Response to Reviewers

Manuscript title: "Benchmarking seismic phase associators: Insights from synthetic scenarios"

Jorge Puente[1,2], Christian Sippl[1], Jannes Münchmeyer[3], and 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.*

June 25, 2025

Cover letter

Dear *Dr. Marlon Ramos*,

We thank you and the reviewers for the thorough evaluation of our manuscript *"Benchmarking seismic phase associators: Insights from synthetic scenarios."* The very constructive feedback helped us to refine the narrative and figures. Below we summarize the principal changes; detailed, point-by-point responses follow in the next sections.

- **Conciseness and active voice**. Re-wrote verbose or conversational sentences, replaced qualitative terms (e.g. "significant") with exact numbers, and converted Results to past tense.
- **Single synthetic catalog clarified**. Added an explicit sentence and notes in the captions of figures S7, S8, S11 & S12 to state that every associator receives the same ground-truth catalog.
- **Potential GENIE bias addressed**. Inserted a short paragraph in Section 3.2 and in the Discussion that states why the shared data-generation routine does not favor GENIE.
- **Broader context**. Added discussion of joint detection–association frameworks (e.g., Poiata et al. 2016) and of the computational cost of training DL associators for practitioners.
- **Station-density robustness**. We now highlight GENIE's parameter stability versus PhaseLink's retraining requirement.
- **Updated GENIE and PhaseLink OON runs**. We re-trained GENIE (using its updated version with improved travel-time calculations) and PhaseLink for the extended out-of-network (OON) test scenario. GENIE improved further under these more challenging conditions. PhaseLink, however, slightly declined in performance, likely because the same number of training samples and epochs were used despite the increased spatial extent. We updated the manuscript and clarified this in the out-of-network section.
- **Figure improvements**. Bar plot figures redrawn as heat-maps added as alternative visualizations; fonts enlarged; redundant axis labels removed; legends repositioned; bar-plot pick statistics moved to the Supplement.
- **Minor fixes**. Corrected "backprojection" typo, added numeric percentages throughout, etc.

We hope these revisions fully address the reviewers' concerns and enhance the manuscript's clarity and practical value.

Thank you for considering the revised submission. We look forward to your decision.

Sincerely,
Jorge Puente (corresponding author)
on behalf of all co-authors

Reviewer 1

Reviewer comment:
There is a lot of information in this manuscript, simplifying and/or removing unnecessary phrasing would help with the flow. Where possible, I recommend using an active voice. I also recommend eliminating conversational phrasing. I have given a few examples but encourage the authors to address this throughout their manuscript.

Author response:
We agree and have streamlined the text throughout:

- Re-wrote passive or conversational sentences in active voice.
- Deleted fillers such as "clearly," "significantly," "reasonably," or replaced them with exact numbers.
- Cut redundant phrases, shortening the manuscript

Reviewer comment:
Question: Your synthetic generation process is based on GENIE's training data generation. Is a risk that you have "gamed" the experiments in GENIE's favor. Can you comment on why this is not the case? Briefly commenting on this/acknowledging this in the paper is all I am looking for.

Author response:
We appreciate the reviewer's concern. While our synthetic data generation process is inspired by the same general principles used for GENIE training (e.g., random event locations, 1D velocity models, varied noise levels), we designed the benchmark pipeline independently and applied it consistently across all associators. Importantly, the synthetic events and noise conditions used for testing differ from those used to train GENIE, and GENIE was not specifically tuned to the benchmark datasets. We now briefly acknowledge this point in the manuscript to clarify and avoid the impression of a strong bias.
We added the statement "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." in lines 191-193
We also added the following statement to lines 534-542:
"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."

Reviewer comment:
There are several places where qualitative words are used to describe increases/decreases: "significantly," "drastically," etc... Instead, list the exact increase or decrease/absolute counts or percentages. This will strengthen the arguments.

Author response:
Done. Throughout the text we replaced qualitative adverbs (e.g., "significantly," "drastically," "clearly") with the corresponding numbers.

Reviewer comment:
Most critical comment about the work: Line 190: "Ground truth picks refers to the picks 191 that the synthetic event actually has..." After reading the full manuscript and seeing slight variability in the average number of GT picks for each algorithm for a given experiment, it raises some concern. It seems as though you have generated a new synthetic event GT bulletin per algorithm for a given experiment, i.e. Low Noise, Low Events GT bulletin used for GENIE is not the same as the Low Noise, Low Events GT bulletin for GaMMA, etc... Assuming I have not misunderstood the setup: is not the case, why not use the same GT bulletin for each experiment? Even though the high-level GT statistics are the same, you have introduced variability that is not necessary and raises the question of how do you know that the results are not because of some other aspect of the dataset? My guess is that your results would still hold, but the added complexity seems unnecessary and raises the question of why do this?

Author response:
We apologize for the ambiguity. For every experiment (e.g., "Low-Noise / 100-events") we generate one synthetic bulletin and pass that identical bulletin to all five associators; no algorithm receives a bespoke dataset.
The small fluctuation you noticed in the "ground-truth-pick" bars stems only from how the mean is computed: the average in Figs 6 & 8 is taken over the events that each associator successfully recovers ($\geq$50% pick overlap). An algorithm that misses many low-pick-count events will, by definition, average over a slightly larger-pick subset, hence its mean bar can differ from its peers even though the underlying bulletin is the same.
We added a clarification in the manuscript : Ground truth picks are the arrivals assigned to each synthetic event; one synthetic catalog is generated per experiment and fed unchanged to all associators
Also we added an extra clarification in the caption of Figures S7, S8, S11, S12: "The slight algorithm-to-algorithm variation in panel (e) arises because averages are taken only over the events each method recovered"

Reviewer comment:
I suggest the Results section be rewritten to use past tense. For example, on line 276, rephrase "GaMMA does not complete the processing due to memory..." to "GaMMA did not complete the processing..."

Author response:
Implemented. We converted the entire Results section from present to past tense.

Reviewer comment:
Line 269: "... its accuracy deteriorates significantly in the subduction zone scenario." Suggested: "... its accuracy degraded by X% in the subduction zone scenario." Remove qualitative words such as "clearly." For example Lines 533-534 "PyOcto and GENIE clearly..." At this point you have made the case for this, just state it. In other cases, describe the "why" behind "clearly," or remove it.

Author response:
Implemented.

Reviewer comment:
Lines 267-269: "While it still performs reasonably (metrics largely above 0.8) in most crustal scenario runs except for the most difficult case of 2000 events and 300% noise, its accuracy deteriorates significantly in the subduction zone scenario." Recommended: "In most crustal scenarios GaMMA performs reasonably well with metrics largely above 0.8. In contrast, for the subduction zone scenario, GaMMA registered a K% reduction in the F1 score between experiments X and Y."

Author response:
Revised as suggested and replaced the qualitative phrase with exact numbers.

Reviewer comment:
Figures 3 and 4: I suggest remaking these figures as heatmaps (formatted the same as Figures 5 and 7). I found the heatmap figures much easier to compare an algorithm's experiment trends and its relation to the other algorithms.

Author response:
Done. We decided to keep the original bar charts in the manuscript, and added an alternative visualization in the supplementary material as heatmaps, following your suggestion (figures S13 and S14). The underlying numerical results are unchanged.

Reviewer comment:
Section 4.2: My overall impression of these pick statistics is that they convey the same information as your event-level stats. This is not too surprising because your definition of a true positive event is one that shares 50% of the associations with the ground truth event. As such, I would recommend moving figures 6 and 8 and analysis text to the supplement. Figures 5 and 7 show the salient differences between the algorithms. This will help keep the work you have done to justify your story while allowing the reader to move on to the "Further Experiments" section with more complex data scenarios.

Author response:
Moved to the supplement as suggested.

Reviewer comment:
Heatmap Figures 5 and 7: These are very good figures. Figure labels should be in a larger font. I would move the text "Precision", "Recall", and F1 Score as Row labels instead of including them in each subplot title. The y-axis label "Events" only needs to be on the left-most column plots. A similar comment for the x-axis labels. Doing so will give you more space to make font sizes bigger. Use the figure caption to describe the Precision, Recall, F1 rows, and to explain the left-most column y-axis ("Event") and bottom most row x-axis label ("Noise) apply to all subplots.

Author response:
Implemented. We redesigned Figs 5 and 7.

- Row labels: "Precision," "Recall," and "F1 score" now appear once, as bold row headers on the left, instead of inside every subplot.
- Shared axis: The y-axis label "Events" is shown only on the left-most column; the x-axis label "Noise (%)" appears only on the bottom row.
- Font sizes: Tick labels, color-bar labels, and row/column headers were enlarged.
- Caption: Rewrote the caption to describe the shared axes and the three metric rows.

Reviewer comment:
Figures 6 and 8: A lot of information in these figures and they are lacking subplot explanations. You have introduced what these subplots are in lines 188-192. You can keep that in the introduction, but before line 325, provide a more in-depth explanation of these subplots and why they matter. Lines 325 – 337 "...show that most associators tend to miss an average...." Instead, I would phrase this as "in subplot (a), the average number of missed picks shows that...." The caption reads: "Mean values of six pivotal metrics..." Describe the plots after this statement. "Subplot (a) shows the ... Subplot (b) shows ... "

Author response:
Thank you for the suggestion.

- To streamline the main text we moved the bar-chart pick-statistics (former Figs 6 & 8) to the Supplement.
- Added a short explanatory paragraph before line 325. We inserted a paragraph that (i) defines the six pick-budget terms, (ii) explains why they matter, and (iii) links omission errors to subplot (b) and commission errors to subplots (c) & (f).
- Re-worded the results paragraph; each statement now cites the relevant panel explicitly.
- Expanded the figure caption.

Reviewer comment:
Figure 6 and 8: In the predicted picks subplot, the legend is plotted on top of the plot and should be moved outside of the plot like you did for Figure 4. Additionally, there is a lot of vertical space between the subplots that can be reduced. Reducing this vertical space will make it easier for the reader to see the x-axis labels on the bottom subplot (which apply to all plots). Perhaps you can remove the subplot titles as the y-axis describes the statistic being measured? Common Picks subplot labels for PhaseLink and PyOcto run together at 500 events, 30% noise and 500 events 100% noise, making them hard to distinguish.

Author response:
All requested layout fixes have been applied to the bar-chart pick-statistics, which now appear in the Supplement.

Reviewer comment:
Section 4.3 Line 355: "... the amount of noise picks does not influence ..." Suggested: "... noise picks do not influence ..."

Author response:
Done.

Reviewer comment:
Figure 9: Include the scenario name in the title. Example Left plot title: "Crustal Scenario: Associator Runtime vs Total Picks" Right plot: "Subduction Scenario: Associator Runtime vs Total Picks"

Author response:
Revised as suggested.

Reviewer comment:
Section 4.4 Further tests Remove conversational phrasing and employing active voice: Lines 359-365: "Although we took care to design our main synthetic experiments in a way that resembles natural use cases in many ways, there are still a few additional sources of complexity that we did not address in those tests. Out-of-network events ... All of these conditions" Suggestion: "In this section we address several real-world data complexities not included in our main synthetic experiments from Section 4.1. First, out-of-network events are a common occurrence in most monitoring environments, especially in subduction zones where most of the plate interface 18 as well as the often seismically active outer rise are located offshore (Stern, 2002). Second, 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. 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..."

Author response:
Implemented. We rewrote the opening of section 4.4 in active voice and removed conversational wording.

Reviewer comment:
Line 375: "In order to gauge the effect of adding more noise onto the utilized picks, . . . " Suggested: "To test travel time noise effects on association, . . . " Another suggestion: You have already introduced the low pick time error idea in the Section 4.4 introduction, possibly just start this Section 4.4.1 with: "Figure 10 shows the pick error for +-0-1% (the same noise level as our synthetic experiments), (+-1-5%), ... of travel time. REAL, GaMMA . . . "

Author response:
Revised. Thanks.

Reviewer comment:
Line 379 – 380 "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" No suggestion to change, I am just noting that this is good information. It hints at "transferability" relative to tuning for the non-data driven associators. Additionally, since GENIE and PhaseLink were not retrained and their results, in some cases, were better than the "algorithmic" associators speaks to the power of DL-based associators.

Author response:
Thank you. We are glad that point was clear and valuable.

Reviewer comment:
Figure 10: This is the type of caption I would like to see on the other figures in the manuscript. Lines 389-390 ". . . 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). . . " Here you have clarified what "significantly" means. I would like to see this type of phrasing for all cases where you are using qualitative descriptions.

Author response:
We agree. We treated Figure 10 as the template and applied the same caption style.

Reviewer comment:
Lines 408-411 "The bar charts reveal that the influence of out-of-network events on the correct association of in-network events is small, with only a slight decrease in performance for the run with the highest amount of out-of-network events being apparent for most associators." Suggestion: Rephrase and list the worst performing associator, with respect to original versus "High" for each bar chart. "Pick association performance was only slightly degraded between the "Original," (i.e. no out-of-network-events) and the "High," (i.e. 450 out-of-network events). For precision, associator X had the largest degradation with K%. "For recall, . . . Finally, associator X has the largest association F1 degradation with K%." Line 414 Rephrase: ". . . algorithms but REAL (which does not feature such a parameter) was set to 71 414 ∘W."

Author response:
Thank you for this suggestion. We have rephrased those lines with specific numbers for the worst performing associator (PhaseLink). We also rephrased the boundary-parameter sentence for clarity.

Reviewer comment:
Line 417: "...of choosing it large enough." Suggested "...setting the appropriate range for a given monitoring scenario."

Author response:
Revised.

Reviewer comment:
Line 419: "...does not have to be defined by the user." Suggested: "...is not user-configured."

Author response:
Revised.

Reviewer comment:
Line 421: "...still retrieves reasonable proportion of events." Suggested: Remove "reasonable" and state the percentage "retrieves K% of the events."

Author response:
Revised.

Reviewer comment:
Line 424: "...the one of REAL..." Suggested: "...that of REAL..."

Author response:
Revised.

Reviewer comment:
Line 425: "GaMMA's results are less convincing..." Suggested: "GaMMA's only retrieves a significant..."

Author response:
Revised.

Reviewer comment:
In the location plots in figure 10, It would help if a set of regions (boxes). Then when referring to a "significant number of events west of the center. . . " they could present absolute counts/percentages etc. . . This would help the reader understand what they define as significant and what the extent of the regions. I am thinking you can divide the plots into a northern, center, and southern region (as you have referred to in the text) and show the boundary lines on the plots and name the regions: R1, R2, etc. . . Your yellow boundary would stay defined as it is. Or, if the sub-regions need to be specific for each algorithm as you described in the text, you can do this as well.

Author response:
Thank you. Rather than overlaying extra latitude–longitude boxes, which we found made the already busy maps harder to read, we have inserted the explicit coordinate limits directly in the text. We understand the reviewer's goal to quantify retrieval rates in well-defined sub-regions. However, experimenting with several "boxed" drafts, we concluded that the extra graphics:

- Did not convey new information: The manuscript already reports the key numbers in the text:". . . REAL still retrieved 80 % of the events. . . "
- Legibility: Overlaying three horizontal boxes (and individual counts for five associators) produced labels that collided with the colored dots, especially in the dense central band.
- Shifted emphasis away from the take-away message: The purpose of Fig. 10 is to illustrate distance-dependent fall-off and the effect of the western search boundary (yellow line). Latitude slicing is orthogonal to that axis and therefore of secondary relevance.

These added numerical bounds now make the spatial reference unambiguous without adding visual clutter. We believe this meets the reviewer's request for clarity while preserving the figure's readability.

Reviewer comment:
Section 5 Section 5.1 Good commentary about the difficulty of tuning the different associators and the monitoring scenarios each may be best suited.

Author response:
Thank you for the positive feedback. We are glad the discussion in Section 5.1 was useful and have retained it, making only minor wording adjustments for clarity.

Reviewer comment:
Line 547: "For the backpropagation-based. . . " Do you mean "backprojection-based?" I know "backpropagation" as the neural network training algorithm.

Author response:
Revised. Thank you for spotting the mix-up with "backpropagation," which indeed refers to neural network training rather than the spatial backprojection approach used by REAL and PyOcto.

Reviewer 2

Reviewer comment:
The paper benchmarks associators that assume pre-picked arrivals (e.g., P and S) as input and focus on grouping them into events. This is a valid and commonly used modular pipeline. However, other methods (especially in full waveform-based frameworks) delay or merge the picking and location stages, detecting events directly via spatially-coherent characteristic functions (e.g., P/S likelihoods, kurtosis, e.g., Poiata et al. 2016). It would strengthen the paper to acknowledge this distinction in the introduction or discussion. While the modular approach remains standard in many workflows, and essential in operational networks, the choice to benchmark associators decoupled from picking is a design decision that contrasts with newer paradigms. Discussing this could help readers situate the scope of this study within the broader ecosystem of seismic event detection and cataloging methods. At the same time, the current focus allows for fair and controlled comparisons under well-defined conditions, which remains valuable and widely applicable.

Author response:
Thanks. We agree and have clarified this in the introduction.

Reviewer comment:
As well explained by the authors, the deep learning-based associators (GENIE, PhaseLink) require significant training using large synthetic datasets. Although the supplementary material explains the training process clearly, it would likely help practitioners if the main text more clearly acknowledged the computational and technical cost of training, particularly in contexts without access to GPUs or large-scale infrastructure.

Author response:
We agree and have added a brief comment at the end of Section 5.1 “Associator configuration or training”.

Reviewer comment:
The supplementary section on station density (Text S3) reveals important differences in associator robustness. For example, GENIE maintains strong performance across high- and low-density networks using the same parameters, while PhaseLink requires retraining. These differences in sensitivity to network geometry and tuning deserve a brief mention in the discussion, as they have real implications for usability across different seismic deployments.

Author response:
Done. We added two sentences to Section 5.1 (Associator configuration or training) that explicitly call out the station-density test and its implications.

Reviewer comment:
Although the figures are comprehensive and informative, the paper could benefit from a high-level summary visualization that captures the key differences (only) between the tested associators. For instance, a radar chart or performance table could provide a compact overview of each method's behavior under different conditions (e.g., noise robustness, scalability, accuracy, and speed). Such a summary would greatly enhance the paper's utility for practitioners.

Author response:
We appreciate the reviewer's wish for a concise, at-a-glance graphic. After careful testing, however, we decided not to distill the results into a single radar plot or condensed table, for three reasons:

- Multi-dimensional design space: Each associator was evaluated along 18 distinct experimental conditions (2 network geometries $\times$ 3 event rates $\times$ 3 noise levels) and 4 orthogonal metrics (event-level and pick-level precision, recall, and F1 plus runtime). Collapsing this rich landscape onto four or five axes would force us to average across very heterogeneous scenarios.
- Non-linear trade-offs: Several algorithms flip rank order depending on the stress test. A radar chart would either (a) show multiple overlapping polygons that would be difficult to read, or (b) use scenario-averaged scores that obscure those crossovers.
- Risk of misleading takeaways: Practitioners often focus on the quadrant most similar to their deployment (aftershock sequences vs. sparse background seismicity, subduction vs. crustal). A grand average would encourage "one-size-fits-all" conclusions that our detailed analysis explicitly warns against (Discussion Section 5.3).

Instead we reshaped the Results section to have key takeaway paragraphs that describe the dominant trends for each scenario. In addition, Discussion Section 5 was reorganised to give explicit, scenario-conditioned recommendations, making the practical guidance clear without collapsing all metrics into a potentially misleading single plot.