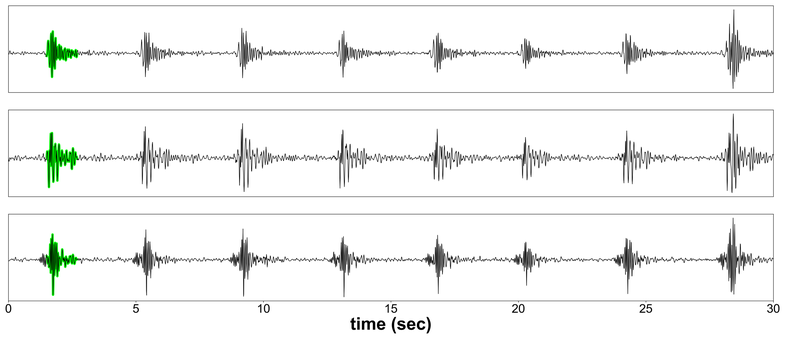
CAUTION: PREDICTION INCONSISTENCY OF NEURAL PHASE PICKERS SHOULD NOT BE OVERLOOKED

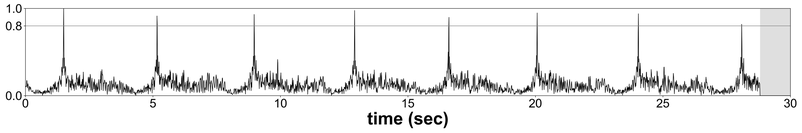
Yongsoo Park (ysp@stanford.edu) | Gregory C. Beroza | William L. Ellsworth | Department of Geophysics, Stanford University

The Problem

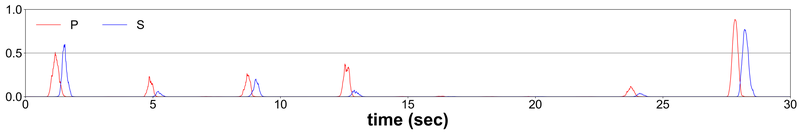
The following trace contains multiple similar looking seismic signals



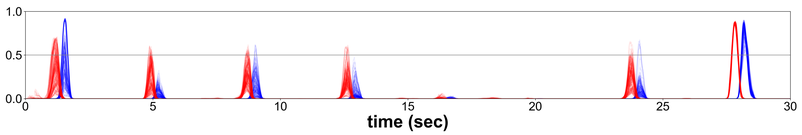
If we use the green as a template and cross-correlate, we can confirm their high similarity



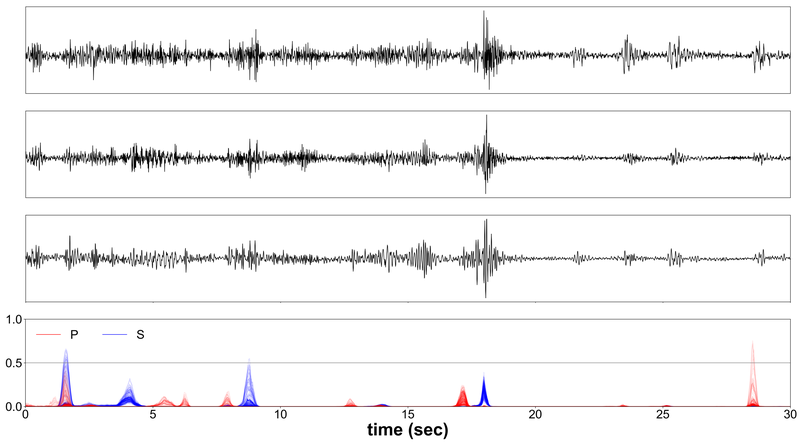
However, if we use a pre-trained neural phase picker[1], we get highly inconsistent results



If we shift the traces from -1 to 1 second and repeat, we can see that the results fluctuate



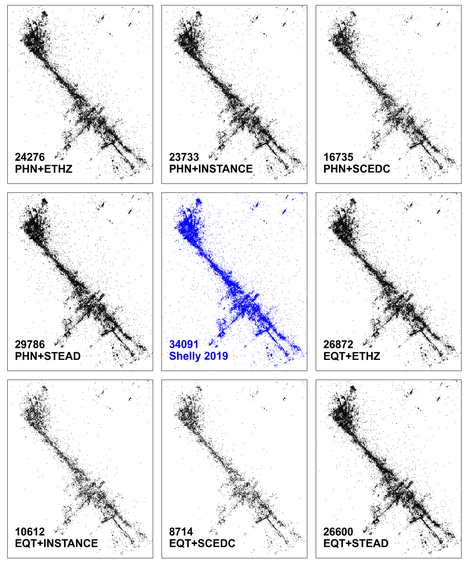
Same thing happens to traces containing arbitrary signals



Let's call this phenomenon **prediction inconsistency**

How serious is this?

How many of the quakes in Shelly 2019 template matching catalog[2] be reproduced from picks predicted from various picker models in SeisBench[3] ?



PHN: PhaseNet[1] EQT: EQTransformer[4]

- There were at least 3 stations within 30 km from every quake
- If arrivals were picked from 3 or more stations within 30 km and their total count was at least 5, consider it as a reproducible quake. Allow up to 1 second of difference between calculated and measured arrival times
- When picking, use a sliding window with a stride length of half the window length and apply a probability threshold of 0.5 like people typically do

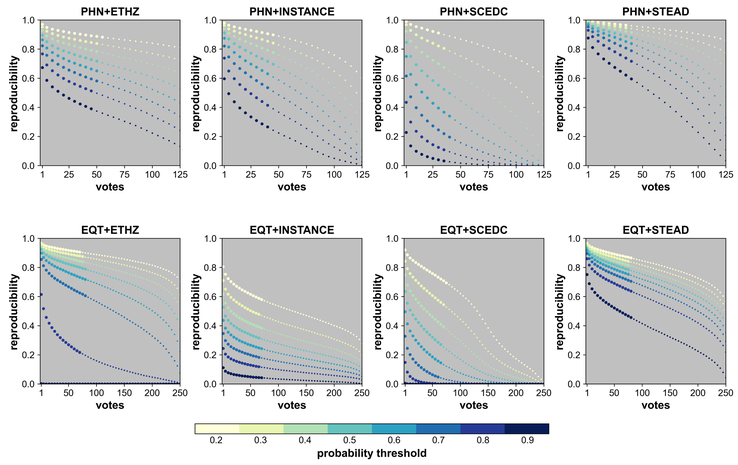
← Many quakes are missing!

Mitigation Strategies

- Ideally: Design a picker that doesn't have this issue (but NNs are meant to be nonlinear)
- Practically: Use a small stride and apply a vote threshold (refer to as small stride approach)

How effective is this?

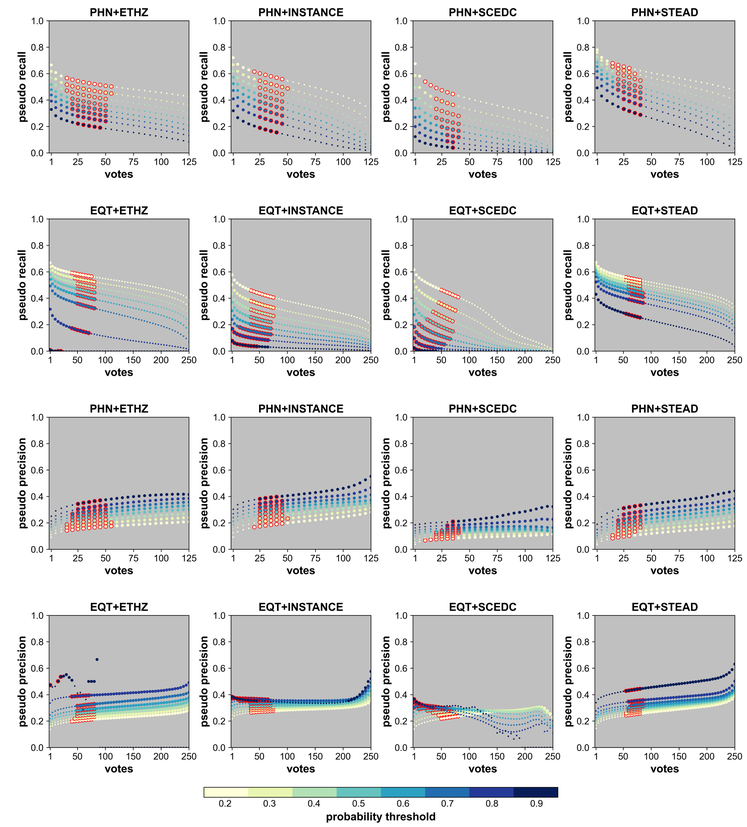
Use a stride length of 0.2 seconds (chosen arbitrarily) and apply a threshold on the number of times each sample point was predicted as an arrival (aka. votes). While doing so, treat the picks within 0.2 seconds as the same picks. How does the 'reproducibility' change with respect to the baseline?



Thick data points indicate that the small stride approach reproduced more quakes than the baseline. We have a nontrivial range of vote thresholds that made the approach better.

What about the commonly used evaluation metrics?

Assume that the quakes in the Shelly 2019 catalog are all true quakes. For each quake, query the stations within 30 km. Calculate arrival times and count the closely matching picks (within 1 second of difference) from these stations as TP and missed arrivals as FN. Count the remaining picks as FP. Then calculate recall and precision.



Thick data points indicate that the small stride approach scored higher on the metric than the baseline. Data points highlighted in red indicate that the approach scored higher on both metrics than the baseline. We have a nontrivial range of vote thresholds that made the approach better. Recall tends to decrease with increasing votes, which means a tighter threshold removes not only FPs but also TPs. However, precision tends to increase with increasing votes, which means more FPs get removed than TPs with a tighter threshold.

Key Takeaways

- Be aware of and do NOT overlook prediction inconsistency
- Having a higher probability (or prediction score) does not necessarily mean a better pick
- Using a small stride and aggregating the picks is highly encouraged

References

1 Zhu and Beroza (2019). Geophys. J. Int. 216(1), 261-273.
2 Shelly (2019). Seismol. Res. Lett. 91(4), 1971-1978.
3 Woollam et al. (2022). Seismol. Res. Lett. 93(3), 1695-1709.
4 Mousavi et al. (2020). Nat. Commun. 11(1), 1-12.

Acknowledgements

This work was supported by the Stanford Center for Induced and Triggered Seismicity, and the Department of Energy (Basic Energy Sciences; DE-SC0020445).