Earth and Space Science

RESEARCH ARTICLE
10.1029/2023EA003332

Special Section:
Advances in Machine Learning for Earth Science: Observation, Modeling, and Applications

Key Points:

- We assembled a database of ocean Bottom Seismometer (OBS) waveforms and manual P and S picks, on which we train PickBlue, a deep learning picker
- Our picker significantly outperforms pickers trained with land-based data with confidence values reflecting the likelihood of outlier picks
- The picker and database are available in the SeisBench platform, allowing easy and direct application to OBS traces and hydrophone records

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:
D. Lange,
dlange@geomar.de

Citation:
Bornstein, T., Lange, D., Münchmeyer, J., Woollam, J., Rietbrock, A., Barcheck, G., et al. (2024). PickBlue: Seismic phase picking for ocean bottom seismometers with deep learning. *Earth and Space Science*, *11*, e2023EA003332. https://doi.org/10.1029/2023EA003332

Received 3 OCT 2023
Accepted 21 NOV 2023

Author Contributions:
Conceptualization: D. Lange, J. Münchmeyer, J. Woollam, A. Rietbrock, F. Tilmann
Data curation: D. Lange, G. Barcheck, I. Grevemeyer

PickBlue: Seismic Phase Picking for Ocean Bottom Seismometers With Deep Learning

T. Bornstein[1,2,3], D. Lange[1], J. Münchmeyer[2,4], J. Woollam[5], A. Rietbrock[5], G. Barcheck[6], I. Grevemeyer[1], and F. Tilmann[2,7]

[1]GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany, [2]GFZ, German Research Centre for Geosciences, Potsdam, Germany, [3]Now at Gempa GmbH, Potsdam, Germany, [4]University Grenoble Alpes, University Savoie Mont Blanc, CNRS, IRD, University Gustave Eiffel, ISTerre, Grenoble, France, [5]Karlsruhe Institute of Technology, Karlsruhe, Germany, [6]Department of Earth and Atmospheric Sciences, Cornell University, Ithaca, NY, USA, [7]Institute for Geological Sciences, Freie Universität Berlin, Berlin, Germany

Abstract Detecting phase arrivals and pinpointing the arrival times of seismic phases in seismograms is crucial for many seismological analysis workflows. For land station data, machine learning methods have already found widespread adoption. However, deep learning approaches are not yet commonly applied to ocean bottom data due to a lack of appropriate training data and models. Here, we compiled an extensive and labeled ocean bottom seismometer (OBS) data set from 15 deployments in different tectonic settings, comprising ~90,000 P and ~63,000 S manual picks from 13,190 events and 355 stations. We propose PickBlue, an adaptation of the two popular deep learning networks EQTransformer and PhaseNet. PickBlue joint processes three seismometer recordings in conjunction with a hydrophone component and is trained with the waveforms in the new database. The performance is enhanced by employing transfer learning, where initial weights are derived from models trained with land earthquake data. PickBlue significantly outperforms neural networks trained with land stations and models trained without hydrophone data. The model achieves a mean absolute deviation of 0.05 s for *P*-waves and 0.12 s for *S*-waves, and we apply the picker on the Hikurangi Ocean Bottom Tremor and Slow Slip OBS deployment offshore New Zealand. We integrate our data set and trained models into SeisBench to enable an easy and direct application in future deployments.

Plain Language Summary Ocean bottom seismometers (OBS) are seismic stations on the seafloor. Just like their counterparts on land, they record many earthquakes on three component sensors but are additionally equipped with underwater hydrophones. To determine the location of an earthquake, seismologists must precisely measure the arrival times of seismic waves. For onshore data, machine learning (ML) has been highly successful in determining earthquake arrival times. However, the noise and the signal are different in the ocean environment. For example, the recordings can contain whale songs and water layer reverberations and are disturbed by ocean bottom currents. We have assembled an extensive database of ocean bottom recordings and trained artificial neural networks to use the underwater hydrophone information and cope with the ocean noise environment. We demonstrate that the resulting ML picker picks are similar to those of human experts and outperform phase pickers based on land data only. We compare earthquake catalogs based on different pickers created from an OBS deployment offshore New Zealand and demonstrate that PICKBLUE outperforms previous pickers. We make the database and ML picker available with a standard interface so that it is easy for other scientists to apply them in their studies.

1. Introduction

Determining the arrival times of seismic phases in seismograms is crucial for detecting and locating earthquakes. Accurate and precise onset times of P and S phases are a precondition for many seismological applications such as analysis of seismicity distribution and travel time tomography. Consequently, automated onset picking of earthquake arrivals has been an active field of research for several decades (Allen, 1982; Diehl, Kissling, et al., 2009; Küperkoch et al., 2010; Leonard & Kennett, 1999; Sleeman & Van Eck, 1999). Until the recent advance of machine learning (ML) pickers, most phase picking algorithms were picking P arrivals (e.g., Baer & Kradolfer, 1987; Lomax et al., 2012; Sleeman & Van Eck, 1999), while S phase picking algorithms were less common (Diehl, Deichmann, et al., 2009; Sleeman & van Eck, 2003) due to the increased complexity of phase determination in the coda of the *P*-wave, and the generally lower frequency content of the *S*-wave. They often

Formal analysis: T. Bornstein, D. Lange, J. Münchmeyer, F. Tilmann
Funding acquisition: D. Lange, A. Rietbrock, F. Tilmann
Investigation: T. Bornstein, D. Lange, I. Grevemeyer
Methodology: T. Bornstein, D. Lange, J. Münchmeyer, A. Rietbrock, F. Tilmann
Project Administration: D. Lange, A. Rietbrock, F. Tilmann
Resources: D. Lange, G. Barcheck, I. Grevemeyer, F. Tilmann
Software: T. Bornstein, J. Münchmeyer, J. Woollam
Supervision: D. Lange, F. Tilmann
Validation: T. Bornstein, J. Woollam
Visualization: T. Bornstein, D. Lange
Writing – original draft: T. Bornstein, D. Lange
Writing – review & editing: T. Bornstein, D. Lange, J. Münchmeyer, J. Woollam, A. Rietbrock, G. Barcheck, I. Grevemeyer, F. Tilmann

require a prior estimate of the hypocenter to rotate horizontal components and to identify the approximate time window where S is expected. Although ML methods have been occasionally used for decades in seismology (e.g., Dai & MacBeth, 1995), only in recent years the training of more complex deep learning models became feasible due to computational and technological breakthroughs, with a performance now at least matching that of human analysts for local seismicity (e.g., Mousavi et al., 2020; Ross et al., 2018; Zhu & Beroza, 2019), see Mousavi and Beroza (2023) for a review. Application of ML algorithms for event detection and phase picking, when coupled with automatic association algorithms has allowed to increase the size of catalogs by up to an order of magnitude (e.g., Cianetti et al., 2021; Liu et al., 2020).

Most classical automated phase picking algorithms are optimized for land stations, and the same is true for ML pickers. However, as many tectonically interesting regions are submarine, accurate phase picking is required for ocean bottom seismometers (OBS) as well. Unfortunately, picking phase onsets of OBS waveforms is more difficult: confounding factors, such as water column reverberations, distant seismic or volcanic signals, sounds of marine mammals and anthropogenic noise from shipping, generally lead to lower quality phase arrivals than on land. In addition, the free-fall deployment procedure leads to tilt noise and unknown orientations of the horizontal channels (Crawford et al., 1998). This makes polarization techniques, commonly employed for S-onset determination (Diehl, Deichmann, et al., 2009), more difficult to apply. On the upside, OBS data often include an additional hydrophone channel not available at onshore seismometers. The hydrophone channel, in combination with the vertical channel, can help to distinguish arrivals traveling through the water column (e.g., water multiples or direct waves from marine mammal soundings) from those traveling through the solid earth.

Most of the time, algorithms developed for land stations are applied to marine seismological data with little modification, thereby foregoing any benefit from hydrophone data. This was true for conventional phase picking algorithms (e.g., Kuna et al., 2019; Lieser et al., 2014) some years ago and, so far, remains true for applications of ML pickers to ocean bottom data. Wu et al. (2022) proposed a workflow for building a high-resolution local submarine earthquake catalog using EQTransformer, the Siamese EQTransformer (Xiao et al., 2021) and PickNet (Wang et al., 2019) for detection and picking. Gong et al. (2022) and Gong and Fan (2022) applied EQTransformer to OBS records at the Quebrada transform fault system with generally good performance, but the picker missed large magnitude events with very emergent *P*-wave arrivals. Ruppert et al. (2022) applied EQTransformer with the original land-station trained weights to the full amphibious AACSE data set, pointing out that the original EQTransformer does not necessarily generalize well to the AACSE OBS. To our knowledge, all of the approaches so far proposed for arrival time picking on OBS disregard hydrophone data.

Here, we train and test different variants of the PickBlue phase picker targeting OBS instruments using deep learning. Our models measure the arrival times of distinct seismic phases (*P*- and/or *S*-waves) with a focus on temporal resolution and precision, minimizing false positive and false negative rates. Following common usage in the ML community, we refer to these as phase picking models. Our models, trained directly on OBS data, are able to learn the characteristics of OBS three-component data and also factor in the hydrophone information.

We base our picker on two recent neural network architectures designed for picking, EQTransformer (Mousavi et al., 2020) and PhaseNet (Zhu & Beroza, 2019). We extend both network designs by adding an additional hydrophone component as a fourth input channel, using the implementation integrated within SeisBench (Woollam et al., 2022), an open-source toolbox and model repository for ML in seismology. To train the OBS picking models, we compiled an extensive database of annotated local event waveforms recorded with free-fall OBS, including magnitudes, locations, and manually picked arrival times. We demonstrate the superior performance of these picking models compared to the equivalent pickers trained without hydrophone traces. Finally, we integrated our models and the trained model weights into SeisBench to enable the straightforward application to new data.

2. Data

We compiled an extensive database of waveforms from local earthquakes in various submarine tectonic environments, mostly four-component data, but also some OBH data, that is, with the hydrophone as the only channel. With each waveform, we provide manually labeled P and/or S phase picks and, for most deployments, station locations and estimated earthquake locations and magnitudes. Figure 1 shows the global distribution of the OBS deployments included and the recorded seismicity. Maps for each OBS network are shown in Figure S1

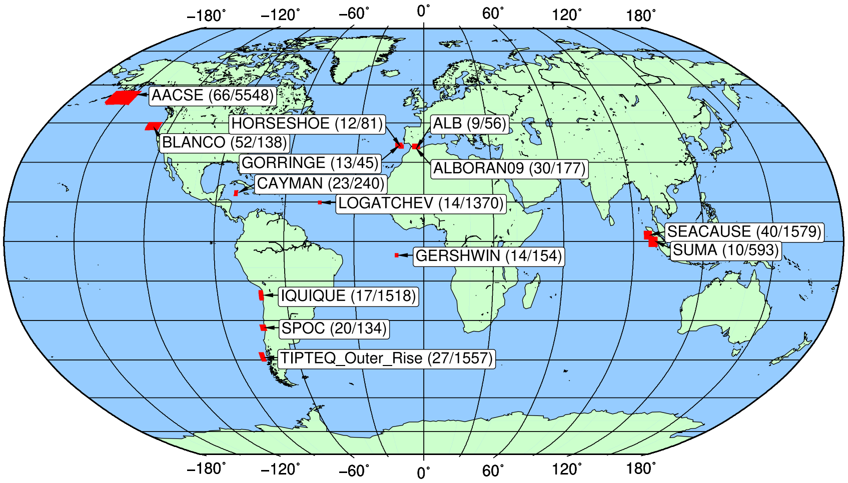


Figure 1. Global map showing the distribution of the ocean bottom seismometer (OBS) networks used for training the PickBlue picking algorithms. Labels indicate the data set name and numbers of OBS stations and events used. Red boxes encircle the OBS networks. ALA is not shown as the station set is identical to that of AACSE. Detailed maps of each deployment are available in Figure S1 in Supporting Information S1.

in Supporting Information S1. Table 1 provides additional information, including references for the contributed deployments.

The complete OBS data set contains manually picked phases from 15 deployments and a total of 355 stations (Table 1). The data set comprises 13,190 events, 109,210 traces and 153,338 picks (about 90,000 P and 63,000 S

Table 1
Table With Ocean Bottom Seismometer Deployments Used for Training the PickBlue Networks

Experiment[a]	Tectonic setting[b]	OBS type[c]	Start date	End date	#wave-forms	#P	#S	Reference
AACSE	S	BB	2018-05-12	2019-08-31	52,645	40,051	39,725	Ruppert et al. (2022),Barcheck (2023), XO (2018–2019)
ALA[d]	S	BB	2018-10-01	2019-02-09	315	302	93	Barcheck et al. (2020), XO (2018–2019)
ALB	I/CR	SP	2016-09-15	2016-12-03	274	201	192	I. Grevemeyer, pers. comm.
ALBORAN2009	I/CR	SP	2009-08-13	2010-01-16	2,252	1,622	1,900	Grevemeyer et al. (2015)
BLANCO	T	BB	2012-09-26	2013-09-23	2,882	2,850	961	Kuna et al. (2019), Nabelek and Braunmiller (2012), X9 (2012–2013)
CAYMAN	R	SP	2015-04-03	2015-04-16	2,302	1,582	1,665	Grevemeyer et al. (2019)
GERSHWIN	R	SP	2000-05-03	2000-05-12	834	811	28	Tilmann et al. (2004)
GORRINGE	I	SP	2013-10-11	2014-03-25	404	343	178	Grevemeyer et al. (2017)
HORSESHOE	I	SP	2012-04-15	2012-10-14	696	677	175	Grevemeyer et al. (2017)
IQUIQUE	S	SP	2014-12-09	2016-10-29	6,913	6,712	1,608	Petersen et al. (2021)
LOGATCHEV	R	SP	2009-01-18	2009-03-26	11,427	9,428	6,571	Grevemeyer et al. (2013)
SEACAUSE	S	SP	2005-10-16	2006-03-02	16,177	15,103	4,104	Tilmann et al. (2010)
SPOC	S	SP	2001-10-13	2001-12-01	1,339	1,332	62	Thierer et al. (2005)
SUMA	S	SP/BB	2008-06-06	2009-02-09	2,468	2,345	954	Lange et al. (2010)
TIPTEQ	OR	SP	2004-12-12	2005-01-27	8,280	6,734	5,025	Tilmann et al. (2008)
total			2000-05-03	2019-02-09	109,208	90,093	63,241	

[a]The OBS data sets are accessible through the SeisBench framework. The waveforms and metadata comprise ~35 GB. [b]I = Intraplate, R = Ridge, S = Subduction Zone, T = Transform Fault, and OR = Outer Rise. [c]BB = Broadband OBS and SP = Short Period OBS. [d]The ALA picks are based on the identical station set as the much larger AACSE data set. A small fraction of picks (272 picks, or 0.18% of the complete data set) were also independently picked for ALA, by a different human analyst. Since the overlap is very small, this will only have a negligible influence on the performance evaluation.

picks, implying that there are 44,128 traces with both P and S picks, ~46,000 with only P and 19,000 with only S picks). The data represent a variety of tectonic environments, seismometer and hydrophone types. The data include 38,419 4-component waveforms; 35,654 waveforms have only seismometer data (no hydrophone), and 8,187 traces are for hydrophone-only instruments (Table S1 in Supporting Information S1). For the remainder of the traces (16%), 1 or 2 seismometer components are missing due to equipment malfunction. Magnitudes range from 0.1 to 5.8 (Figure S2 in Supporting Information S1).

We split each deployment into a training set (65%), a development set (10%) and a holdout test set (25%). When we added the AACSE subset, we kept its pre-defined split ratio (70%/15%/15%). The effective split ratio for the whole data set was 66.8%/12.8%/20.4%. We randomly distributed entire events over the splits to ensure that all traces belonging to an event are part of the same split.

The OBS data is provided through SeisBench (Woollam et al., 2022) and can be used for other ML learning tasks. No explicit noise traces are included but noise samples can be generated by extracting the waveform ahead of the P arrival.

3. Methods

Several deep-learning models for seismic phase picking have recently been published. Generalized phase detection (GPD) (Ross et al., 2018) is a phase identification model based on a convolutional network and a point-wise fully connected network. GPD takes a short input window of 4 s at 100 Hz sampling rate as input and outputs for each window a prediction for P, S or noise. PhaseNet (Zhu & Beroza, 2019) is an arrival time picker based on a U-net (Ronneberger et al., 2015) architecture taking 30 s waveforms at 100 Hz as inputs and returning probability curves for P and S arrivals. BasicPhaseAE is another U-net based model for phase detection and onset picking, which takes 6 s waveforms at 100 Hz as input. In contrast to PhaseNet, BasicPhaseAE uses smaller filter sizes and more filters and eschews residual connections. Earthquake transformer (EQTransformer) (Mousavi et al., 2020) combines event detection, phase detection and onset time picking. It takes 60 s waveform windows at 100 Hz as input and outputs probability traces of the same length for detection, P and S for each point in time. EQTransformer uses CNNs, long short-term memory cells (LSTMs) and self-attention layers. DeepPhasePick (DPP) (Soto & Schurr, 2021) is a collection of models for event detection and phase picking. For detection, DPP uses CNNs to denote 5 s windows with probabilities for P, S and noise. For onset time picking, DPP uses LSTMs and fully connected layers. ARRU (Liao et al., 2021) is a phase picking model which also follows the U-net approach while expanding the architecture by adding adoptions of attention gates (Schlemper et al., 2019) to increase the weights of seismic phases and recurrent-residual convolution units (Liang & Hu, 2015) to strengthen temporal linkages of features at multiple scales. We base PickBlue on these previously published models. For the application to OBS data, we focus on PhaseNet and EQTransformer. We chose these models because of their favorable performance identified in the recent benchmark study of Münchmeyer et al. (2022). For ease of reading, in the following, we refer to our adapted versions of the models as BluePhaseNet and BlueEQTransformer, and to those other versions that we also studied simply as PhaseNet and EQTransformer.

Our training and evaluation procedure follows the steps outlined in Münchmeyer et al. (2022). The neural networks are trained with manually picked phase arrival times of known earthquake waveforms and noise samples taken from the same data set. We train the networks to reproduce a characteristic function where Gaussian peaks with an amplitude 1 and a half-width 0.2 s are centered on the manual picks of P and S arrivals, respectively, identically to the procedure in the benchmarking study. For those traces where either only a P or only an S manual pick is present, the characteristic function for the non-existent phase is set to zero for the whole record.

For evaluation, the ML picker is provided with a longer waveform segment (30 or 60 s, depending on the picker architecture), which contains the manual pick at a random position. Based on this waveform sample, two characteristic functions are calculated, which can be thought of as (non-calibrated) measures of the probability of a sample to be a *P*-wave or *S*-wave, respectively. We then examine a 10 s window from the output that contains the manual pick at a random location within the window. We evaluate both the phase type classification, P or S, and the accuracy of the onset determination, with a focus on the latter. The onset time is given by the time of the peak value of the characteristic function for the respective phase. The target is to be as close as possible to the manual prediction. Of course, particularly for low signal-to-noise ratios or otherwise ambiguous scenarios, the manual pick might be inexact. Therefore, it is not to be expected that a deep-learning picker can perfectly reproduce the

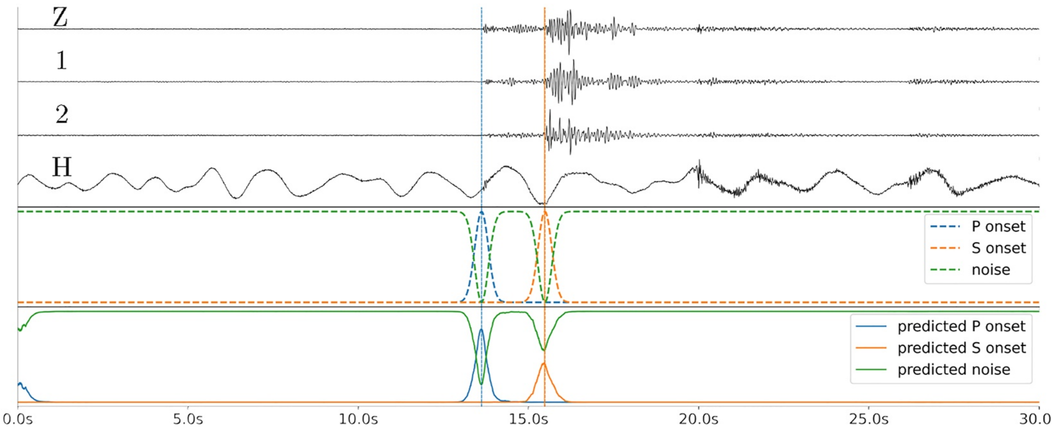


Figure 2. Example of an input waveform sample for BluePhaseNet. From top to bottom: trace with four components after preprocessing (resampling, cutting of waveform, and normalization); ground truth characteristic functions (P, S phase arrivals and noise) used for training the models; characteristic function predicted by BluePhaseNet prediction. Z: vertical component; 1 and 2: two horizontal components; and H: Hydrophone channel.

manual pick times. Nonetheless, the average difference between manual and deep learning picks is a solid indicator of the model performance.

3.1. Models

We trained two models for picking the OBS data. Both models have originally been trained on land station data.

Earthquake Transformer (*EQTransformer*): EQTransformer is a model for joint event detection, phase detection, and onset picking (Mousavi et al., 2020). EQTransformer uses a stack of convolutional layers (LeCun & Bengio, 1995), LSTMs (Hochreiter & Schmidhuber, 1997) and self-attention layers (Luong et al., 2015; Yang et al., 2016). It consists of a down-sampling section of CNNs and max-pooling layers, followed by an encoder using residual CNNs and LSTMs. Then, self-attention layers add contextual information, enabling the model to focus on the most important parts of the sequence. Subsequently, three separate decoders map the information to three probability sequences (detection, P phase, S phase). EQTransformer consists of 378,928 trainable parameters. It expects 60 s long input traces for all channels.

PhaseNet: PhaseNet (Zhu & Beroza, 2019) is a U-Net (Ronneberger et al., 2015), consisting of a convolutional and a deconvolutional branch. During the down-sampling process, four convolutional stages condense the information. Then, in the up-sampling process of deconvolutions, the model expands and converts this information into probability distributions (Zhu & Beroza, 2019). PhaseNet consists of 268,499 trainable parameters, about 30 percent less than EQTransformer.

3.2. Training Workflow

As discussed above, the data set contains traces with missing components due to data gaps, too large timing offset between the components and different hydrophone and seismometer types (Table S1 in Supporting Information S1). For these traces, missing data and components are replaced with zeros. This ensures that the model input always contains four channels and at the same time, makes the models robust to missing data in future application scenarios.

The traces are resampled to 100 Hz. We ensure that our data selection leads to a uniform distribution of pick locations within the input windows. Where traces are too short, missing samples were padded with zeros. We normalize the amplitude by dividing each component by its maximum amplitude independently. By normalizing each component individually, we avoid either hydrophone or velocity (seismometer) amplitudes being reduced to values close to zero, as the data were not corrected to physical units and absolute amplitudes depend on digitizer gain settings. Last, we subtract the mean and the linear trend for all channels. Figure 2 shows an example input after the preprocessing steps used for PhaseNet.

The original EQTransformer makes intensive use of augmentations during training. Therefore, we also apply the following augmentations to the EQTransformer training process to each sample, with probability p: (a) adding a secondary earthquake signal into the empty part of the trace ($p = 0.3$), (b) adding a random level of Gaussian noise ($p = 0.5$), (c) randomly inserting gaps ($p = 0.2$), and (d) dropping one or two channels ($p = 0.3$). We do not apply the additional random wrap-around shifting of the traces, as employed in the original implementation. This is not required, as the lengths of the original waveforms (at least 60 s noise before the P pick is available for 7% of the traces and for 36% in case of S picks) ensures that the pick can naturally occur at any location in the trace (see also Figure S2 in Supporting Information S1).

While we input the three seismometer components unfiltered to the networks, we high-pass filter (0.5 Hz) the hydrophone data before feeding it to the network as a fourth component. Filtering proves beneficial due to strong low-frequency noise on the hydrophone traces resulting from the broadband characteristics of hydrophones.

We experiment with two strategies for model initialization.

1. Standard training: initializing the models with random weights and training them on the OBS data set tailored to the OBS domain;
2. Transfer learning: initializing the models with weights from the same models but pre-trained on STEAD and INSTANCE data sets, which both consist of land station data only; the hydrophone related channels are still initialized with random weights. Transfer learning takes advantage of the model performance in a similar domain, thus resulting in faster convergence and often higher generalizability, particularly where the number of training data samples used for the pre-training significantly exceeds that available for the specialized training (Jozinović et al., 2021; Münchmeyer et al., 2020; Pan & Yang, 2009).

For cross-domain applications, the data sets used for training or pre-training must be of the same distance range as the later application domain (Münchmeyer et al., 2022). Furthermore, source models trained on large data sets generally yield the best performance after fine-tuning. As the OBS data contains local earthquakes, we choose the PhaseNet and EQTransformer models trained with INSTANCE and STEAD data sets for initializing the weights.

We tested different hyperparameters and determined the following set as best suited: batch size 2048 (EQTransformer)/1024 (PhaseNet); learning rate 0.001/0.01; epochs 200/400. We use the Adam optimizer (Kingma & Ba, 2014). The models with minimum loss of the development subset were found after 185 (EQT) and 240 epochs (PhaseNet), respectively. Those models were subsequently used in the evaluation. The training of the models took about 19 hr on an NVIDIA A100-40 GPU with 40 GB memory.

4. Results

In the following, we compare the performance of the different models. In order to retain clarity of presentation, we start with a discussion of the preferred models, that is, the models showing the best overall performance. These are the EQTransformer and PhaseNet models using transfer learning with pre-training on INSTANCE, which we refer to as BlueEQTransformer and BluePhaseNet, respectively. However, the PhaseNet model trained without pre-trained weights effectively achieves the same performance. We then provide a more detailed analysis and comparisons to other model variants. Unless specified otherwise, all analyses were conducted for the preferred models.

4.1. Onset Time Determination

We use the residuals, that is, the differences between the ML pick and the onsets picked by human analysts (as stored in the data set metadata), to evaluate the quality of onset time picking. We analyze the modified root mean squared error (RMSE), the mean absolute deviation (MAD), and the modified mean absolute error (MAE). Specifically, we calculate the RMSE and MAE by (arbitrarily) defining outliers as residuals outside the interval ±1.0 s, and taking them into account with a value of 1 s. In this way, we avoid overly strong influence of the outliers on the metrics. We separately report the outlier fraction according to this definition.

Figure 3 shows the residual distribution of *P*- and *S*-wave picks using the preferred BlueEQTransformer and BluePhaseNet models. For P onsets, BluePhaseNet outperforms BlueEQTransformer with MAEs of 0.23 versus 0.32 s and RMSEs of 0.30 versus 0.33 s, whereas both models show comparable results for S onsets. The median is very close to 0.0 s for both models, showing that there is virtually no bias with respect to the manual picks. The central distribution resembles a Laplacian. For P picks, 90% of picking errors are smaller than 0.62 and 0.46 s for

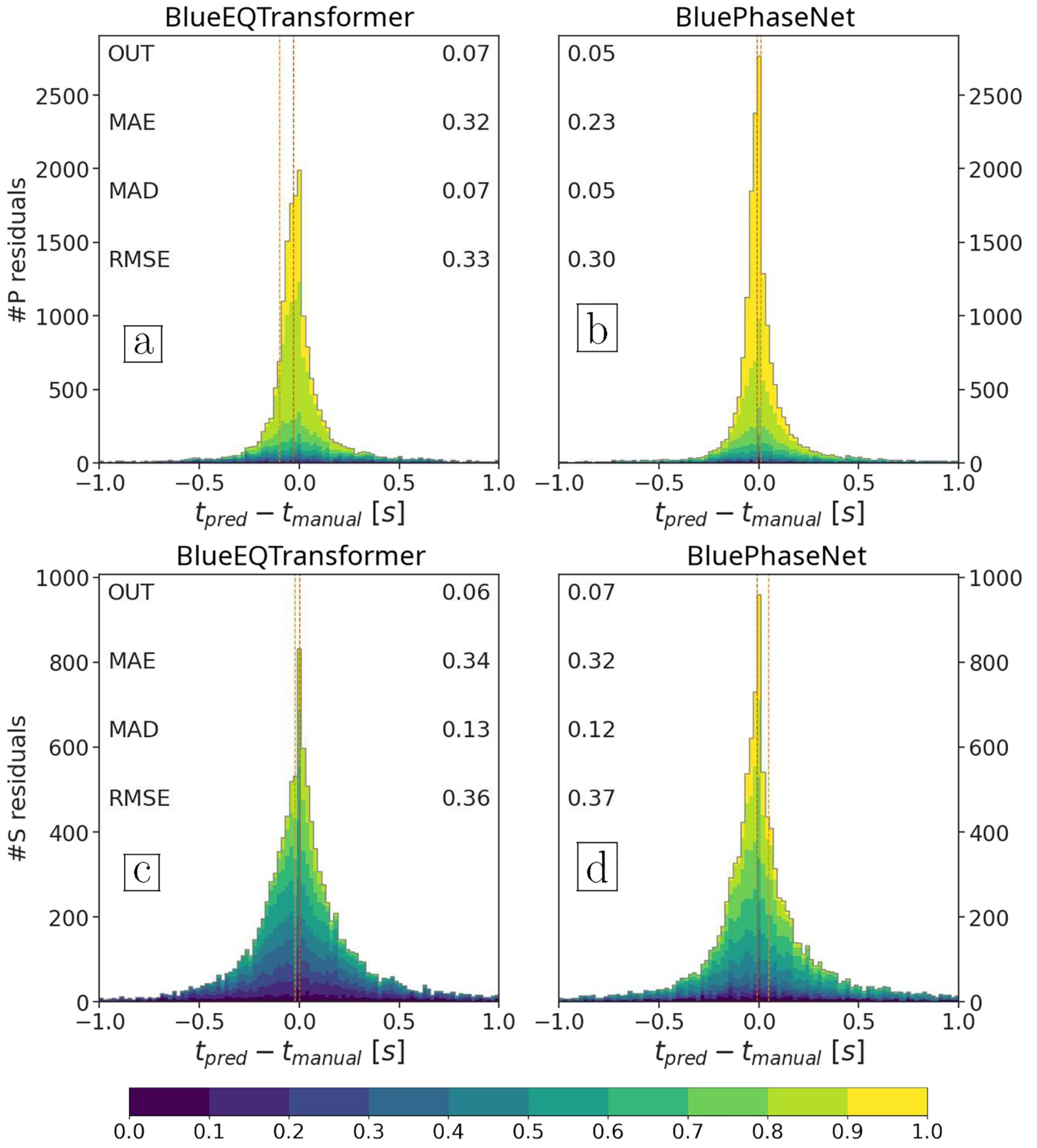


Figure 3. Histogram of residuals between the manual picks and picks by BlueEQTransformer and BluePhaseNet (pre-trained on INSTANCE) for P phases (panels (a, c)) and S phases (panels (b, d)); the bin width is 0.02 s. The vertical dashed lines mark the median (orange) and mean (yellow) residual, but note that the mean is strongly influenced by outliers. The histogram columns are subdivided by color-coded pick confidence (peak of the characteristic function for P or S arrivals); each segment length corresponds to the frequency of residuals with the respective confidence. OUT: Fraction of outliers (residuals with absolute value > 1.0 s, i.e., outside the window shown), MAE: Mean absolute error, MAD: Median absolute deviation. RMSE: Root mean square error. Note that the *y*-axis differs for the lower and upper panels. Note that in the calculation of MAE and RMSE, residuals with absolute values exceeding 1 s were set to ±1 s in order to reduce the dependence of these measures on outliers. See Figure 4 for a view of the same distribution but extending to ±10 s.

BlueEQTransformer and BluePhaseNet; 95% are within 1.33 s/0.99 s. For S picks, 90% are below 0.67 s/0.71 s, 95% below 1.17 s/1.2 s.

Comparing P residuals to S residuals, we confirm that the determination of S arrivals is more difficult, as can be seen, for example, in the nearly double MAD for S compared to P. We attribute the lower performance for S picks to two main factors. First, S arrivals usually have lower signal-to-noise ratios due to the typically much higher noise levels of OBS horizontal components. Also, overlap with the P coda and precursors from a basement or Moho Sp conversion can create ambiguous onsets. Furthermore, our training data set contained substantially fewer examples of S arrivals, giving the models less possibility to learn the characteristics of S arrivals.

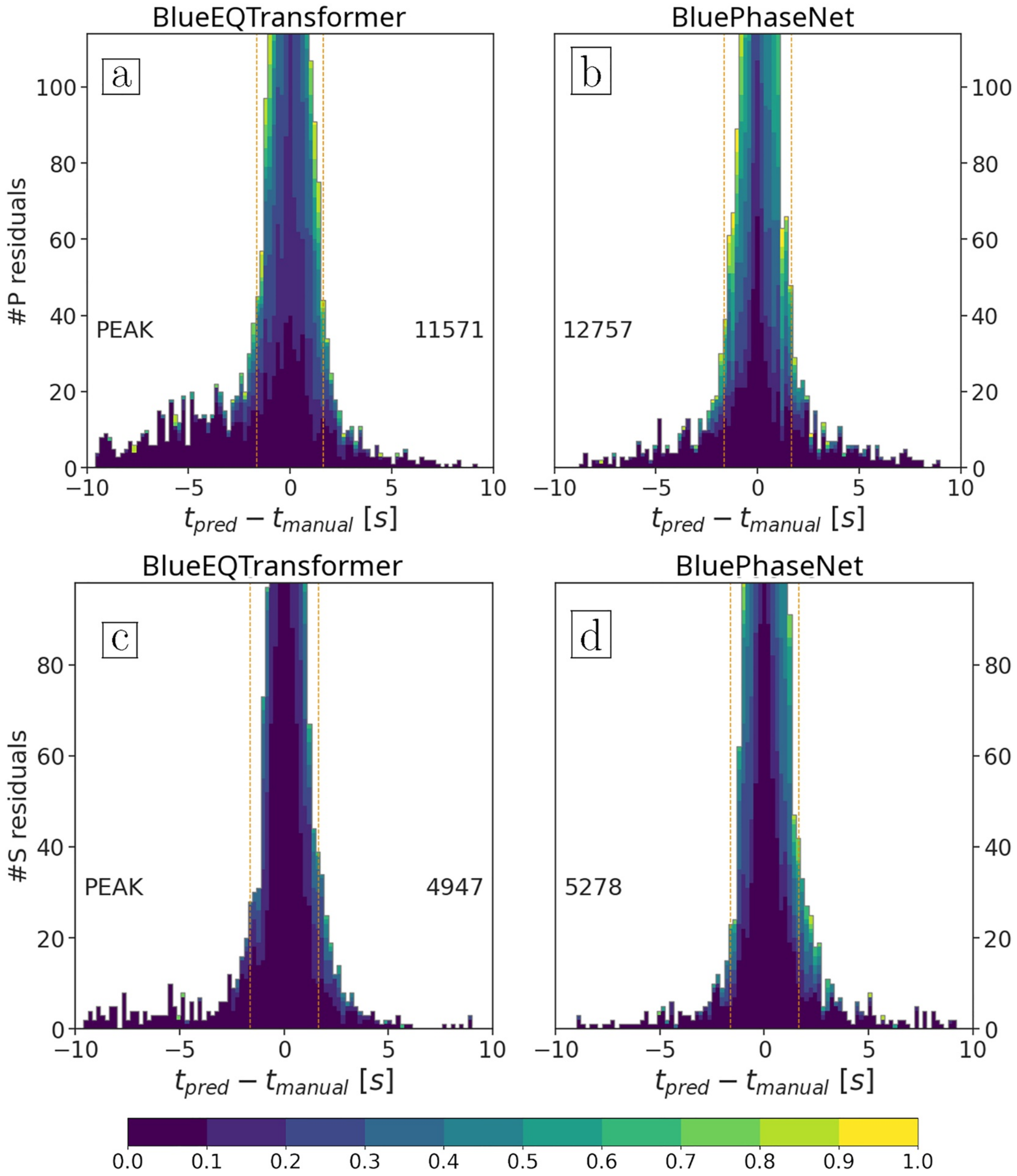


Figure 4. Histogram of residuals between the manual picks (as Figure 3 but showing the full range of possible residuals from −10 to +10 s and with a bin width of 0.2 s) and picks by BlueEQTransformer and BluePhaseNet (pre-trained on INSTANCE) for P phases (panels (a, c)); and S phases (panels (b, d)). The *y*-axis range is chosen to emphasize the tails of the distribution, with PEAK showing the peak value of the histogram, which will be near zero on the *x* axis but be far outside the *y* range shown. Ninety percent confidence intervals are marked with yellow vertical lines.

Nonetheless, the performance for S arrivals is still excellent, in particular considering that traditional pickers are usually unable to detect S arrivals on OBS recordings at all.

Figure 4 shows a zoomed out version of the error distribution, focusing on the outliers. Up to about ±2.5 s, the residuals follow a Laplacian distribution. Outside this interval, they follow a more uniform or low-sloped triangular distribution. We will refer to these picks as blunders in line with previous usage for describing analyst picks (Diehl et al., 2012). Although both BlueEQTransformer and BluePhaseNet show blunders, there are some subtle differences. For BlueEQTransformer, the central part (i.e., up to ±2.5 s) appears fairly symmetric, but blunders earlier than the manual pick occur more often than blunders in the coda of the *P*-wave. This results in a mean value shifted to early picks (−0.07 s). Conversely, blunders appear fairly symmetric with respect to the manual pick for BluePhaseNet, whereas, regarding S onset residuals, the central part is very slightly skewed toward later arrivals, resulting in too late average picks (mean 0.05 s).

Figures 5 and 6 break down the performance by the individual deployments. The overall performance varies strongly across data sets, likely due to differences in noise environment, magnitude distribution, network geometry, and station types (broadband or short period). Notably, the slight superiority of BluePhaseNet relative to BlueEQTransformer is valid across most data sets. On data sets where BlueEQTransformer outperforms BluePhaseNet, the performance difference is minor. Nonetheless, both pickers are viable across many different settings and return consistently small errors. The great majority of (BlueEQTransfomer/BluePhaseNet) picks have errors below 0.2 s: 77%/81% for P, 63%/65% for S for the whole data set; the ranges for individual deployments are 53%–96%/56%–98% for P, 36%–94% 33%–93%/52%–95% for S. While, for each deployment, every second S pick of BluePhaseNet has an absolute error below 0.2 s, BlueEQTransformer achieves the same only with every third pick.

4.2. Classification of Phase Types

As mislabeled phase types are challenging for association algorithms and result in large residuals during earthquake location, we investigated the reliability of the P versus S classification. For the comparison, we used 14,404 traces from the test set, which have any combination of P and S picks. We classify a predicted P pick as misclassified if it has higher confidence than the predicted S pick on the same trace and if it is closer to the manual S pick than to the manual P pick; vice versa for predicted S picks. Picks with confidence below 0.1 were ignored. In general, phase identification was very accurate (see Table S2 in Supporting Information S1 for the misclassification matrix). Only 11 P picks (0.25%) were wrongly classified as S arrivals from BluePhaseNet. In turn, 20 (0.4%) S phases were wrongly classified as P phases. In sum, BlueEQTransformer has even fewer misclassifications. Only 16 P picks (0.32%) were classified as S arrivals, and 6 S picks (0.14%) were classified as P arrivals. Figures S5–S8 in Supporting Information S1 show examples of misclassified events for P and S phases for both BluePhaseNet and BlueEQTransformer.

4.3. Confidence and Picking Quality

Both models provide time series of probabilities for P and S phase onsets. We interpret the peak values of these time series as the confidence of the model in the pick. Here, we try to understand how this value relates to the accuracy of the pick with respect to the manual pick. In general, we expect smaller residuals for higher pick confidence. This way, confidence values could serve as proxies for pick quality. Depending on the use case, a picker that refuses to return a pick for a few cases might be preferable over a picker that returns more, but low quality picks.

Each of the bins in Figures 3–6 consists of a sorted stack of confidences, giving an impression of how the confidences map to picking errors. The length of each segment corresponds to the frequency of residuals with the respective confidence. Three major observations can be made. First, both models pick P waves with higher confidence than S waves. This becomes even more obvious when considering individual experiments (Figure 5 vs. Figure 6). Second, even though there is a considerable difference between the experiments, high confidence values cluster around 0 residual, that is, the higher the confidence, the lower the expected residual. Third, picks with low confidence still follow an approximate Laplacian distribution centered near 0.0 s but with higher uncertainties (Figure 3). Crucially, almost all of the outliers have low confidence (Figure 4).

Figure 7 explores the relationship between confidence and picking errors in a systematic manner. We plot MAE, MAD and the outlier fraction, depending on which fraction of the most confident picks are considered. For example, the MAE drops significantly for both P and S picks if the least confident 10% are omitted. For S picks, we observe a knee in MAE and OUT curves at about 5%. Further reduction leads to a further but more gradual reduction in MAE. On the other hand, the MAD for P waves decreases only very marginally when the least confident picks are omitted, whereas a strong reduction is seen in the number of outliers. The drop in MAE is thus almost exclusively driven by a reduction in the number of outliers. As the MAD is a robust measure of the spread of the central peak, this implies that there is hardly any dependence of the confidence value on the pick quality as long as the correct phase has been identified; low confidences instead indicate a much higher chance of having misidentified another waveform feature, for example, a later arrival or a local maximum in the noise time series as a phase, and thus produced an outlier.

For P arrivals, the number of outliers decreases until only 60% of the picks are retained and stays almost constant afterward (Figures 7a and 7c). For S arrivals, no such clear cutoff can be identified, with outlier fractions decreasing steadily until only 30%–40% of the picks are retained. *P*- and *S*- waves also differ in the distribution of confidence values. While S confidences are distributed almost uniformly, the P confidences tend toward higher

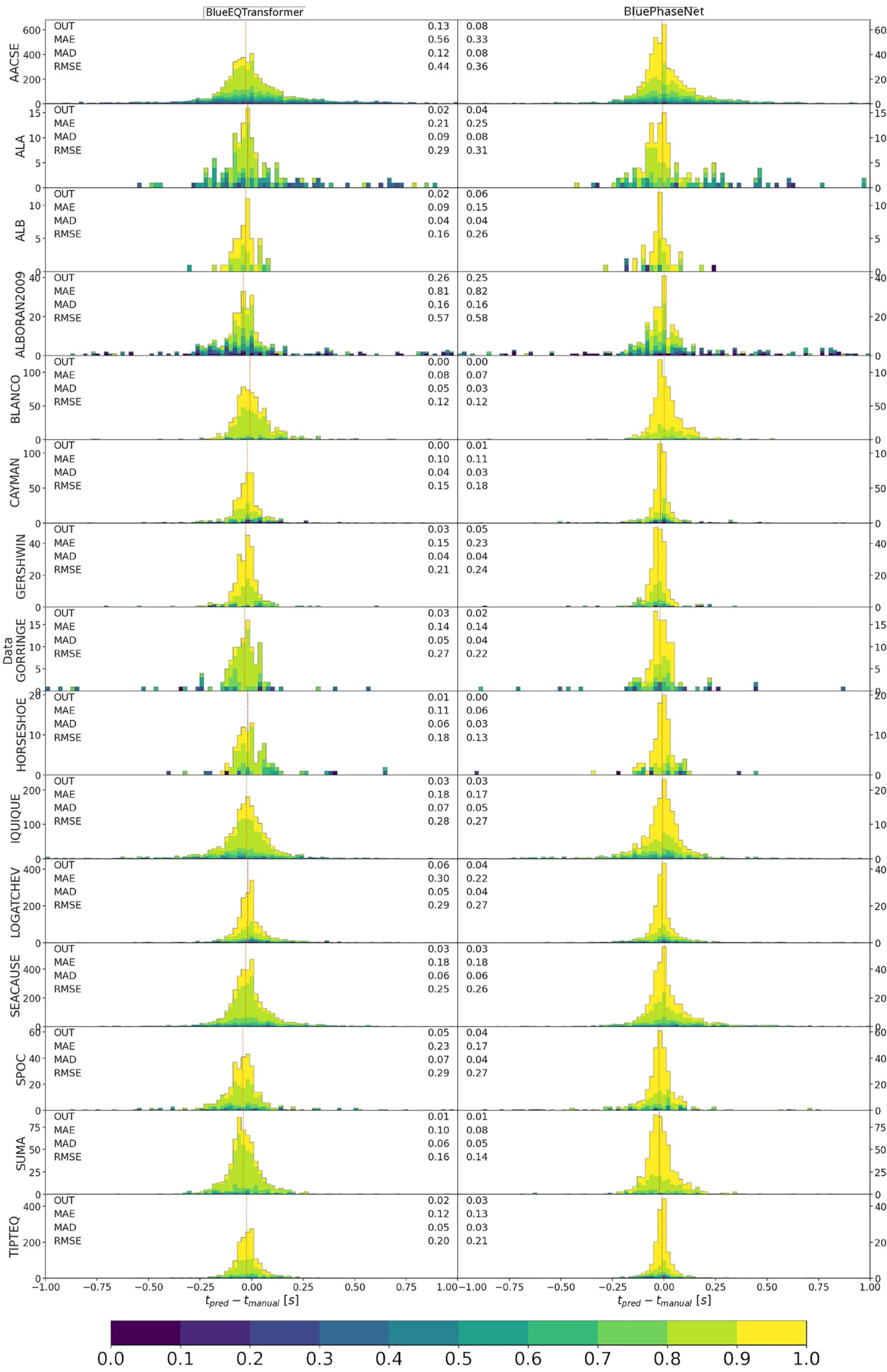


Figure 5. P residuals for the individual experiments between −1.0 and +1.0 s. BlueEQTransformer (BluePhaseNet), pre-trained on INSTANCE, are shown in the left (right) column, respectively. The panels are labeled to the right with the name of the data set. Figure S3 in Supporting Information S1 shows the same data with the range extended to ±10 s.

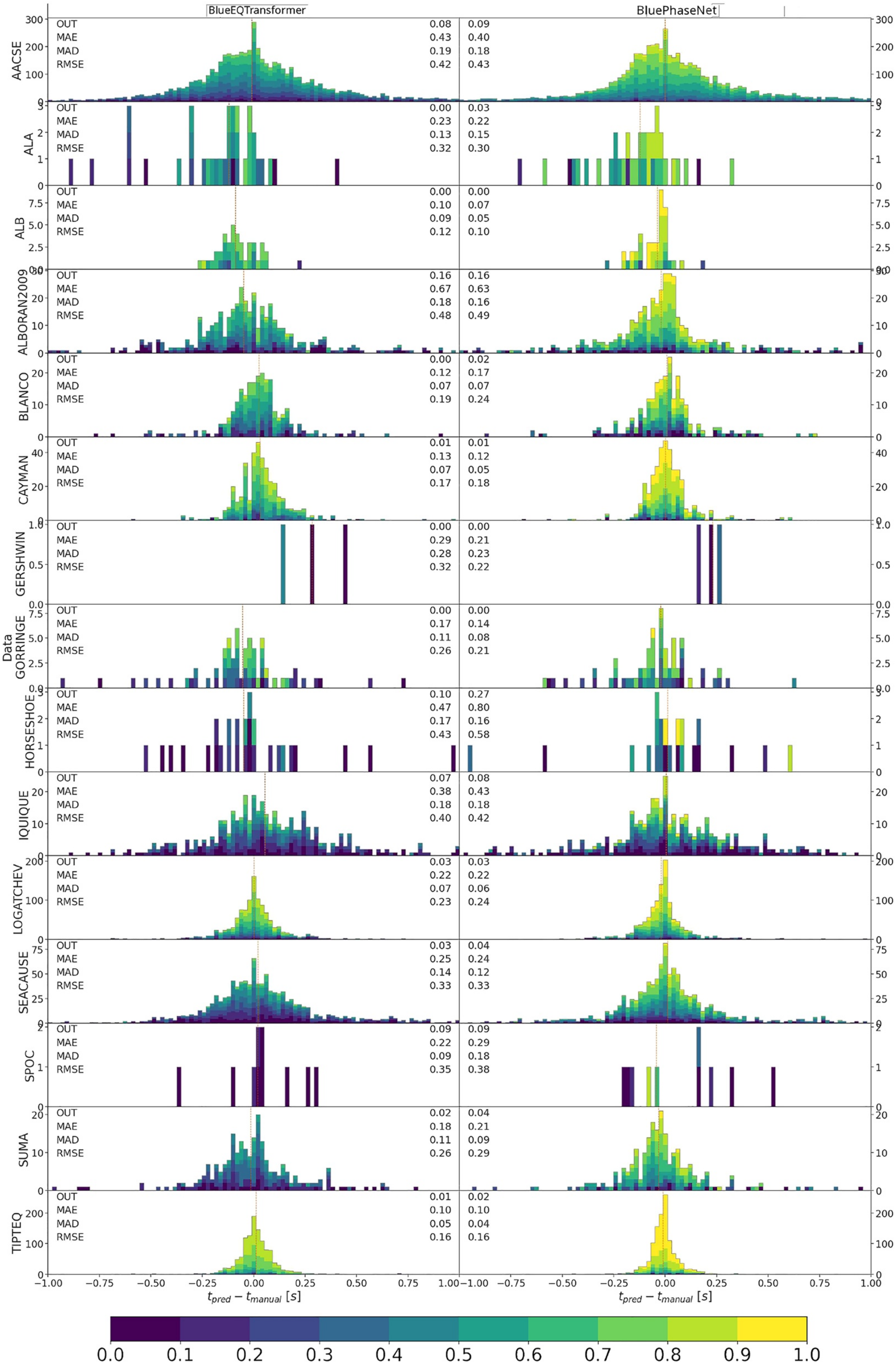


Figure 6. As Figure 5 but for S residuals. Figure S4 in Supporting Information S1 shows the same data with the range extended to ±10 s.

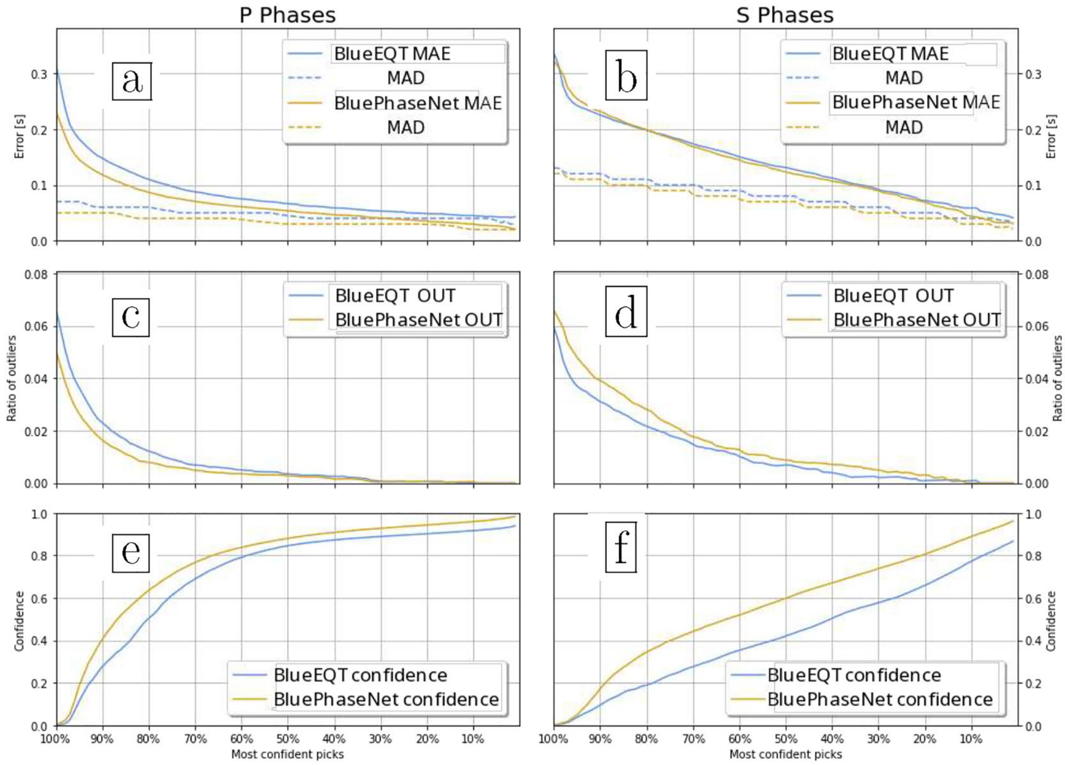


Figure 7. Dependency of mean absolute error, mean absolute deviation (panels (a, b)) and number of outliers (panels (c, d)) for different subsets of the data set, sorted by confidence value, such that the value on the left side of each plot corresponds to the full data set (as shown in Figure 3), whereas subsequently stricter confidence thresholds are applied to select only a set fraction of the whole data set corresponding to the most confident picks. (e, f) Shows percentiles of confidence value. Model training based on INSTANCE pre-training.

values, for example, the 60% most confidence picks already have confidence values around 0.8. These observations should be taken into account when selecting confidence thresholds in applications. They suggest that P confidence values at higher confidences might be less reliable than their S counterparts; this is also reflected in the more continuous degradation in performance for S arrivals compared to their P counterparts. Interestingly, confidence curves are almost identical across the two different models.

Following the finding that residual error statistics are largely driven by outlier fractions with only minor variations of residuals in the central block, the choice of threshold in an application should primarily be guided by the relation of outliers, hit rate and miss rate. Whether to prioritize minimizing missed picks or avoiding outliers will depend on the downstream workflow and the specifics of associator and location programs. Figure 8 visualizes the tradeoff of these parameters for different thresholds for both models. For this analysis, we set a tighter threshold for P pick outliers (0.5 s) to account for the typical accuracy expected for P picks and to make the outlier fraction more visible in the plot. Due to the more steady decrease in hit rate for S arrivals than for P arrivals, in general, a lower threshold is advisable for S arrivals than for P arrivals. The exact choice of parameters depends on the data set and the planned downstream analysis.

5. Discussion

5.1. Effectiveness of Transfer Learning

For all results presented so far, we used the preferred models, BluePhaseNet and BlueEQTransformer. These models were pre-trained on INSTANCE and then fine-tuned on the OBS data. In the following, we discuss the impact of this transfer learning and compare the results to transfer learning on STEAD.

Figure 9 shows the performance in terms of *P*-wave residuals of models without transfer learning and with transfer learning from STEAD. EQTransformer shows substantially worse performance without transfer learning, whereas PhaseNet is not only easy to train with randomly initialized weights but also does not show any

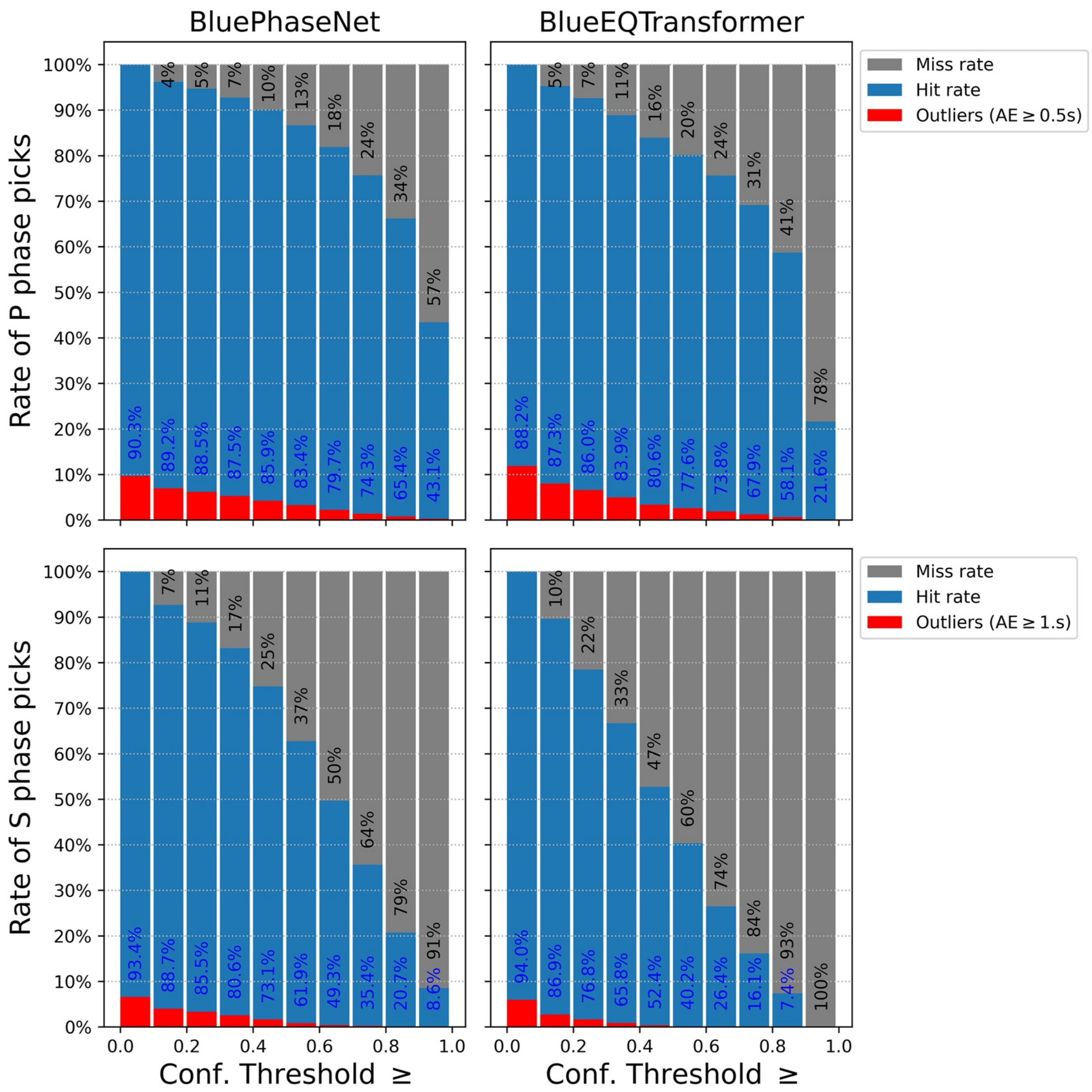


Figure 8. Hit rate, miss rate, and number of outliers for different confidence thresholds for models with INSTANCE pre-training. Black numbers specify the number of misses, whereas blue numbers mark the number of hits that are not considered outliers. Note the different absolute error thresholds for picks to be considered outliers, noted in the legend.

benefit from transfer learning. We hypothesize that the impact of transfer learning on EQTransformer is larger than on PhaseNet due to the deeper and more complex architecture. In particular, PhaseNet might benefit from the implicit parameter sharing across positions through the fully convolutional architecture and from the residual connections. In addition, the EQTransformer without transfer learning generally exposes lower confidence than the one using transfer learning. These patterns are similar for *S*-waves.

Regarding the appropriate source data set, we observe differences between EQTransformer and PhaseNet. For EQTransformer, differences between pre-training on STEAD or INSTANCE are small, and in both cases, pre-training improves performance. For PhaseNet, pre-training on INSTANCE leads to better performance than on STEAD. In fact, the performance of the STEAD pre-trained PhaseNet is less good than that of the non-transfer learned PhaseNet. We hypothesize that This lack of data diversity in STEAD might have led to an overly adapted version of PhaseNet because the STEAD data set contains exclusively recordings of local earthquakes at distances less than 350 km, which implies a narrower frequency range on average. Furthermore, STEAD recordings consist of only 60 s of waveforms with limited time ranges in which P or S arrivals can occur.

Comparing the results on the individual deployments (Figure 5, Figures S12–S14 in Supporting Information S1), there is substantial variability about the best performing model and training scheme. While it is difficult to derive

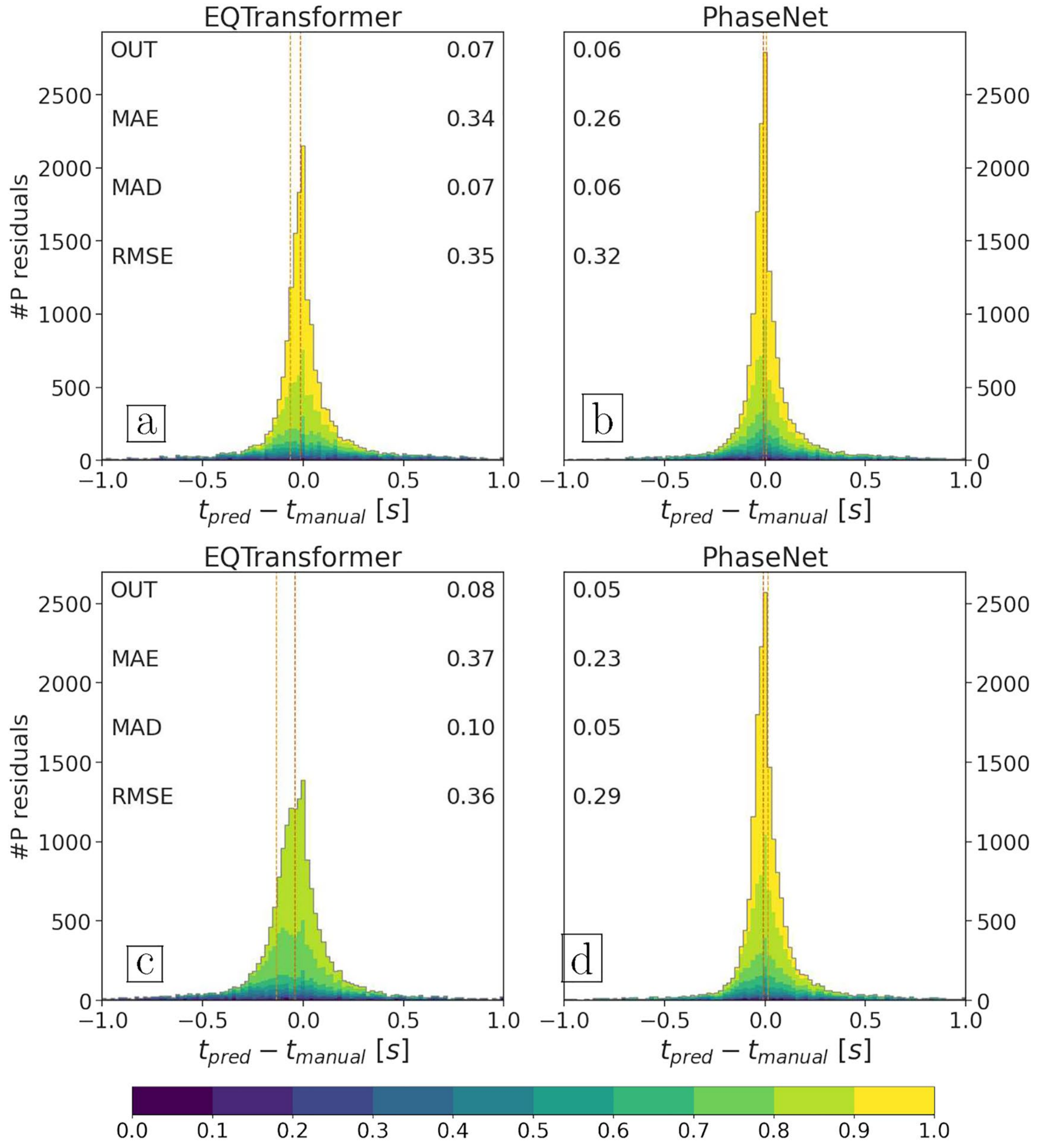


Figure 9. P residuals with confidence stacks for different pre-training set-ups. (a, b) Models pre-trained on STEAD, then trained on ocean bottom seismometer (OBS) data set. (c, d) Models trained on OBS data set without pre-training. The full range to ±10 s is shown in Figure S9 in Supporting Information S1 and the equivalent plots for S residuals are shown in Figures S10 and S11 in Supporting Information S1 for pre-training with STEAD and without pre-training, respectively.

clear insights due to the diverse behavior, for some data sets, a certain training scheme, that is, pre-training either STEAD or INSTANCE or no pre-training at all, seems to be the best option, irrespective of the model.

In conclusion, the analysis for individual data sets suggests that transfer learning often only yields moderate benefits and that its benefits depend on the targeted data set. The already considerable number of training examples in the data set we assembled might explain the measurable but overall limited benefits of transfer learning.

5.2. Significance of the Hydrophone Component and Training on OBS Data

A key difference of OBS recordings to typical land-based recordings is the existence of an additional hydrophone component. Our models incorporate this component as an additional input channel. In addition, the noise environment in the oceans is different from that on land, as described in the Introduction. In this section, we compare the performance of PickBlue to that of the equivalent pickers trained on land data, and to three-component EQTransformer and PhaseNet networks trained on the OBS data set.

In Figures 10a–10d we compare the performance of the OBS picker against EQTransformer and PhaseNet models trained on STEAD and INSTANCE (the performance of the preferred 4-component pickers, BlueEQTransformer and PhaseNet, on the same subset of picks is shown in Figures 10g and 10h). We chose these models as baseline comparisons, as they have been identified as the best performing models in the benchmark study of Münchmeyer et al. (2022). As the land-based models would not be applicable to traces with only hydrophone components, we ignored those traces for the evaluation, that is, removing about a sixth of the data set.

The results show a clear ranking among the models. Best performing are the PickBlue models, which use the hydrophone components. In particular, the outlier fraction doubles for EQTransformer for some configurations when omitting the hydrophone components. Most interestingly, the width of the central peak of the residual distribution gets considerably wider. For *P*-waves on EQTransformer, the MAD increases by 37% for OBS-trained models without hydrophones and increases three- to seven-fold when using the land-based models. For *P*-waves on PhaseNet, the decrease in performance when omitting the hydrophones is less drastic; MAD and outlier fractions barely increase. However, the MAD for land-based models drastically increases five- to eleven-fold with respect to the models incorporating hydrophone data.

Results for the S arrivals closely resemble the results from P waves: the PickBlue models with hydrophones perform considerably better than those not making use of the hydrophone data, even though one might have expected the hydrophone data to make very little difference for *S*-waves. In addition, training on OBS data substantially improves performance in terms of both the number of outliers and the residuals. Interestingly, all three-component land-trained PhaseNet models tend to pick slightly too late for both *P*- (Figure S16 in Supporting Information S1) and *S*-waves (Figure S17 in Supporting Information S1), whereas the EQTransformer equivalents tend to pick too early, except for the STEAD pre-trained variant on P onsets.

Looking into the performance of the individual experiments (Figures S19 and S20 in Supporting Information S1) ALBORAN2009 and LOGATCHEV, we find EQTransformer performing remarkably worse than PhaseNet. The outlier plots (Figures S21 and S22 in Supporting Information S1) underline this impression: Almost all EQTransformer's picks are too early. The difference between those two experiments, in contrast to the others, is caused by the fact that dropping the hydrophone component only leaves data with a single horizontal component. This suggests that PhaseNet is better prepared to deal with such degraded data.

In conclusion, these results highlight that the hydrophone component is essential for optimally picking OBS data. Consequently, pickers trained exclusively on data from land-based stations show clearly inferior performance to those trained on OBS data.

5.3. Improved Seismicity Catalogs With PickBlue

As a final validation of PickBlue, we test it on the Hikurangi Ocean Bottom Investigation of Tremor and Slow Slip (HOBITSS) data set (Wallace et al., 2014). This data set has not been used for model training, so the deployment conditions, region and instruments are unknown to the models. This is a realistic test case as PickBlue is intended to be used on new deployments without model retraining.

We prepared an earthquake catalog spanning April 2014 to May 2015 based on the deployment using a simple workflow. First, we picked the continuous OBS traces using the PickBlue model based on PhaseNet. We used a picking threshold of 0.1 for *P*-waves and 0.15 for *S*-waves. Second, we associated the resulting picks using GaMMA (Zhu et al., 2022). Third, we performed an absolute relocation based on the picked phase arrival times using NonLinLoc (Lomax et al., 2000) with a layered velocity model from Yarce et al. (2019). We note that this workflow only yields a preliminary catalog and is not comprehensive as further steps such as relocation with station residuals, relative relocation, or magnitude estimations are omitted. We only retained events that fulfilled three quality control criteria: at least 10 phase picks in total, at least three stations with both P and S picks, and residual root mean square (RMS) value below 0.3 s.

Figure 11 compares a seismicity catalog generated using PickBlue and using PhaseNet trained on INSTANCE. With the same quality control criteria, the PickBlue catalog contains almost twice as many events (6396 events) as the PhaseNet catalog (3556 events). Both of these substantially surpass the 2313 events reported by Yarce

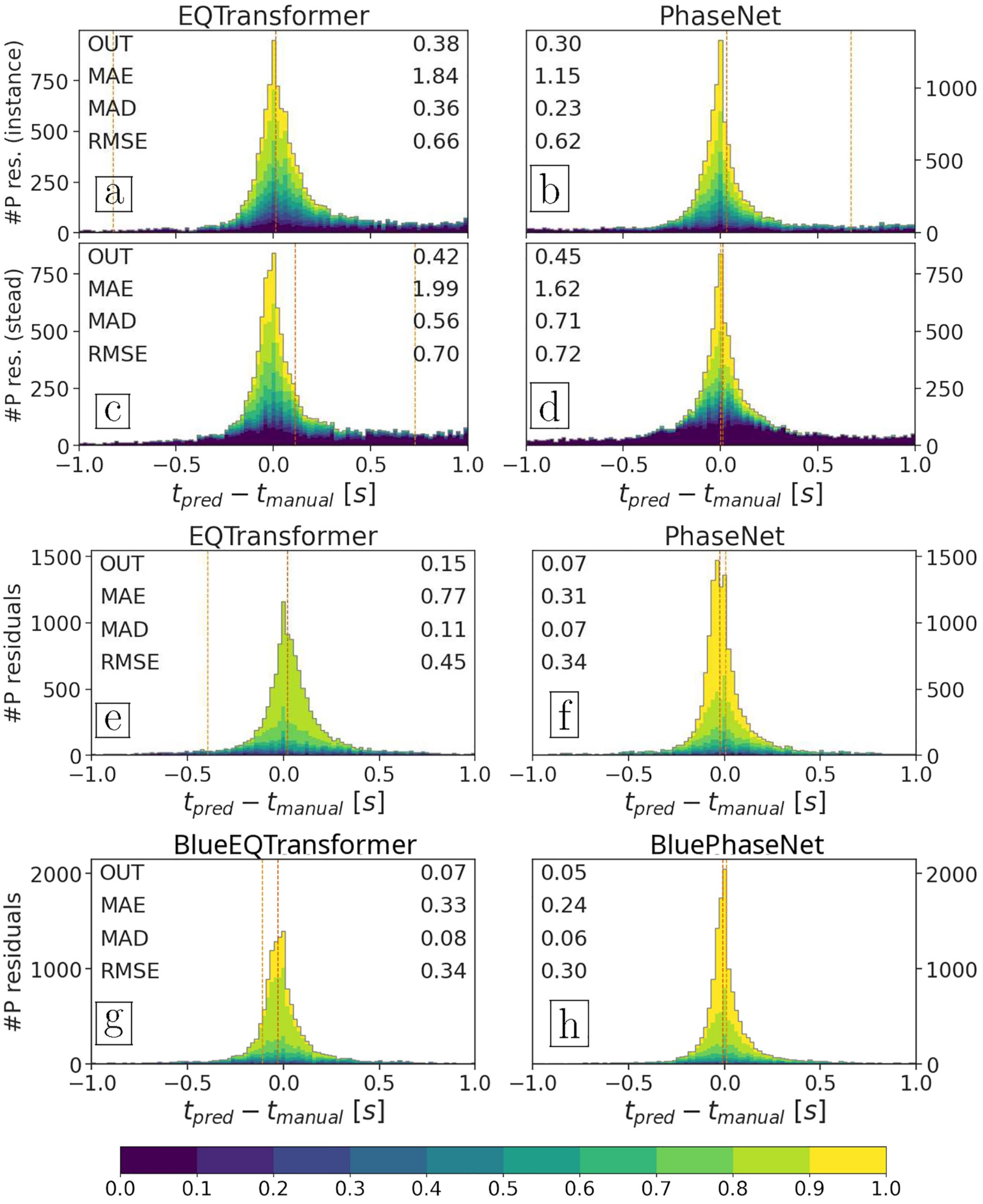


Figure 10. P residuals with confidence stacks for different three-component models, that is, pickers do not use the hydrophone channel. (a–d) Are for pickers trained with onshore station data only, specifically using the models downloaded from the Seisbench platform and trained on the (a, b) INSTANCE and (c, d) STEAD data sets, respectively. (e, f) Show the performance of the three-component model trained with the ocean bottom seismometer data set, that is, only using the three seismometer components (pre-training with INSTANCE). The full range to ±10 s is shown in Figure S16 in Supporting Information S1 and the equivalent plots for S residuals are shown in Figures S17 and S18 in Supporting Information S1. (g, h) Show how the preferred models perform on the same reduced data set (i.e., with all four components present but only those traces with at least one seismometer observation).

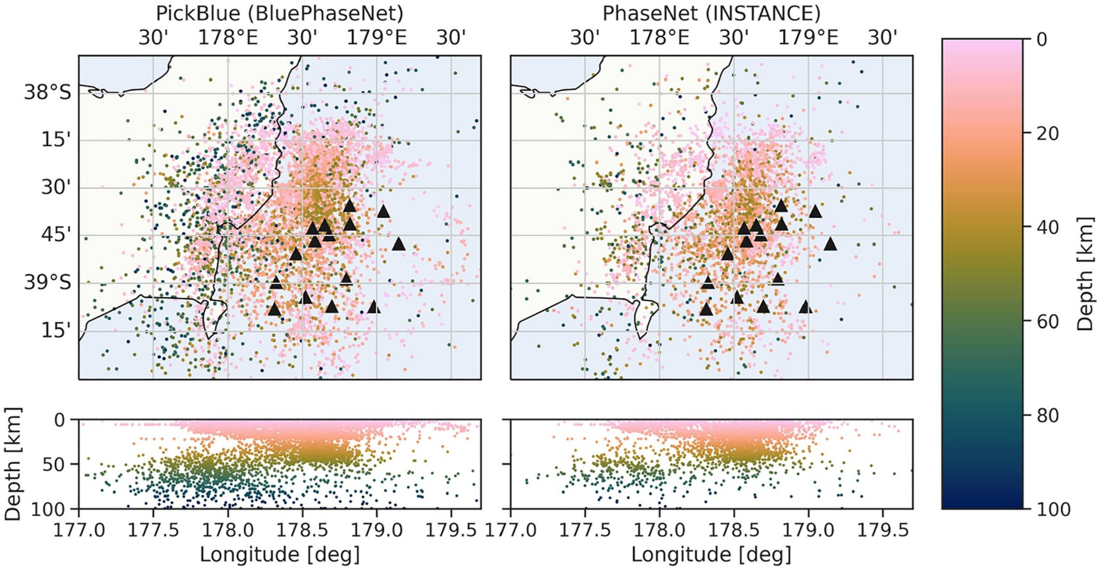


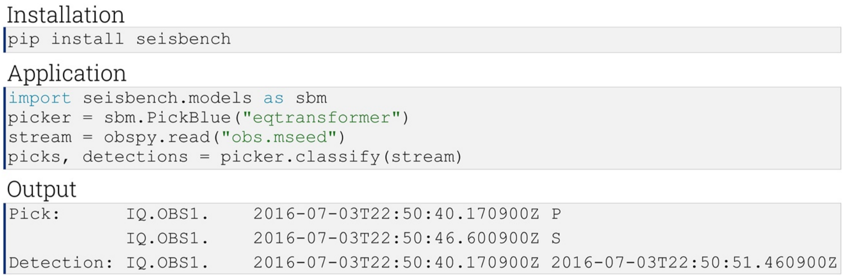
Figure 11. Seismicity catalogs for the Hikurangi Ocean Bottom Investigation of Tremor and Slow Slip ocean bottom seismometer deployment offshore New Zealand. The catalog shown in the left panel has been produced using the PickBlue model based on PhaseNet (6396 events). The catalog shown in the right panel (3556 events) has been produced using PhaseNet trained on INSTANCE, a land data set from Italy. The bottom panels show a longitudinal cross-section including all events; note that this is oblique to the subduction direction.

et al. (2019) for the same time period, even though this catalog used more than 20 land stations in addition to the OBS instruments. The larger catalog size can be attributed both to a higher picking sensitivity (required number of picks) or a higher picking precision (required residual RMS). On average, the PickBlue catalog has 14.8 picks per event, the PhaseNet catalog has only 13.7. The seismicity catalog is consistent with the known seismicity in the region and the large scale pattern reported by Yarce et al. (2019). In particular, the north-westward dipping subduction is clearly visible. For the PickBlue based catalog a faint band to the east and southeast of the OBS array can be discerned, well separated from the main group, which is presumably indicating outer rise seismicity and only has a few isolated events in the PhaseNet (INSTANCE) based catalog.

This test highlights that PickBlue can be applied to new deployments and yields substantial improvements in catalog completeness compared to traditional methods and land-based deep learning pickers.

5.4. Access to Data and Models

We integrated PickBlue into the SeisBench package for easy and direct application. SeisBench is available through PyPI/pip and is licensed under the open GPLv3 license. PickBlue can directly be applied to obspy stream objects to obtain phase picks or full characteristic functions. Figure 12 shows an example of how to install and apply PickBlue, as well as an example output. The implementation automatically applies all necessary preprocessing steps and handles the windowing and reassembling of windows for applying PickBlue to long input streams. It uses SeisBench's efficient implementation and comes with GPU support and parallelization options to be applied to large-scale data sets.



Installation

```
pip install seisbench
```

Application

```
import seisbench.models as sbm
picker = sbm.PickBlue("eqtransformer")
stream = obspy.read("obs.mseed")
picks, detections = picker.classify(stream)
```

Output

```
Pick:      IQ.OBS1.    2016-07-03T22:50:40.170900Z P
           IQ.OBS1.    2016-07-03T22:50:46.600900Z S
Detection: IQ.OBS1.    2016-07-03T22:50:40.170900Z 2016-07-03T22:50:51.460900Z
```

Figure 12. Example code for installing and applying PickBlue within SeisBench. The lower panel shows the example output.

The OBS reference data set compiled for this study is also available through SeisBench. It is accessible through the module *seisbench.data* and comes in the standard SeisBench format. This enables the use of built-in access and filtering methods, as well as the possibility of using the data set with the SeisBench data generation pipelines. We hope this access will stimulate the future development of ML models for OBS data for picking but also for other tasks, such as source parameter estimation (Münchmeyer et al., 2021).

6. Conclusion

In this study, we trained existing deep learning models for OBS data using four components. For training and evaluating the models, we compiled a large scale reference data set of labeled OBS waveforms, encompassing data from 15 deployments with more than 150,000 manually labeled phase arrivals of local earthquakes. Our results show that deep-learning pickers can provide high-quality picks for both P and S phase arrivals. In particular, for *S*-waves, this is a step-change compared to traditional pickers that often can only pick P phases.

We based our model, PickBlue, on EQTransformer and PhaseNet but added an additional input channel for the hydrophone component. Overall, the version based on PhaseNet showed slightly better performance than the version based on EQTransformer. In both cases, the addition of a hydrophone component provides very significant performance improvements. Even excluding this component, targeted models trained on OBS data substantially improve picking performance compared to models trained exclusively on data from land stations.

Using data from 15 independent deployments allowed us to study performance differences in different settings. While the number of deployments was insufficient to infer relations between performance and specific deployment conditions, such as tectonic setting or instrument type, we could show that the models expose performance variability between deployments. This suggests that in application scenarios, both picker versions provided by this study should be tested and carefully evaluated.

We applied PICKBLUE to a deployment of 15 OBS stations offshore New Zealand and showed that PickBlue yields substantial improvements in catalog completeness compared to traditional methods and land-based deep learning pickers.

We make our data set and models available through the SeisBench library to allow easy access and application. We hope this will foster application to new and existing OBS data sets and thereby contribute to seismological analysis, such as automated earthquake catalogs.

Data Availability Statement

The continuous BLANCO (Nabelek & Braunmiller, 2012), HOBITSS (Wallace et al., 2014), and AACSE (Barcheck, 2023; Ruppert et al., 2022) data sets are archived at the IRIS Data Management System (http://www.iris.edu) and are accessible using the network code X9 (2012–2013), YH (2014–2015), and XO (2018–2019), respectively. The waveform of the events and metadata used in this study, together with the picks, are available through the SeisBench platform (https://github.com/seisbench/seisbench) and are archived at https://doi.org/10.5281/zenodo.10277799 (Lange et al., 2023).

Acknowledgments
The authors gratefully thank the Impuls- und Vernetzungsfonds and the Helmholtz Artificial Intelligence Cooperation Unit (HAICU) of the HGF for supporting the REPORT-DL project under the Grant ZT-I-PF-5-53. JM acknowledges the support of the Helmholtz Einstein International Berlin Research School in Data Science (HEIBRiDS). We thank Yu Ren from GEOMAR for providing the manual picks of the BLANCO data set. This work was supported by the Helmholtz Association Initiative and Networking Fund on the HAICORE@KIT partition. Open Access funding enabled and organized by Projekt DEAL.

References

Allen, R. (1982). Automatic phase pickers: Their present use and future prospects. *Bulletin of the Seismological Society of America*, *72*(6B), S225–S242. https://doi.org/10.1785/BSSA07206B0225

Baer, M., & Kradolfer, U. (1987). An automatic phase picker for local and teleseismic events. *Bulletin of the Seismological Society of America*, *77*(4), 1437–1445. https://doi.org/10.1785/BSSA0770041437

Barcheck, G. (2023). Dataset: Ocean-bottom *P* and *S* arrival waveform dataset from the Alaska amphibious community seismic experiment, 2018-2019. https://doi.org/10.7298/01da-ka24

Barcheck, G., Abers, G. A., Adams, A. N., Bécel, A., Collins, J., Gaherty, J. B., et al. (2020). The Alaska amphibious community seismic experiment. *Seismological Research Letters*, *91*(6), 3054–3063. https://doi.org/10.1785/0220200189

Cianetti, S., Bruni, R., Gaviano, S., Keir, D., Piccinini, D., Saccorotti, G., & Giunchi, C. (2021). Comparison of deep learning techniques for the investigation of a seismic sequence: An application to the 2019, Mw 4.5 Mugello (Italy) Earthquake. *Journal of Geophysical Research: Solid Earth*, *126*(12), e2021JB023405. https://doi.org/10.1029/2021JB023405

Crawford, W. C., Webb, S. C., & Hildebrand, J. A. (1998). Estimating shear velocities in the oceanic crust from compliance measurements by two-dimensional finite difference modeling. *Journal of Geophysical Research*, *103*(B5), 9895–9916. https://doi.org/10.1029/97JB03532

Dai, H., & MacBeth, C. (1995). Automatic picking of seismic arrivals in local earthquake data using an artificial neural network. *Geophysical Journal International*, *120*(3), 758–774. https://doi.org/10.1111/j.1365-246x.1995.tb01851.x

Diehl, T., Deichmann, N., Kissling, E., & Husen, S. (2009). Automatic S-wave picker for local earthquake tomography. *Bulletin of the Seismological Society of America*, *99*(3), 1906–1920. https://doi.org/10.1785/0120080019

Diehl, T., Kissling, E., & Bormann, P. (2012). Tutorial for consistent phase picking at local to regional distances. In P. Bormann (Ed.), *New manual of seismological observatory practice* (2nd ed.). IASPEI, GFZ German Research Centre for Geosciences. (chap. IS 11.4). https://doi.org/10.2312/GFZ.NMSOP-2_IS_11.4

Diehl, T., Kissling, E., Husen, S., & Aldersons, F. (2009). Consistent phase picking for regional tomography models: Application to the greater Alpine region. *Geophysical Journal International*, *176*(2), 542–554. https://doi.org/10.1111/j.1365-246X.2008.03985.x

Gong, J., & Fan, W. (2022). Seismicity, fault architecture, and slip mode of the westernmost Gofar transform fault. *Journal of Geophysical Research: Solid Earth*, *127*(11), e2022JB024918. https://doi.org/10.1029/2022JB024918

Gong, J., Fan, W., & Parnell-Turner, R. (2022). Microseismicity indicates atypical small-scale plate rotation at the Quebrada Transform Fault system, east Pacific rise. *Geophysical Research Letters*, *49*(3), e2021GL097000. https://doi.org/10.1029/2021GL097000

Grevemeyer, I., Gràcia, E., Villaseñor, A., Leuchters, W., & Watts, A. B. (2015). Seismicity and active tectonics in the Alboran Sea, Western Mediterranean: Constraints from an offshore-onshore seismological network and swath bathymetry data. *Journal of Geophysical Research: Solid Earth*, *120*(12), 8348–8365. https://doi.org/10.1002/2015JB012073

Grevemeyer, I., Hayman, N. W., Lange, D., Peirce, C., Papenberg, C., Van Avendonk, H. J., et al. (2019). Constraining the maximum depth of brittle deformation at slow- and ultraslow-spreading ridges using microseismicity. *Geology*, *47*(11), 1069–1073. https://doi.org/10.1130/G46577.1

Grevemeyer, I., Lange, D., Villinger, H., Custódio, S., & Matias, L. (2017). Seismotectonics of the Horseshoe Abyssal plain and Gorringe bank, eastern Atlantic ocean: Constraints from ocean bottom seismometer data. *Journal of Geophysical Research: Solid Earth*, *122*(1), 63–78. https://doi.org/10.1002/2016JB013586

Grevemeyer, I., Reston, T. J., & Moeller, S. (2013). Microseismicity of the mid-Atlantic ridge at 7°S–8°15′S and at the Logatchev Massif oceanic core complex at 14°40′N–14°50′N. *Geochemistry, Geophysics, Geosystems*, *14*(9), 3532–3554. https://doi.org/10.1002/ggge.20197

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Jozinović, D., Lomax, A., Štajduhar, I., & Michelini, A. (2021). Transfer learning: Improving neural network based prediction of earthquake ground shaking for an area with insufficient training data. *Geophysical Journal International*, *229*(1), 704–718. https://doi.org/10.1093/gji/ggab488

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv. https://doi.org/10.48550/ARXIV.1412.6980

Kuna, V. M., Nábělek, J. L., & Braunmiller, J. (2019). Mode of slip and crust–mantle interaction at oceanic transform faults. *Nature Geoscience*, *12*(2), 138–142. https://doi.org/10.1038/s41561-018-0287-1

Küperkoch, L., Meier, T., Lee, J., Friederich, W., & Working Group, E. (2010). Automated determination of P-phase arrival times at regional and local distances using higher order statistics. *Geophysical Journal International*, *181*(2), 1159–1170. https://doi.org/10.1111/j.1365-246X.2010.04570.x

Lange, D., Bornstein, T., Grevemeyer, I., Tilmann, F., Barcheck, G., & Münchmeyer, J. (2023). Database of local seismicity registered on ocean bottom seismometers (OBS) [Dataset]. https://doi.org/10.5281/zenodo.10277799

Lange, D., Tilmann, F., Rietbrock, A., Collings, R., Natawidjaja, D. H., Suwargadi, B. W., et al. (2010). The fine structure of the subducted investigator ridge in Western Sumatra as seen by local seismicity. *Earth and Planetary Science Letters*, *298*(1–2), 47–56. https://doi.org/10.1016/j.epsl.2010.07.020

LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. *The Handbook of Brain Theory and Neural Networks*, *3361*(10), 1995.

Leonard, M., & Kennett, B. (1999). Multi-component autoregressive techniques for the analysis of seismograms. *Physics of the Earth and Planetary Interiors*, *113*(1–4), 247–263. https://doi.org/10.1016/S0031-9201(99)00054-0

Liang, M., & Hu, X. (2015). Recurrent convolutional neural network for object recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3367–3375).

Liao, W.-Y., Lee, E.-J., Mu, D., Chen, P., & Rau, R.-J. (2021). Arru phase picker: Attention recurrent-residual U-Net for picking seismic P- and S-phase arrivals. *Seismological Research Letters*, *92*(4), 2410–2428. https://doi.org/10.1785/0220200382

Lieser, K., Grevemeyer, I., Lange, D., Flueh, E., Tilmann, F., & Contreras-Reyes, E. (2014). Splay fault activity revealed by aftershocks of the 2010 Mw 8.8 Maule earthquake, central Chile. *Geology*, *42*(9), 823–826. https://doi.org/10.1130/G35848.1

Liu, M., Zhang, M., Zhu, W., Ellsworth, W. L., & Li, H. (2020). Rapid characterization of the July 2019 Ridgecrest, California, earthquake sequence from raw seismic data using machine-learning phase picker. *Geophysical Research Letters*, *47*(4), e2019GL086189. https://doi.org/10.1029/2019GL086189

Lomax, A., Satriano, C., & Vassallo, M. (2012). Automatic picker developments and optimization: FilterPicker–a robust, broadband picker for real-time seismic monitoring and earthquake early warning. *Seismological Research Letters*, *83*(3), 531–540. https://doi.org/10.1785/gssrl.83.3.531

Lomax, A., Virieux, J., Volant, P., & Berge-Thierry, C. (2000). Probabilistic earthquake location in 3D and layered models: Introduction of a metropolis-Gibbs method and comparison with linear locations. *Advances in Seismic Event Location*, 101–134.

Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025.

Mousavi, S. M., Ellsworth, W. L., Zhu, W., Chuang, L. Y., & Beroza, G. C. (2020). Earthquake transformer—An attentive deep-learning model for simultaneous earthquake detection and phase picking. *Nature Communications*, *11*(1), 3952. https://doi.org/10.1038/s41467-020-17591-w

Mousavi, S. M., & Beroza, G. C. (2023). Machine learning in earthquake seismology. *Annual Review of Earth and Planetary Sciences*, *51*(1), 105–129. https://doi.org/10.1146/annurev-earth-071822-100323

Münchmeyer, J., Bindi, D., Leser, U., & Tilmann, F. (2020). The transformer earthquake alerting model: A new versatile approach to earthquake early warning. *Geophysical Journal International*, *225*(1), 646–656. https://doi.org/10.1093/gji/ggaa609

Münchmeyer, J., Bindi, D., Leser, U., & Tilmann, F. (2021). Earthquake magnitude and location estimation from real time seismic waveforms with a transformer network. *Geophysical Journal International*, *226*(2), 1086–1104. https://doi.org/10.1093/gji/ggab139

Münchmeyer, J., Woollam, J., Rietbrock, A., Tilmann, F., Lange, D., Bornstein, T., et al. (2022). Which picker fits my data? A quantitative evaluation of deep learning based seismic pickers. *Journal of Geophysical Research: Solid Earth*, *127*(1), e2021JB023499. https://doi.org/10.1029/2021JB023499

Nabelek, J., & Braunmiller, J. (2012). Plate boundary evolution and physics at an oceanic Transform Fault System [Dataset]. International Federation of Digital Seismograph. https://doi.org/10.7914/SN/X9_2012

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359. https://doi.org/10.1109/tkde.2009.191

Petersen, F., Lange, D., Ma, B., Grevemeyer, I., Geersen, J., Klaeschen, D., et al. (2021). Relationship between subduction erosion and the up-dip limit of the 2014 Mw 8.1 Iquique earthquake. *Geophysical Research Letters*, *48*(9), e2020GL092207. https://doi.org/10.1029/2020GL092207

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241).

Ross, Z. E., Meier, M., Hauksson, E., & Heaton, T. H. (2018). Generalized seismic phase detection with deep learning. *Bulletin of the Seismological Society of America*, *108*(5A), 2894–2901. https://doi.org/10.1785/0120180080

Ruppert, N. A., Barcheck, G., & Abers, G. A. (2022). Enhanced regional earthquake catalog with Alaska amphibious community seismic experiment data. *Seismological Research Letters*, *94*(1), 522–530. https://doi.org/10.1785/0220220226

Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., & Rueckert, D. (2019). Attention gated networks: Learning to leverage salient regions in medical images. *Medical Image Analysis*, *53*, 197–207. https://doi.org/10.1016/j.media.2019.01.012

Sleeman, R., & Van Eck, T. (1999). Robust automatic P-phase picking: An on-line implementation in the analysis of broadband seismogram recordings. *Physics of the Earth and Planetary Interiors*, *113*(1–4), 265–275. https://doi.org/10.1016/S0031-9201(99)00007-2

Sleeman, R., & van Eck, T. (2003). Single station real-time P and S phase pickers for seismic observatories. In T. Takanami & G. Kitagawa (Eds.), *Methods and applications of signal processing in seismic network operations* (pp. 173–194). Springer. https://doi.org/10.1007/BFb0117702

Soto, H., & Schurr, B. (2021). Deepphasepick: A method for detecting and picking seismic phases from local earthquakes based on highly optimized convolutional and recurrent deep neural networks. *Geophysical Journal International*, *227*(2), 1268–1294. https://doi.org/10.1093/gji/ggab266

Thierer, P. O., Flueh, E. R., Kopp, H., Tilmann, C., & Contreras, S. (2005). Local earthquake monitoring offshore Valparaiso, Chile. *Neues Jahrbuch für Geologie und Paläontologie - Abhandlungen*, *236*(1–2), 173–183. https://doi.org/10.1127/njgpa/236/2005/173

Tilmann, F., Craig, T. J., Grevemeyer, I., Suwargadi, B., Kopp, H., & Flueh, E. (2010). The updip seismic/aseismic transition of the Sumatra megathrust illuminated by aftershocks of the 2004 Aceh-Andaman and 2005 Nias events. *Geophysical Journal International*, *181*, 1261–1274. https://doi.org/10.1111/j.1365-246X.2010.04597.x

Tilmann, F., Flueh, E., Planert, L., Reston, T., & Weinrebe, W. (2004). Microearthquake seismicity of the mid-Atlantic ridge at 5°S: A view of tectonic extension. *Journal of Geophysical Research*, *109*(B6), B06102. https://doi.org/10.1029/2003JB002827

Tilmann, F., Grevemeyer, I., Flueh, E. R., Dahm, T., & Goßler, J. (2008). Seismicity in the outer rise offshore southern Chile: Indication of fluid effects in crust and mantle. *Earth and Planetary Science Letters*, *269*(1–2), 41–55. https://doi.org/10.1016/j.epsl.2008.01.044

Wallace, L., Sheehan, A., Schwartz, S., & Webb, S. (2014). *Hikurangi ocean bottom investigation of tremor and slow slip*. International Federation of Digital Seismograph Networks. https://doi.org/10.7914/SN/YH_2014

Wang, J., Xiao, Z., Liu, C., Zhao, D., & Yao, Z. (2019). Deep learning for picking seismic arrival times. *Journal of Geophysical Research: Solid Earth*, *124*(7), 6612–6624. https://doi.org/10.1029/2019JB017536

Woollam, J., Münchmeyer, J., Tilmann, F., Rietbrock, A., Lange, D., Bornstein, T., et al. (2022). SeisBench—A toolbox for machine learning in seismology. *Seismological Research Letters*, *93*(3), 1695–1709. https://doi.org/10.1785/0220210324

Wu, X., Huang, S., Xiao, Z., & Wang, Y. (2022). Building precise local submarine earthquake catalogs via a deep-learning-empowered workflow and its application to the challenger deep. *Frontiers in Earth Science*, *10*, 817551. https://doi.org/10.3389/feart.2022.817551

Xiao, Z., Wang, J., Liu, C., Li, J., Zhao, L., & Yao, Z. (2021). Siamese earthquake transformer: A pair-input deep-learning model for earthquake detection and phase picking on a seismic array. *Journal of Geophysical Research: Solid Earth*, *126*(5), e2020JB021444. https://doi.org/10.1029/2020JB021444

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 1480–1489).

Yarce, J., Sheehan, A., Nakai, J., Schwartz, S., Mochizuki, K., Savage, M., et al. (2019). Seismicity at the northern Hikurangi Margin, New Zealand, and investigation of the potential spatial and temporal relationships with a shallow slow slip event. *Journal of Geophysical Research: Solid Earth*, *124*(5), 4751–4766. https://doi.org/10.1029/2018jb017211

Zhu, W., & Beroza, G. C. (2019). Phasenet: A deep-neural-network-based seismic arrival-time picking method. *Geophysical Journal International*, *216*(1), 261–273. https://doi.org/10.1093/gji/ggy423

Zhu, W., McBrearty, I. W., Mousavi, S. M., Ellsworth, W. L., & Beroza, G. C. (2022). Earthquake phase association using a Bayesian Gaussian mixture model. *Journal of Geophysical Research: Solid Earth*, *127*(5), e2021JB023249. https://doi.org/10.1029/2021jb023249