doi:10.26443/seismica.v4i2.1559

SEISMICA

Benchmarking seismic phase associators: Insights from synthetic scenarios

Jorge Puente Huerta * [1,2], **Christian Sippl** [1], **Jannes Münchmeyer** [3], **Ian W. McBrearty** [4]

[1]Institute of Geophysics, Czech Academy of Sciences, Prague, Czech Republic, [2]Charles University, Faculty of Mathematics and Physics, Department of Geophysics, Prague, Czech Republic, [3]Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, IRD, Univ. Gustave Eiffel, ISTerre, Grenoble, France., [4]Department of Geophysics, Stanford University, Stanford, California, U.S.A.

Author contributions: *Conceptualization*: Christian Sippl. *Methodology*: Jorge Puente Huerta. *Software*: Jorge Puente Huerta, Jannes Münchmeyer, Ian W. McBrearty. *Formal Analysis*: Jorge Puente Huerta, Christian Sippl. *Writing - Original draft*: Jorge Puente Huerta. *Writing - Review & Editing*: Christian Sippl, Jannes Münchmeyer, Ian W. McBrearty. *Visualization*: Jorge Puente Huerta.

Abstract Reliable seismicity catalogs are fundamental for seismological analysis. Following phase picking, phase association groups arrivals into sets with consistent origins (i.e., events), determines event counts, and identifies outlier picks. To handle the substantial increase in the quantity of seismic phase picks from improved picking methods and larger deployments, several novel phase associators have recently been proposed. This study presents a detailed benchmark analysis of five seismic phase associators, including classical and machine learning-based approaches: PhaseLink, REAL, GaMMA, GENIE, and PyOcto. We use synthetic datasets mimicking real seismicity characteristics in crustal and subduction zone scenarios. We evaluate performance for different conditions, including low- and high- noise environments, out-of-network events, very high event rates, and variable station density. The results reveal notable differences in precision, recall, and computational efficiency. GENIE and PyOcto demonstrate robust performance, with almost perfect performance for most scenarios, but under the most challenging conditions with high noise levels and event rates, performance drops while F1 scores still remain above 0.8. PhaseLink's performance declines with noise and event density, particularly in subduction zones, dropping to near zero in the most complex cases. GaMMA outperforms PhaseLink but struggles with accuracy and scalability in high-noise, high-density scenarios. REAL performs reasonably but loses recall under extreme conditions. PyOcto and PhaseLink show the quickest runtimes for smaller-scale datasets, while REAL and GENIE are more than an order of magnitude slower for these cases. At the highest pick rates, GENIE's runtime disadvantage diminishes, matching PyOcto and scaling effectively. Our results can guide practitioners compiling seismicity catalogs and developers designing novel associators.

Production Editor:
Andrea Llenos
Handling Editor:
Marlon Ramos
Copy & Layout Editor:
Sarah Jaye Oliva

Received:
January 21, 2025
Accepted:
July 31, 2025
Published:
September 11, 2025

1 Introduction

High-quality and reliable seismicity catalogs are an essential resource in seismology and fundamental for understanding earthquake processes. They form the basis for a wide range of studies in seismology and beyond, including travel time tomography (White et al., 2021), statistical seismology (Hainzl et al., 2019; Xiong et al., 2023), hazard assessment (Mancini et al., 2022), as well as research into tectonic processes (Maharaj et al., 2023; Sippl et al., 2019). Commonly, earthquake detection is performed with a two-step approach: phase picking and phase association. During phase picking, the task is to identify the onset of seismic phases, usually P- and S- waves, at individual seismic stations. Once phase picking is complete, the next fundamental step is phase association, which groups the seismic phases that were detected at different stations into common seismic events. A group of picks belongs to an event if all of them originate from the same location at the same time, i.e., a distinct hypocenter. The accuracy of phase association is essential for determining earthquake location, depth, and magnitude, and hence forms the backbone of subsequent seismological analyses. In addition, phase association allows discarding spurious phase picks, as these will usually not be consistent across stations.

In the broader landscape of seismic event detection and cataloging, integrated full-waveform paradigms have emerged that combine phase detection and association in a single step. Such approaches leverage network-level coherence metrics (i.e., measures of waveform similarity or energy across stations) to detect and locate earthquakes directly, without relying on individual P- and S-phase picks. For example, Poiata et al. (2016) introduced an array-based scheme that images coherent seismic energy across the network, enabling simultaneous detection and location of earthquakes without intermediate picking. While such joint detection-association approaches are gaining traction, the classical modular pipeline remains the norm in operational earthquake monitoring, with phase association still underpinning most seismicity catalogs and offering practical advantages in routine monitoring (Ross et al., 2019). Accordingly, this study focuses on bench-

*Corresponding author: puente@ig.cas.cz

marking the performance of phase associators within a modular pipeline, using synthetic yet realistic conditions to enable fair and controlled comparisons of different algorithms.

Historically, both phase picking and phase association were performed manually. However, to keep up with the rapidly growing data availability, automatic methods were developed (Allen, 1978). For phase association, early automated approaches were grid-based, involving the creation of a grid over a region of interest and associating phases based on the best-fitting grid points, using travel time tables (Johnson et al., 1995; Ringdal and Kværna, 1989). However, the runtime of such approaches becomes prohibitive when faced with a high number of picks. While historically issues were most commonly encountered with dense seismic activity such as aftershock sequences, the growing size of seismic networks and the advent of novel picking methods now routinely leads to vast quantities of picks that produce challenges for association even during background seismicity rates. In particular, the advent of machine learning techniques in phase picking has increased the volume of picks of small earthquakes to an unprecedented level, posing a new challenge to the phase association process (Zhu and Beroza, 2018; Ross et al., 2018; Zhu et al., 2019; Mousavi et al., 2019, 2020; Yang et al., 2021; Münchmeyer et al., 2022; Woollam et al., 2022; Zhu et al., 2022b).

Given these developments, the performance of phase associators has become increasingly important in the pursuit of building accurate earthquake catalogs with ever lower magnitudes-of-completeness. Consequently, there are now significant efforts to improve seismic phase association using a range of modern approaches. These approaches build on modified traditional techniques (Zhang et al., 2019; Münchmeyer, 2024), or use machine learning (Zhu et al., 2022a) and deep learning (Ross et al., 2019; McBrearty and Beroza, 2023) techniques and represent a significant advancement in the field. In addition to their different conceptual approaches, each algorithm's performance is dependent on the specific configuration of parameters used, and different associators may behave differently under different conditions (e.g., picks, number of stations, and noise density). Here, we conduct an in-depth benchmarking study to understand how different phase associators perform in a range of different scenarios. In addition, we provide insights into effective parameter choices for each associator. As establishing a "ground truth" catalog in a real scenario is nearly impossible, we use synthetic scenarios for our benchmark. This allows us to create "ground truth" datasets, with which the performance of each associator can be determined based on event and pick-level metrics. In addition, synthetics allow us to evaluate the impact of aspects such as event density or noise levels. Such a controlled environment is an effective way to systematically compare the methods and identify their strengths and limitations.

2 The algorithms

We evaluate five different algorithms for seismic phase association, with each of them taking labeled arrival times of P and S phases as input.

PhaseLink (Ross et al., 2019) is a deep learning (DL) approach for seismic phase association that uses a recurrent neural network with long short-term memory units to process a sliding window of phase picks. The input to PhaseLink is a fixed length sequence of picks from multiple stations, and the network predicts which picks belong to the same source. The framework aggregates predictions over time to form clusters, identifying individual earthquakes. The network is trained using a supervised learning approach, with the loss function optimized to minimize the misclassification of picks. PhaseLink requires training that can use real or synthetic data. The use of synthetic training data is crucial, as it allows exposing the network to a large range of seismicity scenarios. For the training step, providing a 1D velocity model of the region of interest is necessary.

REAL (Rapid Earthquake Association and Location; Zhang et al., 2019) is an optimized grid search-based algorithm. It is designed to rapidly and simultaneously associate seismic phases and locate seismic events. REAL performs a grid search in three dimensions around each station, with the earliest P arrival determining potential event locations. This reduces the search space from the entire study area to a smaller volume and eliminates the time dimension from the search, as the approximate origin time for each potential event can be inferred from the initial pick. The theoretical P and S travel-time tables are pre-calculated using a given homogeneous or 1D velocity model. The initial event location is determined at the grid point with the most associated P and S picks. If multiple grid points have the same maximum number of picks, the grid point with the smallest travel-time residuals is selected. REAL implements parallelization to reduce runtime.

GaMMA (Gaussian Mixture Model Association; Zhu et al., 2022a) treats the phase association problem as an unsupervised clustering problem within a probabilistic framework. It models each seismic event as a mixture component within a Gaussian Mixture Model (GMM; Bishop, 2006). It uses an expectation-maximization algorithm for optimizing the clusters. This iterative process can identify optimal phase associations by maximizing the likelihood of the observed data, considering both arrival time and amplitude. GaMMa employs DBSCAN (Ester et al., 1996) to segment phase picks into sub-windows prior to running the GMM for association. Each cluster can be associated in parallel to maximize CPU usage. GaMMA identifies "core points" based on the density of neighboring points to form clusters around them. This preprocessing step helps to manage the computational complexity and increase the scalability and efficiency by dividing the data into smaller, manageable segments, making the subsequent Expectation-Maximization algorithm more efficient. GaMMA can model travel-times with homogeneous and 1D models. In addition, it can incorporate amplitude decay relationships. We do not use amplitude information in this study for consistency with the other methods.

GENIE (Graph Earthquake Neural Interpretation Engine; McBrearty and Beroza, 2023) employs a graph neural network (GNN) to predict earthquake source locations and the likelihood of phase associations. GENIE constructs two graphs: one representing the seismic stations (station graph) and another representing the potential source locations (source graph). The source graph's nodes span the source region of interest, with edges connecting nearby spatial elements. Similarly, the station graph links nearby stations. Both graphs enable transfer and sharing of information between the connected elements to help the GNN identify likely source hypocenters and association assignments. Training GENIE involves generating synthetic data that covers a wide range of station configurations, source distributions, and pick sets. Synthetic catalogs are created by sampling network realizations, computing arrival times, corrupting data with noise by a certain percentage, and adding a percentage of false picks to the dataset. The generation of training data can make use of homogeneous or 1D velocity models. This diverse approach to training ensures that the model is exposed to a wide range of scenarios. GENIE supports both CPU and GPU processing.

PyOcto (Münchmeyer, 2024) employs a 4D space-time partitioning strategy inspired by the Oct(o)tree data structure. This way, PyOcto focuses computational resources on promising origin regions and reduces complexity. To minimize runtime, PyOcto discards event-free nodes early and uses a priority queue to scan promising nodes first. Once a node has reached a critically small size, PyOcto locates and outputs the event. PyOcto removes picks associated with the event from the input set to avoid duplicate associations. To model travel times, PyOcto supports homogeneous and 1D velocity models. PyOcto uses parallelization across different time blocks, to optimize CPU usage.

3 Benchmarking approach

3.1 Event-station scenarios

We conduct our benchmark study with two typical examples of seismic network geometry and seismicity depth range: a crustal seismicity scenario and a subduction zone scenario. Both scenarios are designed to replicate real-world conditions in terms of station density and distribution as well as the range of hypocentral depths. Note that we do not use real seismicity distributions but prefer events that are randomly distributed in space to test the algorithms' ability to detect arbitrarily located events (see Section 3.2). The station distributions and 1D velocity models for both scenarios are based on existing seismic network deployments (Figure 1) and geological settings.

For the crustal seismicity scenario, we use a set of stations from the Southern California Seismic Network (California Institute of Technology and United States Geological Survey Pasadena, 1926) and the 1D velocity model of Hadley and Kanamori (1977). Seismic events are randomly generated within the region depicted in Figure 1 (right) following a uniform distribution and covering the depth range of 0-30 km.

The subduction zone scenario employs the station distribution of the IPOC (Integrated Plate Boundary Observatory Chile; GFZ German Research Centre for Geosciences and Institut des Sciences de l'Univers-Centre National de la Recherche CNRS-INSU, 2006) CX seismic network in Northern Chile and the 1D velocity model of Graeber and Asch (1999). As for the crustal scenario, we generate seismic events randomly distributed in space and time in a uniform way, but cover a much larger range of hypocentral depths, from shallow crustal to intermediate-depth intraslab earthquakes (0-250 km; see Figure 1, left).

3.2 Synthetic pick/event generation

We create our synthetic benchmark datasets by partially following the approach outlined by McBrearty and Beroza (2023). From randomly generated origin times and hypocentral locations, we generate labeled P and S arrival times at the different stations with the respective 1D velocity models and station distributions. The dataset construction process comprises the following steps:

1. **Event Location and Timing Selection**: Event locations are randomly generated within the station network of the scenario being simulated. Origin times are arbitrarily assigned within a 24-hour time span.

2. **Arrival Time Calculation**: For each event, at all stations, we compute P- and S-wave arrival times using the NonLinLoc raytracer (http://alomax.free.fr/nlloc/) and a 1D velocity model

3. **Arrival-Time Data Corruption**: To simulate real-world arrival time heterogeneity due to 3D velocity structure as well as picking errors, arrival times are perturbed by adding random noise. While picking errors are often modeled as Gaussian (Diehl et al., 2009), velocity model uncertainties can lead to higher-tailed distributions of error, so we perturb the arrival time data by uniform random noise proportional to travel time (-1 to 1% of travel time).

4. **Application of Distance Threshold**: We assign a randomly determined cutoff distance threshold (uniformly between 70 km and 150 km for the crustal scenario and between 160 km and 500 km for the subduction scenario) to each event. All arrivals from source-station paths exceeding this limit are deleted. This can be seen as a rough approximation of event magnitude.

5. **Station Dropout**: A percentage of the stations (20%) are randomly deleted for each event to introduce operational variability.

6. **False Pick Integration**: The process concludes with the incorporation of a predefined percentage (30%, 100% and 300%) of additional false picks (or "noise picks"), effectively simulating automatic

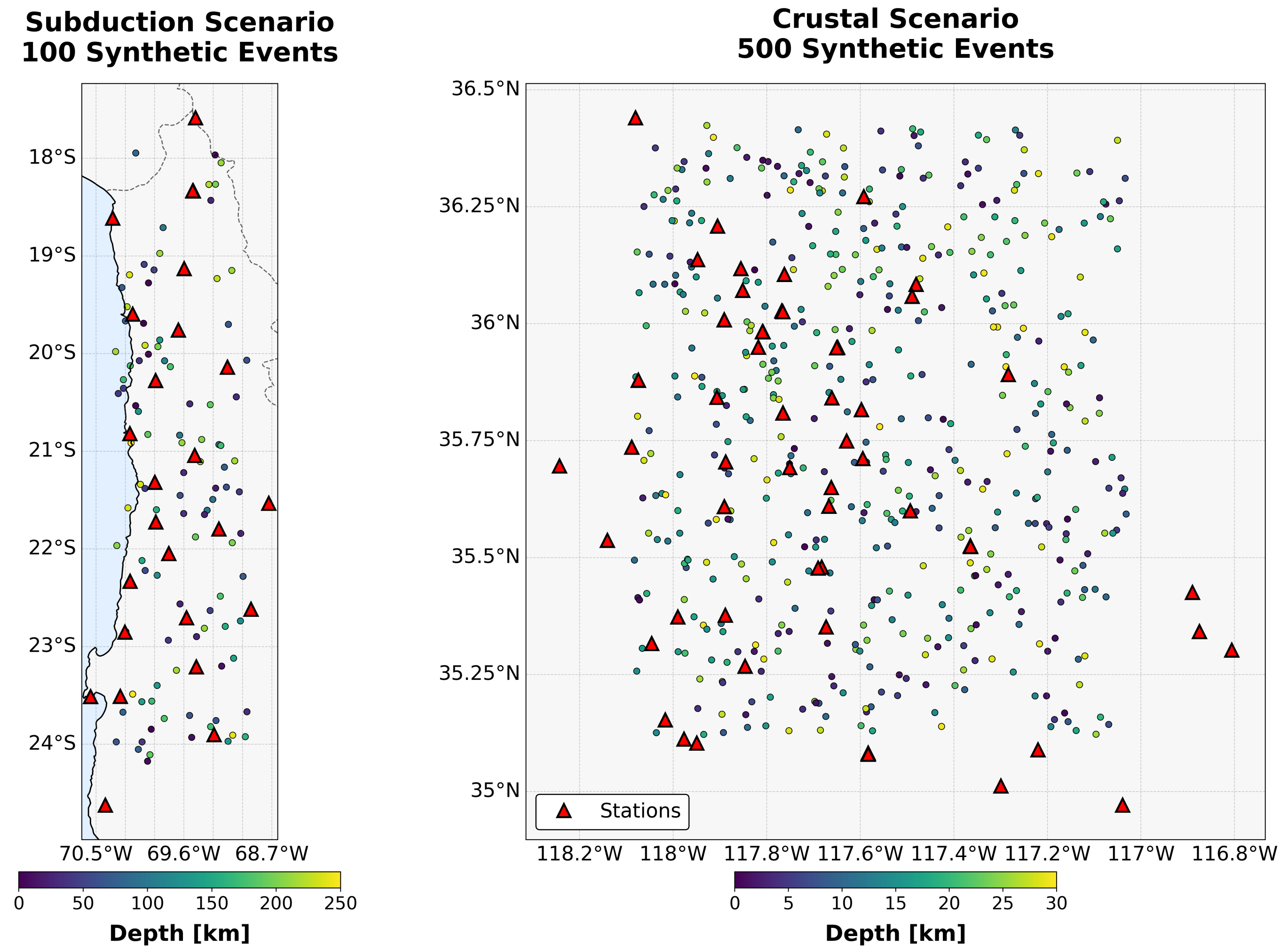

Figure 1 **Left**: Station configuration and synthetic seismic event distribution example for the subduction zone scenario. Red triangles represent the IPOC network's seismic stations, colored dots an example set of 100 synthetic events. The fixed station layout, combined with variable event densities and noise levels, forms the basis of our associator evaluation. **Right**: Station configuration and synthetic seismic event distribution example for the shallow seismicity scenario. We show an example realization with 500 events.

picker outputs that often contain many picks that do not belong to actual earthquakes. These false picks are randomly uniformly distributed over time, stations, and phase type.

Details of the resulting event and pick distributions are provided in Tables S1 and S2. The distance threshold ensures the generation of a diverse set of events, including "large-moveout" events, that are detected across the majority of the seismic network, as well as "small-moveout" events, that are detected by a limited amount of stations (see pick count distributions in Figure S1). The synthetic seismicity we use is randomly distributed across the regions, different from real-world patterns where seismicity is concentrated near active faults, or inside the downgoing slab in subduction zones. However, for purposes of performance evaluation, the approach of evaluating all possible event locations, whether they are tectonically likely or not, has the advantage that it ensures the associators can also detect events in areas that have not previously had seismicity.

Although we attempt to design our synthetic scenarios in a realistic way, a number of complications that exist in real-world applications are still neglected. For instance, a real-world subduction zone dataset will most likely contain out-of-network events offshore. The level of arrival time noise we assume may easily be exceeded in real applications, and we unrealistically assumed that a station always has both a P- and S-pick. We thus perform our synthetic experiments in two main steps. The main set of experiments is performed with the above approach for creating synthetic datasets, and performance is evaluated for different amounts of events within 24 hours as well as different proportions of noise picks. After this evaluation, we perform a suite of tests where we introduce additional real-world problems such as having different proportions of out-of-network events, higher travel-time noise levels, and increased rates of missed picks. We evaluate the effect on performance each of these complications has one-by-one (see Section 4.4). The same synthetic-data generator, implemented independently of any individual associator's training pipeline (PhaseLink, GENIE), is applied to all algorithms during testing. None of the associators were trained or fine-tuned on these benchmark datasets, ensuring that every method is evaluated on previously unseen inputs.

3.3 Performance evaluation approach

To assess the performance of the seismic phase associators, we employ a set of evaluation metrics at both the event level and the pick level. This means that we first check how many events were correctly retrieved, how many were missed and how many false events were created from noise picks. We consider an event correctly retrieved if the associator yields an event that contains at least 50% of the picks originally created for the synthetic event. In this way, we ensure that the original set of picks can not create more than one real event, and the loss of a fraction of real picks does not affect whether or not the event is correctly retrieved.

On the pick level, we then evaluate how many picks are correctly associated to an event (commonly associated picks), how many are missed (missed picks), how many are wrongly associated (i.e., picks from one event that get associated to a different one) and how many false picks are added to an event. Ground truth picks are the arrivals assigned to each synthetic event; we generate one synthetic catalog per experiment and feed it unchanged to all associators. Predicted picks are the picks retrieved for this event (may contain correctly associated, wrongly associated, and false picks).

We employ the following set of metrics:

- **Precision**: Measures the proportion of true positives (TP) in the entire output. High precision indicates few false positives (FP).

$$Precision = TP/(TP + FP) \tag{1}$$

  On the event level, TP corresponds to correctly identified events, FP to false or additional events that were associated from ground truth or noise picks. For the pick level analysis, TP marks the amount of picks correctly associated to an event, whereas FP is the sum of the number of noise picks added to the event and the number of wrongly associated picks that stem from other events.

- **Recall**: Measures the proportion of true positives compared to all ground truth correct associations. High recall indicates few false negatives (i.e., most actual events or picks were detected).

$$Recall = TP/(TP + FN) \tag{2}$$

  On the event level, TP again corresponds to the correctly identified events, while FN are the ground truth events that are missed. At pick level, TP are the picks correctly associated to an event, and FN are the ground truth picks from that event that are missing in the associated event.

- **F1 Score**: The harmonic mean of precision and recall, providing a balance between sensitivity and accuracy. A high F1 score indicates strong overall performance in correctly determined detections while minimizing false detections.

$$F1 = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall} \right) \tag{3}$$

- **Runtimes**: Runtime is a crucial metric for practical implementations. Here, we evaluate the processing speed of each associator. We measure the time from the initiation of the model to the generation of its outputs. It is important to note that our measurement does not take into account any preprocessing steps such as the construction of the velocity model and travel time tables, or the model training for the DL-based associators, because these are processes usually executed only once within a given application framework. Our experiments are conducted using a consistent computational environment: all associators are run on systems utilizing 25 CPU threads, with access to 200 GB of RAM. For DL-based associators that leverage GPU acceleration, such as GENIE and PhaseLink, we use an NVIDIA A40 GPU for both training and inference.

Finally, for completeness we repeat the entire analysis without the 50% match threshold, i.e., including all false-positive events. Resulting pick-level heat maps (Figures S9 and S10) and extended pick budget bar plots (Figures S11 and S12) are provided in the Supplement.

3.4 Parameter optimization approach

The performance of each association algorithm is heavily dependent on the choice of tuning parameters. Except for the association threshold, which defines the minimum number of picks needed to define an event, the different algorithms have very different parameter sets, a consequence of their quite different approaches. In order to provide a fair comparison between the different algorithms, we have to optimize the parameter choices for each of them, which is a time-consuming activity. For the sake of comparability and also to mimic real-world applications, we chose an association threshold of 10 picks for declaring an event (without specification how many of them have to be P or S) for all associators.

For REAL, GaMMA and PyOcto, we then conduct a large series of runs, changing parameters one by one and evaluating the change in performance metrics in response to these changes. While varying parameters individually is not the optimal approach, conducting a complete grid search would be computationally prohibitive. Where available, we used published parameter choices from an earlier associator comparison (Münchmeyer, 2024) or application studies (Becker et al., 2024) as an initial parameter guess. An example of optimizing a single parameter for PyOcto is shown in Figure 2: we systematically vary the parameter *pick match tolerance*, and determine metrics like precision, recall, F1 score (on event and pick level) as well as runtime for each of these trial runs. The parameter choice with the overall

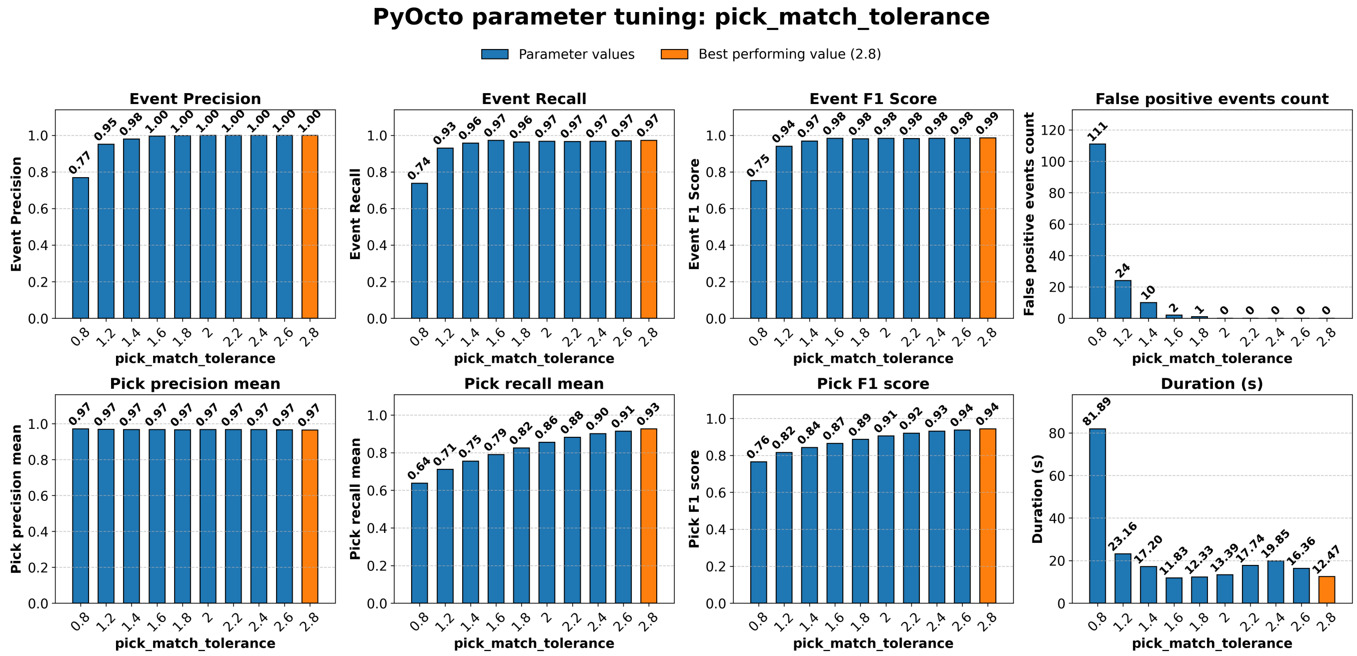

Figure 2 Example of our parameter optimization approach, here for parameter *pick match tolerance* of PyOcto. The metrics event- and pick-level precision, recall and F1 score, as well as runtime and false positive count, are monitored against a systematic change of this parameter. For the run shown here, the choice marked in orange is evaluated to perform best. Note that we do not show the entire extent of the utilized search space here, values >2.8 were also tested. The finally chosen optimum parameters are determined by comparing performance for all nine runs (with 100, 500 and 2000 events as well as 30, 100 and 300% of noise picks) that we evaluate in Section 4.4.

best performance, as indicated by these different metrics, is chosen (here highlighted in orange color). We optimized two separate sets of parameter choices for the crustal and subduction zone scenario. For each of these sets, the final parameter choice is a compromise between the optimizations on all nine different runs (all combinations of 100, 500 and 2000 events as well as 30, 100 and 300% noise picks). That is, once selected, the same set of parameters is used for all tests, regardless of the number of noise picks and event rates.

The neural network based algorithms, PhaseLink and GENIE, require a training step before application, in which the majority of parameter optimization occurs. Because this step is time-consuming, the iterative tuning strategy as used for the traditional associators is not possible, and only a minimal amount of parameter tuning was possible for these methods. To create the training datasets for these algorithms we used the codes available with each method, which follow a similar approach of synthetic pick and event creation as outlined in Section 3.2. For details of the training process for PhaseLink and GENIE, please refer to Text S2 and Figure S4 in the Supplementary Material, as well as the descriptions supplied in the original publications. All our final parameter choices for each associator and scenario are listed in Tables S3–S7 in the Supplementary Material.

4 Results

We evaluated the performance of the five seismic phase associators — PhaseLink, REAL, GaMMA, GENIE, and PyOcto — in the two different event-station scenarios introduced in Section 3.1. GaMMA, REAL and PyOcto offered the possibility of using either a homogeneous seismic velocity (0D model) or a 1D velocity model for the association process. We tested the different configurations and here only used their best-performing configurations as identified by our analysis (see Text S1, Figures S2 and S3, and Table S8 in the Supplementary Material). For each of the two event-station scenarios, we performed a total of 9 different runs, which feature different event numbers (100, 500 and 2000 events within 24 hours) as well as different proportions of randomly distributed “noise picks” (30, 100 and 300% of the true picks).

4.1 Event-Level Performance Metrics

The event-level results are presented in Figure 3 for the subduction scenario and Figure 4 for the crustal scenario. The full numerical results are available in Tables S8 and S9. At low noise levels (30% noise) and small event counts (100 events), all associators demonstrated high event-level precision and recall, with relatively minor differences between different scenarios and associators. While GENIE, PyOcto and REAL showed values above 0.97 for precision, recall and F1 score in both scenarios, GaMMA obtained lower scores around 0.9 for the subduction zone scenario, and PhaseLink scored around or even below 0.9 for both scenarios. With higher noise levels and event counts, the performance of the different associators diverged significantly. This was especially true for the subduction zone scenario, where the performance drops for the more difficult

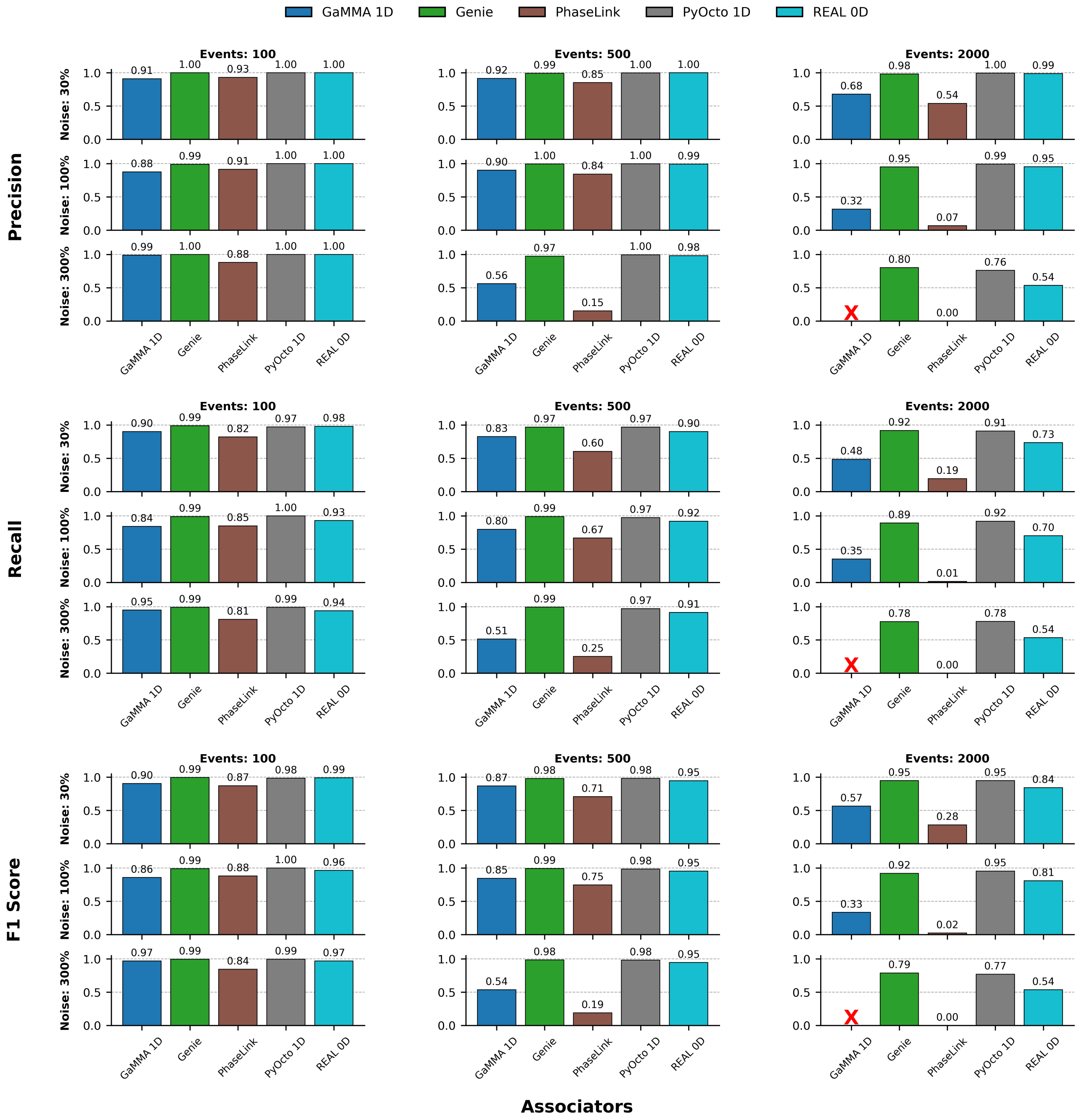

Figure 3 Event-level performance in the subduction zone scenario. The figure is a 3 × 3 grid. Columns share the subduction zone scenario event rate (100, 500, 2000 earthquakes; titles above each column). Rows share the evaluation metric—Precision, Recall, F1-score (labels on the left). Within every small panel three horizontal bar clusters track increasing catalogue noise (30%, 100%, 300% false picks; annotated on the left axis of each panel). Each cluster contains the five associators: GaMMA 1D, Genie, PhaseLink, PyOcto 1D, REAL 0D. The numeric value printed on each bar is the mean metric over the events recovered in that run. A red "×" replaces a bar where the algorithm could not finish the run (out-of-memory or other error). Alternative visualization using heatmaps is supplied in Figure S13 in the Supplementary Material.

runs were more pronounced than for the crustal scenario.

Adding more noise picks to the smallest run with only 100 events had no major impact on performance, whereas increasing event numbers deteriorated performance values more clearly. Of all associators, PhaseLink exhibited the most drastic performance drops with increasing event numbers and noise percentages. In most crustal scenarios, except for the most difficult case of 2000 events and 300% noise, PhaseLink performs reasonably well with metrics largely above 0.8. In contrast, for the subduction zone scenario, PhaseLink registered a sharp decline from an F1-score of ≈0.87 in simpler runs down to ≈0.1 in the most challenging scenarios. There, it already had low precision, recall and F1 score values below 0.3 for the run with 500 events and 300% noise as well as for all runs with 2000 events. For 2000 events and more than 100% noise, its F1 score was nearly zero. Overall, PhaseLink's precision results were higher than its recall values. GaMMA

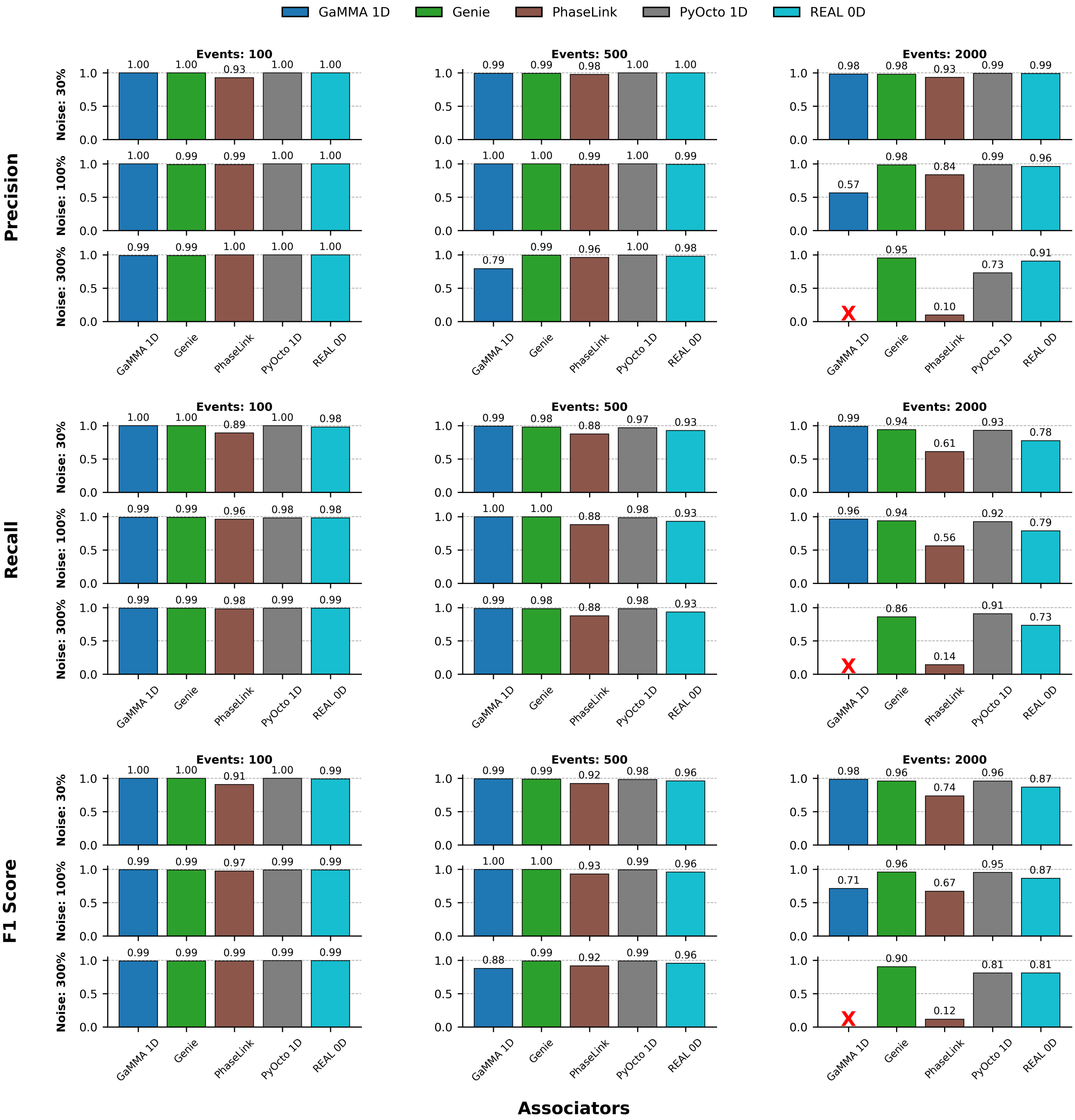

Figure 4 Event-level performance in the crustal zone scenario. The figure is a 3 × 3 grid. Columns share the crustal scenario event rate (100, 500, 2000 earthquakes; titles above each column). Rows share the evaluation metric—Precision, Recall, F1-score (labels on the left). Within every small panel three horizontal bar clusters track increasing catalogue noise (30%, 100%, 300% false picks; annotated on the left axis of each panel). Each cluster contains the five associators: GaMMA 1D, Genie, PhaseLink, PyOcto 1D, REAL 0D. The numeric value printed on each bar is the mean metric over the events recovered in that run. A red "×" replaces a bar where the algorithm could not finish the run (out-of-memory or other error). Alternative visualization using heatmaps is supplied in Figure S14 in the Supplementary Material.

performed markedly better than PhaseLink overall, but also exhibited a performance drop of F1 values to between 0.5 and 0.55 already in the high-noise case of 500 events for the subduction zone scenario. In the crustal scenario, it achieved clearly better results than in the subduction zone scenario, with a clear performance drop only for the case with 2000 events. For the most complex runs (2000 events with 300% noise), GaMMA did not complete the processing due to memory allocation issues. The high computational demands of clustering large volumes of data with significant noise led to excessive memory usage for GaMMA and exceeded the available RAM (200 GB). There is a clear tendency of reduced precision with more stable recall for GaMMA when moving to the more challenging runs in the crustal scenario, while no such systematic tendency could be seen for the subduction zone scenario.

REAL achieved overall good results in the crustal scenario, with metrics above 0.9 everywhere except for the runs with 2000 events. There, its recall dropped

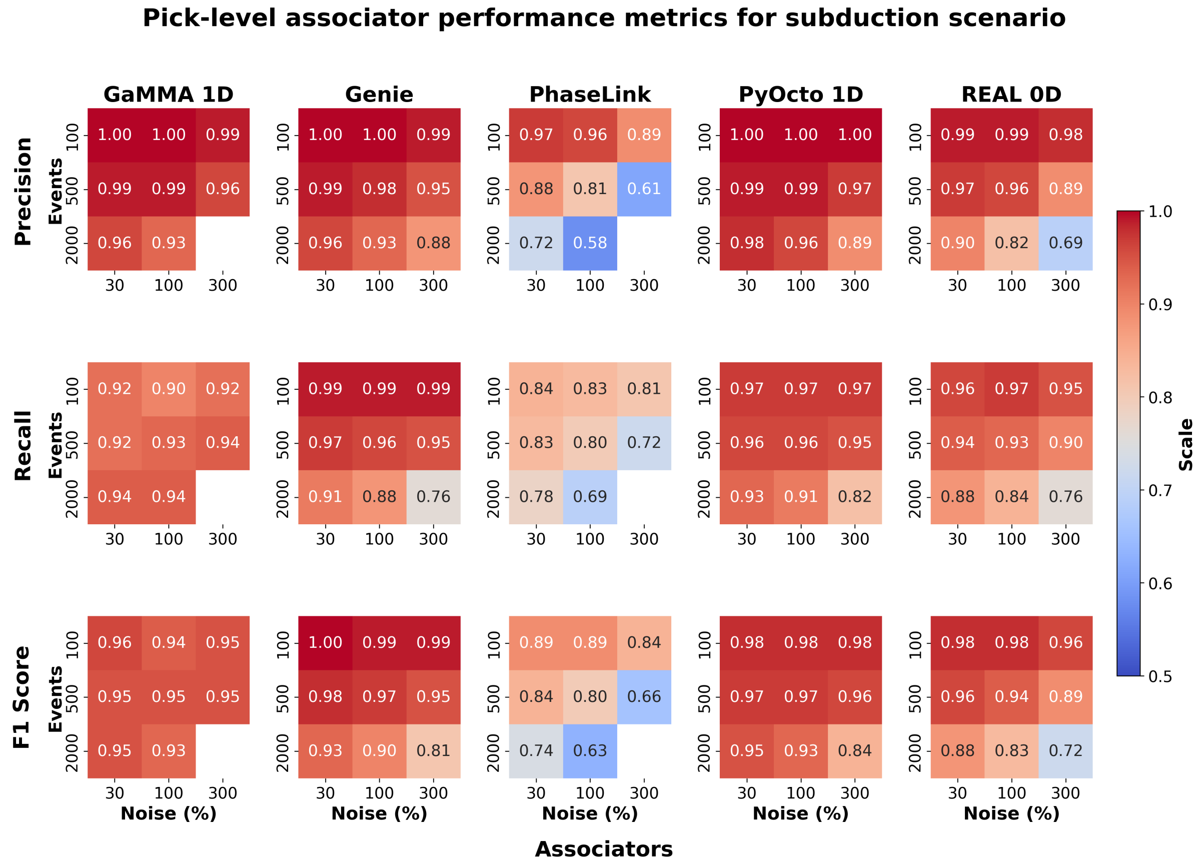

Figure 5 Pick-level performance of five associators in the subduction scenario. Columns group the associators (GaMMA 1D, GENIE, PhaseLink, PyOcto 1D, REAL 0D). Rows give the three metrics: Precision, Recall, and F1 score, shown once as row labels on the left. Within each heat-map the x-axis steps through increasing catalogue noise (30%, 100%, 300% false picks) and the y-axis through higher pick rates (100, 500, 2000 events). The three rows display, from top to bottom, precision, recall, and F1-score; warmer colors indicate better performance according to the shared scale bar. Blank cells mark runs that were not completed. Each panel shows the mean performance derived from events that matched the synthetic ground truth.

more significantly (to values around 0.75) than its precision (still above 0.9) for the most challenging runs. This constitutes a significantly better performance than GaMMA. In the subduction zone scenario, REAL likewise only experienced a significant performance drop for the runs with 2000 events, but here it performed worse than in the crustal scenario, with the recall dropping to around 0.7 already for low-noise conditions. Again, REAL's precision is generally higher than recall, but both decrease to 0.54 for the most challenging run. Finally, GENIE and PyOcto achieved the highest scores throughout the different runs, with only very minor differences between the two algorithms. Their metrics were above 0.97 for all runs with 100 or 500 events, in both scenarios. For the runs with 2000 events, precision and recall stayed above 0.9 except for the very last run with 300% noise. There, they both dropped just under 0.8 for precision and recall in the subduction scenario, whereas GENIE obtained a higher precision than PyOcto (0.95 vs. 0.73) at similar recall (0.86 vs. 0.91) in the crustal scenario.

4.2 Pick-Level Performance Metrics

Due to the previously used definition of an event being correctly identified if it contained at least 50% of the original (ground truth) picks, event-level metrics did not fully indicate which associator had a tendency to miss picks or to incorporate "noise picks" into correctly retrieved events. Such information became apparent when analyzing the performance on the pick level. Here, we perform this pick-level analysis. Picks were classified as correctly associated (CA), wrongly associated (WAP) or false as described in Section 3.3 and precision, recall and F1 scores were calculated similarly to the event-level metrics. While the presented differences in pick-level performance did not have direct consequences on event retrieval (only correctly retrieved events were evaluated in Section 4.1), missing picks and especially the incorporation of erroneous picks could have had a large impact on the quality of the final seismicity catalog, leading to wrong and more uncertain hypocentral locations and magnitudes if no additional post-processing was applied. It should be noted that these pick-level results were derived only from the events that were successfully retrieved, i.e., that exceeded the threshold of 50% matching picks to the ground truth event. This criterion ensured that only events with a significant overlap between the predicted dataset and the ground truth synthetic dataset were considered. This implies that the set of events considered differed between the different associators, and it also means that additional false events that may have

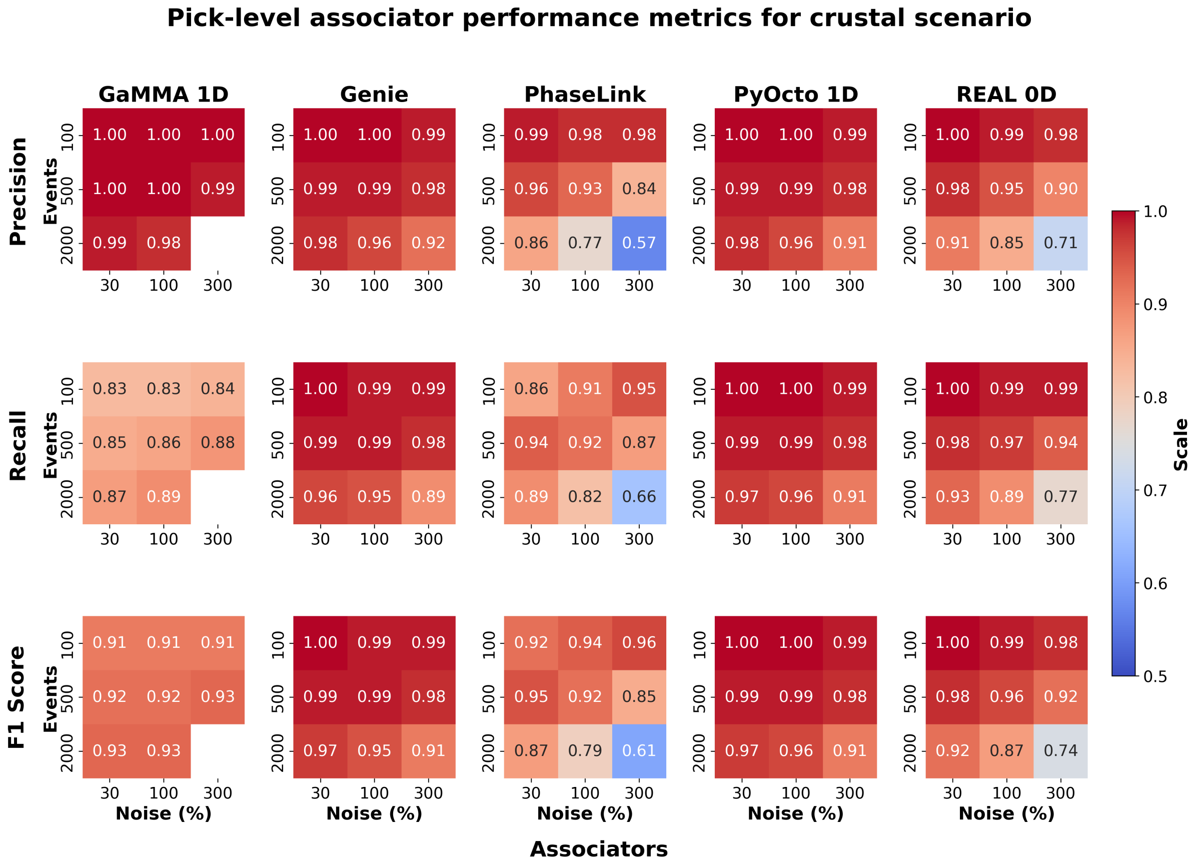

Figure 6 Pick-level performance of five associators in the crustal scenario. Columns group the associators (GaMMA 1D, GENIE, PhaseLink, PyOcto 1D, REAL 0D). Rows give the three metrics: Precision, Recall, and F1 score, shown once as row labels on the left. Within each heat-map the x-axis steps through increasing catalogue noise (30%, 100%, 300% false picks) and the y-axis through higher pick rates (100, 500, 2000 events). The three rows display, from top to bottom, precision, recall, and F1-score; warmer colors indicate better performance according to the shared scale bar. Blank cells mark runs that were not completed. Each panel shows the mean performance derived from events that matched the synthetic ground truth.

been created from the remainder of ground truth picks, noise picks or a mixture of the two, did not impact the pick-level metrics. Hence, these pick-level metrics did not take into account event-level precision, which decreased proportional to the extent that false events were created, and which could be highly variable between different algorithms, as shown in Figures 3 and 4.

Heat maps in Figures 5 and 6 show the mean values of precision, recall, and F1 score at pick level. Figures S7 and S8 show the mean values for the pick-level results per event (ground truth picks, predicted picks, commonly associated picks, missed picks, false picks, and wrongly associated picks) across the different associators and runs. All values from these figures are also provided numerically in Tables S11 and S12 in the Supplementary Materials. The observed general performance trends are largely similar to the event level ones. At low noise levels and smaller event counts, all associators demonstrated high pick-level accuracy, which deteriorated with increasing event numbers and noise picks. GENIE and PyOcto again showed the highest accuracy, with performance metrics above 0.9 in nearly all cases, and retained values above 0.8 even under the most adverse conditions. REAL nearly matched their performance in the smaller-scale runs, but performed worse in the runs with 2000 events, where it obtained values around 0.7 for the most challenging run with 2000 events and 300% noise. GaMMA featured high precision, but recall did not exceed 0.92 even for the smallest and simplest runs, and it failed to finish the hardest case due to memory issues. Moreover, low event-level precision for GaMMA, especially in the subduction zone scenario, implies that it created many secondary events with falsely associated picks. Lastly, PhaseLink had the weakest overall results, with performance deteriorating (values below 0.8) already at the intermediate-difficulty runs, and nearly zero performance for the hardest runs.

The detailed pick statistics (Figures S7 and S8) can be read as a six-subplot pick budget for every run. Subplot (e) lists the ground truth picks available for each event, while subplot (a) lists the predicted picks returned by the associator. Their intersection is plotted in subplot (d) as common picks—the part of the catalog that is matched correctly. Picks present in the ground truth but absent from the prediction appear in subplot (b) as missed picks and represent omission errors. Picks present in the prediction but absent from the ground truth are divided between two commission-error classes: false picks created from noise (subplot c) and wrongly associated picks that originate from a different event (subplot f). Together, the six subplots expose whether performance losses under increasing catalog noise and event rate originate primarily from dropping real information (subplot b) or from injecting spurious information (subplots c and f). Subplot (a) shows that most associators tended to miss an average of one or two picks per event even for the easiest runs, whereas the incorporation of false or wrongly associated picks is virtually zero there (subplots c and f). As the runs became more demanding, more picks were missed, but this was largely compensated by also incorporating more false or wrongly associated picks, so that the average total number of picks per event did not change significantly. PhaseLink started to miss large amounts of picks already in the intermediate difficulty scenarios and at the same time incorporated many false or wrongly associated picks. For the hardest test case, PhaseLink did not retrieve any correct events in the subduction scenario, which is why missed, false, and wrongly associated picks for PhaseLink were zero for this case. For the other associators, the tendency to miss or wrongly incorporate picks was less strong than for PhaseLink. GaMMA missed a substantial amount of picks (an average of 14 per event in the crustal scenario) in the easier runs, and this proportion of missed picks stayed relatively stable across the different runs. REAL's performance was close to the level of GENIE and PyOcto throughout most of the runs but deteriorated faster for the highest event rates, where it missed more picks and incorporated more noise or wrongly associated picks than these algorithms.

A complementary "all-events" assessment retaining every prediction, including pure false positive events, is shown in Figs. S9 and S10. As expected, the inclusion of these extra events depresses overall precision (and thus F1), while recall is affected to a lesser extent, because most true events are still recovered. However, the qualitative behaviour visible in the matched-event analysis persists: algorithms that already tended to tolerate more noise or cross-associated picks now manifest that tendency through a larger population of spurious events, whereas methods that were more conservative at the pick level continue to return cleaner catalogs. In other words, the shape of the performance curves remains the same, only their absolute scale shifts downward once false positives are counted. Detailed pick budgets for the matched-event view—already discussed above in Figures S7 and S8 are complemented by full-catalog budgets in Figures S11 and S12, where the same six-panel breakdown is applied after all predicted events (including false positives) have been retained.

4.3 Processing runtimes

The last performance metric we evaluated is processing runtime, as introduced in Section 3.3. Figure 7 shows summaries of runtimes for all different associators and evaluated runs (values are also listed in Tables S9 and S10 in the Supplementary Material), which were represented by the total number of picks (ground truth plus noise). Although the crustal scenario featured a larger amount of stations (Figure 1) and thus more picks by a factor of 3–4, runtimes were generally slower for the subduction scenario. This was likely a consequence of events being distributed over a larger spatial region, as well as extending to much deeper depths. This increased the search space for potential sources and may also necessitate more complex travel time calculations. Processing times generally increased with scenario size, but the different associators showed very different scaling behavior. While PyOcto, PhaseLink and partially also GaMMA finished the smaller scenarios in less than or around 10 seconds, REAL and especially GENIE were slower by an order of magnitude or more. For higher pick rates, it was apparent that the neural network-based associators (PhaseLink and GENIE) had better scalability than the other methods, in that their runtimes grew less strongly with an increasing number of picks. PyOcto, REAL and GaMMA had more significant processing time growth with total pick numbers, with GaMMA's curve being the steepest. However, since GENIE was quite slow for small scenarios, this flatter curve only meant that its processing time is similar to PyOcto and somewhat faster than REAL for the largest scenarios we evaluated. PhaseLink, on the other hand, clearly processed large-scale problems fastest, but due to its near-zero performance for such cases (Section 4.1) it is still not an effective choice for processing such datasets. One could also observe that for GENIE and PhaseLink, which are based on neural networks, the amount of noise picks did not influence the total processing time significantly, whereas it played a major role for the other, more classical associators.

4.4 Further tests

In this section we address several real-world data complexities not included in our main synthetic experiments from Section 4.1. First, the signal detection time error was less than ±1% of the predicted travel time, thus the results did not address the effects of increasing pick errors. Second, out-of-network events are a common occurrence in most monitoring environments, especially in subduction zones where most of the plate interface as well as the often seismically active outer rise are located offshore (Stern, 2002). Third, each station's associations always included both the P arrival and S arrival. Fourth, we did not characterize small magnitude event performance as the synthetic events were detected on most of the monitoring network. All of these conditions are typically not present in real-world applications; instead, travel time noise levels may be higher, and most events will be of small magnitude and hence only observed on a small fraction of the network. Thus, we also evaluated the deterioration in accuracy that occurs when these complications are increased to more challenging real-world levels. We conducted these additional tests on the intermediate subduction zone scenario with 500 events and 100% noise picks. In a first set of runs, we systematically increased travel time noise levels (see Figure 8), then moved on to introduce different proportions of out-of-network events to the west of the station network (see Figure 9), and finally removed a higher proportion of P- or S-phases (see Figure 10), which emulates the creation of smaller magnitude events. The evaluation of these additional com-

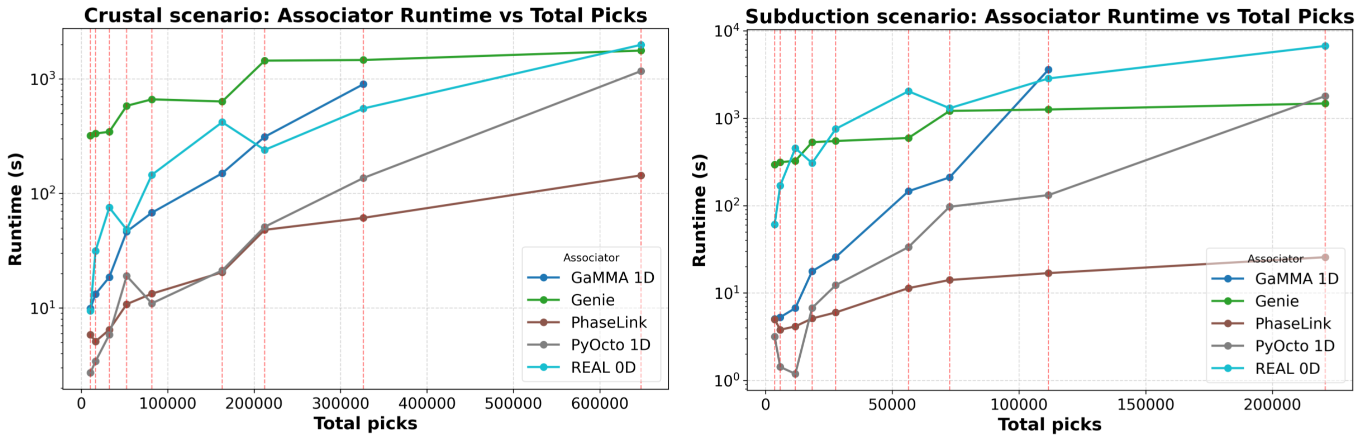

Figure 7 Left: Logarithmic scale comparison of processing duration against the total number of picks for various seismic phase associators for the crustal seismicity scenario. Right: Logarithmic scale comparison of processing duration against the total number of picks for various seismic phase associators for the subduction scenario. Note that GaMMA 1D did not complete processing for the most complex case in both scenarios due to memory allocation issues, and thus its runtime is not plotted for those cases.

plications complements the main analysis presented in Sections 4.1 through 4.3.

4.4.1 Travel time noise

To test travel time noise effects on association, we conducted three additional runs with noise added from random uniform distributions of ±1–5%, ±5–10% and ±10–15% of travel time. Results from these runs, in addition to the one with the original ±0–1% noise, are shown in Figure 8. REAL, GaMMA and PyOcto have tolerance-type parameters (REAL: *nrt*; GaMMA: *max_sigma11*; PyOcto: *pick_match_tolerance*) that put an upper bound on what misfit between predicted and observed arrival times was permissible. In a first series of runs, we kept the tolerance parameters fixed at the same values as determined in our previous optimizations. We then re-optimized these single parameters for each of these associators and runs, which in some cases yielded significantly better results (see hatched and filled bars in Figure 8). The neural network based associators, PhaseLink and GENIE, were kept with their originally chosen parameters, as these methods appeared less sensitive to travel time noise levels. However, if these models were re-trained for the higher expected noise levels this would likely increase performance further.

We observed a general performance decay of all associators with increasing travel time noise level, with a clearer decrease of recall values compared to precision. PhaseLink appeared to be least affected by increasing travel time noise levels, but since its performance was already relatively low for the low noise levels of the original application, it was still not among the best performing algorithms for the highest noise levels. PyOcto and GaMMA showed a large dependence on the re-optimization, with the sometimes very low event recall levels of the original tolerance parameter choices (below 0.25 for the high-noise case) significantly improving to around 0.6 (GaMMA) or 0.85 (PyOcto) with new, more adequate choices of tolerance parameters. For REAL, in contrast, re-optimizing the tolerance parameter only brought a subtle performance increase even with high noise, as it already performed quite well (recall above 0.75) with the original setting. While increasing the tolerance parameter led to better metrics for GaMMA and PyOcto, it also caused a notable increase in false and wrongly associated picks that were incorporated into the retrieved events. GENIE maintained >0.75 F1 scores for both event-level and pick-level metrics even at the highest travel time noise levels, despite not being re-optimized for these higher noise level cases.

These results underscore the importance of parameter optimization, and illustrate the inherent tradeoff between robustness against travel time noise and the incorporation of noise picks that REAL, GaMMA and PyOcto exhibited. The two associators using a neural network approach, PhaseLink and GENIE, are more flexible with respect to travel time noise and largely did not need to be re-optimized once they were properly trained.

4.4.2 Out-of-network events

The correct identification and accurate location of out-of-network events represents a major challenge in seismology (Williamson et al., 2023). We thus conducted three additional runs where we added an additional 150, 300 and 450 out-of-network events to the subduction zone scenario run with 500 events and 100% noise picks. These events were randomly placed up to 200 km west of the network and at depths of 0–40 km. We did not perform any additional parameter optimization for these runs, but used the previously determined optimal parameters. In Figure 9, we show the performance for the in-network events with bar charts in the left panel, and the retrieval of out-of-network events (only for the case with 450 such events) in the map plots on the right. Pick association performance of in-network events was only slightly degraded between the “Original” (no out-of-network events) and the “High” (450 out-of-network events) runs. For precision, PhaseLink experienced the largest drop (12.99%), while the re-

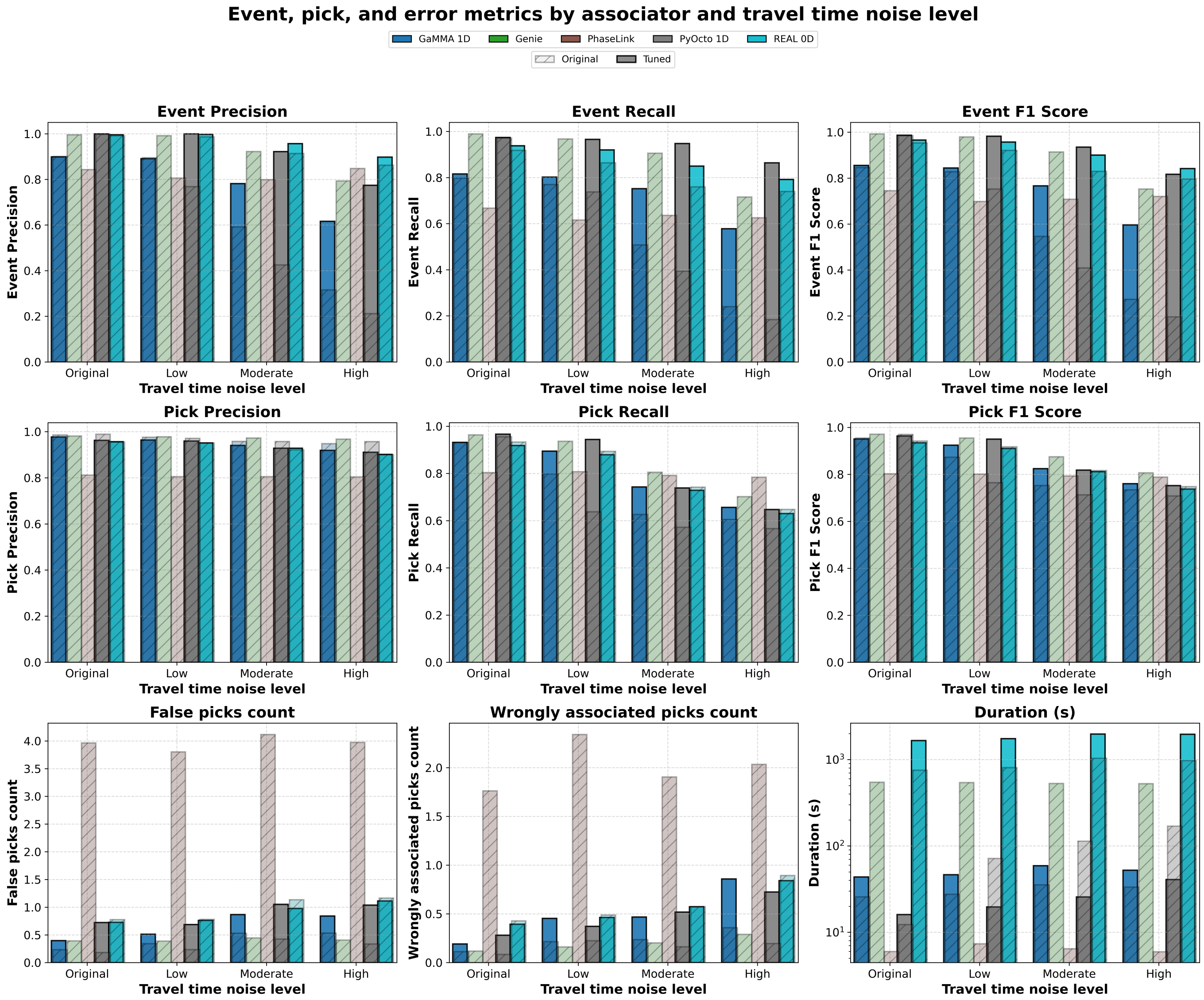

Figure 8 Event-level (upper row), pick-level (middle row) and other (lower row) metrics for runs with systematically changed levels of travel time noise. The original configuration corresponds to the run with 500 events and 100% noise picks of the subduction zone scenario. To simulate low, moderate and high noise conditions, ±1–5%, ±5–10% and ±10–15% of the travel time were added as noise to the picks. Note that for REAL, GaMMA and PyOcto, two different runs are shown, one with the original parameter optimization (hatched bars) and one with their tolerance parameters re-optimized for each noise level (solid bars).

maining associators saw only modest declines (all under about 4%). For recall, PhaseLink again fell the most (3.65%), with the other algorithms losing no more than roughly 2.5%. Finally, in F1 score, PhaseLink's performance decreased by 8.49%, whereas all other methods dropped by less than about 3%. However, the algorithms differed markedly in how well they retrieved out-of-network events. In all cases, the event retrieval rate declined with distance from the network, but the nature of this decline was different among all associators. In our original parameter optimization, out-of-network events had not been expected, so that the permissible search area for all algorithms was set to 71°W (Note: as REAL does not incorporate this parameter, it was not affected). When keeping this choice, PyOcto and GENIE performed very well for closeby out-of-network events, but then showed a sharp decline in retrieval rate in the close vicinity of this boundary. This means that events only slightly outside this search space limit would have been missed, highlighting the importance of setting the appropriate range for a given monitoring scenario. For GaMMA and PhaseLink, a substantial amount of closeby out-of-network events was missed, but a small proportion of events beyond the search area limit were retrieved as well. For REAL, the search space is not user-configured. Our results show that it had a high event retrieval rate that declined with distance from the network. At distances that roughly correspond to the location of the seismically active outer rise in a subduction zone, REAL still retrieved ≈80% of the events.

When re-configuring GaMMA and PyOcto to include all out-of-network events in the search space, PyOcto's performance surpassed that of REAL, although it also started to miss events towards the western edge of the event cloud. GaMMA only retrieved a significant proportion of events west of the network center (71° W),

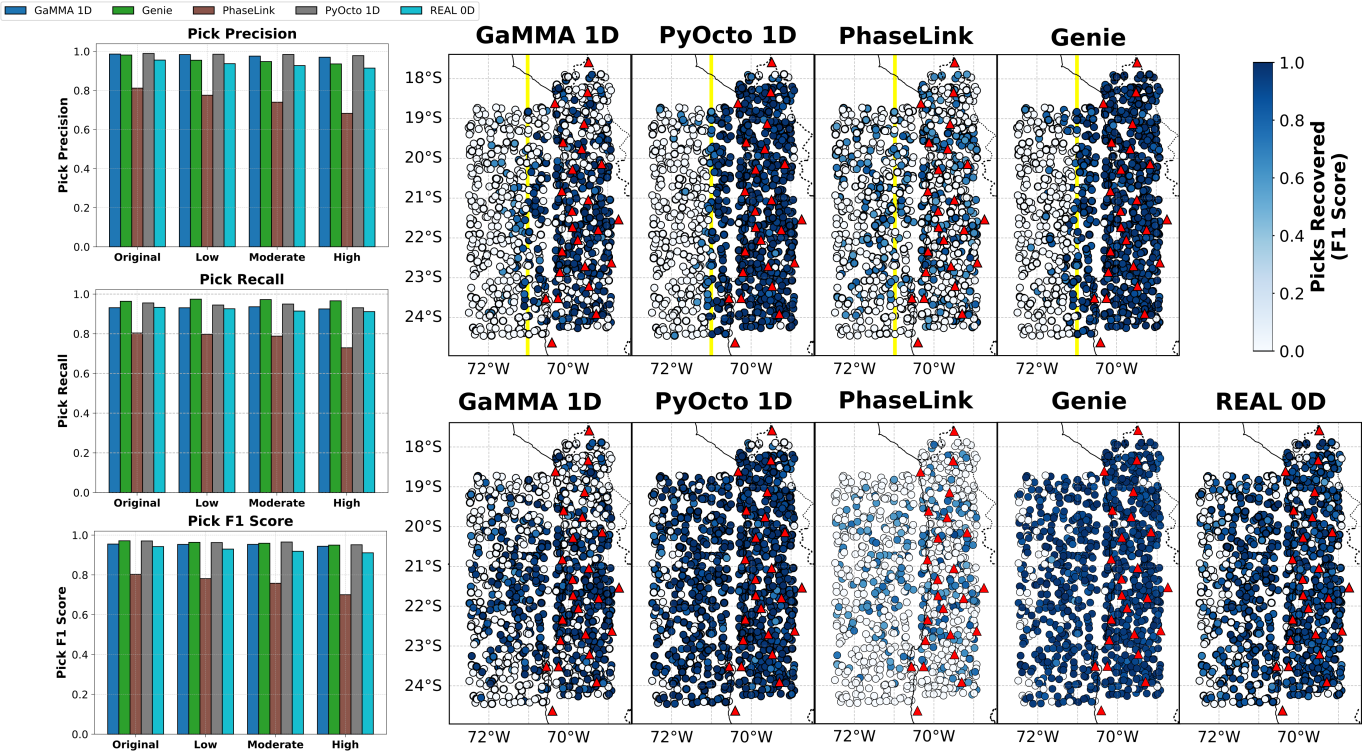

Figure 9 Test results for including 150, 300 and 450 out-of-network events placed up to 200 km west of the network, at depths between 0 and 40 km. The left panel shows performance metrics for the in-network events for the different runs, on the right map view plots for the run with 450 out-of-network events are shown that indicate the performance of the different associators for the single events that are colored by pick-level F1 score. The upper row contains runs with a yellow line that shows the western edge of the search area in those runs where it falls within the out-of-network events. The lower row contains the runs with extended longitude, which include all out-of-network events.

while very few events were found at the northern (19°S) and southern (24°S) ends of the out-of-network event cloud. For the deep learning-based associators, GENIE and PhaseLink were re-trained specifically for this extended scenario. In particular, GENIE was re-trained using the latest public version of the model, which includes improved travel time computation. The updated GENIE model not only maintained high performance for in-network events but also achieved strong pick-level recall and F1 scores for out-of-network events, even exceeding the original model's performance under more complex conditions with added out-of-network events. While we retained the original model version for consistency across all associators in the main analysis, these results indicate that GENIE's performance has further improved in recent versions available on GitHub, highlighting the continuous development of associators. PhaseLink, on the other hand, showed a slight decrease in performance after retraining. This may be due to the larger spatial extent introduced in the out-of-network runs: although the model was retrained with the same number of epochs and synthetic samples as in the original setup, the expanded search space likely requires a larger or more diverse training dataset to maintain performance.

4.4.3 Removal of P- or S-phases

In a last additional test, we removed different amounts (20, 40 and 60%) of each event's picks, randomly between stations and P or S phases. This was meant to investigate the associators' performance in case many stations only had one pick, and not both paired P and S picks. At the same time, this run modified the original distribution of pick numbers (Figure S1) in which only a relatively small proportion of events had pick numbers close to the association threshold of 10, creating a more realistic configuration in which most events emulate small magnitude events and only slightly exceed the association threshold.

Results from this test are shown in Figure 10. The average number of ground truth picks per event, shown in the lower left panel, demonstrated the decrease of total picks for the different runs, where the last run of 60% removed picks only had about 15 picks per event on average, which is close to the association threshold of 10 picks. The scores for event precision and recall show that event precision actually increased with pick removal for most associators, most clearly for GaMMA and PhaseLink. At the same time, event recall deteriorated for all associators. As in most other runs, GENIE and PyOcto showed the best overall performance, with scores >0.8 throughout all runs, and only minor differences between them. Here, PyOcto was slightly better for the run with the highest pick removal rate, whereas GENIE had minimally higher scores for the runs with fewer picks removed. REAL showed effective results, but its performance was systematically inferior to GENIE and PyOcto by about 0.1 in precision and recall. While PhaseLink performed poorly throughout

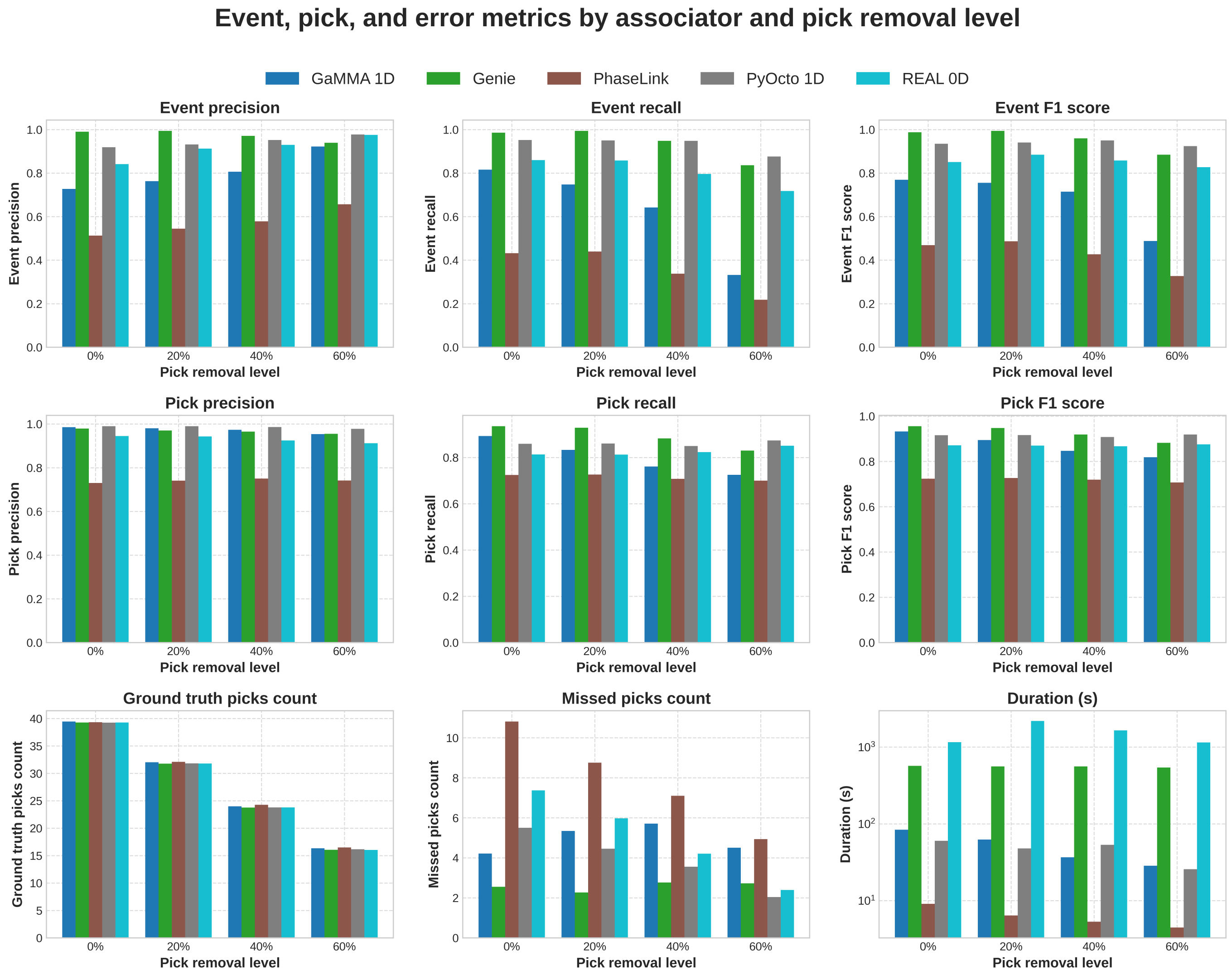

Figure 10 Results of the random pick removal test. Starting from the subduction zone scenario with 500 events and 100% noise picks, we remove 20, 40 and 60% of picks (randomly P or S) to simulate more sparsely detected events. Results are provided in the same form as in previous figures.

all runs, GaMMA's recall also decreased significantly (to only about 0.3) for the highest rate of pick removal. For this run, the average event only comprised 15 picks, and as GaMMA missed more than 4 picks per event on average, this led to many events moving below the association threshold and thus not being detected.

5 Discussion

5.1 Associator configuration or training

As briefly outlined in Section 3.4, each of the used associators requires the tuning of a number of parameters. Parameter choices are specific to the setting, to conditions such as station distribution or the amount and quality of input picks. This means that in all cases, a certain amount of tuning to the setting at hand is required, and none of the algorithms can be generalized to perform well 'out-of-the-box'. In real-world use cases, ground truth catalogs in the form of verified picks and events may not exist, so users must rely on experience and conducting and analyzing test runs to configure parameters. The amount of effort and expertise that is required to properly configure and apply the different algorithms also differs widely. In this Section, we discuss some of the tradeoffs that are inherent to the parameter optimization and comment on the practical use of the different algorithms. We acknowledge that any synthetic-data protocol can, at least in principle, favor certain algorithmic features. What guards us against that here is both the breadth of our generator and the diversity of methods we tested. All five associators, ranging from grid-search (REAL) and Gaussian-mixture clustering (GaMMA), through recurrent DL sequence models (PhaseLink) and graph-based deep learning (GENIE), to purely back-projection (PyOcto), saw identical event locations, station-dropout patterns, velocity models (0D vs. 1D), noise levels, and distance thresholds. Despite this uniform test bed, two fundamentally different schemes, GENIE's GNN and PyOcto's 4D octree search, both sustain near-perfect F1 scores across all test cases, while the other methods show more variable performance. That consistency across two very distinct algorithms suggests that any residual bias in our benchmark is likely small rather than fatal to the comparison.

For the backprojection-based associators (REAL, PyOcto) as well as GaMMA, a tolerance-type parameter defines the allowed arrival time misfit for picks. This parameter has to be adapted to the expected travel time noise level (see Section 4.4.1) due to pick uncertainties or the deviation of the used velocity model from reality, and a suboptimal choice can have severe consequences for associator performance (e.g., Figure 8). A too high value will lead to the incorporation of noise picks and thus decreased precision, whereas a too small value will lead to many true picks being missed (lower pick-level and event-level recall). While such a tolerance-type parameter also exists implicitly in the training process of the DL algorithms and the level of travel time noise added to training picks, its role in defining the quality of achievable phase associations is less prominent. Secondly, a choice of grid density or refinement level is required, which leads to a second substantial tradeoff. A very fine parameterization will typically lead to improved performance, though it can severely increase runtimes, while a too coarse parameterization will lead to quick runtimes but inferior results.

While the previous parameter tradeoffs have to be addressed when configuring PyOcto, REAL or GaMMA, these algorithms nevertheless feature a relatively limited set of parameters that need tuning, which means that finding a suitable (while maybe not optimal) configuration is not very time-intensive. Training the neural networks for PhaseLink and GENIE needs a higher amount of effort and expertise, and it could take a substantial amount of time to find a well-working setup for new users. In particular, the choice of the parameters used for the generated synthetic picks in training (e.g., proportion of noise picks, event density, levels of travel time noise, etc.) are important, yet may be hard to tune. The choice of the level of noise and event rates during training the DL associators implicitly affects the precision and recall tradeoffs, however directly assessing this tradeoff is difficult as it requires multiple rounds of training. For GENIE, it is required to set a few scale-dependent parameters such as the maximum moveout distance of sources, and the label kernel widths. The level of travel time noise and event rates can also be chosen to roughly reflect realistic conditions. In case of a real-world application, it is important to use the real data characteristics to guide the choice of training data parameters, but this process may necessitate some trial-and-error until a working configuration is found.

The advantage offered by the DL algorithms is flexibility, which is illustrated in the test runs with different noise levels (Figure 8). Once properly trained, PhaseLink and GENIE generally do not require parameter adaptation to perform well in a wide range of settings, while REAL, GaMMA or PyOcto have to be adjusted in case different conditions are encountered. Thus, the higher amount of initial investment in training the network can result in increased flexibility. For GENIE, since it relies on graph neural networks, this flexibility also extends to handling significantly different station configurations, e.g., if a seismic network is heavily modified over time by adding or removing stations, robustness can be maintained without requiring re-training. For example, GENIE performs well when trained on a dense network of 91 stations and then applied to a much smaller subnetwork of 21 stations, while PhaseLink has to be re-trained for such an application (see Text S3 and Figures S5 and S6 in the Supplementary Material). Supplementary tests that vary station density (Text S3 and Fig. S6) reinforce this contrast. With exactly the same hyper-parameter set, GENIE maintains F1 scores above 0.9 as the California network is thinned from 91 to 21 stations, whereas PhaseLink's F1 drops from ≈0.9 to ≈0.4 unless the model is retrained on the sparse geometry. This robustness stems from GENIE's graph-based architecture, which generalizes across networks of different aperture and density without parameter retuning. In practice, that translates to lower maintenance overhead in deployments where stations are frequently added, removed, or temporarily out of service. By contrast, PhaseLink and to a lesser extent, the classical back-projection methods, benefit from dedicated retraining or re-optimization whenever network geometry changes markedly.

Training GENIE and PhaseLink is a one-off but nontrivial step: for each scenario we generated ≈1 million synthetic samples (Section 3.2; Text S2) and optimized the networks over many epochs on a modern GPU. Although this requires several GPU-hours, subsequent inference is lightweight. Practitioners without dedicated GPUs can still deploy these methods by starting from publicly available pretrained weights, or by retraining with much smaller synthetic datasets (50–100 k samples); our pilot runs retained ≥90% of the full-training accuracy while cutting compute cost by an order of magnitude.

5.2 Event duplicates and multiplets

An issue we did not analyze in detail is the possible creation of duplicate or multiplet events by phase associators. As none of the associators allows a single pick to be used by more than one event, our event definition of ≥50% of ground truth picks ensures that only one output event per ground truth event is analyzed. Whether additional false events with smaller amounts of ground truth picks, possibly mixed with noise picks, are created was not evaluated independently. However, this effect is encoded in the statistics for event-level precision, pick-level recall as well as missed picks (Figures 3, 4, 5, 6, S7 and S8).

In the subduction zone scenario, both GaMMA and PhaseLink have decreased event-level precision even for the simplest runs, which does not occur in the crustal scenario. In both cases, both algorithms also show lower values for pick-level recall, which is due to missed picks. This likely implies that while picks are simply missed in the crustal scenario, they are at least sometimes combined to secondary events in the subduction case (thus the lower event-level precision). This may be due to the larger spatial search space in this scenario, which allows more possibilities for a secondary event to achieve a consistent source location with several picks "by chance". Interestingly, decreasing the number of constituent picks per event (Section 4.3 and Figure 10) increases the event-level precision of GaMMA and PhaseLink substantially. We interpret that this observation implies that if a smaller total number of picks are available, producing a false secondary event with more than 10 arrivals is less likely.

5.3 Runtime trends and applicability

The runtime trends observed across the different associators (Section 4.3) show significant differences in scalability and computational efficiency. The DL-based associators GENIE and PhaseLink here demonstrate superior scalability compared to more traditional methods, but in the case of GENIE this is coupled with much slower runtimes especially in smaller scenarios. Runtimes of PhaseLink and GENIE mostly scale with the number of identified events and are largely independent of the number of noise picks, whereas an increase of the total number of picks drastically increases runtimes of REAL, GaMMA and PyOcto.

Comparing directly between the best-performing algorithms PyOcto and GENIE, PyOcto is substantially faster (factor of 100 or more) for the smaller-scale applications we tested, while for the largest runs that encompass $>$100k picks, runtimes are roughly similar between these two algorithms. While the largest scenario we tested, which contains 2000 events within 24 hours on a network of $\sim$50 stations, is already quite extreme in terms of event rate (likely corresponding to the aftershock series of a large earthquake), many current (or future) seismic networks can include 100s or even 1000s of stations. In such cases, an algorithm such as GENIE may be beneficial, and the advantage in runtime for such large datasets as well as its flexibility towards network geometry changes over time may easily outweigh the larger effort in initially training the model. For seismic networks of more limited scale, i.e., many regional and local as well as temporary deployments of $\sim$dozens of stations, PyOcto is potentially the most appropriate choice of associator, as it achieves similar performance as GENIE, is much faster, and requires relatively limited parameter configuration before application.

6 Conclusions

We evaluated five phase association algorithms with scenarios of synthetic picks and events that were designed to approximate real-world conditions. We find that GENIE and PyOcto show the overall best performance across all tested scenarios and runs. These two algorithms are the most recently published algorithms, and are also based on very different techniques: one uses neural networks, while the other uses an efficient back-projection based search scheme. Our results indicate distinct advantages and tradeoffs of each algorithm and do not allow a decision of which fundamental phase association approach (classical or DL) is superior.

While GaMMA and especially PhaseLink showed significant problems in more challenging conditions, REAL exhibited robust performance overall, but has slow runtimes due to its grid search-like approach. PyOcto and GENIE performed best, with only small differences between the two algorithms except for runtimes. There, PyOcto is substantially faster (factor of $\sim$100) for smaller-scale problems, whereas GENIE catches up for larger problems due to better scalability. For the largest problems we tested, their runtimes were comparable. However, greater differences between the two algorithms may appear for larger seismic network applications, and are also indicated by the need for different levels of re-tuning based on observed seismicity characteristics and noise levels.

Taking into account additional considerations such as parameter tradeoffs and ease of configuration, we conclude that PyOcto is well suited for most phase association problems today, unless they feature very high seismicity rates coupled with more than hundreds or thousands of seismic stations. In this latter case, the better runtime scaling as well as the flexibility towards network geometry changes can be significant assets for GENIE. Such applications may become more commonplace in the future, as instrumentation is ever increasing globally.

Acknowledgements

C.S. received funding from the European Research Council (ERC) through the Horizon 2020 program (ERC Starting Grant MILESTONE; StG2020-947856). J.P. and C.S. were supported by the Czech Academy of Sciences through a Mobility Plus (MPP) Grant, grant number ANID-23-08. I.M. was supported by AFRL under contract PA04-S-D00243-12-TO-03-Stanford. J.M. was supported by the European Union under the grant agreement n°101104996 (“DECODE”). This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (ID:90254).

Data and code availability

No actual data was used in this article, only synthetic experiments were conducted, which can be repeated based on the information given in the paper. The five tested phase association algorithms, PhaseLink (https://github.com/interseismic/PhaseLink), REAL (https://github.com/Dal-mzhang/REAL), GaMMA (https://github.com/AI4EPS/GaMMA), GENIE (https://github.com/imcbrearty/GENIE) and PyOcto (https://github.com/yetinam/pyocto), are all freely available for download under the provided links, and installation instruction as well as documentations are provided. The optimal sets of tuning parameters we derived are given in the Supplementary Material (Tables S3–S7).

The utilized raytracer is contained in the NonLinLoc software package (http://alomax.free.fr/nlloc/), the 1D velocity models can be found in the respective publications (Graeber and Asch, 1999; Hadley and Kanamori, 1977). For our different station configuration scenarios, we used real station locations from the networks CX in Chile (GFZ German Research Centre for Geosciences and Institut des Sciences de l’Univers-Centre National de la Recherche CNRS-INSU, 2006), and networks CE (California Geological Survey, 1972), CI (California Institute of Technology and United States Geological Survey Pasadena, 1926), GS (Albuquerque Seismological Laboratory (ASL)/USGS, 1980), NN (University of Nevada, Reno, 1971), NP (U.S. Geological Survey, 1931), PB (https://www.fdsn.org/networks/detail/PB/) and ZY (https://www.fdsn.org/networks/detail/ZY_1990/) in California.

Competing interests

Authors have no competing interests.

References

Albuquerque Seismological Laboratory (ASL)/USGS. US geological survey networks, 1980. doi: 10.7914/SN/GS.

Allen, R. V. Automatic earthquake recognition and timing from single traces. *Bull. Seismol. Soc. Am.*, 68(5):1521–1532, 1 Oct. 1978. doi: 10.1785/BSSA0680051521.

Becker, D., McBrearty, I. W., Beroza, G. C., and Martínez-Garzón, P. Performance of AI-based phase picking and event association methods after the large 2023 MW 7.8 and 7.6 Türkiye doublet. *Bull. Seismol. Soc. Am.*, 29 May 2024. doi: 10.1785/0120240017.

Bishop, C. M. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New York, NY, first edition, 17 Aug. 2006.

California Geological Survey. California strong motion instrumentation program, 1972. doi: 10.7914/B34Q-BB70.

California Institute of Technology and United States Geological Survey Pasadena. Southern California seismic network, 1926. doi: 10.7914/SN/CI.

Diehl, T., Kissling, E., Husen, S., and Aldersons, F. Consistent phase picking for regional tomography models: application to the greater Alpine region. *Geophys J Int*, 176(2):542–554, 1 Feb. 2009. doi: 10.1111/j.1365-246X.2008.03985.x.

Ester, M., Kriegel, H., Sander, J., and Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD*, pages 226–231, 2 Aug. 1996.

GFZ German Research Centre for Geosciences and Institut des Sciences de l'Univers-Centre National de la Recherche CNRS-INSU. IPOC Seismic Network, 2006. doi: 10.14470/PK615318.

Graeber, F. M. and Asch, G. Three-dimensional models ofP-wave velocity andP-to-Svelocity ratio in the southern central Andes by simultaneous inversion of local earthquake data. *J. Geophys. Res.*, 104(B9):20237–20256, 10 Sept. 1999. doi: 10.1029/1999jb900037.

Hadley, D. and Kanamori, H. Seismic structure of the Transverse Ranges, California. *GSA Bulletin*, 88(10):1469–1478, 1 Oct. 1977. doi: 10.1130/0016-7606(1977)88<1469:SSOTTR>2.0.CO;2.

Hainzl, S., Sippl, C., and Schurr, B. Linear relationship between aftershock productivity and seismic coupling in the northern Chile subduction zone. *J. Geophys. Res. Solid Earth*, 124(8): 8726–8738, Aug. 2019. doi: 10.1029/2019jb017764.

Johnson, C. E., Bittenbinder, A., Bogaert, B., Dietz, L., and Kohler, W. Earthworm: A flexible approach to seismic network processing. *Iris newsletter*, 14(2):1–4, 1995.

Maharaj, A., Roecker, S., Alvarado, P., Trad, S., Beck, S., and Comte, D. Are volatiles from subducted ridges on the Pampean flat slab fracking the crust? Evidence from an enhanced seismicity catalog. *Geochem. Geophys. Geosyst.*, 24(10), Oct. 2023. doi: 10.1029/2023gc011021.

Mancini, S., Segou, M., Werner, M. J., Parsons, T., Beroza, G., and Chiaraluce, L. On the use of high-resolution and deep-learning seismic catalogs for short-term earthquake forecasts: Potential benefits and current limitations. *J. Geophys. Res. Solid Earth*, 127(11):e2022JB025202, Nov. 2022. doi: 10.1029/2022JB025202.

McBrearty, I. W. and Beroza, G. C. Earthquake Phase Association with Graph Neural Networks. *Bull. Seismol. Soc. Am.*, 113(2): 524–547, 1 Apr. 2023. doi: 10.1785/0120220182.

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. *Sci. Rep.*, 9(1):1–14, 16 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. *Nat. Commun.*, 11(1):3952, 7 Aug. 2020. doi: 10.1038/s41467-020-17591-w.

Münchmeyer, J. PyOcto: A high-throughput seismic phase associator. *Seismica*, 3(1), 29 Jan. 2024. doi: 10.26443/seismica.v3i1.1130.

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. *J. Geophys. Res. [Solid Earth]*, 127(1), Jan. 2022. doi: 10.1029/2021jb023499.

Poiata, N., Satriano, C., Vilotte, J.-P., Bernard, P., and Obara, K. Multiband array detection and location of seismic sources recorded by dense seismic networks. *Geophys. J. Int.*, 205(3): 1548–1573, 1 June 2016. doi: 10.1093/gji/ggw071.

Ringdal, F. and Kværna, T. A multi-channel processing approach to real time network detection, phase association, and threshold monitoring. *Bulletin of the Seismological Society of America*, 79 (6):1927–1940, 1 Dec. 1989.

Ross, Z. E., Meier, M., Hauksson, E., and Heaton, T. H. Generalized Seismic Phase Detection with Deep Learning. *Bull. Seismol. Soc. Am.*, 108(5A):2894–2901, 1 Oct. 2018. doi: 10.1785/0120180080.

Ross, Z. E., Yue, Y., Meier, M.-A., Hauksson, E., and Heaton, T. H. PhaseLink: A deep learning approach to seismic phase association. *J. Geophys. Res. [Solid Earth]*, 124(1):856–869, Jan. 2019. doi: 10.1029/2018jb016674.

Sippl, C., Schurr, B., John, T., and Hainzl, S. Filling the gap in a double seismic zone: Intraslab seismicity in Northern Chile. *Lithos*, 346-347(105155):105155, Nov. 2019. doi: 10.1016/j.lithos.2019.105155.

Stern, R. J. Subduction zones. *Rev. Geophys.*, 40(4):3–1–3–38, Dec. 2002. doi: 10.1029/2001rg000108.

University of Nevada, Reno. Nevada Seismic Network, 1971. doi: 10.7914/SN/NN.

U.S. Geological Survey. United States national strong-motion network, 1931. doi: 10.7914/SN/NP.

White, M. C. A., Fang, H., Catchings, R. D., Goldman, M. R., Steidl, J. H., and Ben-Zion, Y. Detailed traveltime tomography and seismic catalogue around the 2019 *Mw*7.1 Ridgecrest, California, earthquake using dense rapid-response seismic data. *Geophys. J. Int.*, 227(1):204–227, 18 June 2021. doi: 10.1093/gji/ggab224.

Williamson, A., Lux, A., and Allen, R. Improving out of network earthquake locations using prior seismicity for use in earthquake early warning. *Bull. Seismol. Soc. Am.*, 27 Jan. 2023. doi: 10.1785/0120220159.

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. *Seismol. Res. Lett.*, 93(3): 1695–1709, 1 May 2022. doi: 10.1785/0220210324.

Xiong, Q., Brudzinski, M. R., Gossett, D., Lin, Q., and Hampton, J. C. Seismic magnitude clustering is prevalent in field and laboratory catalogs. *Nat. Commun.*, 14(1):2056, 12 Apr. 2023. doi: 10.1038/s41467-023-37782-5.

Yang, S., Hu, J., Zhang, H., and Liu, G. Simultaneous earthquake detection on multiple stations via a convolutional neural network. *Seismol. Res. Lett.*, 92(1):246–260, 1 Jan. 2021. doi: 10.1785/0220200137.

Zhang, M., Ellsworth, W. L., and Beroza, G. C. Rapid Earthquake Association and Location. *Seismol. Res. Lett.*, 90(6):2276–2284, 1 Nov. 2019. doi: 10.1785/0220190052.

Zhu, L., Peng, Z., McClellan, J., Li, C., Yao, D., Li, Z., and Fang, L. Deep learning for seismic phase detection and picking in the aftershock zone of 2008 M7.9 Wenchuan Earthquake. *Phys. Earth Planet. Inter.*, 293(106261):106261, Aug. 2019. doi: 10.1016/j.pepi.2019.05.004.

Zhu, W. and Beroza, G. C. PhaseNet: a deep-neural-network-based seismic arrival-time picking method. *Geophys. J. Int.*, 216(1): 261–273, 13 Oct. 2018. doi: 10.1093/gji/ggy423.

Zhu, W., McBrearty, I. W., Mousavi, S. M., Ellsworth, W. L., and Beroza, G. C. Earthquake phase association using a Bayesian Gaussian mixture model. *J. Geophys. Res. [Solid Earth]*, 127(5), May 2022a. doi: 10.1029/2021jb023249.

Zhu, W., Tai, K. S., Mousavi, S. M., Bailis, P., and Beroza, G. C. An end-to-end earthquake detection method for joint phase picking and association using deep learning. *J. Geophys. Res. Solid Earth*, 127(3), Mar. 2022b. doi: 10.1029/2021jb023283.