doi:10.26443/seismica.v3i1.1164

# Effects on a Deep-Learning, Seismic Arrival-Time Picker of Domain-Knowledge Based Preprocessing of Input Seismograms

**Anthony Lomax** *[1], **Matteo Bagagli** [2,3], **Sonja Gaviano** [4,5,7], **Spina Cianetti** [4], **Dario Jozinović** [6], **Alberto Michelini** [3], **Christopher Zerafa** [4], **Carlo Giunchi** [4]

[1]ALomax Scientific, Mouans-Sartoux, France, [2]Dipartimento di Scienze della Terra, Università di Pisa, Via Santa Maria, 53, 56126 Pisa, Italy, [3]Istituto Nazionale di Geofisica e Vulcanologia, Osservatorio Nazionale Terremoti, Via di Vigna Murata, 605, Roma, Italy, [4]Istituto Nazionale di Geofisica e Vulcanologia, Sezione di Pisa, via Cesare Battisti, 53, Pisa, Italy, [5]Dipartimento di Scienze della Terra, Università degli Studi di Firenze, Via La Pira 4, Florence, Italy, [6]Swiss Seismological Service (SED), ETH Zurich, Zurich, Switzerland, [7]Now at [2]

Author contributions: Investigation: M. Bagagli, A. Lomax, S. Gaviano. Methodology: All authors. Software and Data Curation: M. Bagagli. Visualization: M. Bagagli, A. Lomax, S. Gaviano. Conceptualization and Supervision: A. Lomax, A. Michelini, C. Giunchi. Funding Acquisition and Resources: C. Giunchi, S. Cianetti, A. Michelini. Writing – original draft: A. Lomax. Writing – additions, review and editing: all authors..

**Abstract** Automated seismic arrival picking on large and real-time seismological waveform datasets is fundamental for monitoring and research. Recent, high-performance arrival pickers apply deep-neural-networks to nearly raw seismogram inputs. However, there is a long history of rule-based, automated arrival detection and picking methods that efficiently exploit variations in amplitude, frequency and polarization of seismograms. Here we use this seismological domain-knowledge to transform raw seismograms as input to a deep-learning picker. We preprocess 3-component seismograms into 3-component characteristic functions of a multi-band picker, plus modulus and inclination. We use these five time-series as input instead of raw seismograms to extend the deep-neural-network picker PhaseNet. We compare the original, data-driven PhaseNet and our domain-knowledge PhaseNet (DKPN) after identical training on datasets of different sizes and application to in- and cross-domain test datasets. We find DKPN and PhaseNet show near identical picking performance for in-domain picking, while DKPN outperforms PhaseNet for some cases of cross-domain picking, particularly with smaller training datasets; additionally, DKPN trains faster than PhaseNet. These results show that while the neural-network architecture underlying PhaseNet is remarkably robust with respect to transformations of the input data (e.g. DKPN preprocessing), use of domain-knowledge input can improve picker performance.

Production Editor: Gareth Funning
Handling Editor: Yen Joe Tan
Copy & Layout Editor: Théa Ragon

Signed reviewer(s): Mostafa Mousavi

Received: December 6, 2023
Accepted: June 2, 2024
Published: June 20, 2024

**Riassunto** Individuare l'arrivo delle fasi sismiche è fondamentale per il monitoraggio e la ricerca dei terremoti. Attualmente, la maggior parte dei programmi di riconoscimento (pickers) utilizza le deep neural network (DNN) su sismogrammi grezzi. Esistono però decadi di ricerche sul rilevamento automatico degli arrivi sismici basate su variazioni in ampiezza, frequenza e polarizzazione dei sismogrammi (domain-knowledge). Sfruttiamo queste conoscenze per pre-processare i sismogrammi grezzi in cinque serie temporali, ottenendo le tre funzioni caratteristiche di un picker multibanda, il modulo e l'inclinazione. Utilizzando questo nuovo input, realizziamo un'estensione di PhaseNet (PN) basata sulla domain-knowledge (DKPN) e confrontiamo i due modelli (PN e DKPN), addestrandoli su stessi dataset di diverse dimensioni. Eseguiamo due test: in-domain (su dati estratti dallo stesso dataset di addestramento) e cross-domain (su dataset diversi). DKPN e PhaseNet mostrano prestazioni quasi identiche per il riconoscimento delle fasi in-domain, mentre DKPN supera PhaseNet per alcuni casi cross-domain, in particolare per dataset di addestramento più piccoli. L'allenamento di DKPN è più veloce di quello di PhaseNet. Questi risultati mostrano che, sebbene l'architettura di rete neurale alla base di PhaseNet sia notevolmente robusta, l'uso di input basati sulla domain-knowledge può migliorare le prestazioni del picker.

**Non-technical summary** Automatic procedures for detecting seismic energy onsets on seismograms are critical for earthquake and environmental monitoring, earthquake and tsunami early-warning, and for fundamental research in seismology and earthquake hazard. Recent, high-performance onset detectors mainly use sophisticated, machine-learning algorithms which are "trained" on large sets of, unprocessed, seismograms. However, there is a long history of rule-based, automated onset detection algorithms in earthquake seismology that efficiently exploit various characteristics of seismogram waveforms. Here we use classical, seismological energy onset detection algorithms to transform seismogram waveforms before input to an established machine-learning onset-detector. We compare this extended detector with the original detector using identical training seismograms and application to diverse test seismograms. We find that the extended detector shows improved performance when applied to seismograms with different characteristics from those used for training, and can allow use of smaller datasets during training. The results show that the established machine-learning detector performs well independent of transformations of the input data, but that such transformations can improve performance and efficiency in some cases.

# 1 Introduction

Automated seismic arrival pickers are algorithms for detection, onset timing, phase type identification, and other characterization of seismic energy arrivals on seismogram waveforms. These pickers are fundamental for earthquake and environmental monitoring, earthquake and tsunami early-warning, arrival-time tomography, subsurface characterization, and for basic research of earthquakes and their hazard. Highly efficient, accurate and consistent automated picking is necessary for processing large to massive datasets with many data channels, high sampling rates or long recording periods, for real-time monitoring and warning, and for analyzing highly productive aftershock sequences and swarms.

For some years, automated seismic arrival picking algorithms have been developed using machine learning (Enescu, 1996; Dai and MacBeth, 1995; Wang and Teng, 1995; Mousset et al., 1996; Gentili and Michelini, 2006; Beyreuther et al., 2012; Kong et al., 2018). Recent, high-performance machine-learning pickers are based mainly on deep-neural-networks (LeCun et al., 2015) and are data-driven—trained and applied to nearly raw, seismogram waveforms as input features (Liao et al., 2021; Mousavi et al., 2019, 2020; Mousavi and Beroza, 2022; Münchmeyer et al., 2022; Ross et al., 2018a; Soto and Schurr, 2021; Woollam et al., 2019; Yu and Wang, 2022; Zhu and Beroza, 2018). There is, however, a long history of automated, seismic arrival detection, onset-timing and phase identification algorithms which efficiently exploit variations in amplitude, frequency and polarization of seismogram waveforms (Stevenson, 1976; Allen, 1978, 1982; Bai, 2000; Bagagli, 2022; McEvilly and Majer, 1982; Baer and Kradolfer, 1987; Withers et al., 1998; Lomax et al., 2012). These "classical" pickers are typically composed of rule-based algorithms defined by experts, using seismological domain knowledge to perform processing of seismogram waveforms, and subsequent analyses for detection, onset-timing and phase identification. Various parameters of the pickers are adjusted through trial-and-error or formal optimization (e.g., machine-learning Vassallo et al., 2012) to produce results best matching manually determined or other reference picks. Detections from classical pickers can also be fed into deep-neural-networks to further refine the arrival timing and characterize the picks (Yeck et al., 2020).

Incorporating expert, domain knowledge in the feature engineering and training of deep-neural-networks has been proposed and shown to improve performance over pure data-driven learning (Marcus, 2018; Borghesi et al., 2020; Jozinović et al., 2021; Kong et al., 2018; Mousavi and Beroza, 2022; Muralidhar et al., 2018), especially when there is limited or poor quality training data, or with difficult learning tasks. A basic question then arises: should the expert, seismological domain-knowledge components of classical pickers be discarded when developing high-performance, deep-neural-network pickers? This may be the case if the deep-neural-networks have a sufficiently large and complex architecture to learn to accurately map key characteristics of seismogram waveforms and phase onset energy into detections and picks, especially for cross-domain application. Such learning requires large, comprehensive and high-quality training datasets, and adequate computing resources to train the network. Otherwise, the use of domain-knowledge for preprocessing input seismogram waveforms may improve the performance, or even make viable, the training and application of deep-neural-network pickers, particularly for small or poor quality training datasets, when computing resources are limited, or for urgent analysis.

Most classical, automated pickers first convert raw seismograms into characteristic function (CF) time-series which greatly amplify the main features of seismic phase arrivals, such as abrupt changes in amplitude, frequency, or polarization of the waveforms. Secondly, these pickers analyze the CFs to detect and determine onset-times, phase types and other characteristics of possible seismic energy arrivals while ignoring background variations in signal. Conversion of raw seismograms into picker CFs is commonly and most basically performed with mean removal and high-pass filtering, followed by sliding-window calculation of a short-term average (STA) and a long-term average (LTA) of the signal amplitude to form the CF based on the ratio STA/LTA (Allen, 1978, 1982; Baer and Kradolfer, 1987). Additional or alternative processing of seismograms for detection, time picking or phase identification include autoregressive analysis (Sleeman and van Eck, 1999), particle-motion and polarization analysis (Vidale, 1986; Bai, 2000; Plešinger et al., 1986; Anant and Dowla, 1997; Ross and Ben-Zion, 2014), vector modulus (Bai, 2000), and time-frequency domain, spectrogram (Lomax et al., 2012; Alvarez et al., 2013; Njirjak et al., 2022) or wavelet analyses (Anant and Dowla, 1997; Zhang et al., 2003; Mousavi et al., 2016). FilterPicker (Lomax et al., 2012) constructs a picker CF through application of an STA/LTA algorithm to a series of band-pass filtered seismograms, equivalent to a simplified spectrogram representation of the raw waveforms. The multi-band nature of FilterPicker enables picking of seismic onsets over a range of dominant frequencies in the presence of signal offset and high noise, improving correct detection of true seismic phase onsets even in complex waveforms.

Here we examine changes in the performance of a deep-learning picker when its raw seismogram input is modified using seismological domain-knowledge from classical pickers. In a manner similar to Gentili and Michelini (2006) for picking and Wang and Teng (1995) and Njirjak et al. (2022) for earthquake detection, we preprocess 3-component, broadband seismograms into a set of 5 input time-series: the 3-component characteristic functions of the multi-band FilterPicker, plus the instantaneous modulus and inclination of the waveforms from particle-motion analysis. This preprocessing increases the dimensionality of the input data. We use these 5 time-series as input features instead of 3-component, raw seismograms to extend the deep-neural-network picker PhaseNet (Zhu and Beroza, 2018) within the SeisBench platform (Woollam et al., 2022).

*Corresponding author: anthony@alomax.net

We compare the original PhaseNet and our extended, domain-knowledge PhaseNet (DKPN) using identical training, validation and test datasets and identical processing pipelines. We train PhaseNet and DKPN on waveforms from the INSTANCE dataset (Michelini et al., 2021). The training is run on 7 subsets with different sizes, leading to a set of trained model variants. We apply this set of trained models to in-domain test datasets from INSTANCE and to cross-domain test data from two different datasets available in SeisBench: ETHZ from the Swiss Seismological Service, and PNW from seismic networks in the US Pacific Northwest (Ni et al., 2023).

The use of domain-knowledge input in DKPN instead of near-raw waveforms for the PhaseNet deep-learning picker requires slightly more computing time for training and application, due to the additional waveform preprocessing, though data preprocessing, code optimization, multi-processing and use of GPUs effectively removes this time penalty. For P and S arrivals the explicit information targeting detection and picking of seismic energy arrivals introduced by the domain-knowledge processing enables DKPN to reach higher performance than PhaseNet for cross-domain application, especially with smaller training datasets. In contrast, DKPN and PhaseNet perform nearly identically for in-domain P and S picking. These results suggest that the underlying PhaseNet architecture can robustly learn arrival detection and pick characterization somewhat independently of the form of the input data, presumably as long as key information relevant to arrival detection and picking remains present in the input, as is the case with DKPN.

## 2 Data and Methods

We use the SeisBench machine learning toolbox (Woollam et al., 2022) to access seismogram waveform datasets and the PhaseNet deep-neural-network picker model, and as a general platform for data processing and augmentation operations.

### 2.1 PhaseNet, a deep-neural-network picker

PhaseNet (Zhu and Beroza, 2018) is a deep-neural-network algorithm for probabilistic detection, onset-timing and phase-type identification of seismic P and S arrivals.A trained deep-neural-network can be interpreted as a very high-dimensional approximation function composed of many, local mappings of input to output (Balestriero and Baraniuk, 2018). These mappings produce increasingly abstract layers which preserve only essential information in the data needed for a target regression or classification task (LeCun et al., 2015). Indeed, in this study we are investigating the effects of using essential information for picking as input and thus potentially reducing the amount of network training needed to identify and isolate essential information.

The input for PhaseNet are 3-component, broadband, seismogram waveforms of 3001 samples (30 sec at 100 Hz sampling) with minimal preprocessing (mean removal and normalization). These input data are processed through a modified U-net (Ronneberger et al., 2015; Zhu and Beroza, 2018, , their figure 5) with 4 stages of down-sampling and reduction in number of nodes based on 1-D convolution followed by 4 stages of near-symmetric up-sampling and expansion based on 1-D deconvolution (Zhu and Beroza, 2018). Direct, skip connections between corresponding down- and up-sampling layers help to improve training convergence. After the last stage of down-sampling, the input is reduced to 22 points x 12 channels, which, considering the size of the convolutional kernel , implies the network has a broad receptive field on the original seismograms (Zhu and Beroza, 2018) which is about 26 sec (Hien, 2018). At the end of up-sampling, PhaseNet outputs 3 channels of 3001 points: prediction probability distributions for P and S arrivals and for noise, time-aligned to the original 3001 input samples. Here we derive arrival picks from peaks in the P and S probabilities through rule-based post-processing to give pick arrival time (at peak maximum), and confidence (peak amplitude)

In Zhu and Beroza (2018) PhaseNet is trained on a dataset of detected earthquakes from Northern California composed of 623,054 3-component recordings, all of which have manually picked, P and S arrival times. Through various experiments, Zhu and Beroza (2018) conclude that PhaseNet achieves much higher picking accuracy and recall rate than a classical STA/LTA plus autoregressive analysis method (Akazawa, 2004) when applied to the waveforms of known earthquakes, with a particularly pronounced improvement in S picking performance.

### 2.2 Modified FilterPicker characteristic functions

FilterPicker (Lomax et al., 2012) applies an adaptation of the STA/LTA algorithm of Baer and Kradolfer (1987) to construct a picker CF from a succession of band-pass filtered seismograms. Seismogram waveforms, with little or no preprocessing, pass through a pipeline of: 1) differentiation, 2) band-pass filtering at a geometric progression of center periods ranging from the sampling interval to the longest period signal to be picked (e.g., for sampling interval 0.01sec, 7 bands at 0.01, 0.02, 0.04, 0.08, 0.16, 0.32, 0.64 sec center periods), 3) squaring of each band-pass series to form a positive, envelope function, 4) forming a CF for each band as the ratio of instantaneous deviation from long-term mean to long-term standard deviation of the band envelope, and 5) forming a definitive, summary CF from the maximum of the band CFs at each sample point. The succession of band-pass filtered seismograms are equivalent to a simplified time-frequency, spectrogram representation of the raw waveforms. The multi-band nature of FilterPicker provides a strong response in the summary CF for seismic onsets with a range of dominant periods, even in the presence of strong, narrow-band noise, of strong microseismic or other, longer period noise, and of offset signals. See Lomax et al. (2012) for details and examples.

We implement FilterPicker within the data augmentation step of SeisBench processing using the FBpicker, Python implementation of FilterPicker from

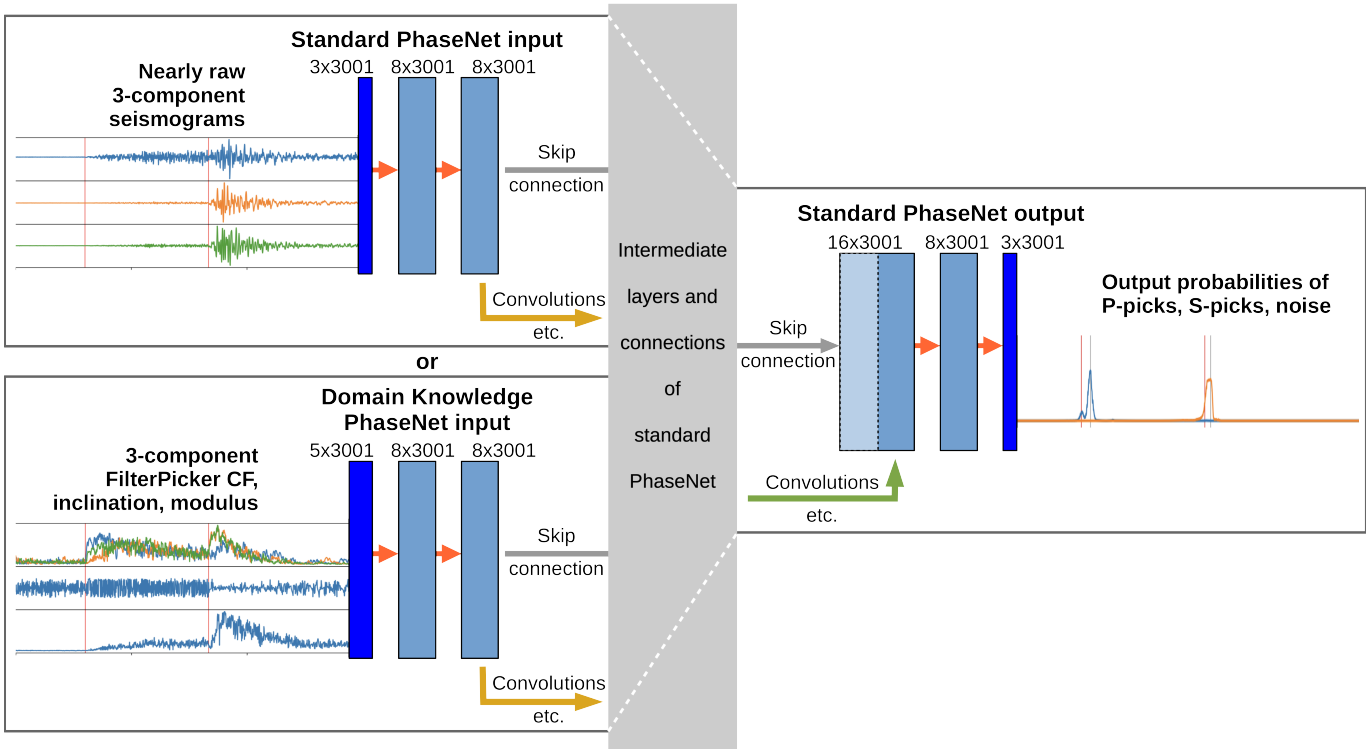


**Figure 1** Standard PhaseNet and modified Domain Knowledge PhaseNet deep-learning architecture. The DKPN model replaces the nearly-raw, 3-component seismogram input of PhaseNet with the 3-component FilterPicker CF plus instantaneous modulus and inclination traces (bottom left panel). Otherwise, the layers, connections and output for both models are identical. For details of the full PhaseNet architecture, symbols and color codes, see Zhu and Beroza (2018, ; their figure 4).

the PhasePApy package (Chen and Holland, 2016). We modify the FBpicker algorithm to replace its sliding, fixed-window procedure for generating band CFs with the decay-constant, recursive filter procedure of the original FilterPicker. We further modify the resulting FilterPicker algorithm by taking the logarithm of the summary CF to compress high amplitudes in the CF at strong arrival onsets and thus avoid the need for a cut-off parameter (Lomax et al., 2012) to limit the highly-variable maximum CF values. Finally, we normalize the CFs with the maximum standard-deviation of the 3-component CFs. As we do not use the detection and pick characterization logic of FilterPicker, there are only two primary picker parameters: a filter window defining the frequency of the longest period band, and the long-term, time-averaging scale for recursive band-pass filtering. In this study we set FilterPicker parameters following the guidelines and defaults in Lomax et al. (2012), with some trial-and-error over a limited range of typical values for broadband, local and regional event picking.

## 2.3 Instantaneous modulus and inclination

In addition to the modified, 3-component CF waveforms from FilterPicker, we also calculate two waveforms consisting of quantities from particle-motion analysis, the instantaneous modulus and inclination, using the FilterPicker, band-pass filtered seismograms. In order to suppress response to background noise, both quantities are calculated independently at each sample point, using the 3-component Z, N, E data values on the band-pass waveform corresponding to the maximum band CF for the sample point. The modulus is the length of the 3-component data vector, $\sqrt{(Z^2 + N^2 + E^2)}$, normalized by dividing by the maximum standard-deviation over all sample points. The inclination or incidence angle is given by $\tan^{-1}[Z / \sqrt{(N^2 + E^2)}] / \pi$, where dividing by $\pi$ normalizes to a range of [-1,1] so -1, 0 and 1 correspond to down, horizontal and up inclination, respectively.

We expect these two additional waveform inputs will help the picker to recognize and discriminate between P and S phases in the stream, especially given that the replacement of raw waveforms with CF's entails a loss of information. Modulus re-introduces absolute amplitude information which might aid in discrimination of, for example, a typically higher amplitude S arrival from a lower amplitude P arrival. Inclination re-introduces polarization information which can indicate the type of arrival since P waves usually exhibit dominantly vertical particle motion and S waves often show stronger horizontal motion.

## 2.4 Domain Knowledge PhaseNet

DKPN is our extension of the PhaseNet picker model to use classical, seismological domain-knowledge as input features. We replace the 3-component, raw seismogram input to the PhaseNet model with the 3-component CFs of the modified FilterPicker and the instantaneous modulus and inclination of the waveforms (Figures 1 and 2-5). This entails that in the first convolution plus rectified linear unit step in the first layer, instead of transforming a 3x3001 input dimension to 8x3001 dimension

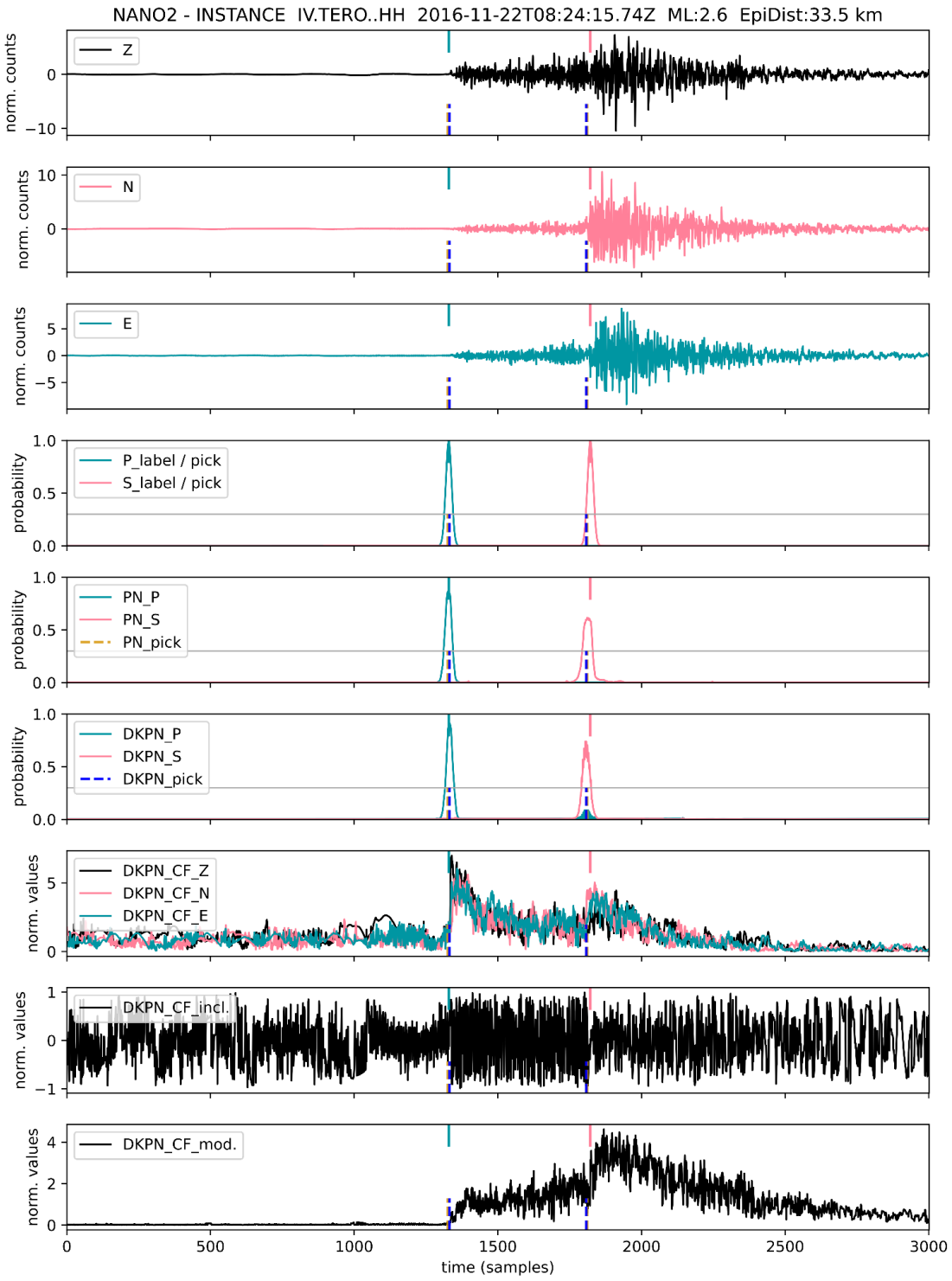


**Figure 2** Example processing and picking results for models trained with the INSTANCE NANO2 dataset and applied to INSTANCE, ETHZ, and PNW trace samples. In each plot, the title indicates INSTANCE training dataset size, test dataset, network-station-location-channel code, event date, magnitude and distance from station. Subplots show: (rows 1-3) de-meaned and normalized Z, N, and E component, observed seismograms; (row 4) P (blue-green) and S (light red) label data; (rows 5-6) PhaseNet (PN) and DKPN P (blue-green) and S (light-red) probabilistic predictions, gray horizontal lines indicate threshold 0.3; (row 7) normalized, DKPN FilterPicker Z, N, E CF's; (rows 8-9) normalized, DKPN instantaneous inclination and modulus. Solid bars at top of each subplots show P (blue-green) and S (light red) label picks. Dashed bars at the bottom of each subplot show PhaseNet (yellow) and DKPN (blue) P and S predicted picks for threshold 0.3. Horizontal axis shows sample count. Seismogram from the INSTANCE dataset (in-domain testing) for which DKPN and PhaseNet both pick P and S strongly near the label times. Note the sharp, strong P onset and emergent S onset in the DKPN CFs, the change in character between P and S of the DKPN inclination, and the clear P and S onsets in the DKPN modulus; these features show the introduced domain knowledge which drives the DKPN picks.

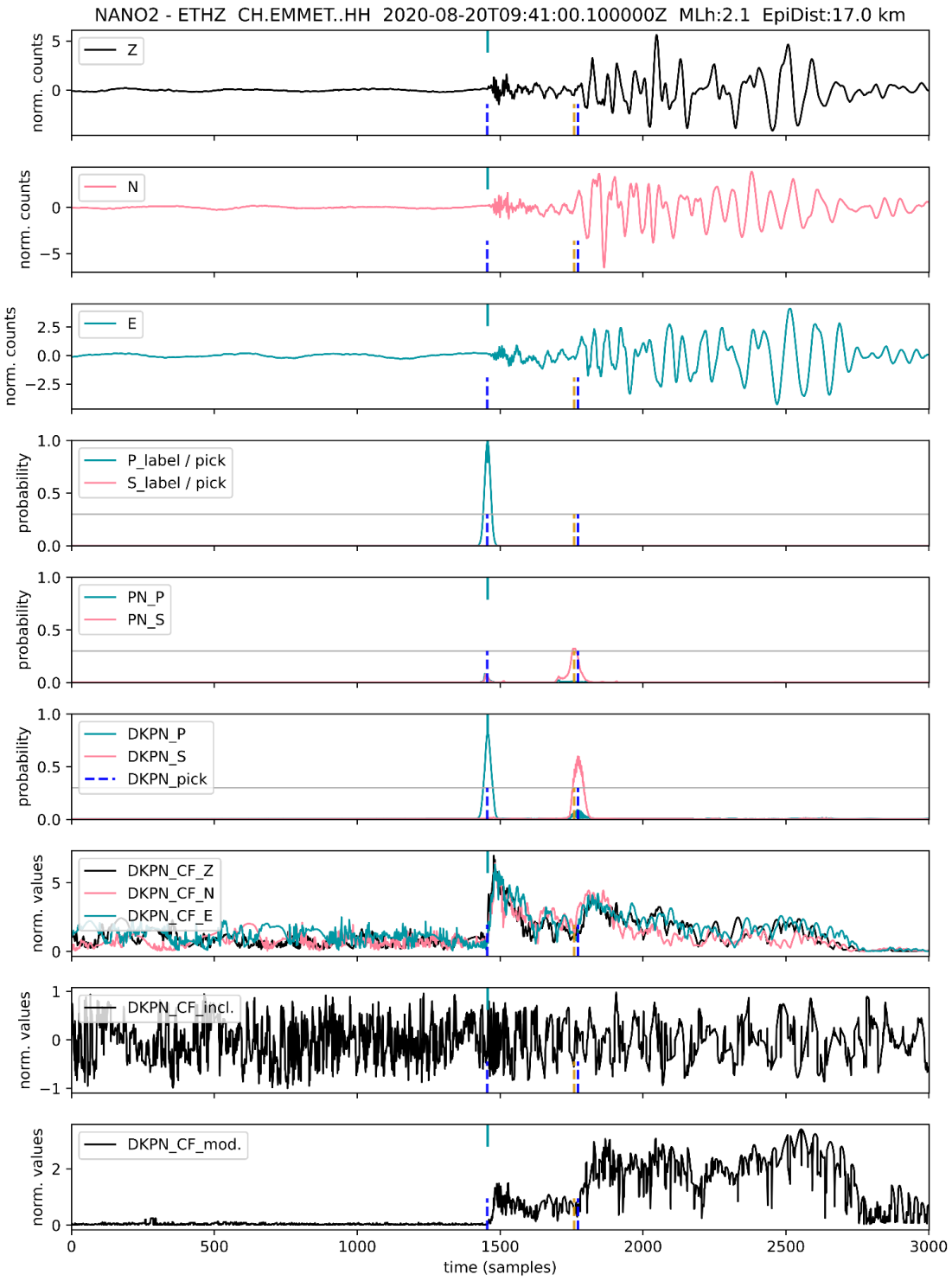


**Figure 3** Same as Figure 2, for a seismogram from the ETHZ dataset (cross-domain testing) for which a clear, high-frequency, labeled P arrival is matched by DKPN but not PhaseNet. DKPN strongly, and PhaseNet more weakly both identify an unlabeled S arrival, which is likely correct given the waveforms and epicentral distance. The relatively poor performance of PhaseNet for this seismogram may be due to the waveform arrivals (e.g. short, high-frequency P signal and long duration, long-period S signal) differing significantly from arrival waveforms in the INSTANCE dataset used for training.

as in standard PhaseNet (Zhu and Beroza, 2018, ; their figure 4), DKPN transforms a 5x3001 input dimension (channels x length) to 8x3001 dimension (Figure 1). We otherwise make no change to the PhaseNet model as implemented in SeisBench (SeisBench v0.3 or later).

## 2.5 Seismogram waveform datasets

In this study we use three benchmark, seismogram waveform datasets provided in SeisBench: INSTANCE (Michelini et al., 2021) composed of ~1.2 million, 3-component waveforms for ~50,000 earthquakes from M 0 to M 6.5 in and around the Italy region (epicentral distances ~0-6°); Eidgenössische Technische

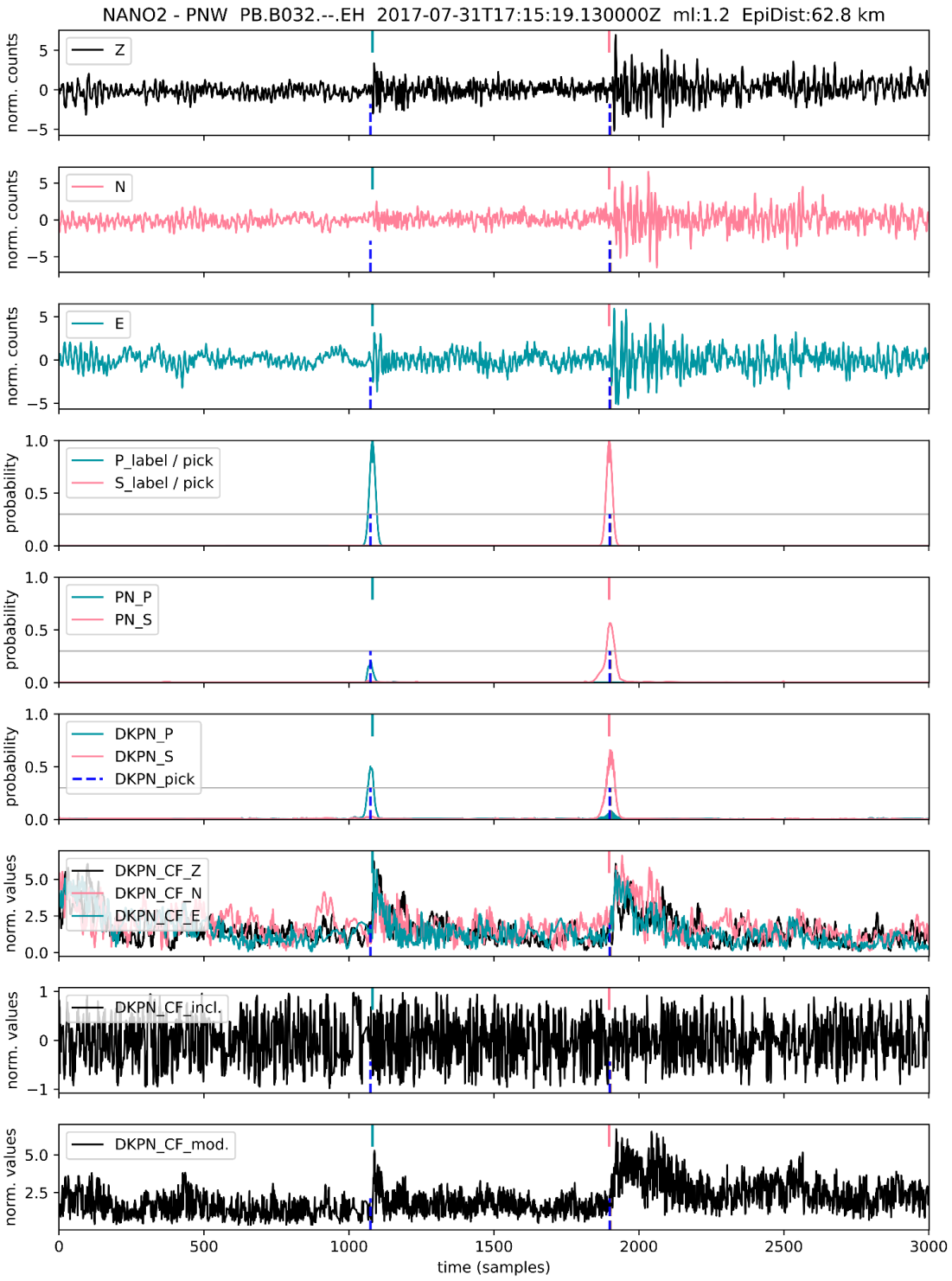


**Figure 4** Same as Figure 2, for a low S/N seismogram from the PNW dataset (cross-domain testing) for which DKPN correctly picks a P arrival with a sharp prediction probability peak, while PhaseNet fails to pick but shows a weak prediction probability peak.

Hochschule Zürich (ETHZ; Woollam et al., 2022) composed of 36,743, 3-component waveforms for 2231 events from M 1.5 to ~M 5 in and around the Switzerland region (epicentral distances ~0-4°); and the Pacific Northwest AI-ready Seismic Dataset (PNW, Ni et al., 2023) composed of ~200,000, 3-component waveforms for ~65,000 events from M 0 to ~M 6.4 in the US Pacific Northwest region with local and regional epicentral distances. A signal-to-noise (S/N) ratio included in the INSTANCE metadata, reported in dB, is calculated (Michelini et al., 2021) from amplitudes in the 5 sec following the S arrival (signal) and 5 sec before the P arrival (noise), though here we do not filter the INSTANCE training datasets on S/N ratio. INSTANCE includes pure noise waveforms, which, following Zhu and Beroza (2018) we do not use for training. For detailed information on filtering criteria for both training and testing stages, the reader is referred to the Supplementary

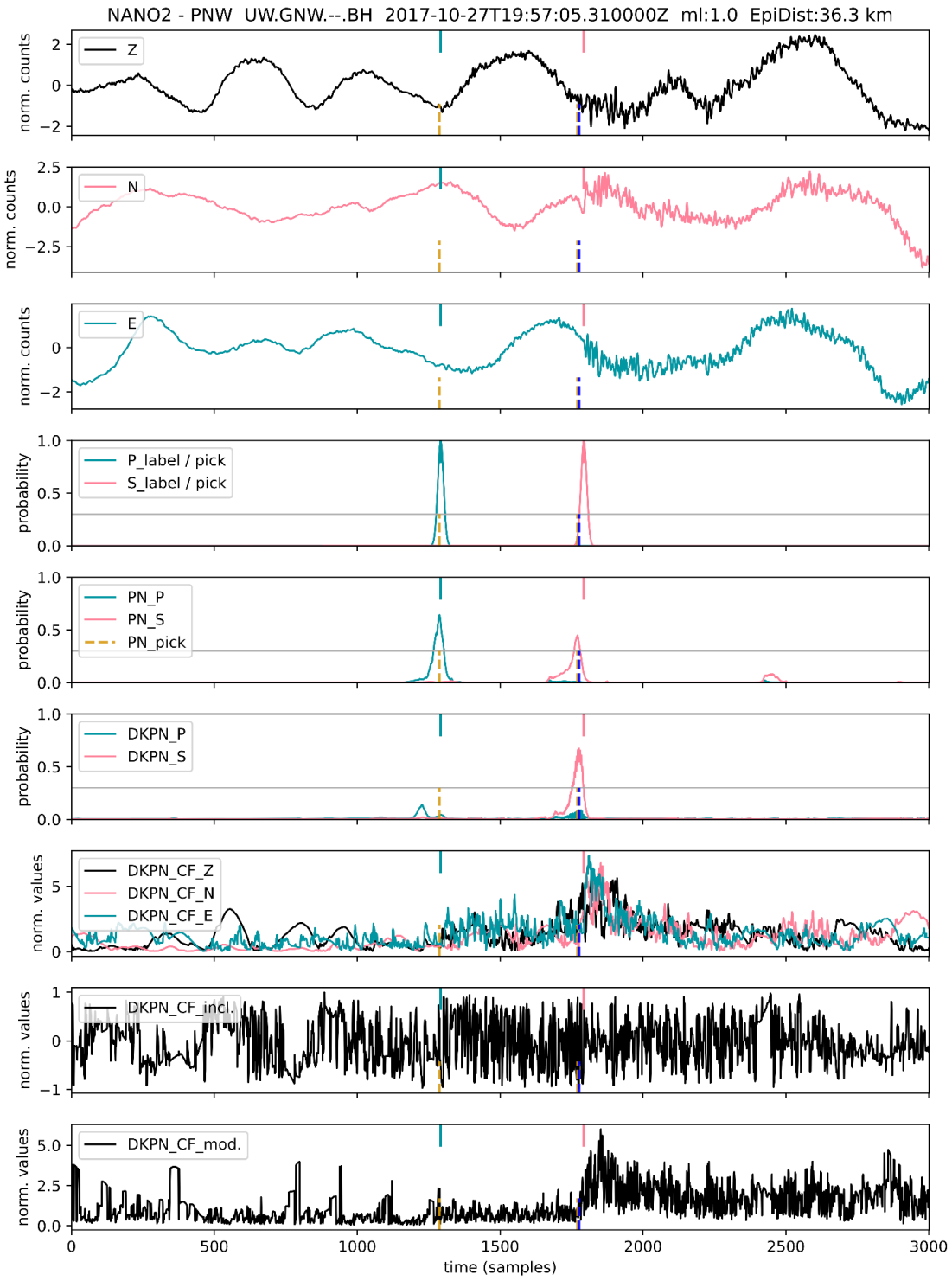


**Figure 5** Same as Figure 2, for a low S/N seismogram from the PNW dataset (cross-domain testing) for which PhaseNet correctly picks a weak P arrival with a sharp prediction probability peak, perhaps by responding to the low amplitude, P coda signal. DKPN fails to pick P, with a very weak prediction probability peak; this failure is likely due to the emergent CF Z component and lack of amplitude increase on the modulus at the P time, both related to the P arrival on the Z seismogram having low amplitude and frequency content similar to the preceding noise.

Text S1 and Table S1.

Visual examination of INSTANCE dataset waveforms and picks shows some traces with: missing S picks; labeled picks on very high or pure noise low-gain data; clearly early or late picks; and unreasonably large pick uncertainty estimates. To mitigate these problems, we filter the dataset metadata to include only high-gain HH channels, events at epicentral distance ≤ 100 km, and, following Zhu and Beroza (2018), traces which have both P and S picks, resulting in ~300,000 3-component trace sets available for training, validation and testing. We also create probabilistic pick labels for training using a fixed sigma of 0.1 sec instead of using the labeled pick uncertainties. See Supplementary File S1 for examples

of filtered INSTANCE waveforms.

In order to evaluate picker performance when different amounts of training data are available, we establish INSTANCE training datasets with different sizes ranging from insufficient for training convergence and stability (NANO3, ~900 samples), to minimal for convergence (NANO2, ~1.6k samples), through intermediate sizes (NANO1, ~3k samples; NANO, ~6k samples; MICRO, ~12k samples; TINY, ~24k samples; SMALL, ~61k samples; MEDIUM, ~163k samples) to much larger than needed for apparent convergence for both pickers (LARGE, ~245k samples). In the following we focus on results for the NANO2, MICRO and MEDIUM training datasets as representative of the main evolutions and features of picker performance across all training dataset sizes.

Missed S picks are also a problem with the ETHZ dataset, but apparently much less so with the PNW dataset. As we do not train on these datasets we do not remove these potential problem traces (examples of used ETHZ and PNW waveforms are shown in Supplementary Files S2 and S3, respectively). However, due to missed S picks, our testing statistics and metrics for S for the ETHZ dataset are likely degraded. Similar and other quality problems are likely present in most seismological waveform datasets for machine-learning, as Münchmeyer et al. (2022) discuss for the datasets provided in SeisBench. Such errors in the data labels will adversely affect the rate and quality of model training and bias the validation and test statistics and metrics, but similarly for the two picker algorithms, thus highlighting their performance in practice.

FilterPicker, like most STA/LTA pickers, requires a minimum length of background data before any phase arrivals for statistical stabilization; this length is controlled by the long-term window parameter. For machine learning training in general, an ample length of background before arrivals is also needed for random window-shift, data augmentation to enhance generalization in the trained model, and, most importantly, to avoid that first arrivals are near the same window position in all or most training samples. Lack of sufficient background data before arrivals can impair classical methods like STA/LTA in comparisons with machine-learning pickers. The processing workflow described below addresses this required minimum length of background data. We note also that lack of sufficient background data before arrivals can preclude pertinent training and evaluation of machine-learning pickers for real-time and early-warning application, where in practice an almost unbounded amount of data before an arrival is available, and very little data may be available after an arrival onset before reaching the last received data sample.

## 2.6 Dataset and model configuration, processing, training and comparison

To compare the performances of PhaseNet and DKPN on different test sets we configure and preprocess datasets and models, train the models and compare PhaseNet and DKPN P and S arrival predictions, statistics and metrics (see Data and code availability). Care must be taken at the data generation stage to ensure that the waveform time-series has sufficient length before the first label pick for FilterPicker stabilization (FPS); this length should be greater than the number of sample points ($N_{\mathrm{FPS}}$) corresponding to the FilterPicker long-term, time-averaging scale. Figures 2-5 show preprocessing, label and pick prediction waveforms for trace samples from the INSTANCE, ETHZ and dataset.

The configuration and processing workflow includes:

1. Load the requested dataset; set the sampling rate to 100 Hz.
2. Optionally select or mask dataset trace sets on presence of P and S picks, channel code, epicentral distance, or other available meta-data. (Supplementary Text T1)
3. For training, randomly split the data into training, validation and test sets according to a target training set size (e.g. for our INSTANCE MEDIUM training dataset: 50% training, 5% validation and 45% remainder for drawing test samples).
4. Define data generators with identical preprocessing and augmentation steps for training, validation and testing, the principal steps are:
   (a) Get a randomly positioned data window in the input, 3-component seismograms starting at least $N_{\mathrm{FPS}}$ points before the first pick label, and with a length of $N_{\mathrm{FPS}}$ plus the 3001 points required for input to the deep-learning models.
   (b) Normalize to the maximum standard-deviation across the 3-component data.
   (c) For DKPN, apply the processing described in the section "Modified FilterPicker characteristic functions" to generate the 3-component, FilterPicker CF time-series, and apply the processing described in "Instantaneous modulus and inclination" to generate the modulus and inclination time-series. In this study the FilterPicker filter window is set so the longest period signal analyzed is 2 sec, and the long-term, recursive-filter time-averaging scale to 4 sec, giving $N_{\mathrm{FPS}} = 401$ point.
   (d) For DKPN, cut the first $N_{\mathrm{FPS}}$ (after FilterPicker stabilization) for all time-series to form a data window of length 3001 points as required for input to the deep-learning models.
   (e) Create probabilistic, P and S label traces with the same 3001 point window from picks specified in the trace metadata using the probabilistic labeler function in SeisBench. Each P or S pick is summed into the corresponding P or S trace as Gaussian function of amplitude 1.0 and with a fixed variance of 5 points (0.05 sec for 100 Hz data).

The training workflow includes:

1. Model and dataset setup following steps 1-4 of the configuration and processing workflow.

2. Run the training on the training dataset for specified optimizer and loss functions, learning rate and number of epochs. In this study the Adam optimizer (Kingma & Ba, 2017), and a cross-entropy loss function are always used. Train models using an early-stopping approach governed by a *"patience"* parameter indicating the number of epochs to tolerate without improvement, and a fixed "*improvement*" threshold. If the mean validation loss over the last two *patience* length epochs does not exceed the *improvement* threshold compared to the mean development loss over the preceding *patience* epochs, the training process is halted. A second condition for halting the training process is if the validation loss over the last *patience* epochs consistently surpasses the training loss, indicating a potential overfitting tendency. In either halting case, the model weights obtained *patience* epochs before the current epoch are utilized. We use this approach to respect the different, natural learning-time behavior of the 2 algorithms, and to smoothly converge to the best possible minima. For details of the training parameters and loss curves comparisons see Supplementary Text S1, Table S1, and File S4).

The comparison workflow includes:

1. Model and dataset setup following steps 1-4 of the configuration workflow.

2. Load a trained model and apply it to traces drawn from the test dataset to obtain prediction probability distributions for P and S arrivals.

3. Post-process the probabilistic, P and S prediction traces with:

   (a) 3-point smoothing to suppress rapid oscillation,

   (b) pick detection at peaks of amplitude greater than specified thresholds and separated by more than 0.5 sec,

   (c) retain the pick time, $T_p$ and amplitude, $A_p$.

4. Accumulation of the P and S, Gaussian labels and prediction picks for multiple traces from the test dataset for calculation of evaluation statistics and metrics as described in the following.

In this study we repeat the training workflow using 7 different, randomly selected training and validation subsets of traces for each experiment. We therefore obtain 7 different models and 7 sets of test results for each individual training dataset size. We merge these test results to reduce the dependence of testing performance statistics with respect to training dataset selection.

### 2.7 Evaluation statistics and metrics

As a basis for comparison of PhaseNet and DKPN performance on test datasets, and following (Zhu and Beroza, 2018), we count, relative to the labeled P or S arrivals in the test dataset, the number for P or S of correct Gaussian predicted arrivals (true positives; TP), incorrect predicted arrivals (false positives; FP), and no prediction of a labelled arrival (false negatives; FN). There may be multiple FP picks for P or for S on each data window. Here, a smoothed, Gaussian predicted P or S arrival is counted as correct when its peak amplitude, $A_p$, is greater than a specified threshold and the difference, $\Delta T_p$, between its peak time and the time of a label arrival of the same phase is less than 0.1 sec for P and 0.2 sec for S, which has typically noisier onsets than P arrivals. Zhu and Beroza (2018) use $\Delta T_p \leq 0.1$ for P and S, and use a threshold $A_p \geq 0.5$. In this work, we examine a range of thresholds $0.1 \leq A_p \leq 0.9$ to find optimal metrics such as F1 score, which vary with training dataset size, test dataset and for PhaseNet versus DKPN. While use of a low amplitude threshold leads to a higher rate of picking, we find that low amplitude predictions generally correspond to correct arrival picks. Additionally, filtering of a limited number of false picks can be done in phase association and hypocenter location processing stages (Kim et al., 2023). For advanced hypocenter location and other algorithms (e.g., Satriano et al., 2008; Lomax et al., 2014) the peak amplitude $A_p$ can also be used for weighting or selection of picks and some measure of the width of the peak (e.g. at half its height) as a pick uncertainty.

From the TP, FP, FN statistics, we form the following metrics for P and for S:

*Precision*, the proportion of positive arrival predictions that are correct,

$$P = \frac{TP}{TP + FP} \tag{1}$$

*Recall*, the proportion of labeled positives that are correctly predicted,

$$R = \frac{TP}{TP + FN} \tag{2}$$

and the *F1 score*, which balances the often opposing, Precision and Recall metrics through their harmonic mean,

$$F1 = 2x\frac{PxR}{P + R} \tag{3}$$

## 3 Results

We present a set of tests to show and compare the performance of PhaseNet and DKPN for different training dataset sizes applied to in-domain (INSTANCE) and cross-domain (ETHZ and PNW) test datasets.

### 3.1 Test 1 – In-domain

We first compare the in-domain performance of PhaseNet and DKPN across different training dataset

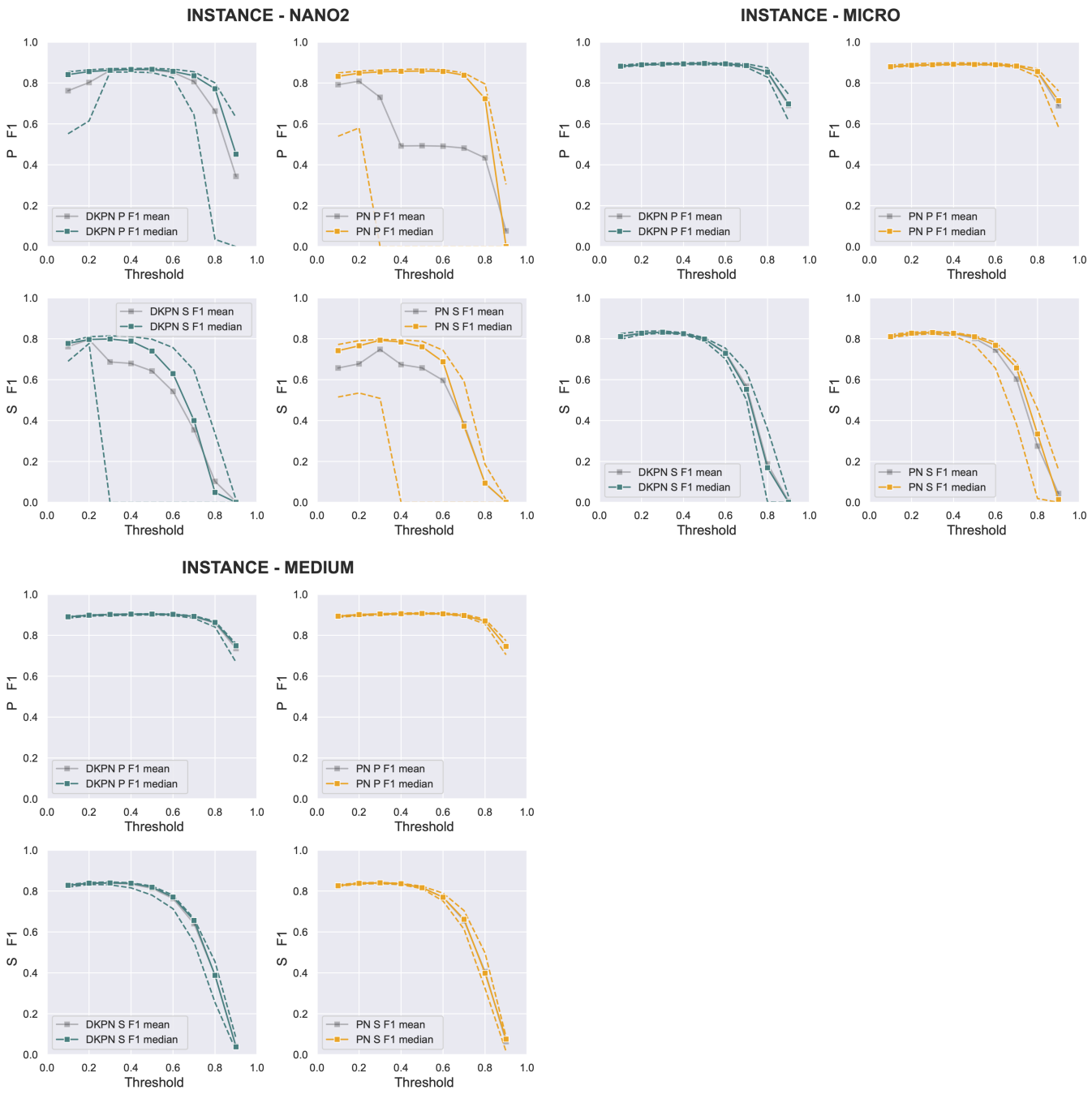


**Figure 6** Comparison of PhaseNet (PN) and DKPN with testing on the INSTANCE dataset (in-domain). F1 score metrics across a range of pick amplitude thresholds for P arrivals and S arrivals for models trained with different size INSTANCE datasets. Mean, median and upper/lower limits of the mean (dashed curves) of F1 for 7 runs each with 5,000 evaluation samples drawn from the test datasets, these samples vary for each training dataset size.

sizes and amplitude thresholds to define correct predicted picks. We train and test with data samples from the INSTANCE dataset. Sample INSTANCE waveforms are presented in Supplementary File S1. The F1 metrics for this test are shown in Figure 6.

For the larger training datasets (e.g. MICRO and MEDIUM) DKPN and PhaseNet show almost identical performance, with F1 scores of about 0.9 for P up to a threshold of about 0.7 and F1 about 0.8 for S up to thresholds of about 0.4 - 0.5. The reduced performance of both pickers for S is almost certainly due to difficulties for both models in detecting and picking the S onset, which is often embedded in the P coda and emergent.

For the smaller NANO2 training dataset, the DKPN and PhaseNet, median P and S F1 scores are slightly reduced relative to those with the larger training datasets. The mean and lower limit of F1 scores, however, show a significant degradation, likely indicating instability in training with small datasets and chance of convergence to an inadequately trained model, even for application to an in-domain dataset.

Histograms of P and S pick residuals (predicted time - label time) for PhaseNet and DKPN for a selection of training data set sizes are shown in Figure 7. For P the total number of predicted picks within twice $\Delta T_p$ is generally similar for PhaseNet and DKPN and independent

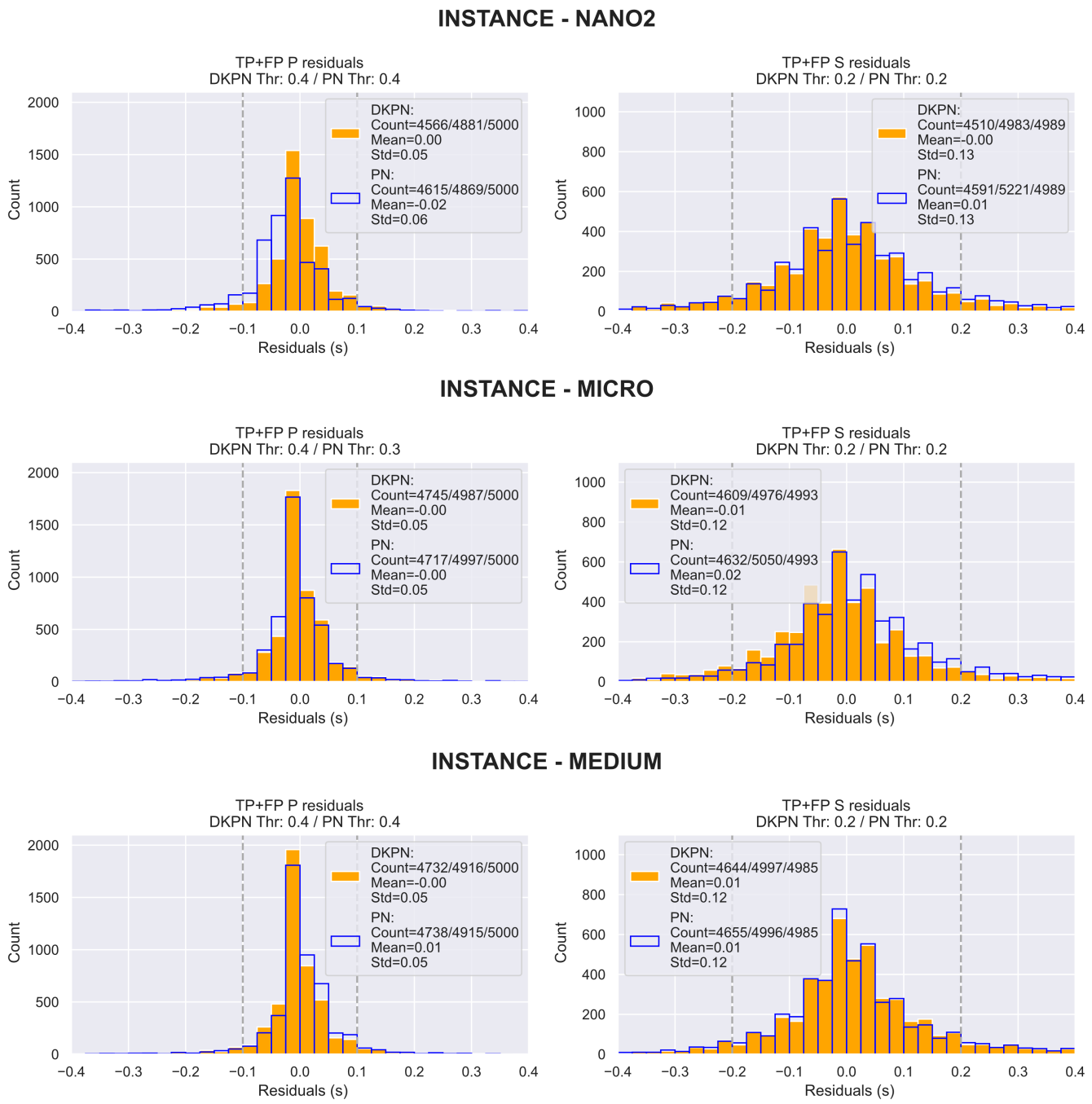


**Figure 7** Histograms of P and S pick residuals for selected INSTANCE training dataset sizes tested on the INSTANCE dataset (in-domain). Results shown for the pick amplitude threshold (Thr) giving the highest F1 score for each case of dataset size and method (DKPN or PhaseNet). Vertical, dashed gray lines show the maximum difference ΔTp between a pick time and the corresponding label time (0.1 sec for P and 0.2 sec for S) to declare a correctly predicted arrival for evaluation statistics (TP). Pick counts show: Number of residuals (number of predicted picks) used in the mean and standard-deviation statistics / Total number of residuals available / Total number of label picks available for the test case. To remove outlier data, the mean and standard-deviation statistics use trimmed residuals, within twice ΔTp: ± 0.2 sec for P and ± 0.4 sec for S.

of training dataset size, and there is little variation in the mean of the residuals, which is always near zero, or for standard deviation with training dataset size. However, the DKPN P residuals are slightly more concentrated and peaked around zero than are the PhaseNet residuals for NANO2. For S the distribution of residuals and statistics for the two pickers are very similar and show little variation with training dataset size, aside from slightly more concentration of residuals around zero for the larger training dataset sizes (e.g. MEDIUM).

Figures 2-5 show test results for PhaseNet and DKPN models trained with the INSTANCE NANO2 dataset and applied to INSTANCE, ETHZ and PNW testing trace samples. These examples illustrate how the FilterPicker CFs, inclination and modulus relate to and help determine the probabilistic P and S predictions and picks for DKPN, and how the amplitude and complexity of probabilistic predictions for PhaseNet and DKPN relate to the trace noise and to the complexity and impulsiveness of arrival onsets. Note in particular how the DKPN CF's for

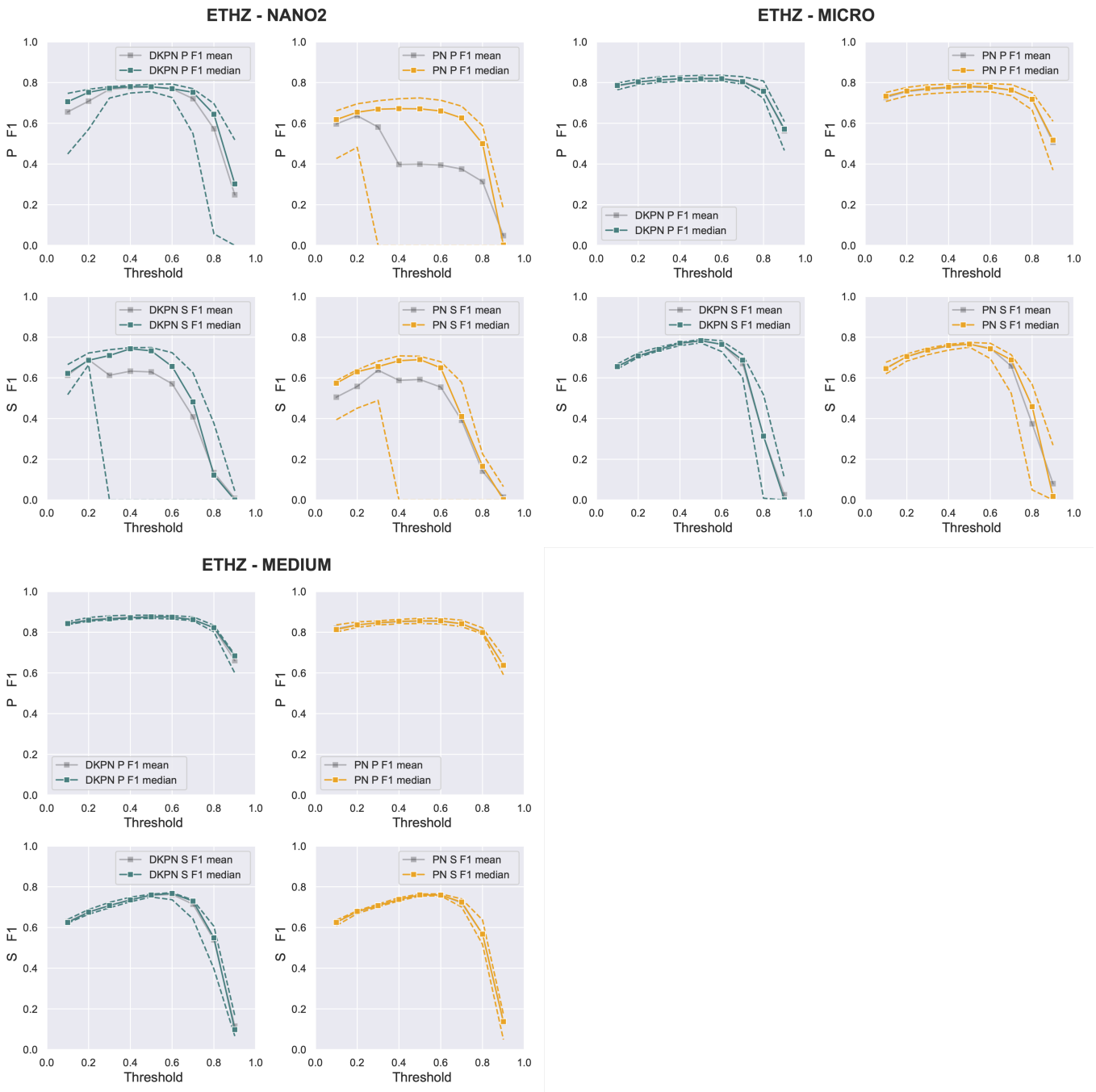


**Figure 8** Comparison of PhaseNet (PN) and DKPN with testing on the ETHZ dataset (cross-domain). F1 score metrics across a range of pick amplitude thresholds for P arrivals and S arrivals for models trained with different size INSTANCE datasets. Mean, median and upper/lower limits of the mean (dashed curves) of F1 for 7 runs each with 5,000 evaluation samples drawn from the test dataset. Because some ETHZ traces have missing P or S picks, FP count may be overestimated.

different events and datasets can have a similar overall form and amplitude even when the corresponding raw seismograms have very different absolute amplitudes and frequency content.

## 3.2 Test 2 – Cross-domain—INSTANCE training and ETHZ test datasets

A most important, general and realistic use case is where a pre-trained picker model will be applied cross-domain—to seismogram waveforms with substantially different characteristics from the training waveforms. The differences may be related to recording instruments, data-loggers, available channel types and gain, wave propagation, site conditions and noise, and the distance, size, stress-drop and other properties of target seismic sources. To examine the cross-domain case, we apply models trained with the INSTANCE datasets of different sizes to testing (i.e. application) on waveforms from the ETHZ and PNW datasets. Sample waveforms are presented in Supplementary Files S2 and S3. The resulting F1 metrics across a range of pick amplitude thresholds for ETHZ test datasets are shown in Figure 8.

For P arrivals, relative to in-domain testing with INSTANCE (Test 1; Figure 6), DKPN shows a small reduction in overall performance and stability (e.g. maxi-

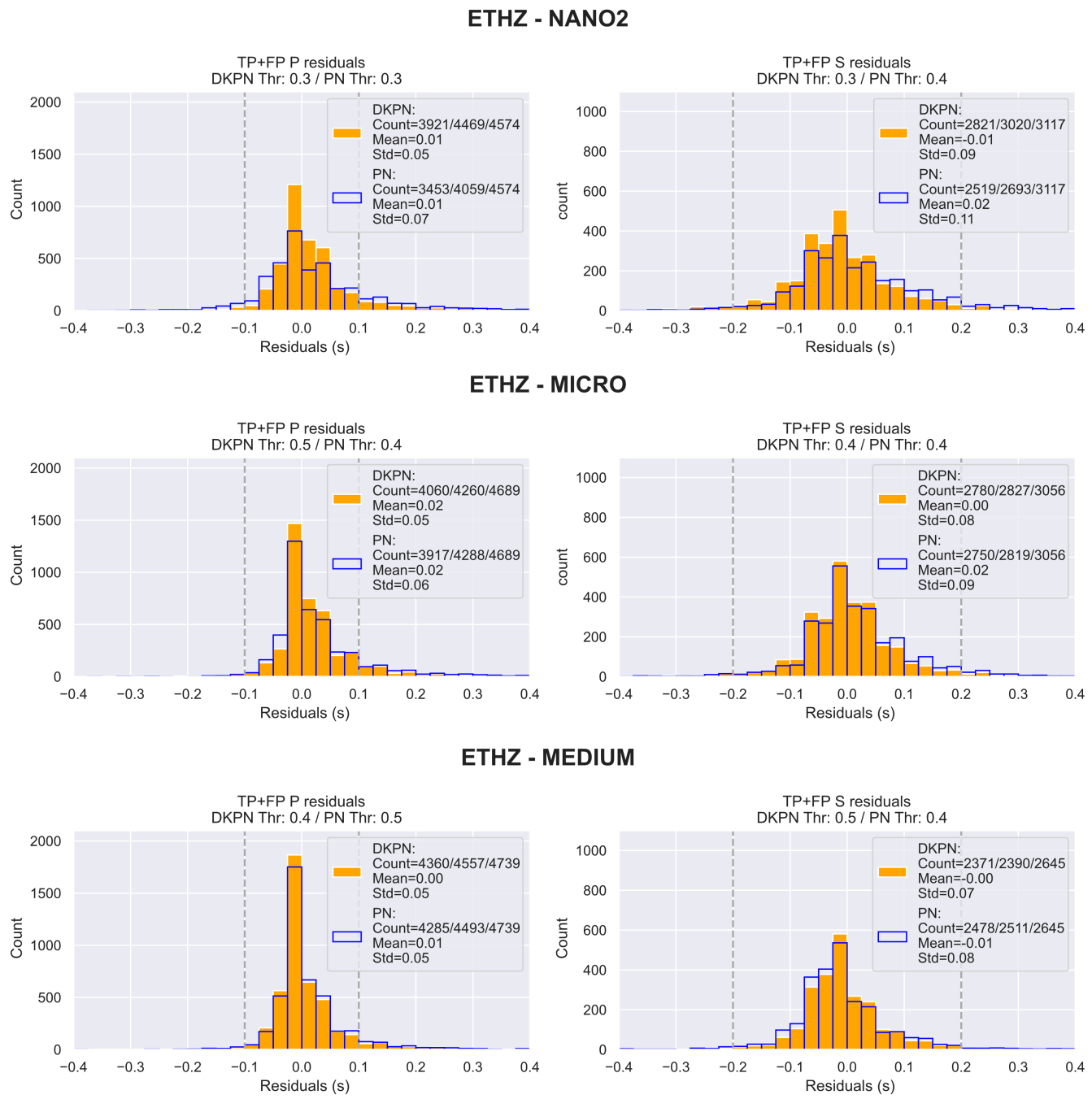


**Figure 9** Histograms of P and S pick residuals for selected INSTANCE training dataset sizes tested on the ETHZ dataset (cross-domain). Results shown for the pick amplitude threshold (Thr) giving the highest F1 score for each case of dataset size and method (DKPN or PhaseNet). Vertical, dashed gray lines show the maximum difference between a pick time and the corresponding label time (0.1 sec for P and 0.2 sec for S) to declare a correctly predicted arrival for evaluation statistics. Pick counts show: Number of residuals (number of predicted picks) used in the mean, mode and standard-deviation statistics (trimmed within twice ΔTp: ± 0.2 sec for P and ± 0.4 sec for S) / Total number of residuals available / Total number of label picks available for the test case.

mum F1 scores almost always ≥ ~0.8 and converging to ~0.9 for larger training datasets) while PhaseNet shows a slightly larger reduction in performance (e.g. maximum F1 scores around 0.7 for the smaller datasets, and converging to ~0.85 for larger datasets). With the smaller training sets (NANO2 and MICRO) DKPN shows better results than PhaseNet, which indicates that, for application to ETHZ data, the DKPN model, with domain-knowledge input processing, has intrinsic properties that improve generalization and effective P picking with cross-domain data, as well as allowing use of smaller training datasets.

For S arrivals (Figure 8), as with Test 1, the performance of DKPN and PhaseNet are notably poorer than for P arrivals, though DKPN performs slightly better for the two smaller datasets. Relative to in-domain testing with INSTANCE (Figure 6), both DKPN and PhaseNet show generally reduced performance (e.g. maximum F1 scores of ~0.7-0.75 instead of ~0.8) except for an increase in Recall, due to a decrease in false negative

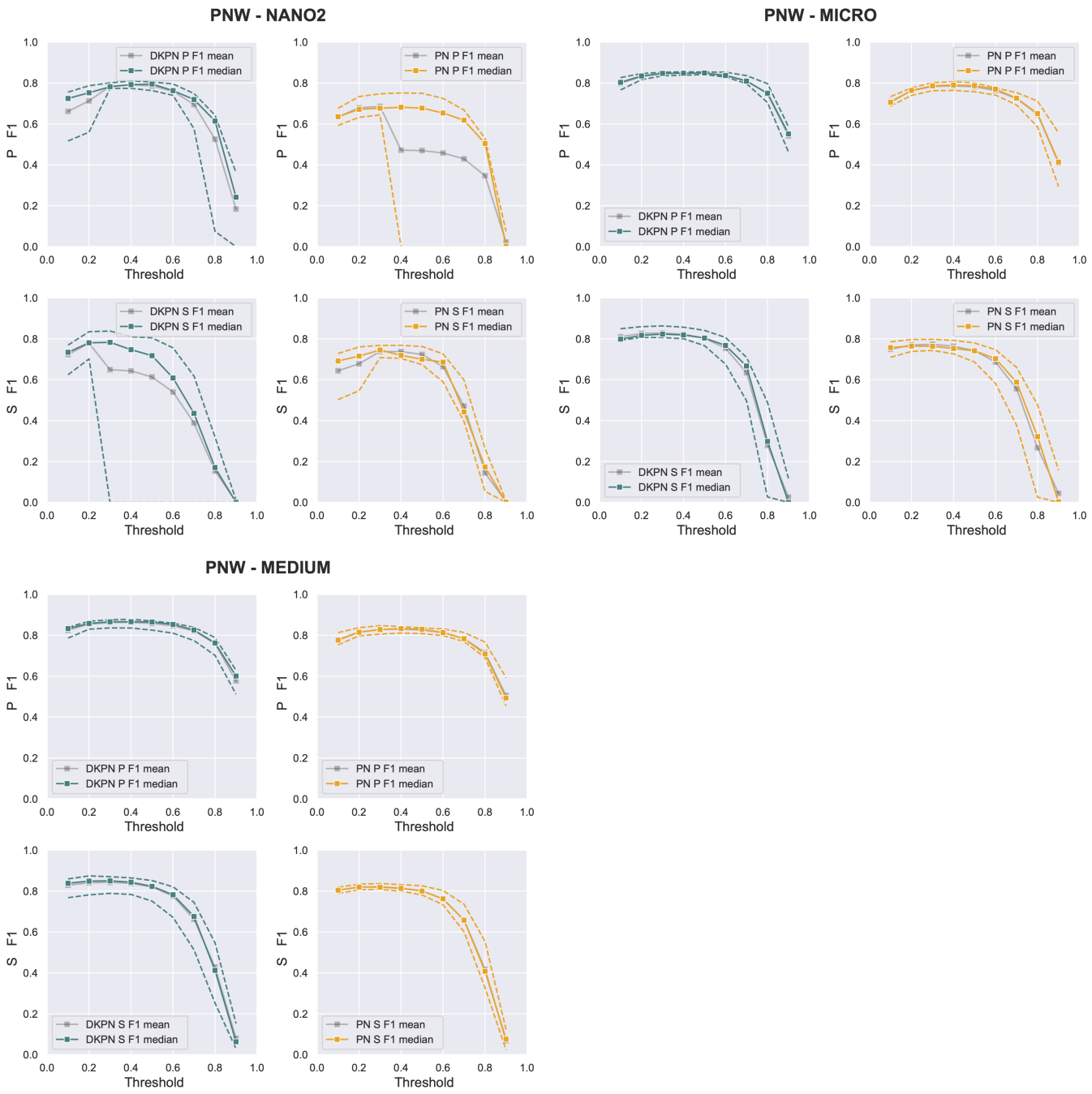


**Figure 10** Comparison of PhaseNet (PN) and DKPN with testing on the PNW dataset (cross-domain). F1 score metrics across a range of pick amplitude thresholds for P arrivals and S arrivals for models trained with different size INSTANCE datasets. Mean, median and upper/lower limits of the mean (dashed curves) of F1 for 7 runs each with 5,000 evaluation samples drawn from the test dataset. Because some PNW traces have missing P or S picks, FP count may be overestimated.

count (Supplementary Text S2) .

Histograms of P and S pick residuals for PhaseNet and DKPN are shown in Figure 9. For P, in contrast to the INSTANCE testing results, the total number of predicted picks within twice $\Delta T_p$ increases with increasing training dataset size and is greater for DKPN than for PhaseNet. There is little change with respect to training dataset size in the mean of the residuals, which is always near zero, or for standard deviation. For P, DKPN always has a higher count of near-zero residual picks than PhaseNet, especially for the smallest training datasets (e.g. NANO2), in agreement with the evolution of F1 score and other statistics discussed above. For S, as with INSTANCE testing, DKPN has a higher count of near-zero residual picks than PhaseNet for the smaller training datasets, while for larger training datasets both models show almost identical performance. Notably, the total number of predicted picks within twice $\Delta T_p$ decreases (DKPN) or is roughly stable (PhaseNet) instead of increasing with increasing training dataset size as for P.

### 3.3 Test 3 – Cross-domain—INSTANCE training and PNW test datasets

We examine a second cross-domain case, applying the models trained with the INSTANCE datasets of different sizes to testing on waveforms from the PNW dataset. Relative to INSTANCE and ETHZ, the PNW waveform dataset is characterized by many trace sets with clear, impulsive S arrivals at larger S-P times (larger epicentral distance), but also trace sets with missing horizontal channels. Since the DKPN processing requires 3 component trace sets, we only use data for PNW which includes all 3 channels (Supplementary File S3). The resulting F1 metrics across a range of pick amplitude thresholds for PNW test datasets are shown in Figure 10.

For P arrivals, the PNW results are similar to those for cross-domain testing with ETHZ (Test 2; Figure 8) with a small reduction in overall performance and stability relative to in-domain testing with INSTANCE (Test 1; Figure 6) (e.g. maximum F1 scores ~0.8 instead of ~0.9 for larger datasets) and a small performance increase of DKPN over PhaseNet for larger training sets (MICRO and MEDIUM), and a more prominent increase for the smallest datasets (NANO2).

For S arrivals (Figure 10), as with ETHZ (Test 2; Figure 8), the performance of DKPN and PhaseNet for the smallest training dataset, NANO2 is slightly poorer than in-domain testing with INSTANCE (Test 1; Figure 6), nearly identical for the MICRO dataset, and, for the largest dataset, MEDIUM, nearly identical for PhaseNet and slightly improved for DKPN. These latter results are surprising for a cross-domain dataset, likely explained by the high rate in the PNW dataset of clear, impulsive S arrivals which may resemble well S arrivals captured most strongly in INSTANCE training, and, for DKPN, by the sensitivity of the introduced domain knowledge to impulsive arrivals.

Histograms of P and S pick residuals for PhaseNet and DKPN for PNW testing are shown in Figure 11. For P, and similar to cross-domain, ETHZ testing, with increasing training dataset size the total pick rate generally increases, there is little evolution for mean (always near zero) and standard-deviation, while DKPN has a higher count of near-zero residual picks than PhaseNet for all training dataset sizes. For S, DKPN shows slightly better statistics and count of near-zero residuals than PhaseNet in agreement with the evolution of F1 score and other S statistics discussed above. However, both PhaseNet and DKPN show a consistent negative mean residual of almost 0.1 sec, perhaps suggesting that the impulsiveness of many PNW S onsets relative to typical S onsets on INSTANCE training waveforms is leading the INSTANCE trained network to bias and advance the pick time predictions relative to those for INSTANCE waveforms.

## 4 Discussion

We use classical picker algorithms as domain-knowledge to transform raw seismogram waveforms into modified input features for the deep-learning PhaseNet picker, without otherwise modifying the picker architecture. We compare the deep-learning picker with modified input, DKPN, with standard PhaseNet when both are trained using the same INSTANCE data, training strategy and hyper-parameters and applied to an INSTANCE in-domain and two cross-domain datasets, ETHZ and PNW.

For P detection and picking, cross-domain application to the ETHZ (Test 2; Figure 8) and PNW datasets (Test 3; Figure 10) shows an improvement in F1 of around 15% for DKPN over PhaseNet for the smallest training dataset NANO2. For larger cross-domain training datasets and for all INSTANCE in-domain testing (Test 1; Figure 6) PhaseNet and DKPN show almost identical performance. These results suggest that with smaller size training datasets the domain-knowledge modified input of the DKPN deep-learning model provides useful prior information for effective and stable seismic phase detection and picking. The DKPN network thus does not need to learn this information during training (Figure 12), though the basic PhaseNet architecture is still capable of efficiently learning equivalent information during training with larger datasets. In histograms of P pick residuals (Figures 7, 9 and 11), DKPN generally shows a higher count of near-zero residual picks than PhaseNet, with slight reduction of this difference for the largest training datasets. This suggests the domain-knowledge modified input of the DKPN provides some improvement in the fine-scale onset timing of picks over PhaseNet. Overall, besides pick detection, much of the training for both PhaseNet and DKPN likely involves refinement of onset timing, phase identification and other pick characterization tasks; these are difficult tasks in manual tuning of classical picker algorithms and perhaps fundamentally better addressed with machine-learning optimization (Vassallo et al., 2012; Yeck et al., 2020).

The improved DKPN performance for P picking relative to PhaseNet for the ETHZ and PNW datasets with smaller training dataset sizes is likely due to the ETHZ and PNW pre-event noise, event waveforms, and P and S phase onsets having greater differences from the INSTANCE training waveforms than can be accommodated by the generalization of the PhaseNet INSTANCE training with small datasets. Important differences in waveforms relative to INSTANCE may include a larger number of regional events with lower frequency waveforms in ETHZ (Supplementary File S2), and the prevalence of sharper S onsets in PNW (Supplementary File S3).

These learning and performance differences indicate that for P picking, relative to purely data-driven deep-learning pickers such as PhaseNet, DKPN or other domain-knowledge machine-learning pickers may be better for small to very small training sets, may generalize better, e.g. when applied cross-domain to waveforms having very different characteristics to the waveforms of the training events, and may provide generally smaller differences relative to manual pick timing.

In addition, for all NANO2 models (Figure 6, 8, 10), DKPN is more stable than PhaseNet in P performances across many threshold levels as indicated by the spread of upper/lower limits of the mean (dashed curves). This

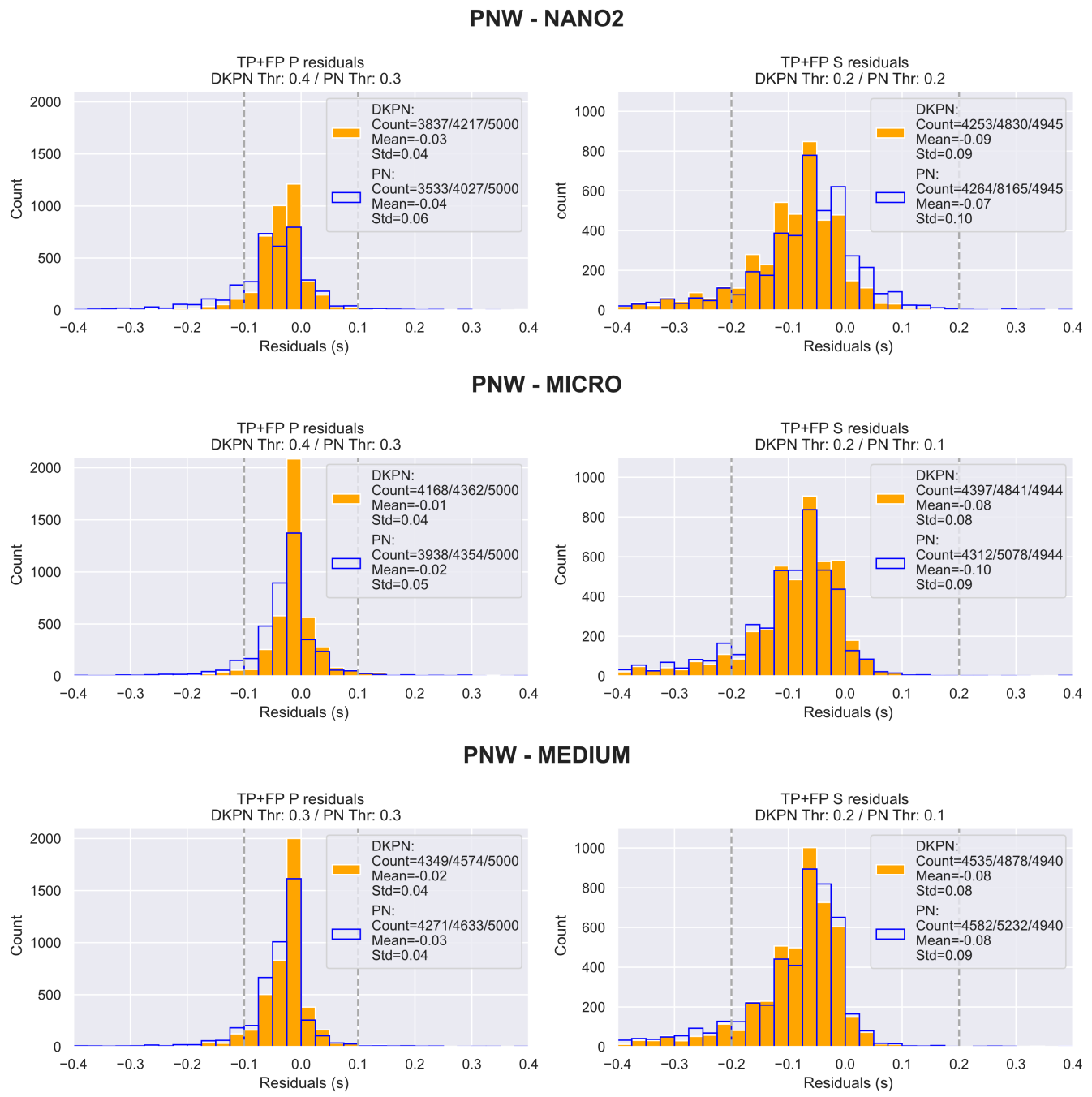


**Figure 11** Histograms of P and S pick residuals for selected INSTANCE training dataset sizes tested on the ETHZ dataset (cross-domain). Results shown for the pick amplitude threshold (Thr) giving the highest F1 score for each case of dataset size and method (DKPN or PhaseNet). Vertical, dashed gray lines show the maximum difference between a pick time and the corresponding label time (0.1 sec for P and 0.2 sec for S) to declare a correctly predicted arrival for evaluation statistics. Pick counts show: Number of residuals (number of predicted picks) used in the mean, mode and standard-deviation statistics (trimmed within twice ΔTp: ± 0.2 sec for P and ± 0.4 sec for S) / Total number of residuals available / Total number of label picks available for the test case.

means that DKPN is less sensitive to changes in pick threshold selection, proving to be more assertive about onset prediction (i.e. sharper prediction probability functions) even when few training data are available. DKPN is also less sensitive to training-data selection as resulting from different random selections across the 7 training-testing experiments, as shown from the upper and lower bounds of F1-scores that better follow the median trends.

For S detection and picking, both PhaseNet and DKPN show lower performance relative to P detection and picking (Figures 6-11), as also found for the deep-learning pickers examined in (Münchmeyer et al., 2022). This reduced performance is most likely due to S arrivals occurring in the P coda, and to the often emergent and complicated form of S arrival onsets, especially at regional distances in areas of complex geology. For in-domain testing on INSTANCE datasets and cross-domain ETHZ testing, PhaseNet and DKPN show nearly identical S performance for all but the smallest

training datasets. However, for the cross-domain PNW test dataset DKPN shows slightly better S picking performance than PhaseNet across all training dataset sizes. This result may be related to the PNW dataset, relative to INSTANCE and ETHZ, having a large number of clear, impulsive S arrivals, which may match well impulsive P arrivals for which classical pickers such as Filter Picker are optimized, and thus more easily detected by DKPN.

Training dataset size is an important issue with deep-learning pickers, as there are few large, well curated and error-free seismic waveform datasets with reliable, manual or other, reference picks. Many studies, such as temporary aftershock monitoring and short-term experiments, may have manually picked datasets that are too small for training with pure, data-driven picker models. Moreover, waveforms for some studies may have specific characteristics (e.g., in frequency content, epicentral distance ranges, noise, distribution of magnitudes) that preclude processing with machine learning methods pretrained with large datasets with different waveform characteristics. Here we have used relatively small to moderate size training datasets (~800 to 245k samples) relative to other key studies (e.g. 11k, 65k, 555k, 780k, 1.3M and 4.5M training and evaluation traces for the 6 picker models examined in Münchmeyer et al., 2022). We have shown that DKPN sometimes outperforms PhaseNet with smaller datasets for P picking, especially for cross-domain picking of the ETHZ and PNW datasets, probably due to the prior, domain-knowledge information on picking inherent in the DKPN input traces (Figure 12). Thus domain-knowledge based methods such as DKPN may be especially useful for studies with smaller datasets, especially those with unusual waveform characteristics which necessitates picker retraining, as well as for when limited computing time or resources are available. The combination of domain-knowledge based methods with transfer learning (e.g., Jozinović et al., 2021) may be particularly useful with small datasets that require retraining of machine learning pickers.

The FilterPicker CF amplifies and transforms energy onsets and changes in frequency content into abrupt, step- or pulse-like waveforms, while remaining fairly insensitive to absolute amplitudes and frequency content which vary between events and datasets (Figs 2-5). Improvements in P picking performance of DKPN over the purely data-driven PhaseNet may primarily be due to the similarity between these CF waveforms and the narrow, Gaussian wavelets of the target, probabilistic, picks (Fig. 12). To help verify this proposition, we ran a version of DKPN which retains the 3 channels of raw waveform input, giving 8 channels total for input. This change gives almost no difference in the picking results such as mean and median F1 scores, except for a degradation of results for S picking with the smallest training dataset NANO2, and the 8 channel input leads to increased spread of the upper/lower limits of the mean for the NANO2 and MICRO training datasets. This reduction in performance suggest the 5 channels of CF's plus inclination and modulus waveforms input to DKPN retain the majority of information relevant to picking effectively contained in the raw waveforms.

The DKPN network thus apparently receives rule-based, *deterministically* modified input that resembles a simple transformation of the required output defining pick detection, potentially simplifying training and improving performance and stability, and also providing an inherent mechanism for generalization. On the other hand, given a classical picker CF, the design and optimization of subsequent algorithms for refining onset-timing, phase identification and other characterization are difficult and somewhat *haphazard* tasks (Lomax et al., 2012; Vassallo et al., 2012). In DKPN and PhaseNet these subsequent tasks are performed by the deep-neural-network; indeed, high-dimensional, *stochastically*-driven machine-learning is eminently suited to such tasks. However, when observations from a network of seismometers are available, a domain-knowledge, rule-based approach may also be valuable for pick characterization tasks such as quality control (Ning et al., 2022). And, in practice, domain-knowledge is used to improve detection and picking even with nominally, data-driven, machine-learning pickers, since many studies apply a high-pass filter to suppress known microseismic noise at longer period and amplify arrivals of interest expected at higher frequencies (Mousavi et al., 2019, 2020; Münchmeyer et al., 2022; Ross et al., 2018b,a; Woollam et al., 2019).

Additional study might investigate the usage of "simpler" and "shallower" model-architectures than that of PhaseNet, while still feeding the DKPN input or similar. Such a configuration could help understand the effects of domain-knowledge on machine-learning model generalization. In particular, less complex architectures may allow easier setting of meta-parameters during the learning stages and better explanation of the machine learning models, and produce more robust models that are easier to debug and improve. However, if pick characterization tasks other than detection account for much of the learning effort during training for both DKPN and PhaseNet, then the use of CF, inclination and modulus waveforms in DKPN is not likely to allow reducing the number of layers or otherwise simplifying the underlying CNN architecture inherited from PhaseNet.

FilterPicker and STA/LTA methods in general require stabilization after the start of a time-series and after impulsive arrivals; the time of stabilization for FilterPicker is proportional to the long-term, recursive-filter time-averaging scale. This stabilization, besides making it necessary to have sufficient background data before the first arrival in a time-series, usually degrades picker sensitivity to arrivals closely following previous arrivals, in particular an S arrival, even when higher amplitude than the preceding P. This is one reason why we include in DKPN the instantaneous polarization modulus time-series, which preserves S amplitude relative to P, and the inclination time-series, which often changes character from predominantly up-down to near horizontal at the S arrival. Future work might investigate if, in the context of a domain-knowledge, machine-learning picker, it is possible to modify the FilterPicker CF algorithm to reduce adverse effects of the stabilization without otherwise adversely affecting the overall picker per-

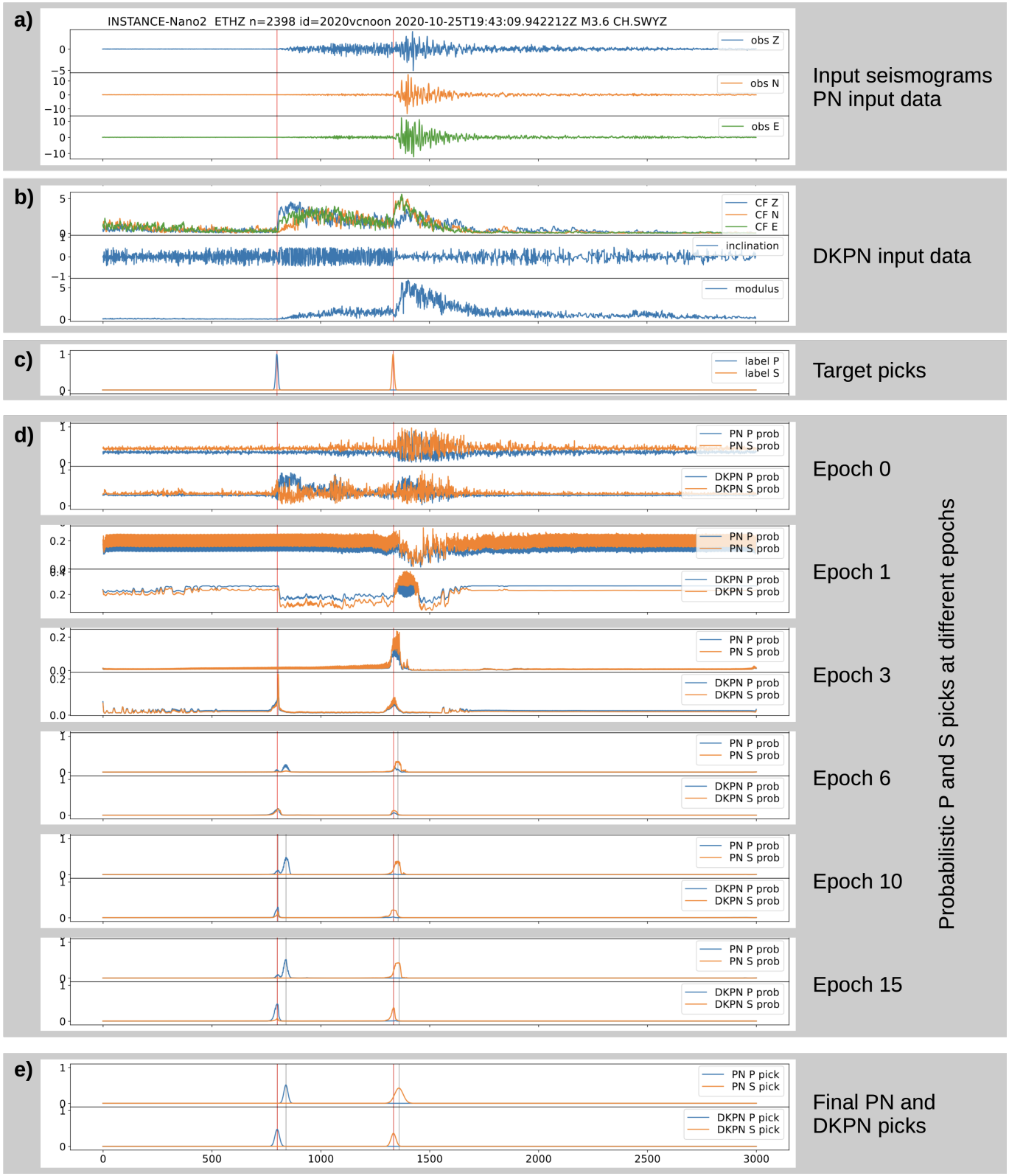


**Figure 12** Results with progression of epoch for models trained with the INSTANCE NANO2 dataset applied to ETHZ trace samples. Panel a) shows de-meaned and normalized Z, N, E component input seismograms; panel b) normalized, DKPN FilterPicker Z, N, E CFs, and normalized instantaneous inclination and modulus; and panel c) P (blue) and S (orange) pick label data (red vertical lines). Panel e) shows PhaseNet (PN) and DKPN Gaussian P (blue) and S (orange) predicted picks after epoch 15 training. Panel d) shows PhaseNet and DKPN probabilistic P (blue) and S (orange) pick predictions for a sequence of training epochs; for clarity, the predictions for epochs 1 and 3 are not normalized. For the untrained model (epoch 0) the predictions are random, non-linear mappings of the input traces. The epoch 0 predictions for DKPN reflect the input CFs and polarization traces and show a distinct P arrival signature which is not present in the predictions for PhaseNet, which reflect mainly amplitudes in the near-raw seismograms (continued).

**Figure 12 (Continued)** The DKPN probabilistic predictions show in epoch 1 the P arrival as a step-like signal and the S arrival as a concentrated prediction, in epoch 3 both P and S arrivals as isolated predictions, and from epoch 6 to 15 as stable predictions, though in epoch 6 the P arrival has both P and S predictions. PhaseNet obtains an S but not P prediction in epoch 3, isolated but noisy P and S predictions starting from epoch 6, and stable predictions between epochs 10 and 15. The slower evolution of the PhaseNet predictions from near random to clear arrivals through epochs 0-6 support that it is learning both arrival detection and picking throughout the training process. P predictions at the S arrival time for both PhaseNet and DKPN, and S predictions at the P arrival for DKPN visible in epochs 3 and 6 are highly suppressed through learning by epoch 10. The final PhaseNet P and S picks are delayed relative to the label picks, likely due to the emergent amplitude of the arrivals which is not represented well in the INSTANCE training dataset. The final DKPN picks do not show this delay, likely due to the high sensitivity of the domain-knowledge preprocessing (panel b) to changes in waveform characteristics besides amplitude, such as frequency content.

formance.

Relative to PhaseNet, DKPN has an increase in overall training and evaluation time due to the preprocessing required to derive the 3-component, FilterPicker CFs, modulus and inclination from the seismogram waveforms. However, our calculations in this study show that the processing time penalty is effectively removed through code optimization and use of parallel, GPU processing. In addition, preprocessing the training dataset once before training and storing it on disk or in memory can remove most of the DKPN training time penalty. Moreover, despite having an increased dimensionality of input data, DKPN training converges faster (requires fewer epochs) than PhaseNet (see Supplementary File S4).

## 5 Conclusions

Using seismological domain-knowledge, we transform 3-component seismograms into the characteristic functions of a classical, multi-band picker, plus instantaneous modulus and inclination. We replace the near-raw seismogram input of the deep-learning picker PhaseNet with these transformed traces, forming DKPN, a modified PhaseNet, and we compare the performance of DKPN and standard PhaseNet with different training and testing datasets. DKPN shows some improvements in performance and generalization over PhaseNet, and may be applicable with smaller training datasets. DKPN requires more computation time than standard PhaseNet due to the additional, domain-knowledge preprocessing. However, this time penalty can be removed with code optimization, GPU use, real-time processing, and storing DKPN preprocessed waveforms for training. Additionally, DKPN training time (number of epochs) may be reduced relative to PhaseNet.

For P arrivals, DKPN shows little or no improvement in performance over PhaseNet in picking the in-domain, INSTANCE dataset for all training dataset sizes, and in picking cross-domain ETHZ and PNW datasets for larger training dataset sizes. These results demonstrate the power and robustness of the PhaseNet architecture for extracting information relevant to pick detection and characterization from near-raw seismogram waveforms. However, DKPN generally shows improved statistics such as true positive rate and increased number of picks with small residuals, and sometimes improved metrics such as F1 score, especially for small training datasets and for cross-domain testing. These improvements can be attributed to the additional information relevant to picking introduced in the DKPN input data by the rule-based, domain-knowledge waveform preprocessing. For the purely data-driven PhaseNet, much of this same picking-specific information must be learned by the network in training; the efficiency and success of this training will depend on the training dataset being sufficiently large and having similar event waveform characteristics to the application datasets.

The performance of both PhaseNet and DKPN is worse for picking S arrivals than for P, likely mainly due to S onsets occurring in the P coda. Both models show similar S performance for in- and cross-domain picking on the ETHZ dataset, but DKPN shows slightly better performance than PhaseNet for cross-domain S picking on the PNW dataset, likely due to the frequent occurrence of sharp S onsets on the PNW waveforms which are less prevalent in the INSTANCE training data.

Overall, our results show that PhaseNet, and perhaps deep-neural-network pickers in general, have a sufficiently large and complex architecture to learn to accurately map key characteristics of seismogram waveforms and phase onset energy into detections and picks, including for cross-domain application. This learning requires comprehensive and high-quality, but not necessarily very large training datasets.

However, given our results, DKPN is of interest for cross-domain picking when retraining on the target dataset is not possible, or for cases where training is needed but can be performed on only a very small dataset, such as when few manual picks are available. Further work with DKPN and other, domain-knowledge augmented machine-learning procedures for picking and other seismological analyses is warranted to investigate performance improvements over pure, data-driven, learning algorithms, especially for small or highly varied training datasets and for strongly cross-domain application.

## Acknowledgments

We gratefully thank Mostafa Mousavi and an anonymous reviewer for extensive and helpful comments, and the editor Yen Joe Tan and all Seismica volunteers for their help and contributions. This research was

supported by the SOME (Seismological Oriented Machine lEarning) project - INGV Pianeta Dinamico 2021 Tema 8 (grant no. CUP D53J1900017001) funded by the Italian Ministry of University and Research "Fondo finalizzato al rilancio degli investimenti delle amministrazioni centrali dello Stato e allo sviluppo del Paese, legge 145/2018". M. Bagagli was partially supported by the European Union's research and innovation program under the Marie Sklodowska Curie Action, grant agreement (101105516).

## 6 Data and code availability

All datasets used in this study are available in SeisBench (https://www.github.com/seisbench/seisbench). Python codes used for training, testing and visualization presented in Tests 1, 2 and 3, and run with PyTorch (https://pytorch.org) and SeisBench (https://www.github.com/seisbench/seisbench), are available at https://github.com/INGV/DKPN.

The supplementary information for this article includes a pdf file containing additional text, table and figures describing the experiment setup, training and testing stages, Supplementary Files S1-S3 contain example waveforms and PhaseNet and DKPN processing for the three datasets, and Supplementary File S4 contain train-validation loss-curves for selected training runs.

All calculations were performed on a GAIA multi GPU server (equipped with 4 80GB NVidia A100) running the ICE4AI software stack by E4 Analytics (https://www.e4company.com/en/gpu-appliance-for-artificial-intelligence) and providing environments with PyTorch (https://pytorch.org), SeisBench (https://www.github.com/seisbench/seisbench), and ObsPy (Beyreuther et al., 2010; Krischer et al., 2015, , http://obspy.org). Word processing and some figures were done with LibreOffice (https://www.libreoffice.org).

## 7 Competing interests

The authors declare no conflicts of interest with respect to the research, authorship, and publication of this article.

## References

Akazawa, T. A technique for automatic detection of onset time of P-and S-phases in strong motion records. In *Proc. of the 13th World Conf. on Earthquake Engineering*, Vancouver, Canada, 2004. http://www.iitk.ac.in/nicee/wcee/article/13_786.pdf.

Allen, R. Automatic phase pickers: Their present use and future prospects. *Bulletin of the Seismological Society of America*, 72 (6B):S225–S242, Dec. 1982. doi: 10.1785/bssa07206b0225.

Allen, R. V. Automatic earthquake recognition and timing from single traces. *Bulletin of the Seismological Society of America*, 68(5): 1521–1532, Oct. 1978. doi: 10.1785/bssa0680051521.

Alvarez, I., Garcia, L., Mota, S., Cortes, G., Benitez, C., and De la Torre, A. An Automatic P-Phase Picking Algorithm Based on Adaptive Multiband Processing. *IEEE Geoscience and Remote Sensing Letters*, 10(6):1488–1492, Nov. 2013. doi: 10.1109/lgrs.2013.2260720.

Anant, K. S. and Dowla, F. U. Wavelet transform methods for phase identification in three-component seismograms. *Bulletin of the Seismological Society of America*, 87(6):1598–1612, Dec. 1997. doi: 10.1785/bssa0870061598.

Baer, M. and Kradolfer, U. An automatic phase picker for local and teleseismic events. *Bulletin of the Seismological Society of America*, 77(4):1437–1445, Aug. 1987. doi: 10.1785/bssa0770041437.

Bagagli, M. *Seismicity and seismic tomography across scales: application to the greater Alpine region*. PhD thesis, 2022. doi: 10.3929/ETHZ-B-000580361.

Bai, C.-y. Automatic Phase-Detection and Identification by Full Use of a Single Three-Component Broadband Seismogram. *Bulletin of the Seismological Society of America*, 90(1):187–198, Feb. 2000. doi: 10.1785/0119990070.

Balestriero, R. and Baraniuk, R. Mad Max: Affine Spline Insights into Deep Learning, 2018. doi: 10.48550/ARXIV.1805.06576.

Beyreuther, M., Barsch, R., Krischer, L., Megies, T., Behr, Y., and Wassermann, J. ObsPy: A Python Toolbox for Seismology. *Seismological Research Letters*, 81(3):530–533, May 2010. doi: 10.1785/gssrl.81.3.530.

Beyreuther, M., Hammer, C., Wassermann, J., Ohrnberger, M., and Megies, T. Constructing a Hidden Markov Model based earthquake detector: application to induced seismicity: Constructing a HMM based earthquake detector. *Geophysical Journal International*, 189(1):602–610, Feb. 2012. doi: 10.1111/j.1365-246x.2012.05361.x.

Borghesi, A., Baldo, F., and Milano, M. Improving Deep Learning Models via Constraint-Based Domain Knowledge: a Brief Survey, 2020. doi: 10.48550/ARXIV.2005.10691.

Chen, C. and Holland, A. A. PhasePApy: A Robust Pure Python Package for Automatic Identification of Seismic Phases. *Seismological Research Letters*, 87(6):1384–1396, Aug. 2016. doi: 10.1785/0220160019.

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

Enescu, N. Seismic Data Processing Using Nonlinear Prediction and Neural networks. In *IEEE NORSIG Symposium*, Espoo, Finland, 1996.

Gentili, S. and Michelini, A. Automatic picking of P and S phases using a neural tree. *Journal of Seismology*, 10(1):39–63, Jan. 2006. doi: 10.1007/s10950-006-2296-6.

Hien, D. A guide to receptive field arithmetic for Convolutional Neural Networks, 2018. https://blog.mlreview.com/a-guide-to-receptive-field-arithmetic-for-convolutional-neural-networks-e0f514068807. Retrieved August 16, 2022, from.

Jozinović, D., Lomax, A., Štajduhar, I., and Michelini, A. 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, Dec. 2021. doi: 10.1093/gji/ggab488.

Kim, A., Nakamura, Y., Yukutake, Y., Uematsu, H., and Abe, Y. Development of a high-performance seismic phase picker using deep learning in the Hakone volcanic area. *Earth, Planets and Space*, 75(1), May 2023. doi: 10.1186/s40623-023-01840-5.

Kong, Q., Trugman, D. T., Ross, Z. E., Bianco, M. J., Meade, B. J., and Gerstoft, P. Machine Learning in Seismology: Turning Data into Insights. *Seismological Research Letters*, 90(1):3–14, Nov. 2018. doi: 10.1785/0220180259.

Krischer, L., Megies, T., Barsch, R., Beyreuther, M., Lecocq, T., Caudron, C., and Wassermann, J. ObsPy: a bridge for seismology into the scientific Python ecosystem. *Computational Sci-*

*ence & Discovery*, 8(1):014003, May 2015. doi: 10.1088/1749-4699/8/1/014003.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature*, 521 (7553):436–444, May 2015. doi: 10.1038/nature14539.

Liao, W.-Y., Lee, E.-J., Mu, D., Chen, P., and Rau, R.-J. ARRU Phase Picker: Attention Recurrent-Residual U-Net for Picking Seismic P- and S-Phase Arrivals. *Seismological Research Letters*, 92(4): 2410–2428, Mar. 2021. doi: 10.1785/0220200382.

Lomax, A., Satriano, C., and Vassallo, M. 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, May 2012. doi: 10.1785/gssrl.83.3.531.

Lomax, A., Michelini, A., and Curtis, A. *Earthquake Location, Direct, Global-Search Methods*, page 1–33. Springer New York, 2014. doi: 10.1007/978-3-642-27737-5_150-2.

Marcus, G. Innateness, AlphaZero, and Artificial Intelligence, 2018. doi: 10.48550/ARXIV.1801.05667.

McEvilly, T. V. and Majer, E. L. ASP: An Automated Seismic Processor for microearthquake networks. *Bulletin of the Seismological Society of America*, 72(1):303–325, Feb. 1982. doi: 10.1785/bssa0720010303.

Michelini, A., Cianetti, S., Gaviano, S., Giunchi, C., Jozinović, D., and Lauciani, V. INSTANCE – the Italian seismic dataset for machine learning. *Earth System Science Data*, 13(12):5509–5544, Nov. 2021. doi: 10.5194/essd-13-5509-2021.

Mousavi, S. M. and Beroza, G. C. Deep-learning seismology. *Science*, 377(6607), Aug. 2022. doi: 10.1126/science.abm4470.

Mousavi, S. M., Langston, C. A., and Horton, S. P. Automatic microseismic denoising and onset detection using the synchrosqueezed continuous wavelet transform. *GEOPHYSICS*, 81 (4):V341–V355, July 2016. doi: 10.1190/geo2015-0598.1.

Mousavi, S. M., Zhu, W., Sheng, Y., and Beroza, G. C. CRED: A Deep Residual Network of Convolutional and Recurrent Units for Earthquake Signal Detection. *Scientific Reports*, 9(1), July 2019. doi: 10.1038/s41598-019-45748-1.

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

Mousset, E., Cansi, Y., Crusem, R., and Souchet, Y. A connectionist approach for automatic labeling of regional seismic phases using a single vertical component seismogram. *Geophysical Research Letters*, 23(6):681–684, Mar. 1996. doi: 10.1029/95gl03811.

Muralidhar, N., Islam, M. R., Marwah, M., Karpatne, A., and Ramakrishnan, N. Incorporating Prior Domain Knowledge into Deep Neural Networks. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, Dec. 2018. doi: 10.1109/bigdata.2018.8621955.

Münchmeyer, J., Woollam, J., Rietbrock, A., Tilmann, F., Lange, D., Bornstein, T., Diehl, T., Giunchi, C., Haslinger, F., Jozinović, D., Michelini, A., Saul, J., and Soto, H. Which Picker Fits My Data? A Quantitative Evaluation of Deep Learning Based Seismic Pickers. *Journal of Geophysical Research: Solid Earth*, 127(1), Jan. 2022. doi: 10.1029/2021jb023499.

Ni, Y., Hutko, A., Skene, F., Denolle, M., Malone, S., Bodin, P., Hartog, R., and Wright, A. Curated Pacific Northwest AI-ready Seismic Dataset. *Seismica*, 2(1), May 2023. doi: 10.26443/seismica.v2i1.368.

Ning, I. L. C., Swafford, L., Craven, M., Davies, K., Earnest, E., and Thornton, D. Automation of passive seismic processing via machine learning and physics-informed methods. In *Second International Meeting for Applied Geoscience & Energy*. Society of Exploration Geophysicists and American Association of Petroleum Geologists, Aug. 2022. doi: 10.1190/image2022-3750116.1.

Njirjak, M., Otović, E., Jozinović, D., Lerga, J., Mauša, G., Michelini, A., and Štajduhar, I. The Choice of Time–Frequency Representations of Non-Stationary Signals Affects Machine Learning Model Accuracy: A Case Study on Earthquake Detection from LEN-DB Data. *Mathematics*, 10(6):965, Mar. 2022. doi: 10.3390/math10060965.

Plešinger, A., Hellweg, M., and Seidl, D. Interactive high-resolution polarization analysis of broad-band seismograms. *Journal of Geophysics*, 59(1):129–139, 1986. https://journal.geophysicsjournal.com/JofG/article/view/203.

Ronneberger, O., Fischer, P., and Brox, T. *U-Net: Convolutional Networks for Biomedical Image Segmentation*, page 234–241. Springer International Publishing, 2015. doi: 10.1007/978-3-319-24574-4_28.

Ross, Z. E. and Ben-Zion, Y. Automatic picking of direct P, S seismic phases and fault zone head waves. *Geophysical Journal International*, 199(1):368–381, Aug. 2014. doi: 10.1093/gji/ggu267.

Ross, Z. E., Meier, M., and Hauksson, E. P Wave Arrival Picking and First-Motion Polarity Determination With Deep Learning. *Journal of Geophysical Research: Solid Earth*, 123(6):5120–5129, June 2018a. doi: 10.1029/2017jb015251.

Ross, Z. E., Meier, M., Hauksson, E., and Heaton, T. H. Generalized Seismic Phase Detection with Deep Learning. *Bulletin of the Seismological Society of America*, 108(5A):2894–2901, Aug. 2018b. doi: 10.1785/0120180080.

Satriano, C., Lomax, A., and Zollo, A. Real-Time Evolutionary Earthquake Location for Seismic Early Warning. *Bulletin of the Seismological Society of America*, 98(3):1482–1494, June 2008. doi: 10.1785/0120060159.

Sleeman, R. and van Eck, T. 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, June 1999. doi: 10.1016/s0031-9201(99)00007-2.

Soto, H. and Schurr, B. 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*, July 2021. doi: 10.1093/gji/ggab266.

Stevenson, P. R. Microearthquakes at Flathead Lake, Montana: A study using automatic earthquake processing. *Bulletin of the Seismological Society of America*, 66(1):61–80, Feb. 1976. doi: 10.1785/bssa0660010061.

Vassallo, M., Satriano, C., and Lomax, A. Automatic Picker Developments and Optimization: A Strategy for Improving the Performances of Automatic Phase Pickers. *Seismological Research Letters*, 83(3):541–554, May 2012. doi: 10.1785/gssrl.83.3.541.

Vidale, J. Complex polarization analysis of particle motion. *Bulletin of the Seismological Society of America*, 76(5):1393–1405, 1986. doi: 10.1785/BSSA0760051393.

Wang, J. and Teng, T.-L. Artificial neural network-based seismic detector. *Bulletin of the Seismological Society of America*, 85(1): 308–319, Feb. 1995. doi: 10.1785/bssa0850010308.

Withers, M., Aster, R., Young, C., Beiriger, J., Harris, M., Moore, S., and Trujillo, J. A comparison of select trigger algorithms for automated global seismic phase and event detection. *Bulletin of the Seismological Society of America*, 88(1):95–106, Feb. 1998. doi: 10.1785/bssa0880010095.

Woollam, J., Rietbrock, A., Bueno, A., and De Angelis, S. Convo-

lutional Neural Network for Seismic Phase Classification, Performance Demonstration over a Local Seismic Network. *Seismological Research Letters*, 90(2A):491–502, Jan. 2019. doi: 10.1785/0220180312.

Woollam, J., Münchmeyer, J., Tilmann, F., Rietbrock, A., Lange, D., Bornstein, T., Diehl, T., Giunchi, C., Haslinger, F., Jozinović, D., Michelini, A., Saul, J., and Soto, H. SeisBench—A Toolbox for Machine Learning in Seismology. *Seismological Research Letters*, 93(3):1695–1709, Mar. 2022. doi: 10.1785/0220210324.

Yeck, W. L., Patton, J. M., Ross, Z. E., Hayes, G. P., Guy, M. R., Ambruz, N. B., Shelly, D. R., Benz, H. M., and Earle, P. S. Leveraging Deep Learning in Global 24/7 Real-Time Earthquake Monitoring at the National Earthquake Information Center. *Seismological Research Letters*, 92(1):469–480, Sept. 2020. doi: 10.1785/0220200178.

Yu, Z. and Wang, W. LPPN: A Lightweight Network for Fast Phase Picking. *Seismological Research Letters*, 93(5):2834–2846, June 2022. doi: 10.1785/0220210309.

Zhang, H., Thurber, C., and Rowe, C. Automatic P-Wave Arrival Detection and Picking with Multiscale Wavelet Analysis for Single-Component Recordings. *Bulletin of the Seismological Society of America*, 93(5):1904–1912, Oct. 2003. doi: 10.1785/0120020241.

Zhu, W. and Beroza, G. C. PhaseNet: A Deep-Neural-Network-Based Seismic Arrival Time Picking Method. *Geophysical Journal International*, Oct. 2018. doi: 10.1093/gji/ggy423.