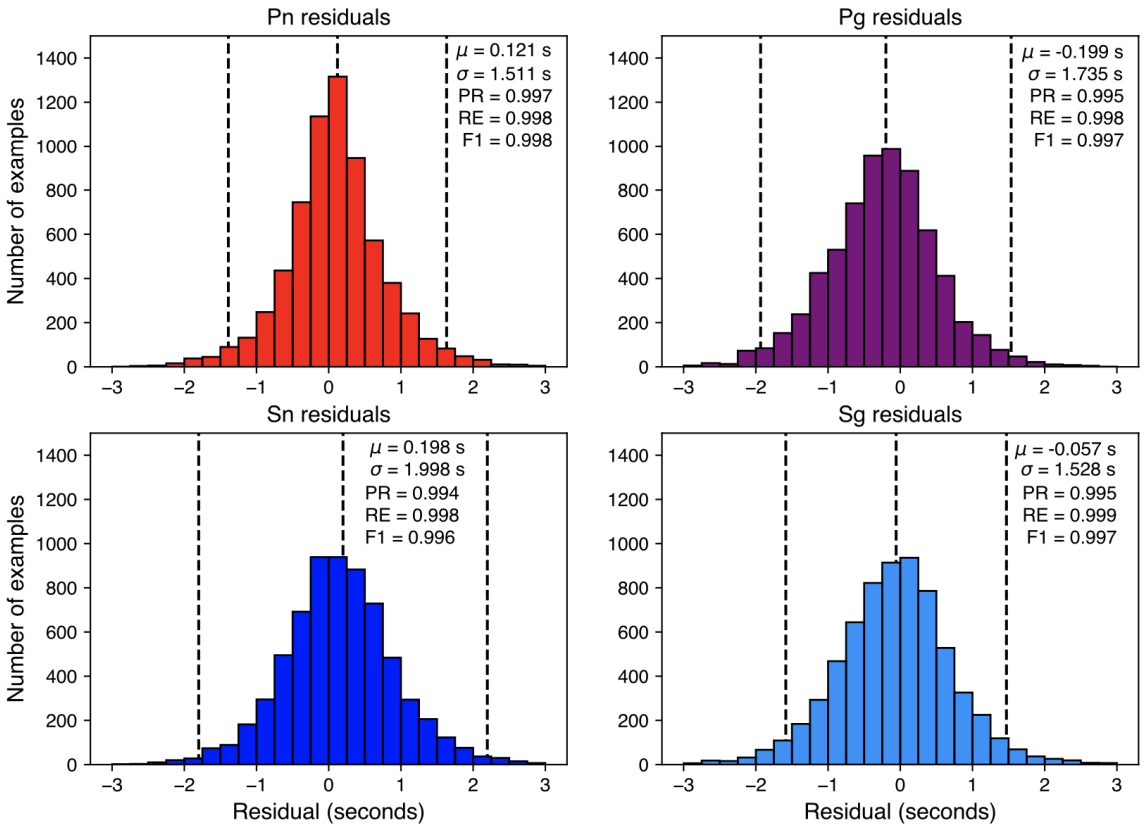
- The training process of the multi-phase picker is a bit confusing for me. Did the author train the model with only 6k samples? Is any data augmentation implemented here?
Yes, that one particular model was trained with 6700 examples as stated, no data augmentation, due to most signals occupying one half or more of the 300 seconds waveforms.

- And again, did the author implement transfer learning here? The SKYNET trained on the whole CREW dataset can be used and fine-tuned with 6k samples.

> No transfer learning here. We experimented with it, but the model that performed the best for the task of picking the multiple phases was the one trained from scratch, with only the 6700 examples with complete labels and synthetic noise.

- The authors should provide more statistics or figures on this small dataset, which are used to train the multi-phase picker. It may be helpful identifying potential overfitting.

We kept this section brief due to the limited amount of data, and the limitations that might stem from the restricted volume of data. Figure S9 (shown below) displays the residuals for each one of the four phases, along with evaluation metrics for the mean residual, the standard deviation, precision, recall and F1-score. These residual distributions are wider than those for the picker shown in Figure 3, as the standard deviations here are a factor of ~2 or ~3 larger. We argue this is evidence that the model is not overfitting.



- L169: “we trained during 12000 stages of samples of 200 randomly chosen examples” is a bit confusing to me. Please explain what “stage” indicates here, and where these 200 samples are chosen from.
It was mentioned that the limited training dataset is 6,700 examples. From the pool of 6,700, 200 randomly chosen are passed in batches of 32, with 8 synthetic noise samples. Stage means there is a forward propagation, an estimation of the loss, a backprop and a model update. It is slightly different to an ‘epoch’, but it represents the same process.

- L177: Only 3.1k out of 22k examples are kept. I am curious of any sign of

overfitting here, as the false positive rate is very high. The authors should better justify the effectiveness of this multi-phase picker, probably experiment with another major events with clear primary and secondary phases and illustrate their results like Figure 3/6/7.
Newly added Figure S9 (see above) shows the residuals for the model predictions on the training dataset. The average error is less than 0.2 seconds, and the standard deviations less than 2 seconds. These distributions are wider than those in Figure 3, which we argue is evidence against over fitting.
Due to the rarity of waveforms with four clear phases we restricted our analysis to an expansion of the labels in the dataset instead of applying to data outside of the CREW dataset. This is a useful model, trained on 6700 examples, used to run predictions on 1.6 million examples, which flagged 22K candidates, which is only 1.4% of the dataset.
Once we had this culled dataset of 22K examples, we visually reviewed these candidates to check two details; first, the presence of clear arrivals for the four phases and second, the accuracy of the model picked times. Figure S10 shows examples that were flagged by the model, but did not pass the visual inspection because either one of the phases is absent or it is inaccurately picked.

I tried to run the code in Listing 1 but failed. The functions here is not synced with the script in your github repository. Please update your scripts (or this manuscript), with appropriate documentations.

The repository has been updated, including a tutorial notebook.

Miscellaneous

Abstract: please be accurate on the use of plural, i.e., models.
We mean we not only provide one model, as mentioned in the model selection section there are different models, with variations on the labels used.

Figure 2: consider add some examples from short-period stations.
There are examples from HH, BH, HN, which are the most common in the dataset. The proportions of the type of instruments are now displayed in Figure S1, where SH and EH instruments are a small fraction. Also, the augmentation examples mix different instrument types, so it is likely that some of the waveforms for which no metadata is displayed correspond to SH instruments.

Figure 4: consider only show the one event that the waveform corresponds to, not all from the swarm, on the map.

We updated the inset map to indicate the USGS catalog displayed in Figure 5, as well as the newly found earthquakes. This way, we show that the newly found events are in the vicinity of the known swarm. The caption has been updated in Figure 4 and 5, besides being commented on in the main text.

Two missing figure labels: Figure 4 caption and L170.

Good eye! It has now been fixed such that the caption of figure 4 references supplementary figure S3. And line 170 references figure 8 (now line 174).

Figure 6: there are several stations shown on the map but has no waveform or picks on the leS. The authors may improve this figure by removing stations that are not used or including these waveforms if they can be picked.
The stations displayed are the CX network, for which we plot all stations. However, HH waveforms are not available at all stations at the given times in Figures 6 and 7. This is the reason why both Figures 6 and 7 display a different number of waveforms. In Figure 6, the stations for which no name is shown are the ones for which no waveforms are available.

The manuscript may benefit from a subplot showing phase arrival times versus source-station distance of the training and testing dataset, probably in the supplementary material.

Newly added supplementary figure S4 displays the residuals for the test set as functions of source-station distance and signal to noise ratio (SNR), separately for P and S arrivals. No apparent increase of the magnitude of residuals with increasing distance or decreasing SNR is observed. The distributions have more scatter at the shorter distances and lower SNR, due to the increase in frequency of examples at lower distances and lower SNR.

Overall, please improve the usages of the inline citation throughout the manuscript, and especially in the introduction section.
Thanks for pointing this out, now all citations are in parenthesis.

Reference

Shi, Q., & Denolle, M. A. (2023). Improved observations of deep earthquake ruptures using machine learning. Journal of Geophysical Research: Solid Earth, 128, e2023JB027334.

D-Review
Review Seismica 1431
Picking Regional Seismic Phase Arrival Times with Deep Learning
November 12, 2024
Dear Authors
First and foremost I want to thank the authors for this interesting work, the curation of the CREW dataset and training of a new ML model beyond regional distances. I have read the manuscript “P Picking Regional Seismic Phase Arrival Times with Deep Learning” with great interest. The work will improve our observational capabilities, which will lead to more complete earthquake catalogs and ultimately contribute to the understanding of the Earth’s dynamics. After carefully reviewing the work I think, that the work would benefit from moderate revisions.
I hope that these comments are helpful, and I look forward to seeing a revised version of the paper.
Please see my commented PDF and comments below.
Sincere regards

Thanks for the thoughtful comments and taking the time to review our work

Comments
General
The synthetic noise is a good approach. However the generation of synthetic noise is not specified. The generation of meaningful noise and training data augmentation is key for a good generalization of the model. Maybe also real noise samples extracted from the most noisy station can be extracted and added to augment the input/training waveforms.

Noise examples are Gaussian noise with some modifications, like adding spikes and offsets. Section 2 of the supplementary material provides more information on noise generation, as well as an expanded figure now displaying 6 noise examples instead of 2 as in the previous version.

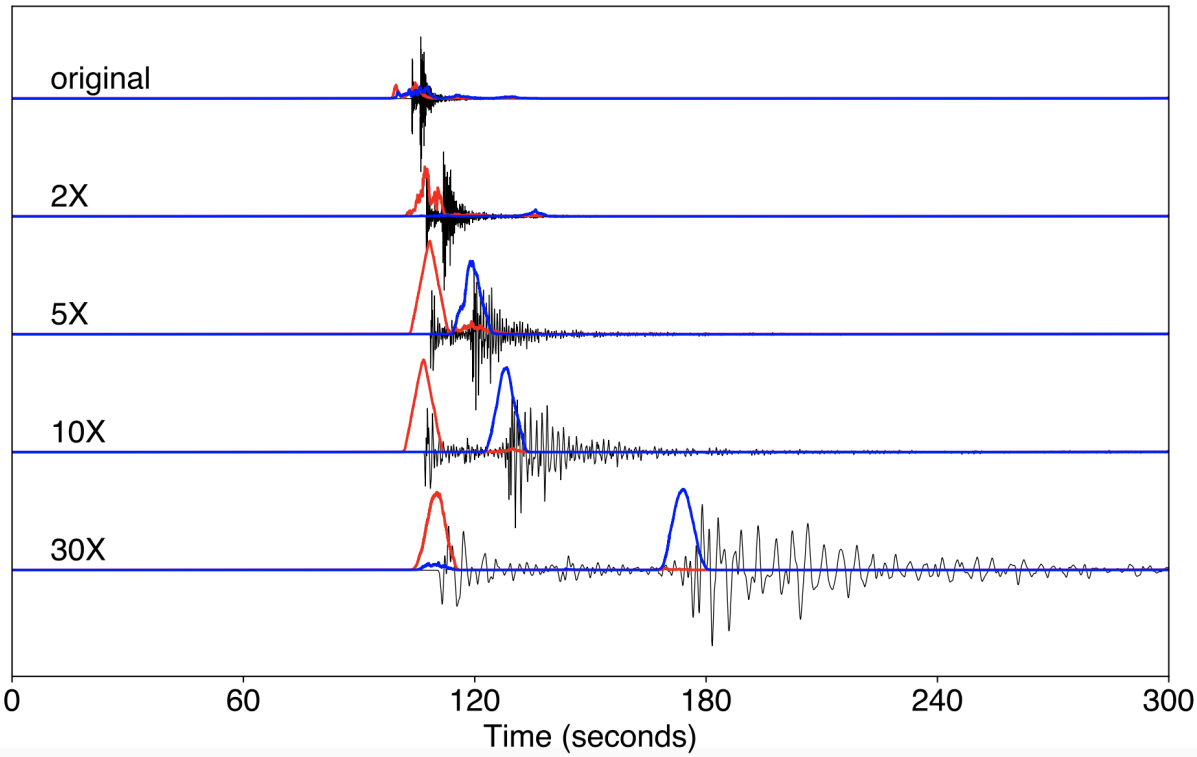
The datasets that contain noise samples are made of waveforms shorter than 300 seconds. Procuring 5 minute long waveforms free of uncataloged earthquakes is a difficult task, for simplicity we opted to generate noise examples instead. This is not optimal, but the results of the model in finding new earthquakes are testament to its performance regardless of the choice of the noise used.

Why were triangular labels chosen? PhaseNet uses Gaussian labels. Please clarify.

When working with our much longer waveforms and much longer labels, the Gaussian becomes too flat and wide on top at the center. The sharper shape of the triangle gives more localized activations that result in more stable picks. For context, EarthquakeTransformer uses triangular shapes too. The choice of label is a heuristic, and for the reasons above we chose to use triangular labels.

Please compare SKYNET against stretched (rescaled) PhaseNet input. It would be interesting to see how this performs. Essentially put another figure 14 into the supplement with 2x stretched PhaseNet input. See Shi et al., 2024 (https://essopenarchive.org/users/551624/articles/740608-from-labquakes-to-megathrusts-scaling-deep-learning-based-pickers-over-15-orders-of-magnitude). Rescaling in code here: https://github.com/pyrocko/qseek/blob/dev/src/qseek/images/seisbench.py#L296

Could not tell which Figure 14 the reviewer is referring to. We run predictions on an earthquake waveform from a near source recording (S minus P time around 3 seconds, same waveform as the top one in Figure S6) stretched by different factors . The predictions from skynet are bad for both the original and twice stretched input. The predictions have a triangular shape for data stretched between 5x and 30x, which would resemble the training data from the CREW dataset.



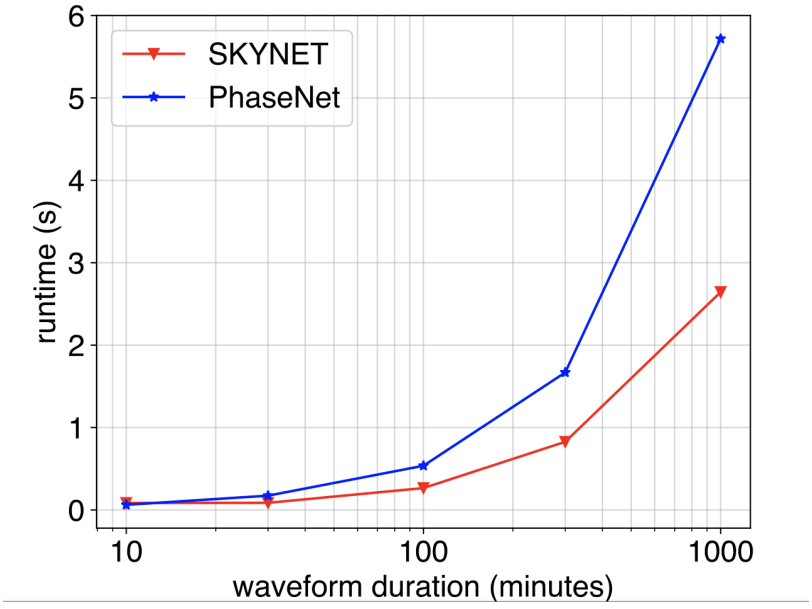
Now the repository at https://github.com/albertleonardo/skynet is a bit hidden. To increase the visibility and the benefit for the community it would be great to include the model into SeisBench. Which is the established library for seismic phase pickers. I am sure the maintainer would welcome the contribution to the open-source framework.
We are in touch with the Seisbench team to integrate our models. However, the policy is that models and datasets are only integrated once the respective papers have been published. Once the manuscript is published, we will move forward with making the models accessible through Seisbench.

Code formatting standards are important for clear communication. Please format the listing in 8.1 according to PEP8 (https://peps.python.org/pep-0008/). Commonly ruff (https://github.com/astral-sh/ruff) is used for this automatic task. Please also format the code in the GitHub repository for better readability and review. This point may seem pedantic, is however crucial for sustainable software development and maintenance.

We are actively working on improving the repository. We have added a tutorial notebook and will continue with the advice of adjusting the formatting and making the package available across platforms.

Please perform computational benchmark of SKYNET/PhaseNet to compare the throughput of the deeper and shallow model, and how the performance could be improved (bloat16 / quantization?). This is an important aspect for scalable analysis of large continuous waveform dataset (e.g. https://github.com/pyrocko/qseek)

Supplementary Figure S7 shows the runtimes on streams of length 10, 30, 100, 300, and 1000 minutes. The last one is close to one day of continuous data. Overall, on the longer streams SKYNET does the prediction in about half the time it takes PhaseNet model. On the shorter end, PhaseNet is faster, but on the longer runs SKYNET is faster. This might be due to the lower number of sliding windows required for SKYNET to complete the computation.
This comparison is between seisbench's -annotate- and skynet's execute function. This is the time required only to compute the forward propagation, it does not include the time it takes to extract phase picks (equivalent to seisbench's -classsify-). These were runtimes with default parameters and averaged over five repetitions of the computations . We don't make any claim of superiority speed wise, as changing the overlap fraction would change these numbers.



Make it crystal clear in the beginning of the manuscript that two models are

presented and how they differ.
The abstract says one model that picks first arrivals and one that can pick first and secondary arrivals

Introduction
L28 small → local and regional distances
Datasets that contain both local (i.e. <1 degree source-receiver distance) and also regional data are dominated by the local recordings, with more than 90% of the labels corresponding to short local distances. This applies to STEAD, INSTANCE, MLAAPDE and NEIC datasets. The wording is meant to convey this.

Fig 1: Add information about what type of instruments are included in the dataset (broadband / short-period?).
Lines 52-53 mention that there is variety of instruments in the dataset, “These waveforms come from a variety of instruments, including high gain seismometers, short period seismometers, and accelerometers”
We have added Figure S1, which displays the fraction of data that comes from these different types of instruments.

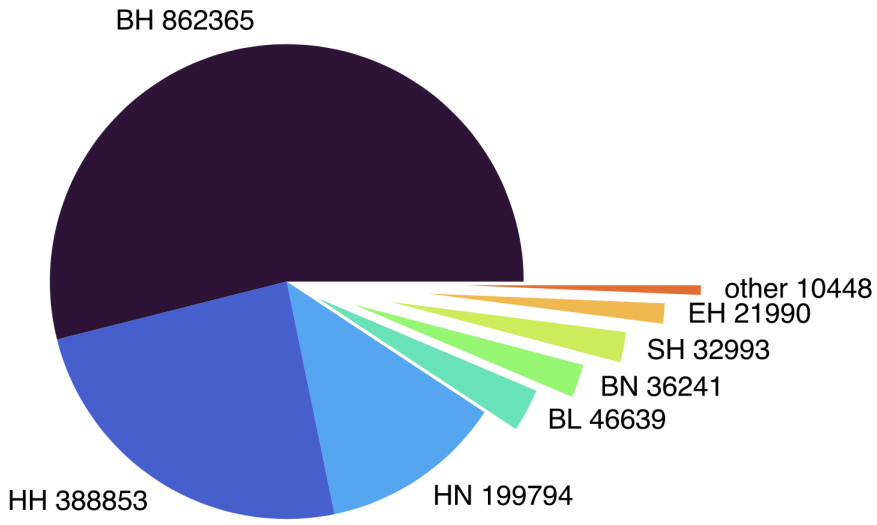


Fig 4 Change color or PhaseNet/SKYNET S arrival. In general the plot is too busy. Replace station names with meaningful distances, add vertical grid to guide the eye. Remove top and right spines. Move map. Add generic axis Label.
The dotted horizontal lines that show the 0.5 threshold are meant to guide the eye vertically. Our overall style is somewhat grid free, which we would like to maintain. We

added the corresponding source to receiver distances for each station. Also, we opted to keep the station names as they are key to identify the stations.

2Fig 7 Here picks are shown as lines. For consistency please show the raw annotations as in Figure 4 and 6. Mention the depth of the earthquake in the caption. Remove the figure title.
Figure 7 has been modified, to add the model predictions on top of the waveforms in a similar manner to the previous figures. The depth has been added to the caption, but we prefer to keep all the origin information in the title, given the depth of the earthquake is central to the message.

Fig 8 Choose better colors for the different arrivals. Remove redundant axis labels. Focus on the key information you want to communicate.
We explored a few color options and put a lot of effort into the choices that lead to these colors, we would like to maintain them. The use of red and blue for Pn and Sn, which are the first P and S arrivals is consistent with all previous figures. While the use of light blue and purple maintain some color similarity for Pn-Pg and Sn-Sg. This is a color scheme corresponding to Figure S9 which displays the residuals for each phase.

L50 … earthquakes recorded from … → earthquakes recorded at distances
Good idea.
Line now reads: “The CREW dataset consists of 5-minute three-component waveforms from earthquakes recorded at distances between 1 and 20 degrees of source-receiver separation”

L93 Be more specific what kind of synthentic noise was generate and added.
Common point with C review.
We have supplemented Figure S1 (now Figure S2) with more synthetic noise examples and more details about their generation in section 2 of the supplement.

L115 Remove clearest use case, this is a judgement.
Reworded line 116 and now reads: ‘A more relevant test for our model’

L136 What magnitudes? ML, MW? How were they estimated for the additionally detected events?

We estimated local magnitudes ML. We removed the instrument response and then simulated a Wood Anderson response. Then, we used the local magnitude equation and calibrated the magnitude scale for the events in the USGS catalog resulting in:

ML = log10(amplitude*1000) + 2.76*log10(distance) -1.48

Which is the equation we used to estimate the ML of the newly found events.

Minor Comments

L6 Sparse instrumental coverage for much of the Earth requires working with regional seismic phase arrivals for effective seismic monitoring.

This line has been reworded to: Sparse instrumental coverage for much of the Earth requires working with regional seismic phases for effective seismic monitoring

L7 Machine learning seismic phase pickers …

L12 Wording, repeated model

Yes, that is intended, as one refers to the first arrival picking model and the other one to the multiphase arrival picking model.

L14 Remove ML abbreviation

Noted and corrected, it now reads machine learning

L181 4 → four

Noted and corrected.

E-Review

Review of Picking Regional Seismic Phase Arrival Times with Deep Learning

by A. L. Aguilar and G. C. Beroza submitted to Seismica

Summary and General Comments

The authors present a new deep learning seismic phase picker to identify P- and S-wave arrivals at regional distances, that is, distances up to 20◦ for which widely used deep learning phase pickers (e.g., Zhu and Beroza, 2019; Mousavi et al., 2020) do not perform well. This work will greatly help improve the quality of seismic monitoring in sparsely instrumented areas and thereby contribute to the better understanding of, for example, subduction zones and stable continental regions. Their application to the picking of secondary phases is also very promising. The manuscript is well organized and written, the figures are clear, I therefore only have minor comments and suggestions.

Thank you for taking the time to review our work.

Comments

Time window probed by each feature of the latent space

At the beginning of section 3 (lines 70-75), the authors mention the dimensions of the feature space in the deepest layer: 32x30. My understanding is that "30" corresponds to the direction of the "transformed" time axis. That is accurate, but the transformed time is a very encoded version of time .So, does that mean that the kernel in the deepest layer probes 300/30=10 s of the original time series? And, ultimately, what is the time duration upon which every sample of the output channels is based? I don't have a good understanding of how the kernel size and stride value can answer my question (although I know they are the key to my question). The reason I'm asking is that this kind of information would be useful for people interested in using your model to analyze time series longer than 300 s without a windowing approach but, instead, taking advantage of the fact that the convolution architecture can slide through any duration.

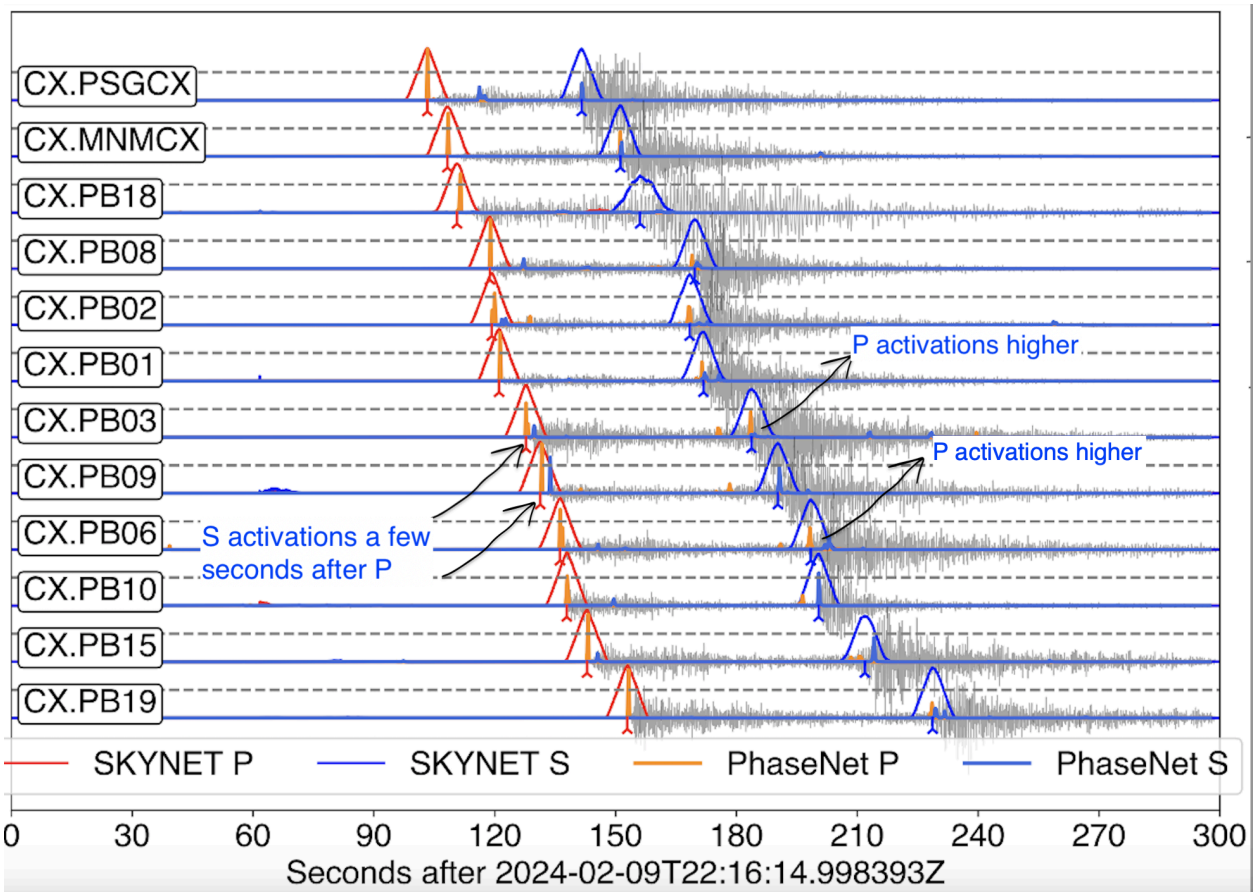
The receptive field of a series of **L** convolution layers can be estimated as:

$$rf = \sum_{l=1}^{L}((k_l - 1)\prod_{i=1}^{l-1} s_i)+1$$

Where layer *l* has kernel size *kl* and stride *s*. In both the original PhaseNet and our version, the kernel size is 7 for all layers, and the stride is 4 and 1 for consecutive convolution layers. Computing this for the 9 and 11 layers that lead to the deepest encoding, PhaseNet has a rf=10,206, whereas SKYNET has rf=40,926. This means that every sample in the deepest layer has access to all of the input.

On the other hand, the question of what is the time duration upon which every sample of the output channels is based is a much harder one to answer, it would require computing saliency maps and still would not answer the specific question, only display which part of the input has more consequence on the output, but not on a sample by sample basis.

To see why the longer waveforms and thus longer context is useful I refer to Figure 6. For stations PB08,PB02,PB01,PB03 PhaseNet produces higher P activations (yellow) than S activations at the times of the actual S arrivals. Also, for stations PB08,PB02, PB03, PB09 there are S activations a few seconds after the inferred P pick, which shows that PhaseNet has a tendency to pick S wave arrivals a few seconds after P arrivals, which is the type of data it was originally trained on. A sliding window of 30 seconds implies that the predictions from PhaseNet around the P arrivals are independent from the predictions around the S arrival, because the S minus P times here are close to 60 seconds.



Line by line comments

- Line 14: "ML" was not defined.
Good catch, noted and corrected, it now reads ‘machine learning’.

- Line 26: It looks like the Park and Schultz references should be in the same parenthesis.
Noted and corrected, alongside many other citations that were missing parenthesis.

- Line 54: Max or standard deviation normalization?

The waveforms have been normalized by the maximum amplitude among the three components but are otherwise in their raw form, now specified in lines 55-57.

- Figure 2: Which earthquake does the label correspond to when several earthquakes are mixed together? The biggest one?
When creating the mixed examples, one random example is chosen as a ‘pivot’, and the other signals are added around this one, estimating delays such that they will not overlap. In the plots, the metadata panel corresponds to the ‘pivot’ earthquake, for which the arrivals are marked by the letters, for the added signals, the arrival times are displayed, but no phase label is marked with letters. The choice of the pivot is not related to magnitude or location or any other parameter. For panels f and h, where long distance signals are mixed with shorter signals, the metadata corresponds to the long distance quake, because for this one augmentation recipe, the pivot is randomly chosen from the subset of available data at distances over 8 degrees.

- 99: "the height of the peak of the predictions" doesn’t read very well, and it’s used again at line 125. Why not talk about probability values instead?

The model prediction shapes are a heuristic and do not have intrinsic statistical value as a pdf along the time axis. This might be a confusion arising from the use of a truncated gaussian in the original PhaseNet. The probabilistic nature of the predictions is only on a point wise basis, due to the last layer being a softmax layer.
We have reworded it to read ‘the peak classification probabilities’

- Line 108: This is an important observation. If the residuals are not gaussian, then using the SKYNET picks to locate earthquakes with least-squares optimization wouldn’t make much sense (although the SKYNET picks will most likely be used like that). Could you comment on that? L1-norm optimization assumes errors are distributed according to the Laplace distribution.
We understand that L1 minimization is the maximum likelihood estimator for a Laplace distribution of errors; however, these residuals are not traveltimes. These residuals are contrasting the model predictions against the dataset labels, so they don’t have a travel time interpretation. It might be related to biases of analyst picking at regional distances, with more emergent arrivals compared to the short local distances with impulsive arrivals. It would be a separate task to associate-locate and estimate the traveltime residuals with respect to a velocity model. That is a separate task that we leave for future work.

- Figure 4 caption and line 170: Figure references are broken.

Good eye! It has now been fixed such that the caption of figure 4 references supplementary figure S3. And line 170 references figure 8 (now line 174).

- Line 159: There’s an extra "arrival".

Noted and corrected

- Line 164: Should be "This type" or "These types".

Now reads ‘one of these examples’

Round 2

Reviewer C

For author and editor

Thank you for your careful revision. I have reviewed your revised manuscript, and my comments and questions have been addressed. I recommend accepting this manuscript in its present form.

Reviewer D

Editor Note: The assignment for Reviewer D was canceled due to an extended delay in providing feedback.

Reviewer E

For author and editor

I thank the authors for taking my comments, as well as the other reviewers', into account in this revised manuscript. As I said in my initial review, I believe that the manuscript is ready for publication. During the review, I had time to test the SkyNet model on my own dataset and obtained very satisfactory results. This new model will greatly help in sparsely instrumented regions and its release is very timely.

A minor comment regarding their response to my comment about residuals. I do understand that the residuals are not travel time residuals but pick residuals. In a Bayesian framework, like that developed by Tarantola and Valette, the l2-norm minimization makes sense when the data (here, the picks), model parameters and theoretical errors are all gaussian. Their figure shows that the data distribution is better described by a Laplace rather than a gaussian distribution, suggesting that an l1-norm minimization would make more sense to handle the type of errors in SkyNet-generated picks. That was my comment.

I look forward to seeing progress in picking depth phases.