SSA2025 Earthquake Catalog Workshop

Contributors: Eric Beauce, Gabrielle Tepp, Clara Yoon, Ellen Yu, Weiqiang Zhu

Topic 2: Building Earthquake Catalog using Machine Learning

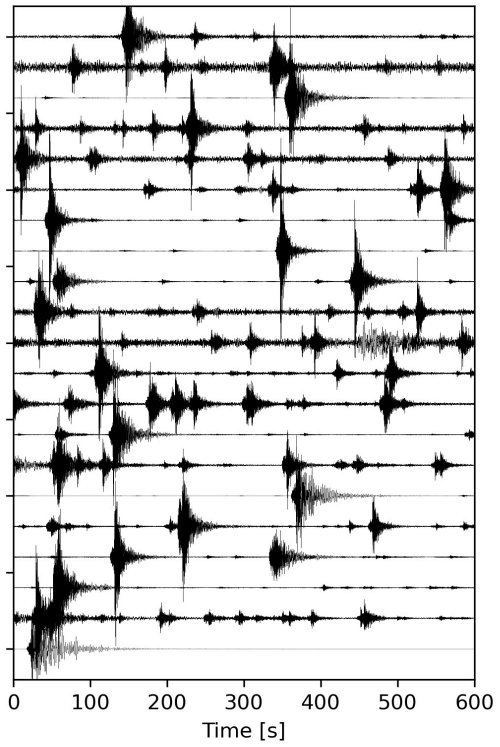
Main contributors: Clara Yoon, Weiqiang Zhu

Notebook example : https://ai4eps.github.io/Earthquake_Catalog_Workshop/notebooks/quakeflow/

Earthquake Deep Catalog

Seismic Data

STEAD/CEED



Phase Detection & Picking

PhaseNet/PhaseNet+

Phase Association

GaMMA

Earthquake Location

ADLoc

Earthquake Magnitude

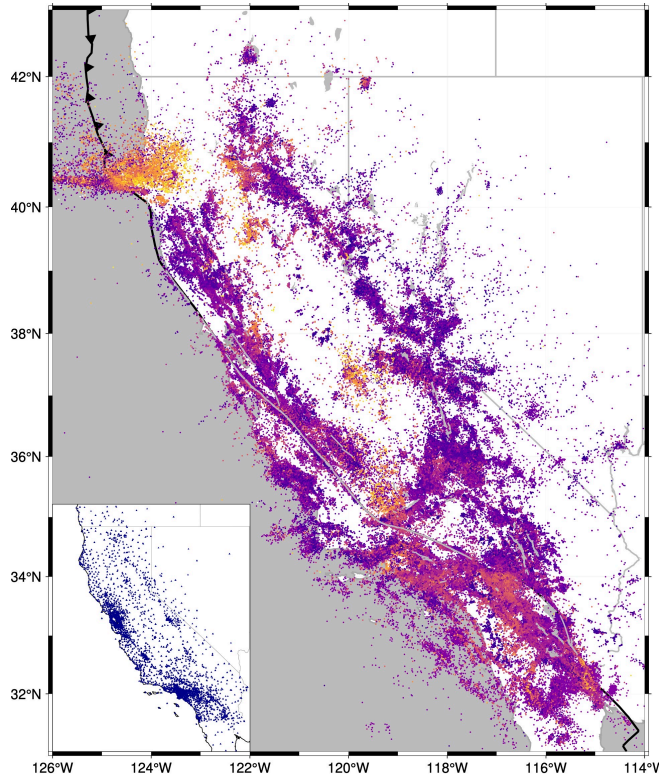
Earthquake Relocation

HypoDD/GrowClust

Earthquake Mechanism

HASH/SKHASH

Earthquake Deep Catalog

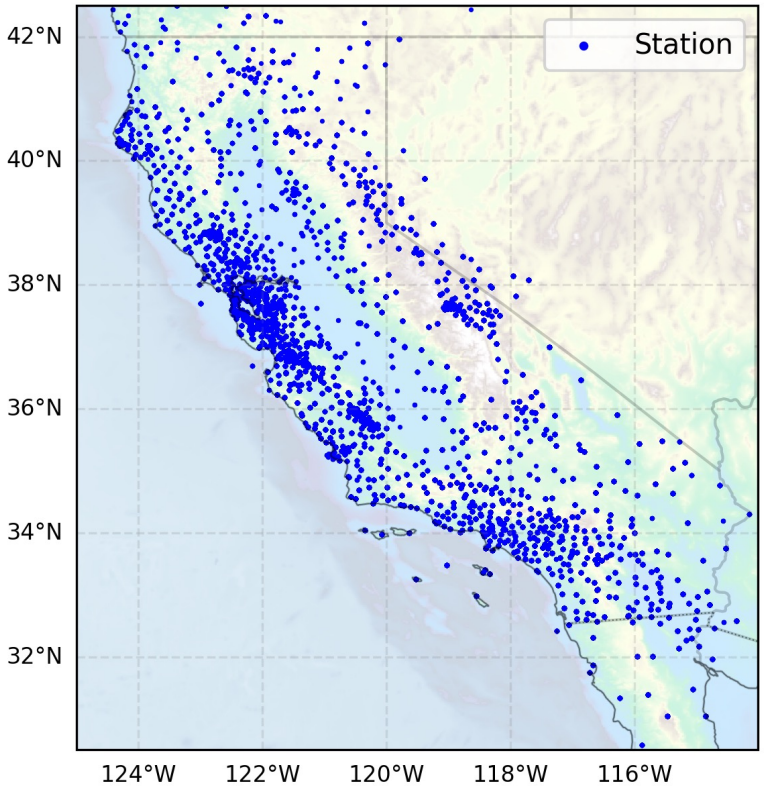


Note: Multiple algorithms exist for each task, with the ones listed here serving as examples.

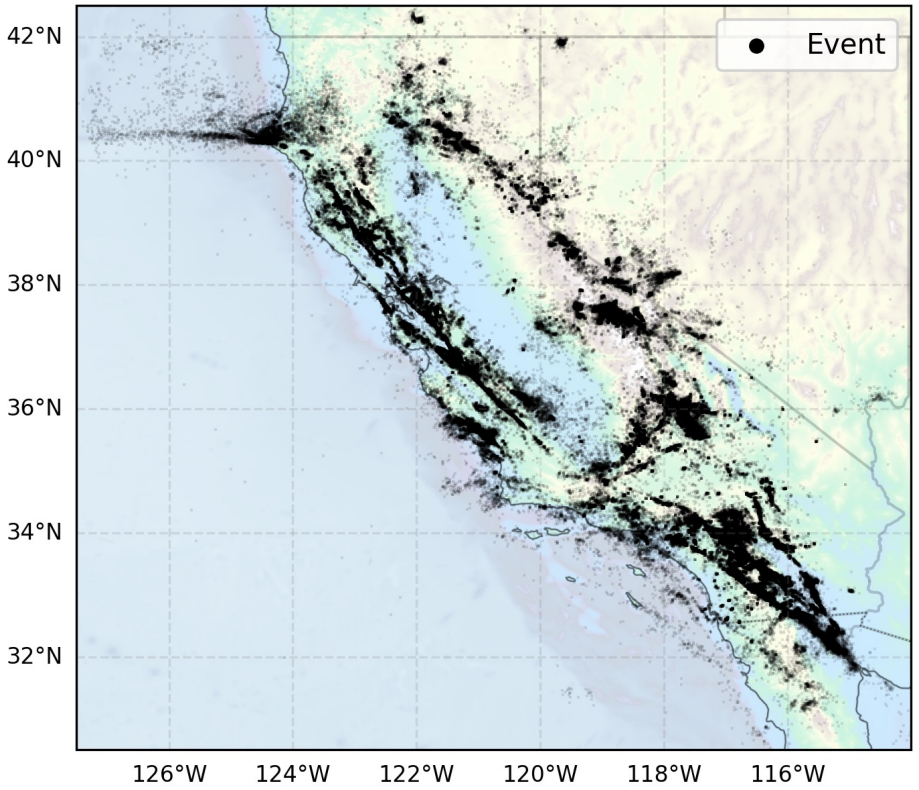
Q0: Seismic datasets for machine learning

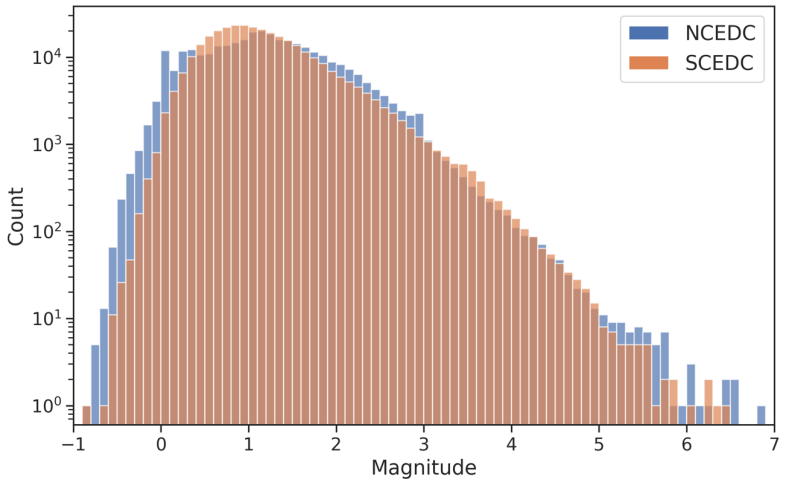
CEED: California Earthquake Event Dataset

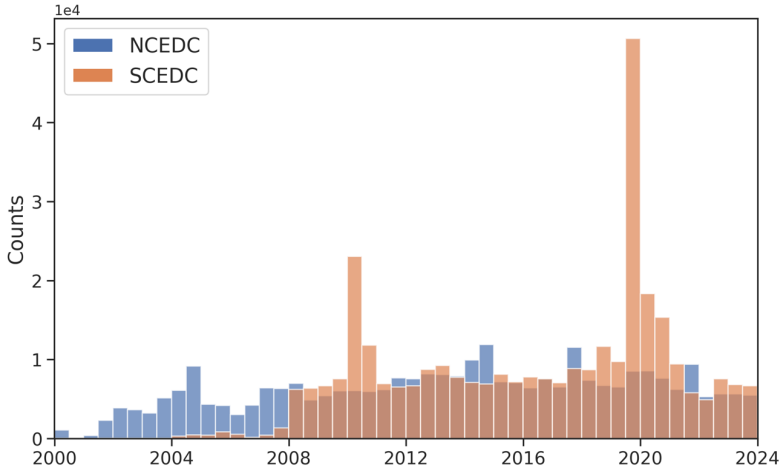
Seismic stations



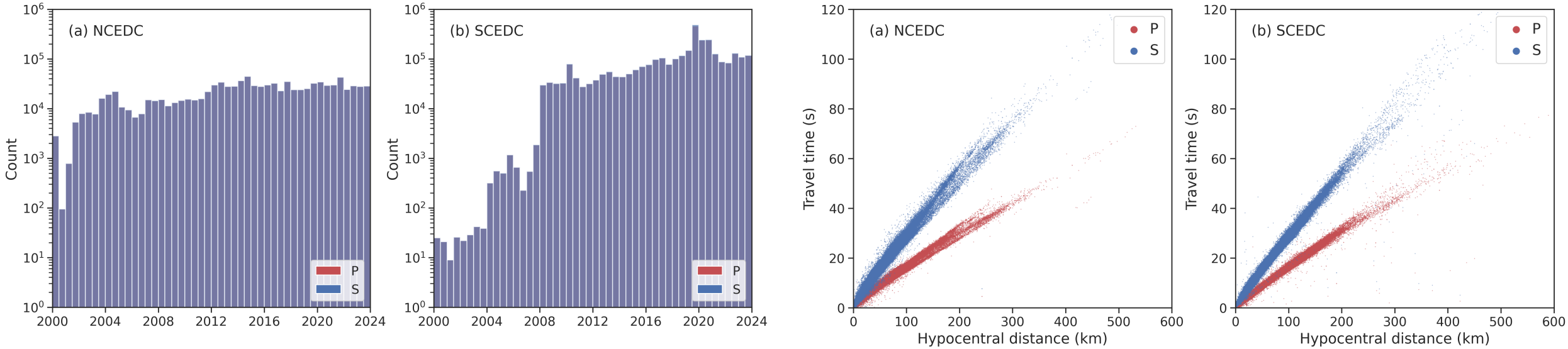
Earthquake catalog





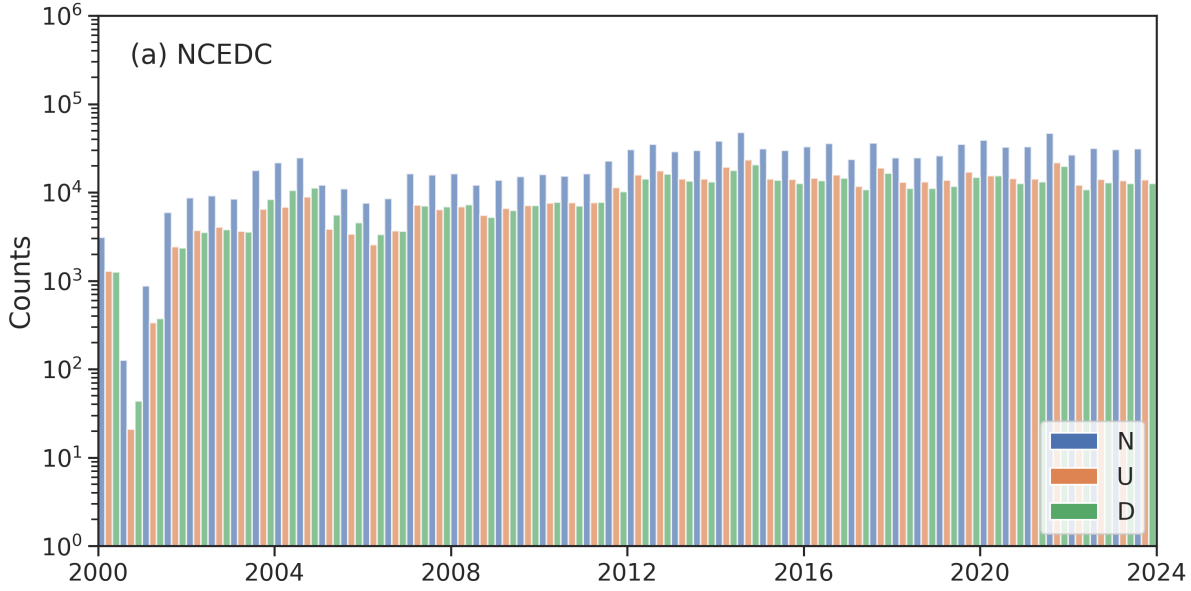


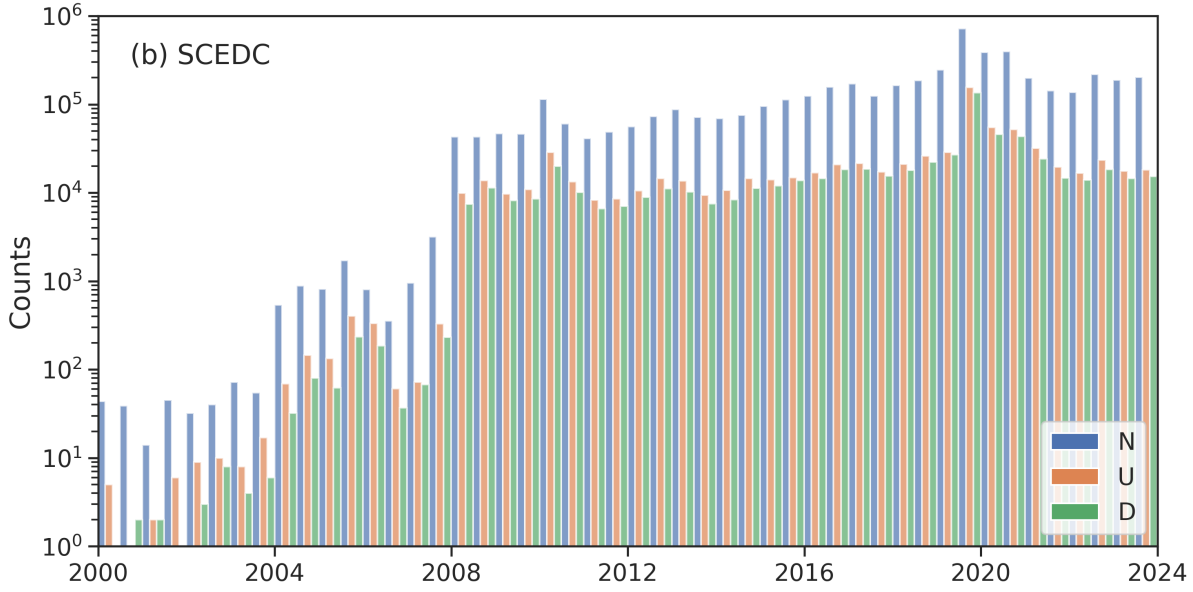
CEED: P&S Phase Picks



Phase arrival picks: 4.14M (NC: 1.09M + SC: 3.05M)

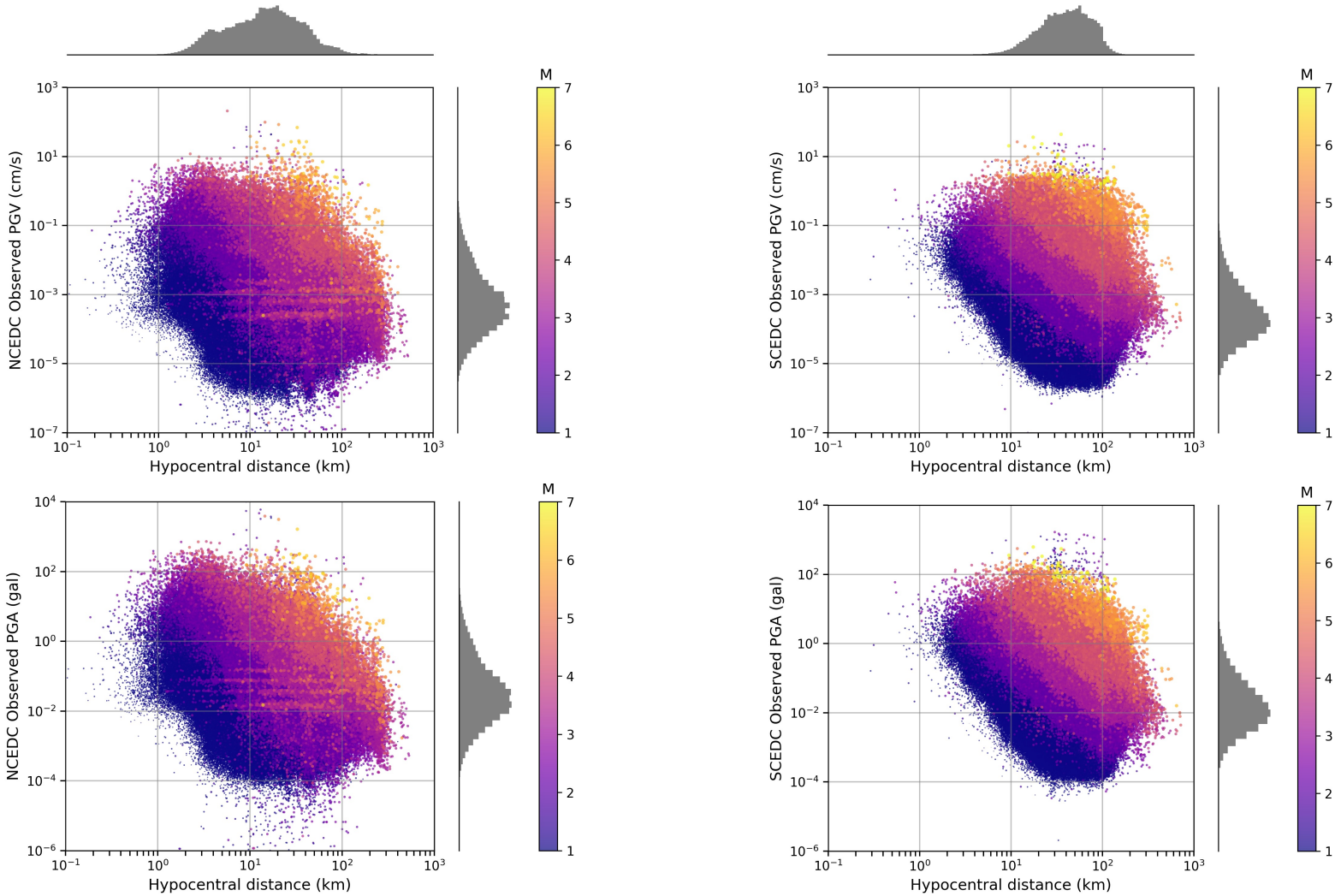
CEED: First-motion Polarity Picks



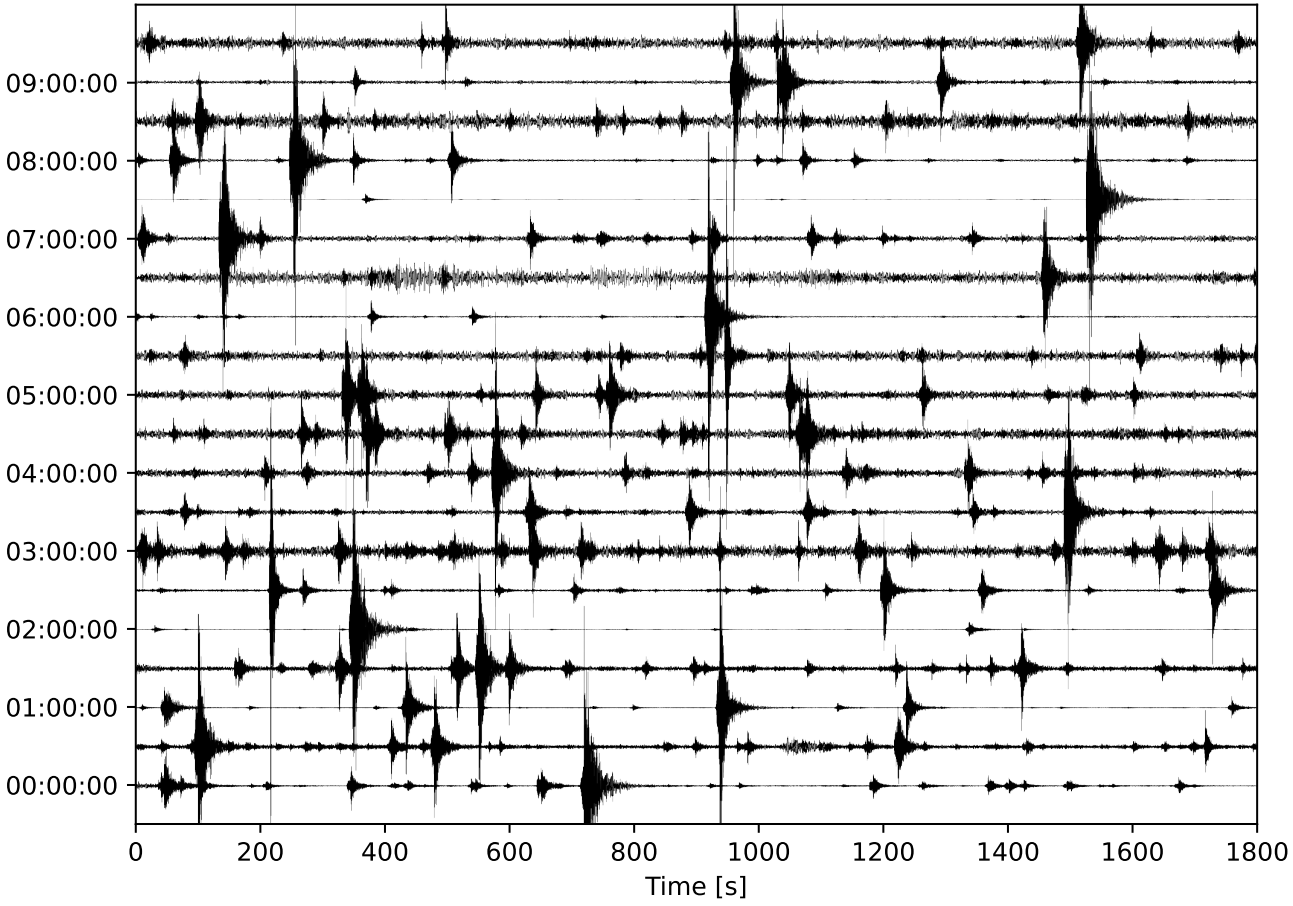


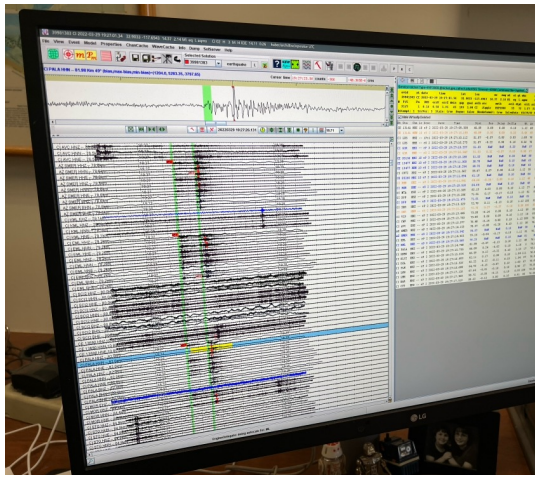
Phase polarity picks: 2.42M (NC: 1.04M + SC: 1.38M)

CEED: Ground Motion Measurements



Q1: How to automatically detect earthquakes?

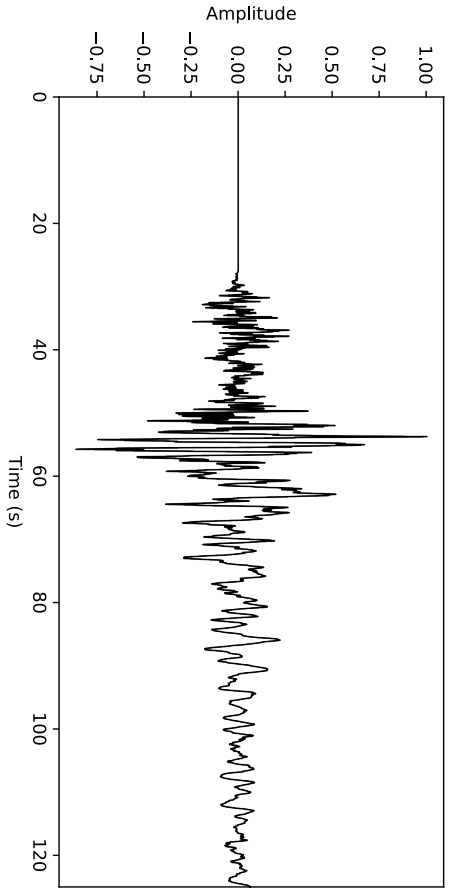


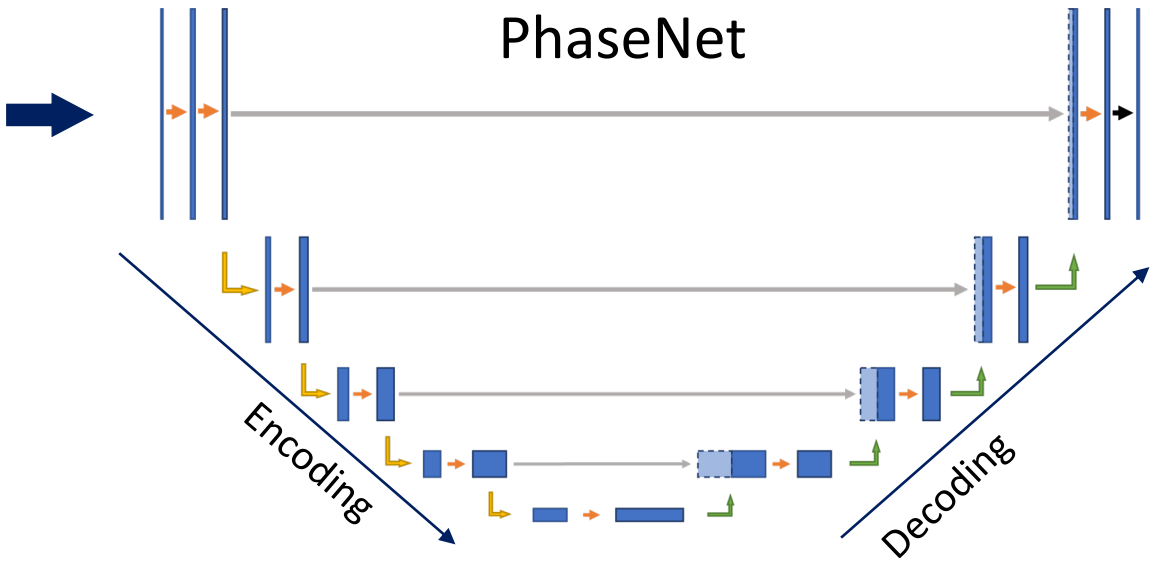


(Photo: Caltech Seismo Lab)

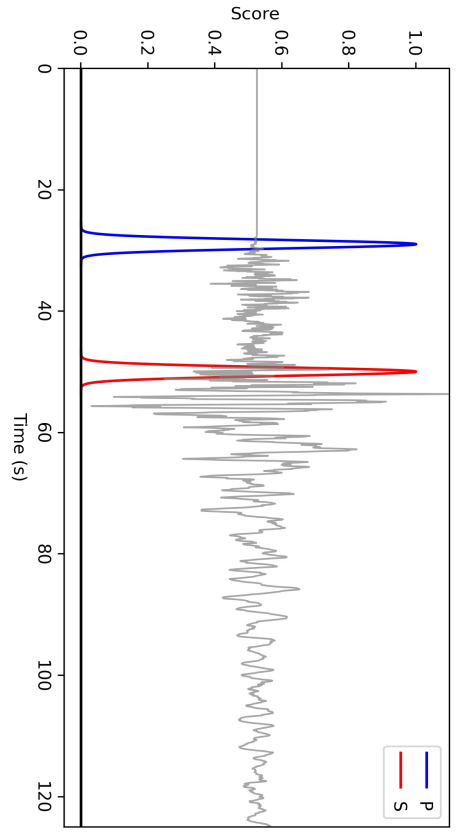
Supervised Learning: Phase Arrival-time Picking

Input waveform





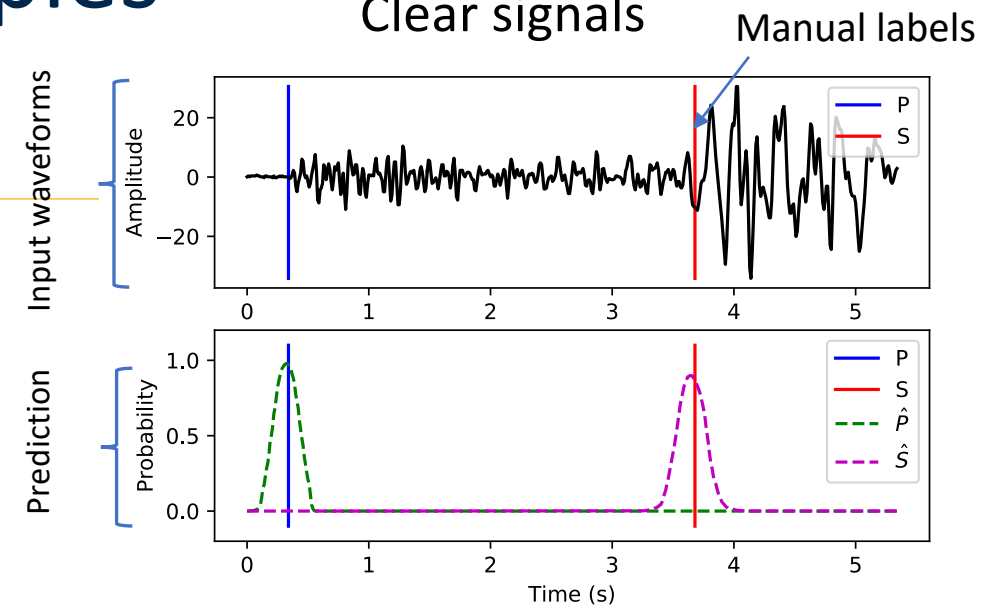
Target characteristic function



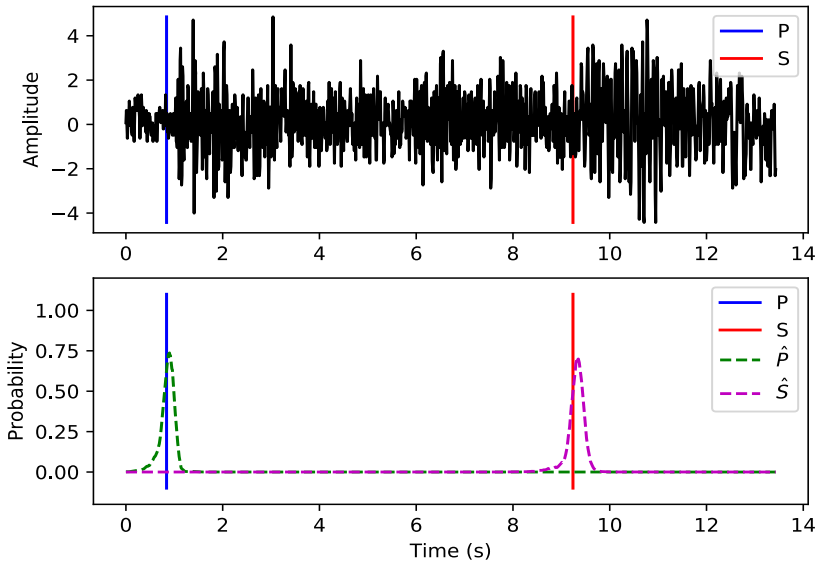
(Zhu and Beroza, 2018)

Examples

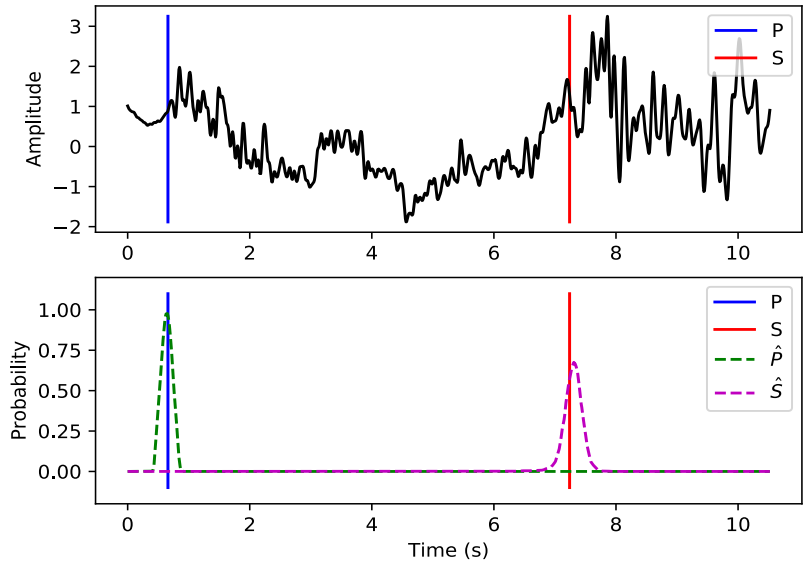
Clear signals



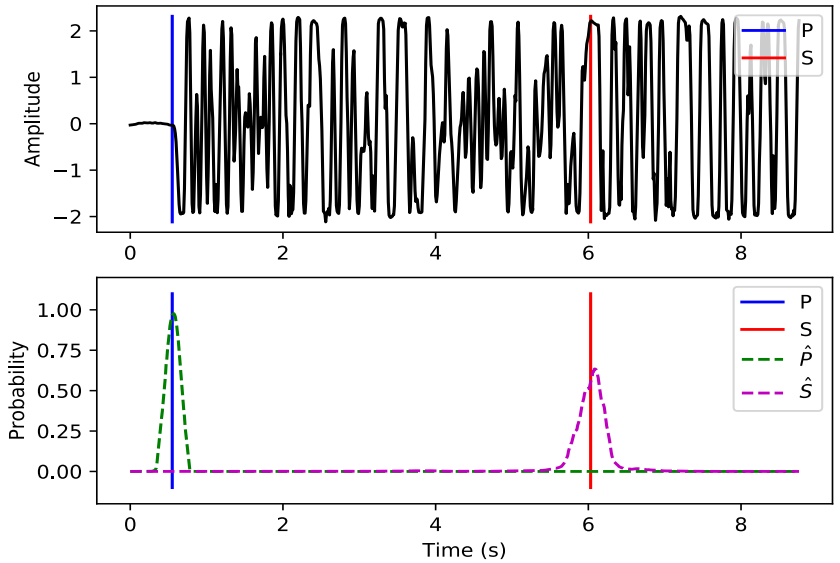
Noisy signals



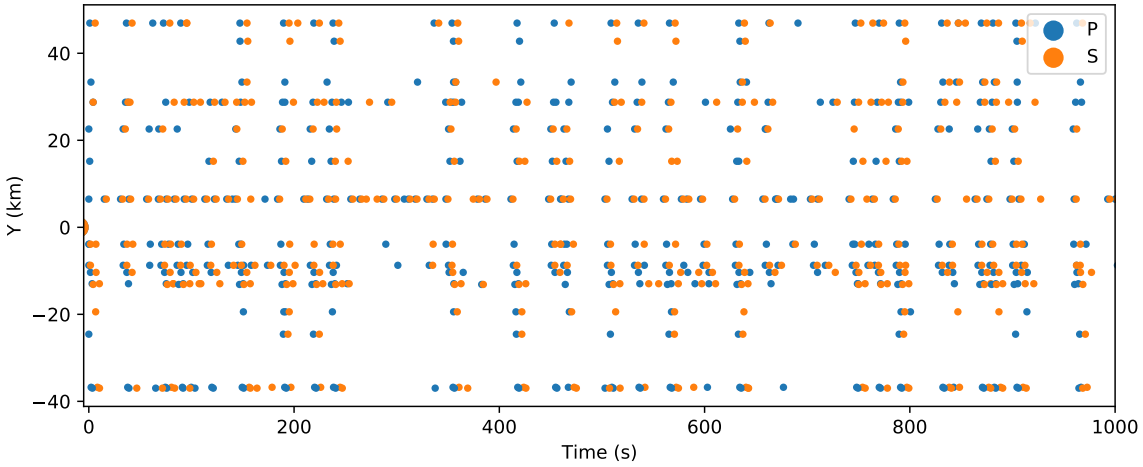
Noisy signals



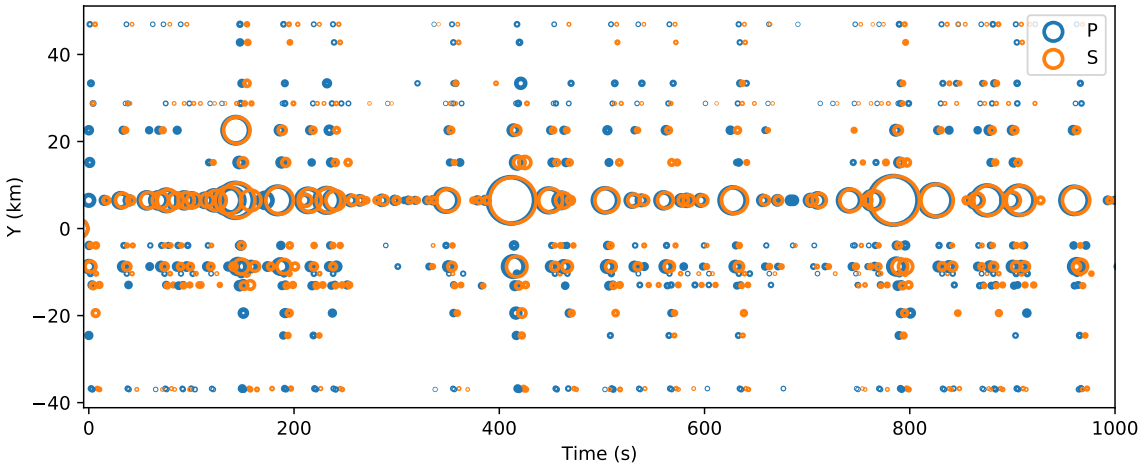
Clipped waveforms



Q2: How to identify underlying events?



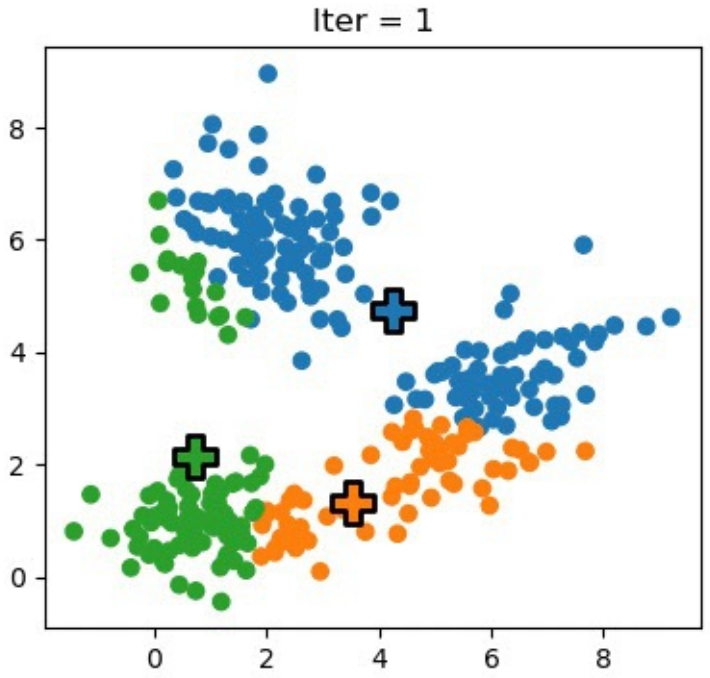
Phase arrival time only



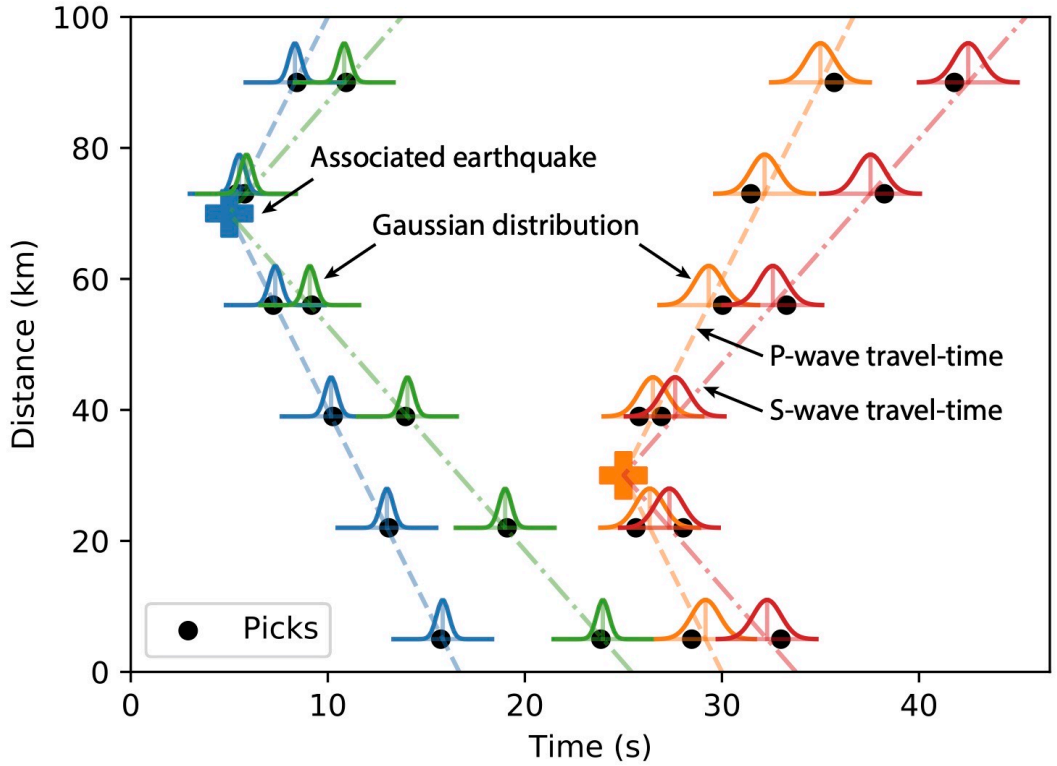
Phase arrival time + amplitude

Unsupervised Learning: Phase Association

K-means clustering

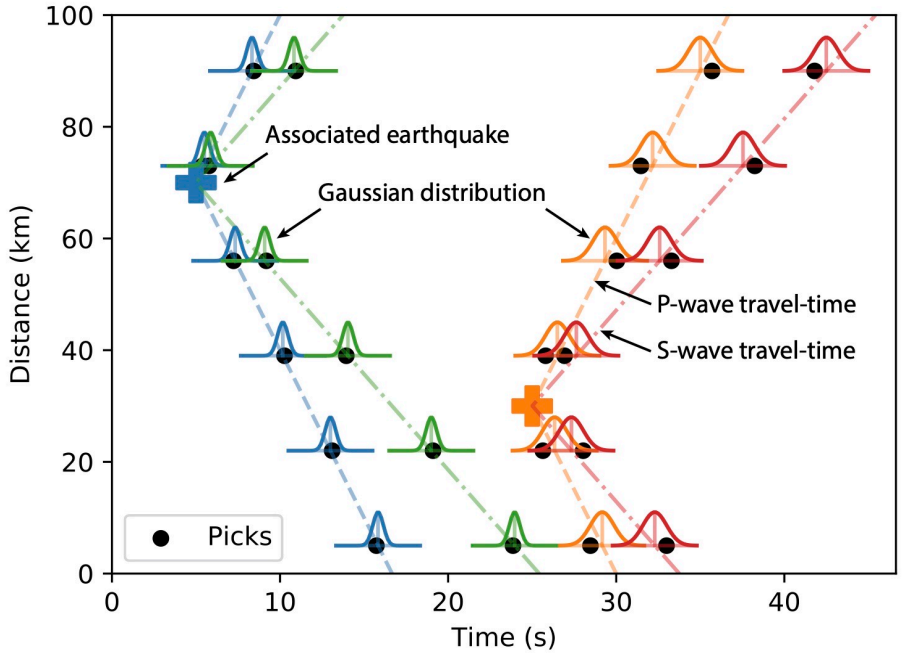


Phase association



GaMMA: Gaussian Mixture Model Association

- Modeling Gaussian distributions based on theoretical travel-time and amplitude
- Solving clustering using the Expectation-Maximization algorithm



- E-step (phase space): Assign picks to each earthquake

$$\gamma_{ik} = \frac{\phi_k \mathcal{N}(\mathbf{x}_i|\mu_k, \mathbf{\Sigma}_k)}{\sum_{k=1}^{K} \phi_k \mathcal{N}(\mathbf{x}_i|\mu_k, \mathbf{\Sigma}_k)}$$

- M-step (earthquake space): Update earthquake parameters

$$\phi_k = \sum_{i=1}^{N} \frac{\gamma_{ik}}{N}$$

$$\underset{(x_k, y_k, z_k, t_k)}{\text{minimize}} \; l(x_k, y_k, z_k, t_k) = \sum_{i=1}^{N} \frac{\gamma_{ik}}{2} \mathcal{L}\left(t_i, \hat{t}_{ik}(x_k, y_k, z_k, t_k)\right)$$

Estimate origin time and location

$$m_k = \frac{\sum_{i=1}^{N} \gamma_{ik} \mathcal{F}'_a(a_i, d_{ik})}{\sum_{i=1}^{N} \gamma_{ik}}$$

Estimate magnitude

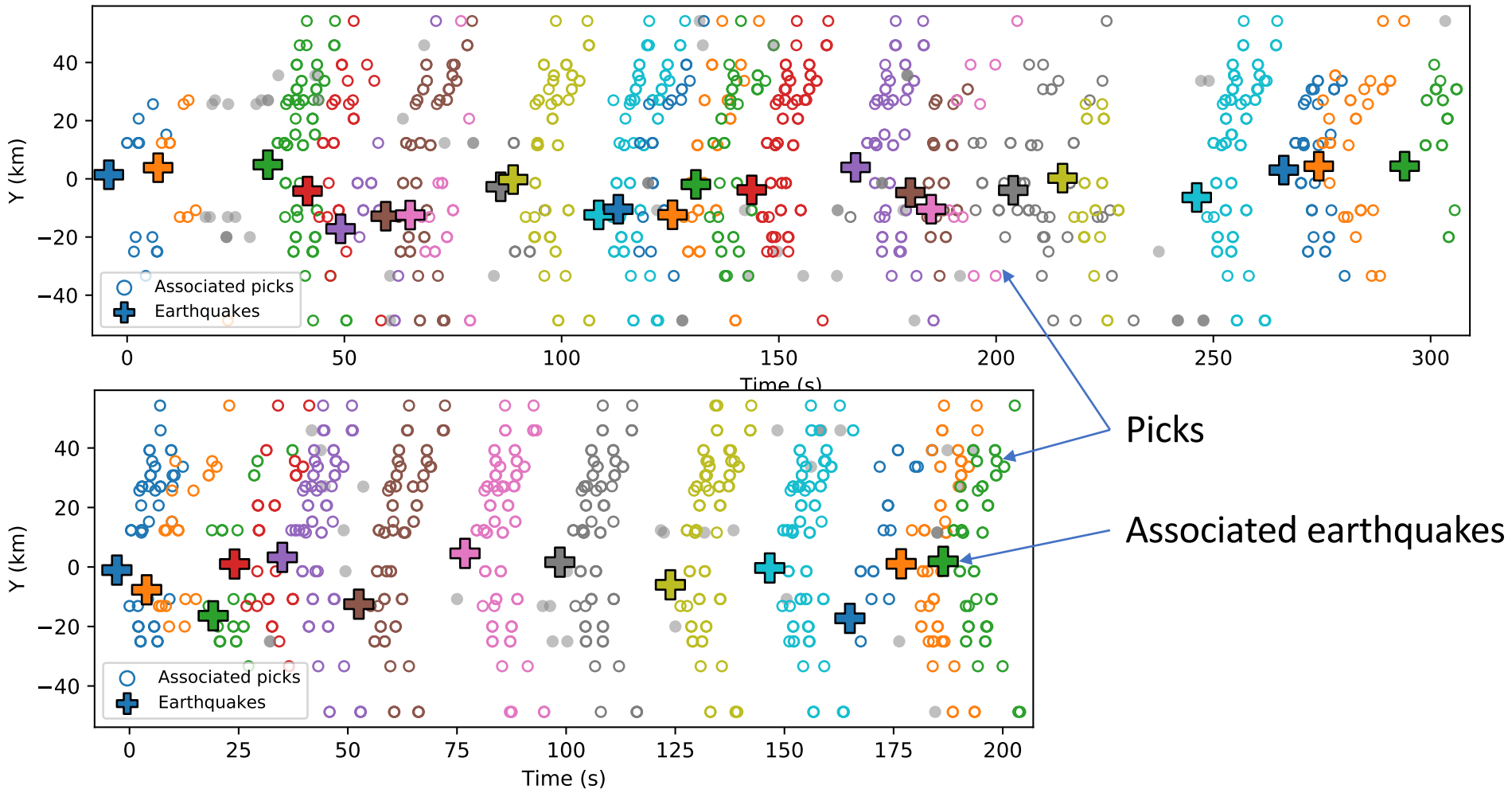
$$\mu_k = \begin{bmatrix} \hat{t}_{ik} \\ \hat{a}_{ik} \end{bmatrix} = \begin{bmatrix} \mathcal{F}_t(x_k, y_k, z_k, t_k) \\ \mathcal{F}_a(m_k, d_{ik}) \end{bmatrix}$$

Update theoretical values

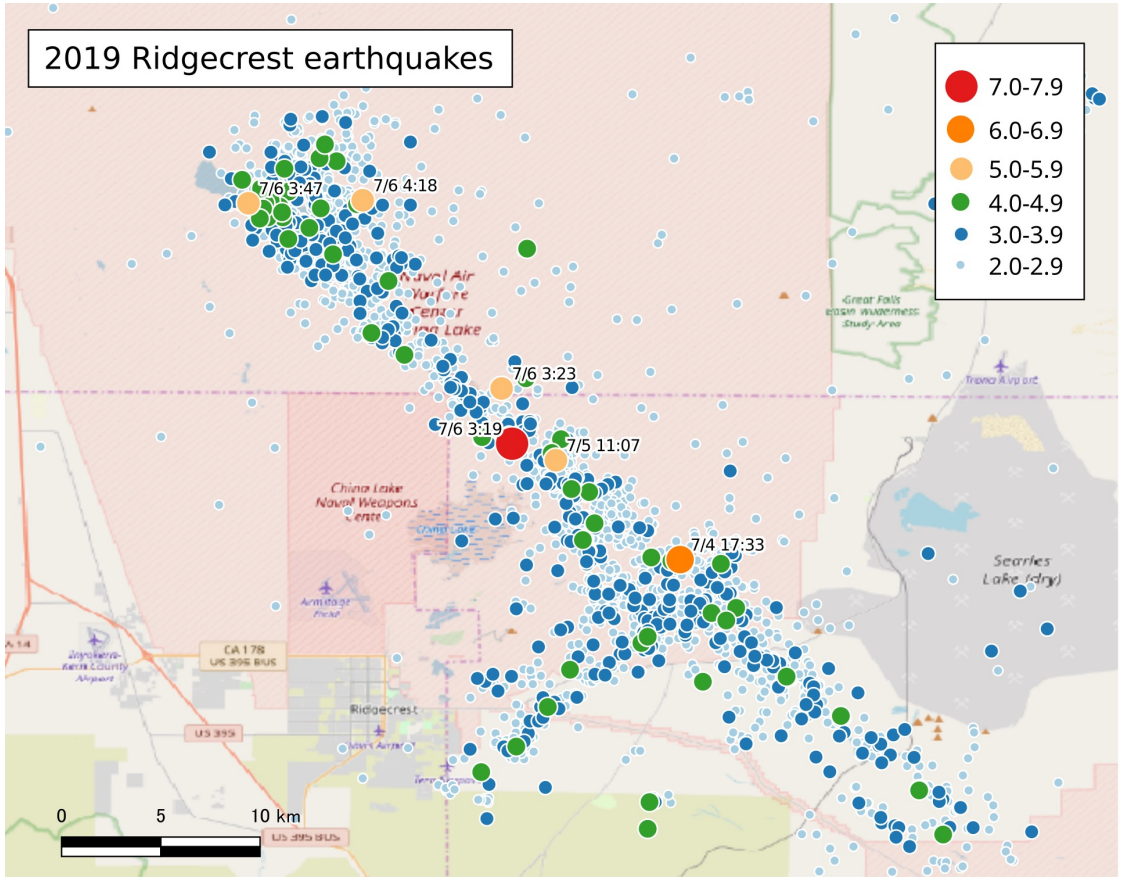
$$\mathbf{\Sigma}_k = \frac{\sum_{i=1}^{N} \gamma_{ik} (\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)^T}{\sum_{i=1}^{N} \gamma_{ik}}$$

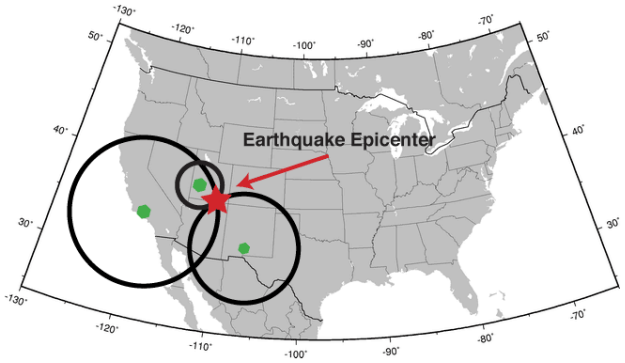
Update covariance matrix

Applications



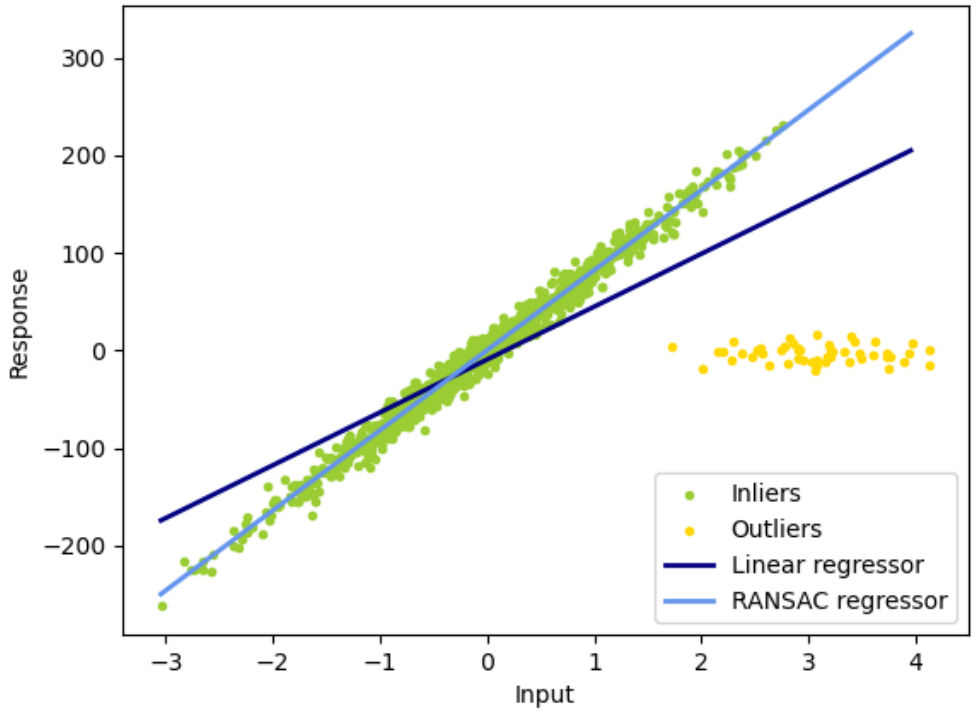
Q3: How determine earthquake locations?





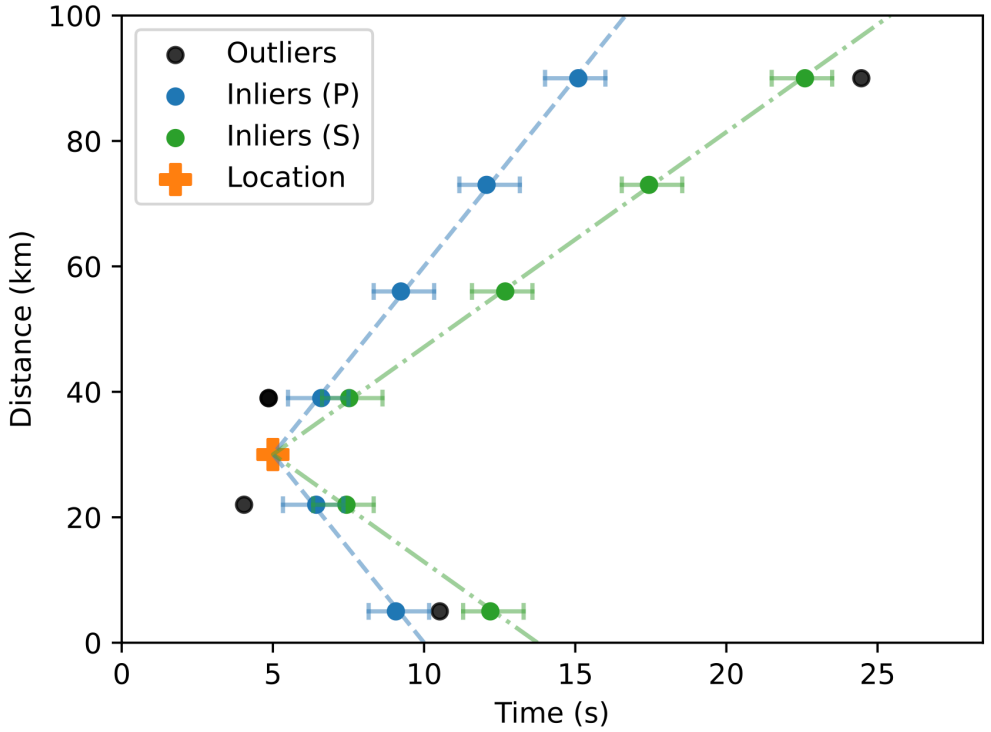
RANSAC: Robust Earthquake Location to Outliers

Robust linear regression



(https://scikit-learn.org)

Robust earthquake location



(Zhu et al. 2025)

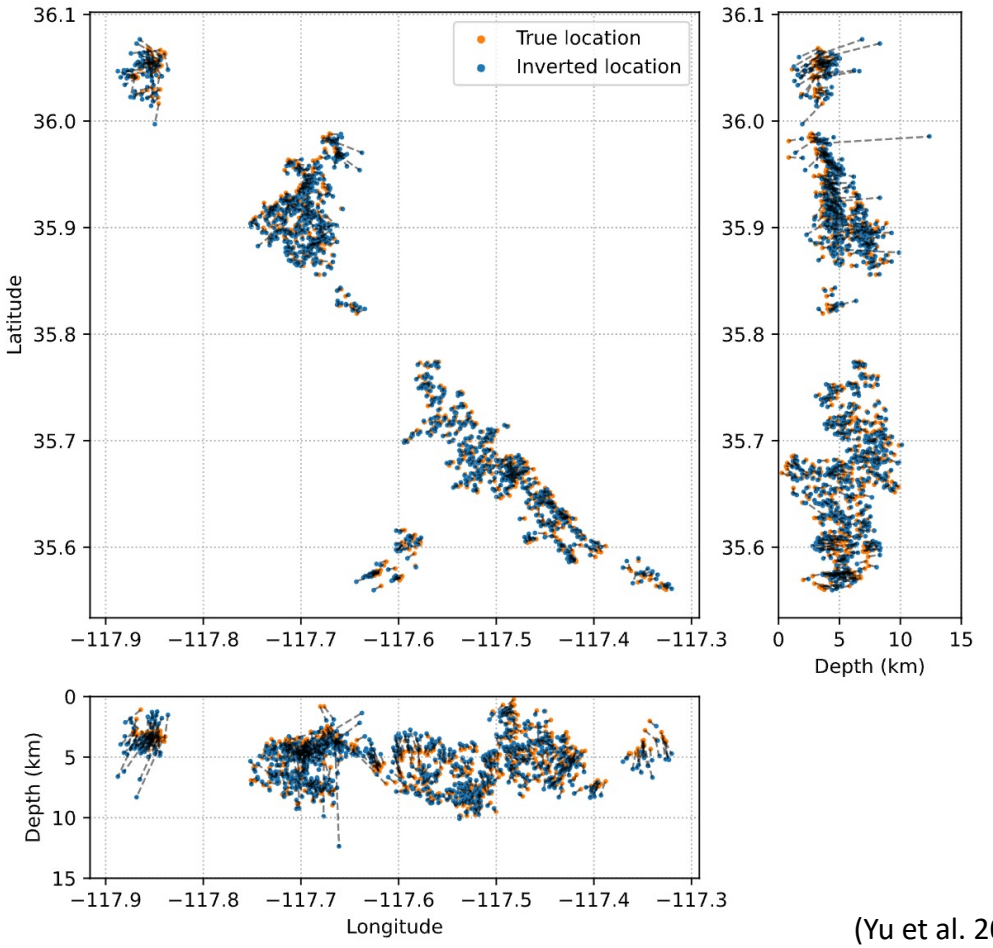
RANSAC (RANdom SAmple Consensus) fits a model from random subsets of inliers from the complete data set.

Earthquake Location + RANSAC

1. Randomly sample a subset of phase picks S
2. Evaluate subset quality (e.g., P&S pick counts)
3. Locate earthquake using the subset S
4. Apply the model to the full dataset G; Compute evaluation metrics (e.g., inliers and residuals)
5. Evaluate model quality; Update the best model if improved;
6. Repeat for K iterations, or until stopping criteria met.
7. Use the best inlier subset to determine the final earthquake location

Benchmarking on synthetic datasets

Stanford Earthquake Location Benchmarks



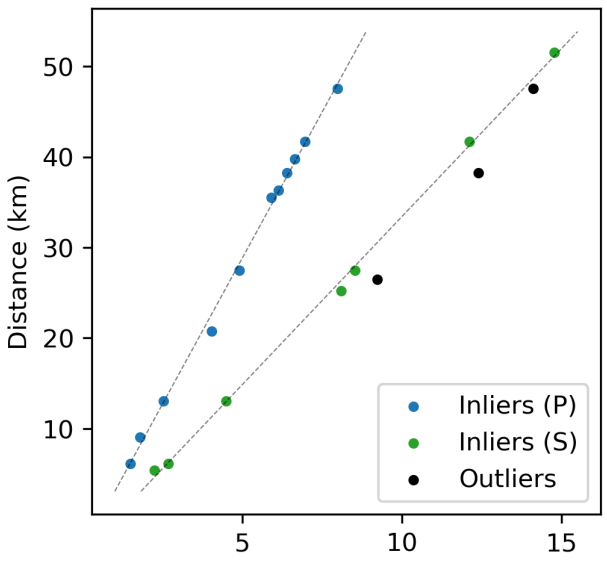
Method	Mean Accuracy Error (km)		Chamfer Distance
	Horizontal	Depth	
HypoInverse	0.824	1.118	1.617
VELEST	0.696	0.559	1.170
NonLinLoc	0.953	0.969	1.626
ADLoc	1.132	1.257	1.802
ADLoc+SST	0.680	0.539	1.080
ADLoc+SST+RANSAC (1.0s)	0.630	0.498	1.035
ADLoc+SST+RANSAC (0.2s)	0.456	0.450	0.915

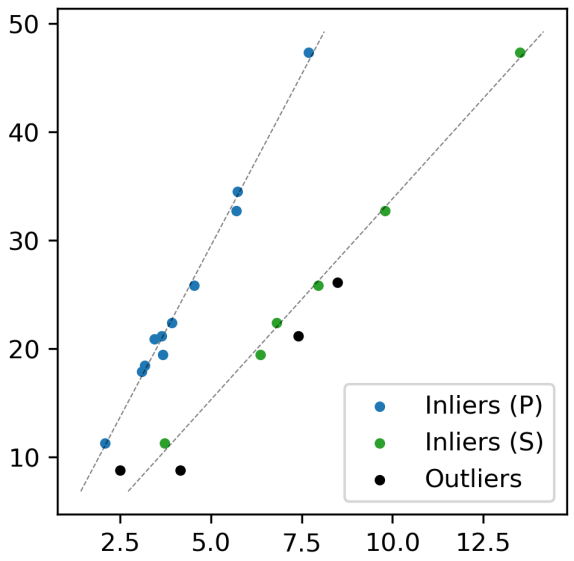
The dataset, comprising 19,994 phase picks, includes:
1% P-phase; 4% S-phase outliers.

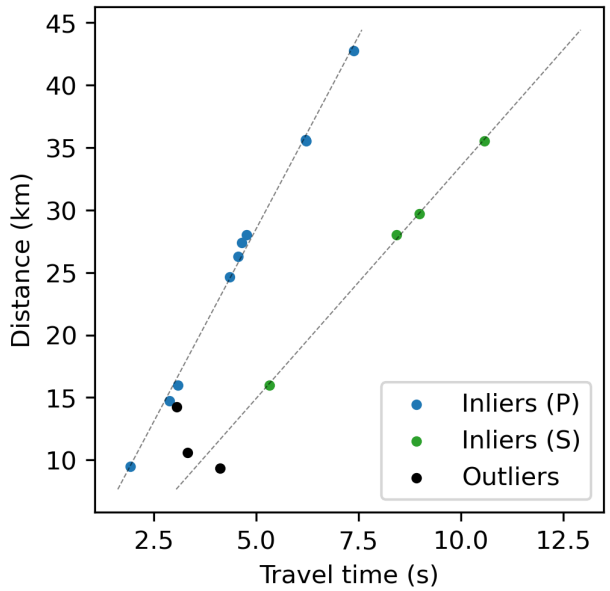
ADLoc, using a 0.2 s threshold, identified:
0.9% (184) P-phase; 2.7% (539) S-phase outliers.

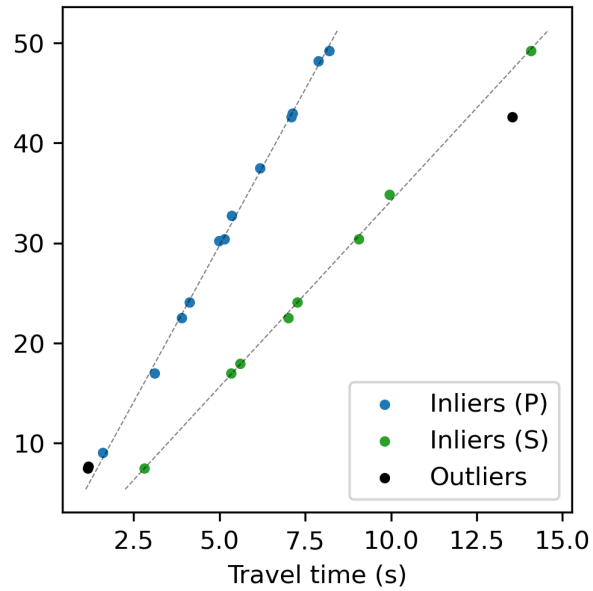
(Yu et al. 2024)

Examples of outliers

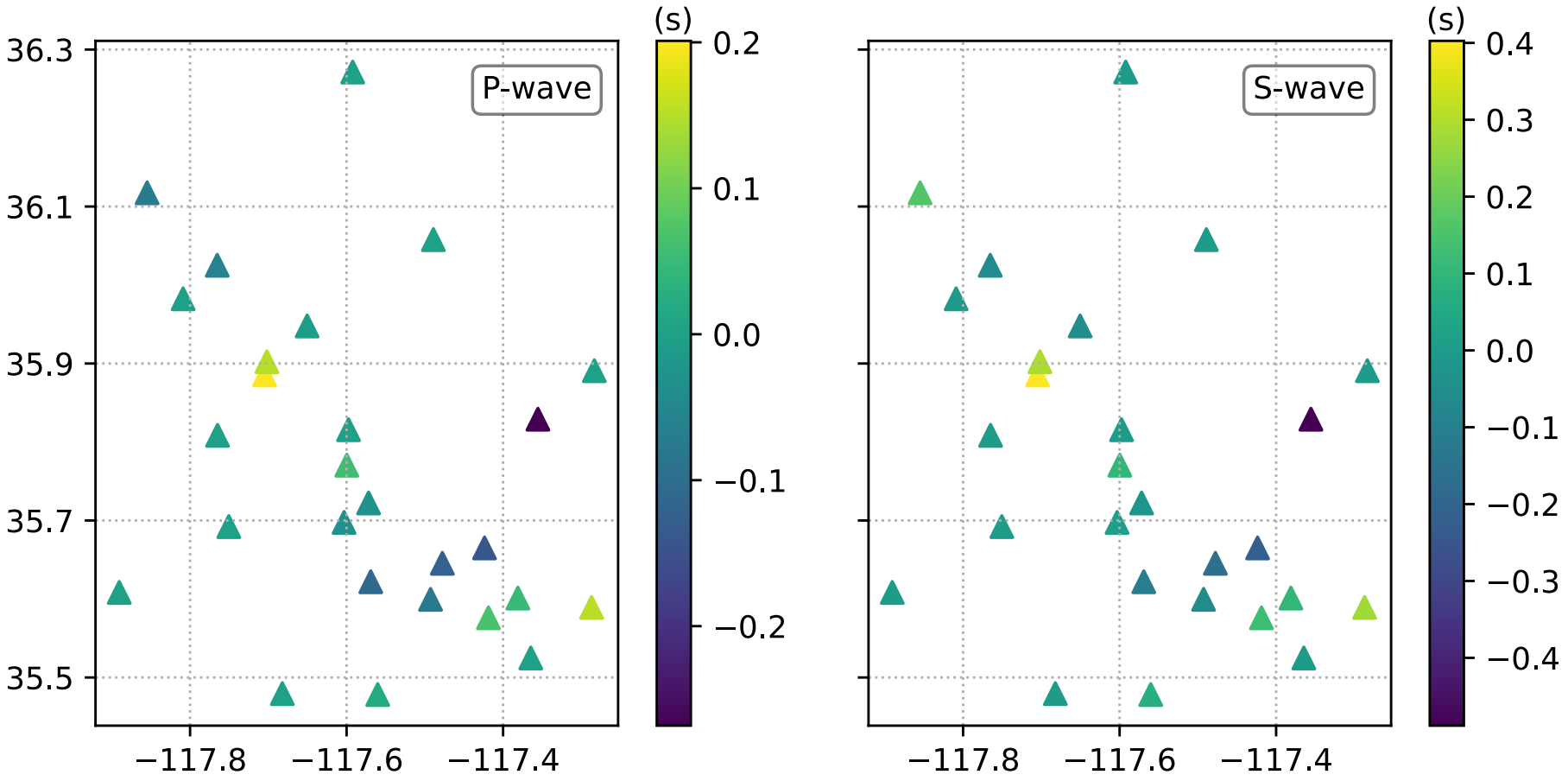






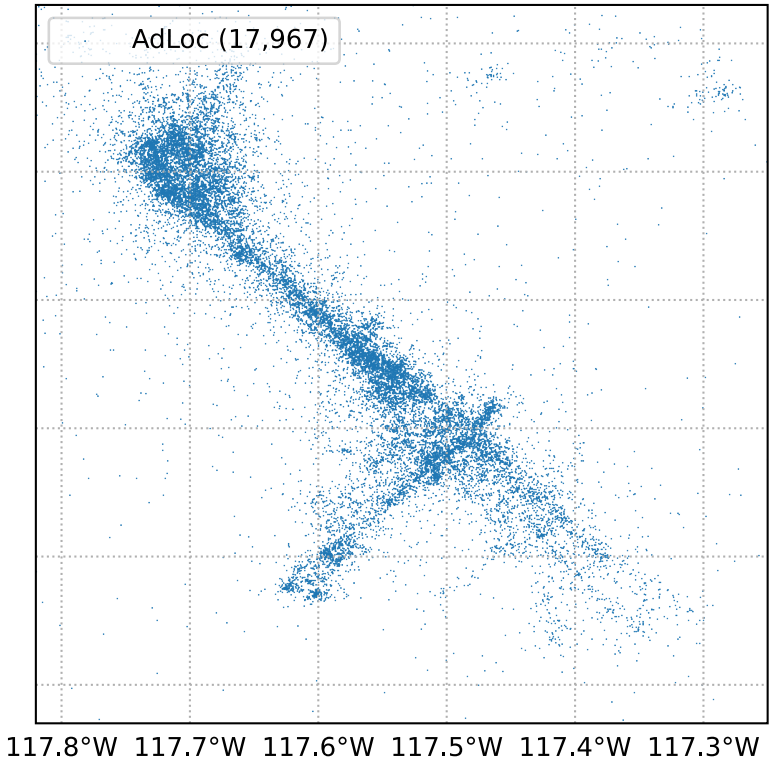


Station correction terms

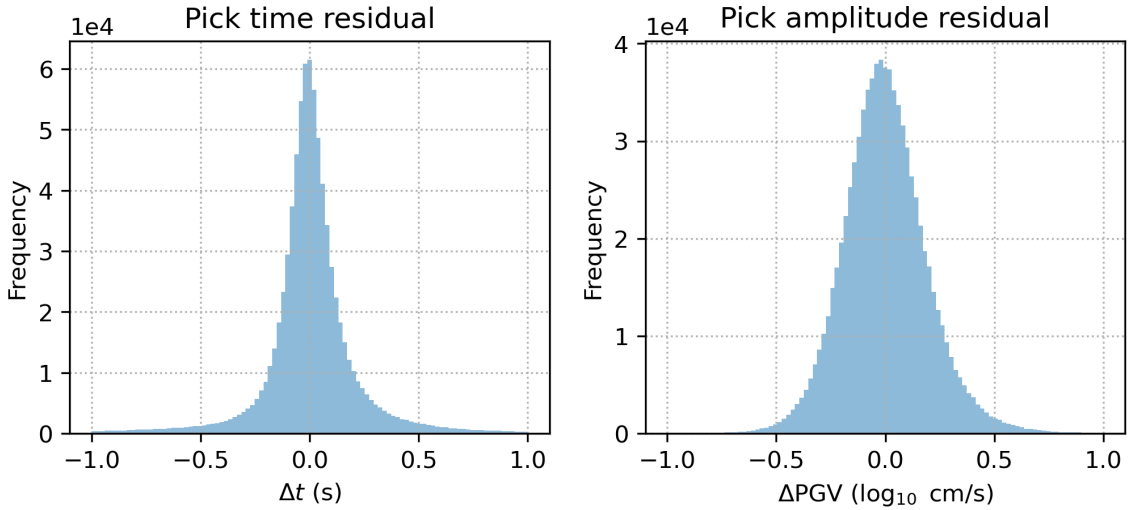


Including amplitudes for outlier detection

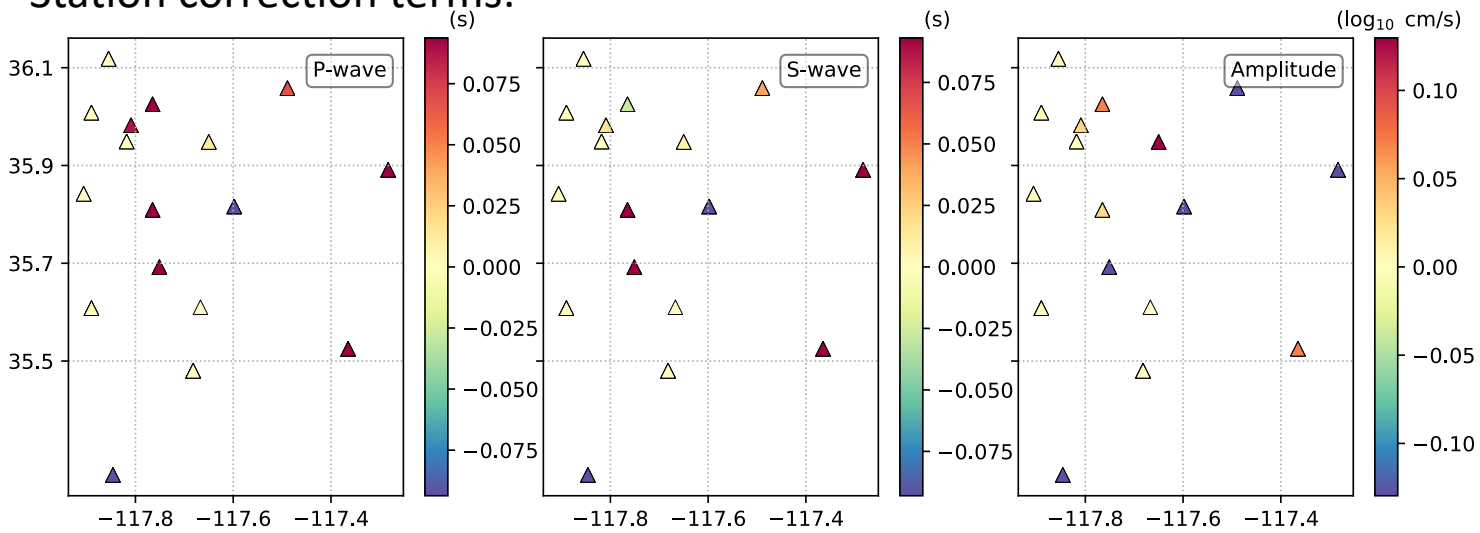
Application to the 2019 Ridgecrest earthquake



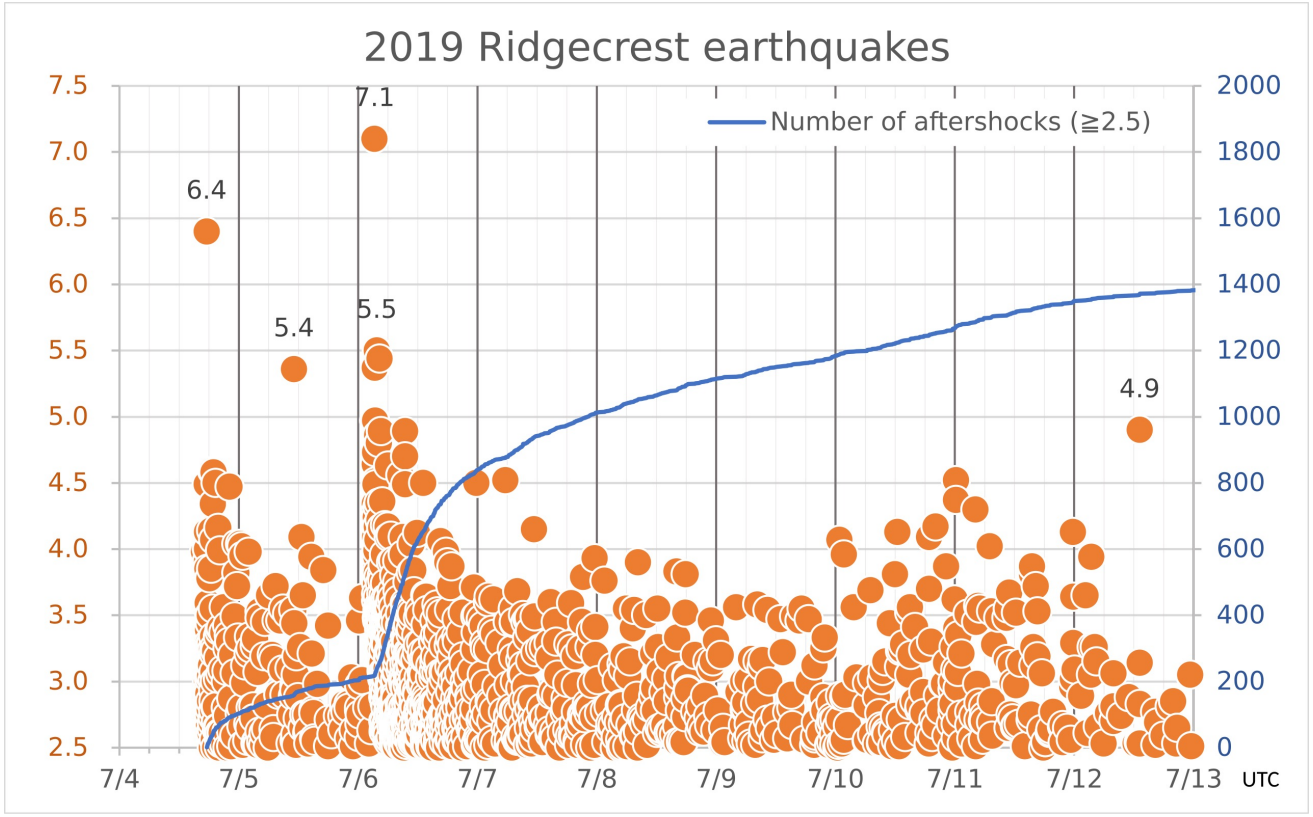
Residuals:



Station correction terms:



Q4: How to determine earthquake magnitudes?



https://en.wikipedia.org/wiki/2019_Ridgecrest_earthquakes

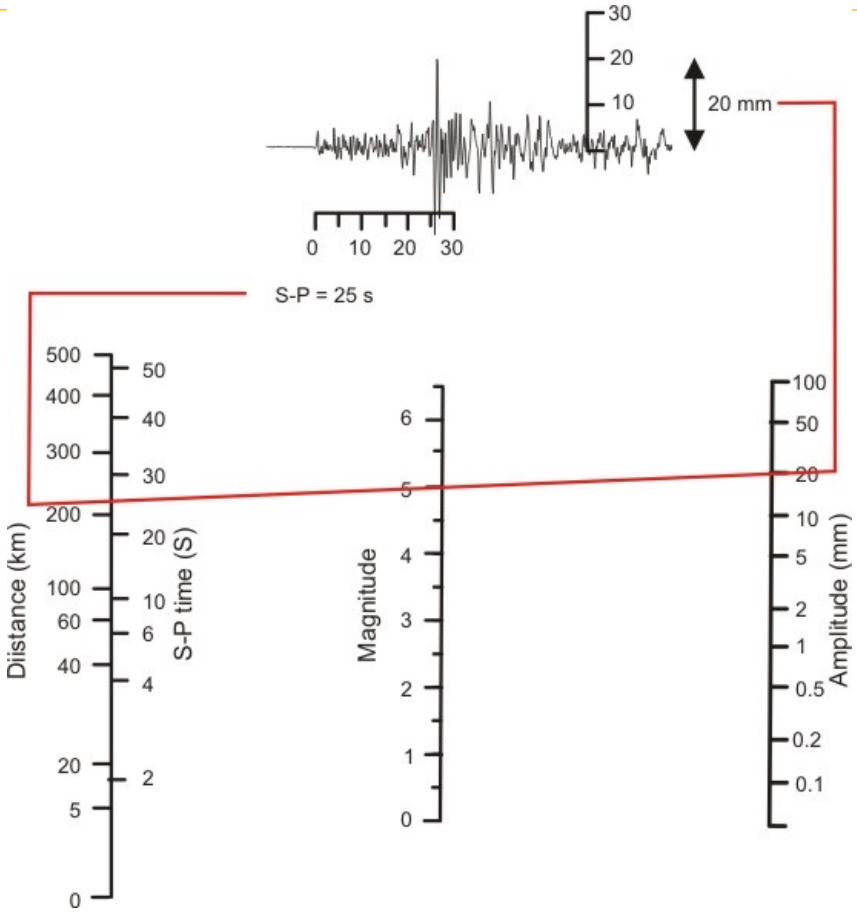
Local Magnitude (M_L)

$$M_L = \log(A) - \log A_0(r) + dM_L,$$

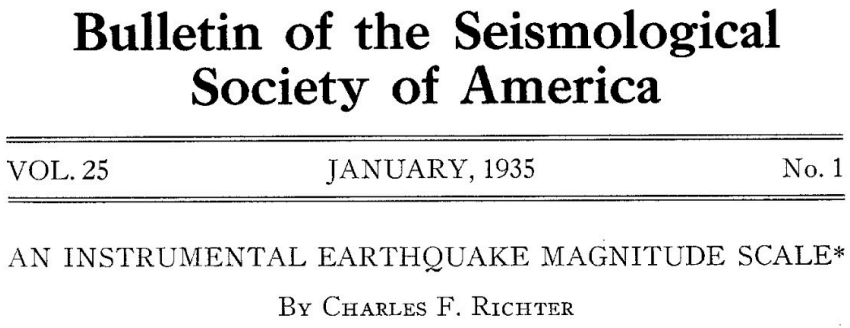
A: **peak amplitude** (mm) on (horizontal) Wood-Anderson seismogram

Attenuation correction: depend on event-station distance *r*; varies by region

dM_L: station correction (different for each station; default value 0)



Adapted from *Richter (1958): Elementary Seismology*



Bulletin of the Seismological Society of America

VOL. 25 JANUARY, 1935 No. 1

AN INSTRUMENTAL EARTHQUAKE MAGNITUDE SCALE*

By Charles F. Richter

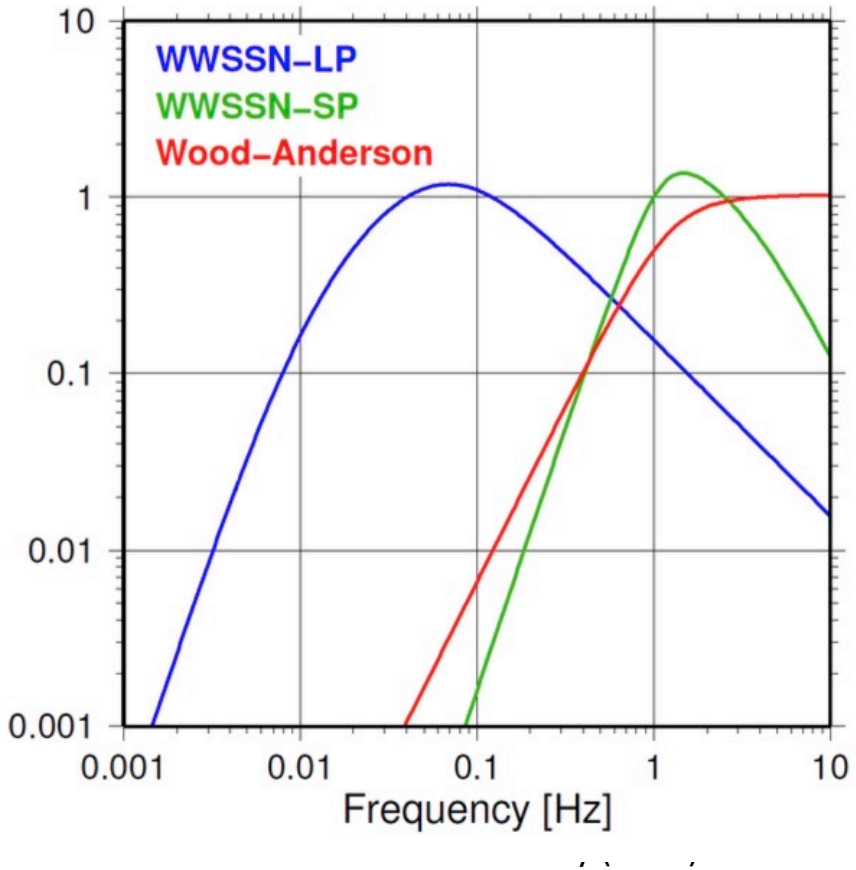
Richter (1935): Southern California local earthquake with amplitude=1 mm at station r=100 km away defined as M_L=3

https://earthquakes.bgs.ac.uk/education/eq_guide/eq_booklet_measuring_size_eqs.htm

Why **Local Magnitude (M_L)** for deep-learning enhanced earthquake catalogs?

- Appropriate for small earthquakes at local distances (Wood-Anderson: freq >1 Hz), newly detected in enhanced catalog

- Quick & easy to compute automatically: get peak amplitude at each station



https://doi.org/10.2312/gfz.nmsop-2_is_3.3

$$M_L = \log(A) - \log A_0(r) + dM_L,$$

peak amplitude | attenuation (distance) correction | station correction

$$-\log A_0(r) = 1.11\log(r) + 0.00189r + 0.591 + \mathrm{TP}(n) \times T(n,z), \quad (4)$$

where n is summed from 1 to 6. The TP(n) coefficients are

$\mathrm{TP}(1) = +0.056$, $\mathrm{TP}(2) = -0.031$,
$\mathrm{TP}(3) = -0.053$, $\mathrm{TP}(4) = -0.080$,
$\mathrm{TP}(5) = -0.028$, $\mathrm{TP}(6) = +0.015$,

and z is the scale transformation of r,

$$z(r) = 1.11366 \times \log(r) - 2.00574, \quad (5)$$

$$T(n,z) = \cos[n \times a\cos(z)]. \quad (6)$$

Uhrhammer et al. (2011), BSSA

$$-\log A_0 = \begin{cases} 2.07 \times \log(R_{hyp}) + 0.0002 \times (R_{hyp} - 100) - 0.72 & R_{hyp} \le 16 \text{ km} \\ 1.54 \times \log(R_{hyp}) + 0.0002 \times (R_{hyp} - 100) - 0.08 & 16 \text{ km} < R_{hyp} \le 105 \text{ km}, \\ 0.29 \times \log(R_{hyp}) + 0.0002 \times (R_{hyp} - 100) + 2.45 & R_{hyp} > 105 \text{ km} \end{cases}$$

Kavoura et al. (2020), SRL

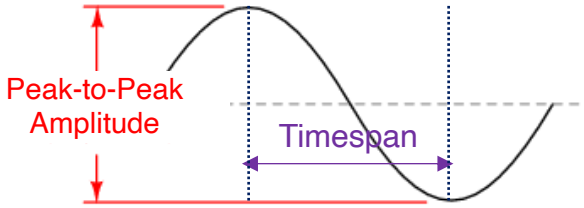
r, R_{hyp}: hypocentral distance to station (km)

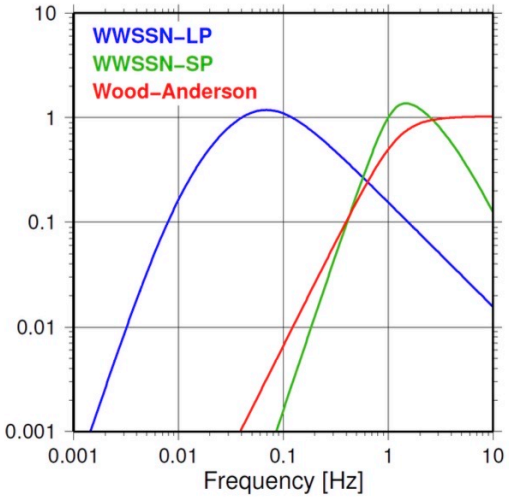
- **M_L equation** (distance & station correction) **varies with region**; often determined empirically with regression or inversion
  - **Preferred**: use existing M_L equation from literature for your region; dM_L usually in a supplement table, or text file from seismic network (e.g. SCSN)
  - **Otherwise**: calibrate M_L yourself with observed peak amplitudes (A) and event-station distances (r) by setting up an inversion (e.g. Yoon et al., 2023, BSSA for southwest Puerto Rico)

M_L: How to get Wood-Anderson seismogram?

$$M_L = \log(A) - \log A_0(r) + dM_L,$$

A: **peak amplitude** (mm) on (horizontal) Wood-Anderson seismogram





- Get StationXML file for each station
  - ObsPy (inside loop over stations):

```
inv = client.get_stations(network=net, station=sta, starttime=start_time,
endtime=end_time, level='response', filename=out_stationxml_file, format='xml')
```

  - https://docs.obspy.org/packages/autogen/obspy.clients.fdsn.client.Client.get_stations.html
- Cut event time window, get peak-to-peak amplitude & timespan, then correct for Wood-Anderson instrument response.
  - ObsPy (inside loop over stations):

```
A_sta_WA = estimate_wood_anderson_amplitude_using_response(inv, A_sta_pkpk, timesp_sta_pkpk)
```

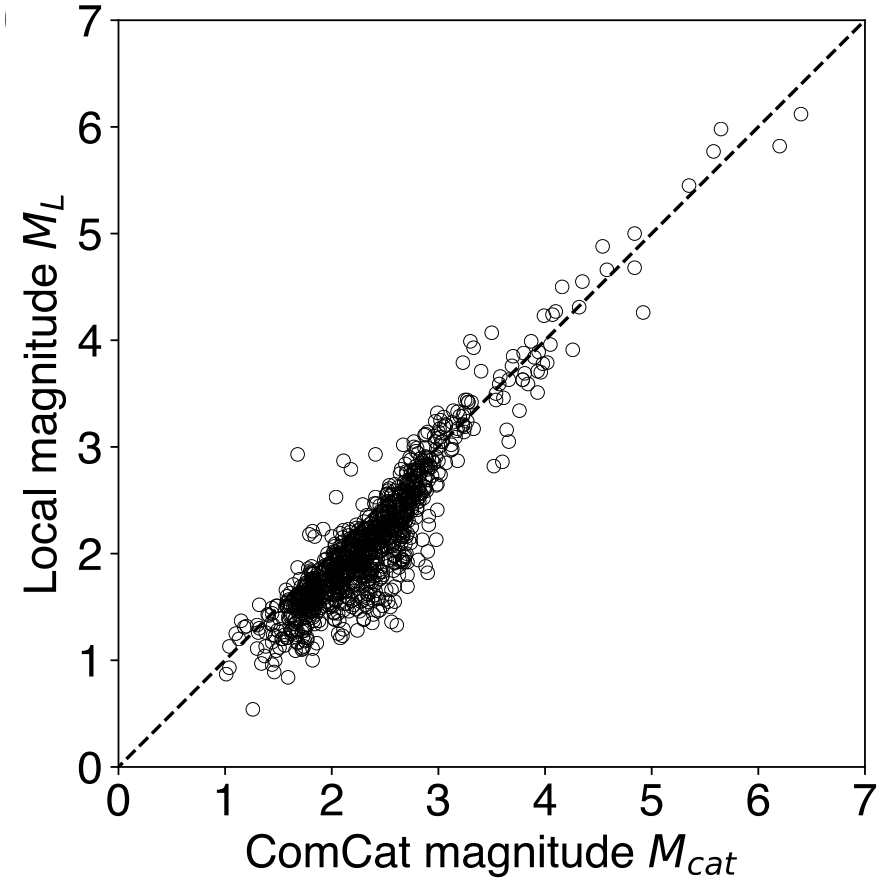
  - https://docs.obspy.org/packages/autogen/obspy.signal.invsim.estimate_wood_anderson_amplitude_using_response.html

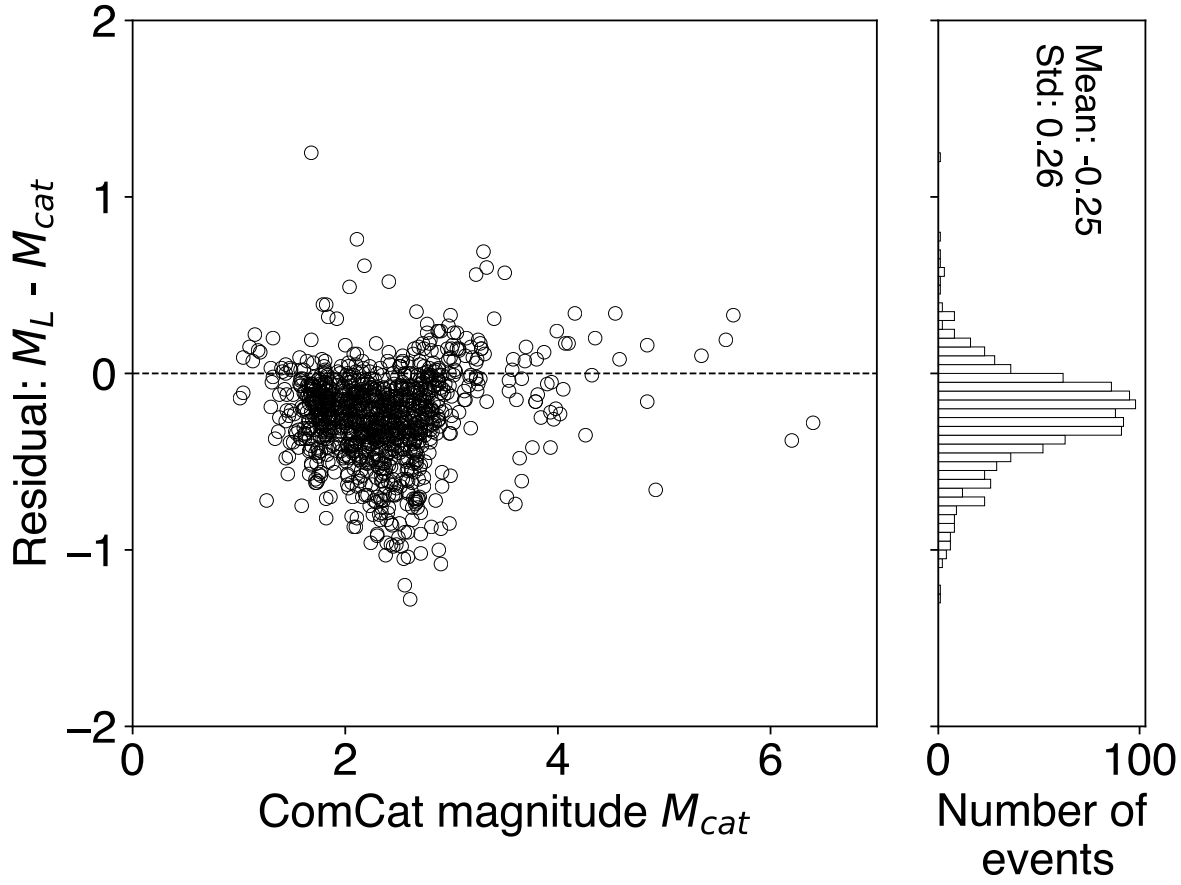
M_L: Calculation steps

- M_L for each station:
  - Get event-station distance *R* [km]; should I use current station to compute magnitude?
    - e.g. use only stations within a maximum distance (R = 100 km?)
    - e.g. use only stations with deep-learning automatic picks?
  - Get event waveform at station within short time window
    - e.g. start time: 5 seconds before P arrival, end time: 4*(S-P time) seconds after S arrival
  - Get peak-peak amplitude & timespan in windowed event waveform
    - 3-component (preferred): average of peak_amp[North] and peak_amp[East]
    - 1-component: peak_amp[Z]
  - Correct for Wood-Anderson instrument response
    - Can use ObsPy **`estimate_wood_anderson_amplitude_using_response()`**
  - Compute station M_L with region-specific equation
- M_L for event: median over all M_L station values
- If possible, compare calculated M_L with reference catalog M_L values for common “MATCH” events

Calculated $\mathbf{M_L}$ vs. reference catalog (ComCat) $\mathbf{M_{cat}}$

Check magnitude type: is it also ML?





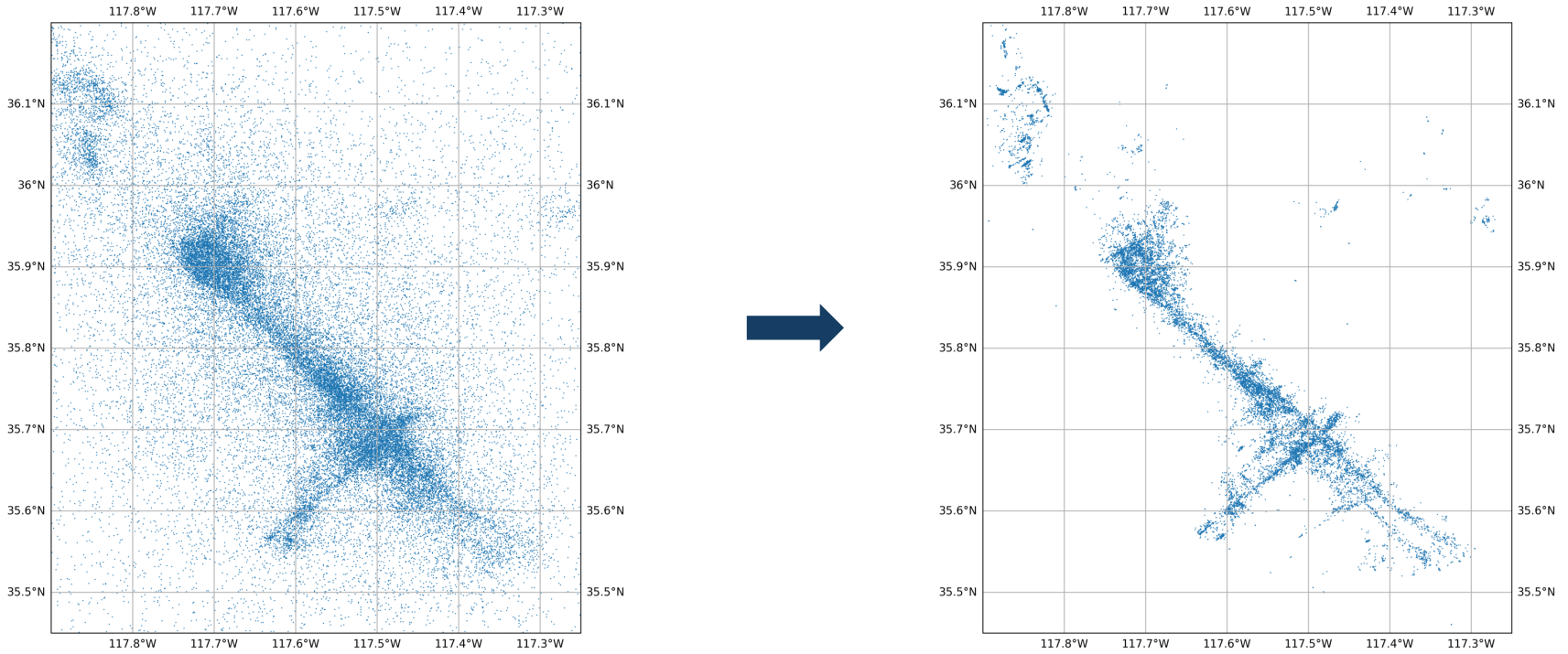
Yoon and Shelly (2024) TSR

Alternatives to M_L (for small local earthquakes)

- Moment magnitude M_w from coda waves
  - Source code: https://github.com/LLNL/coda-calibration-tool
  - Docs: https://software.llnl.gov/coda-calibration-tool/
  - Examples: Holt et al. (2021) SRL; Shelly et al. (2021) BSSA; Patton et al. (2025) BSSA
  - Requires more computation and calibration – I have not tried it myself

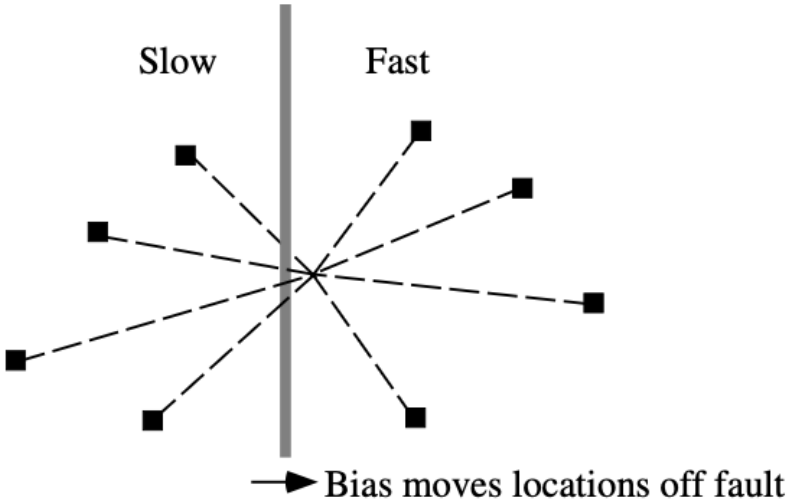
- Relative magnitude based on event-pair cross-correlation or amplitude-ratios: good choice for template-matching (Eric's talk)
  - Calibrate against known magnitudes from reference (template) events
  - Examples: Cleveland and Ammon (2015) BSSA; Shelly et al. (2016) JGR; Gable and Huang (2024) BSSA

Q5: How to constrain earthquake relative locations?

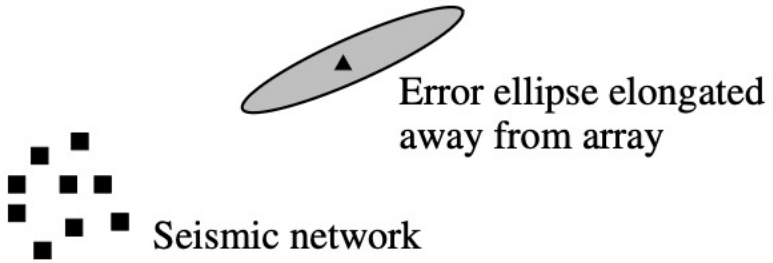


Challenges in absolute locations

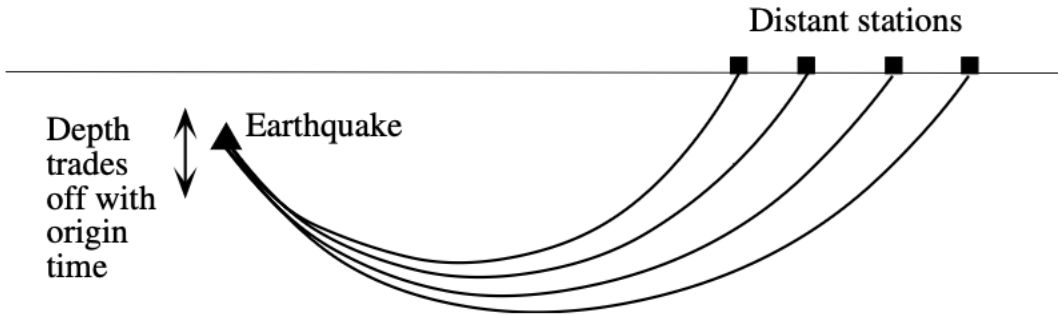
- Unmodeled velocity heterogeneity



- Station coverage

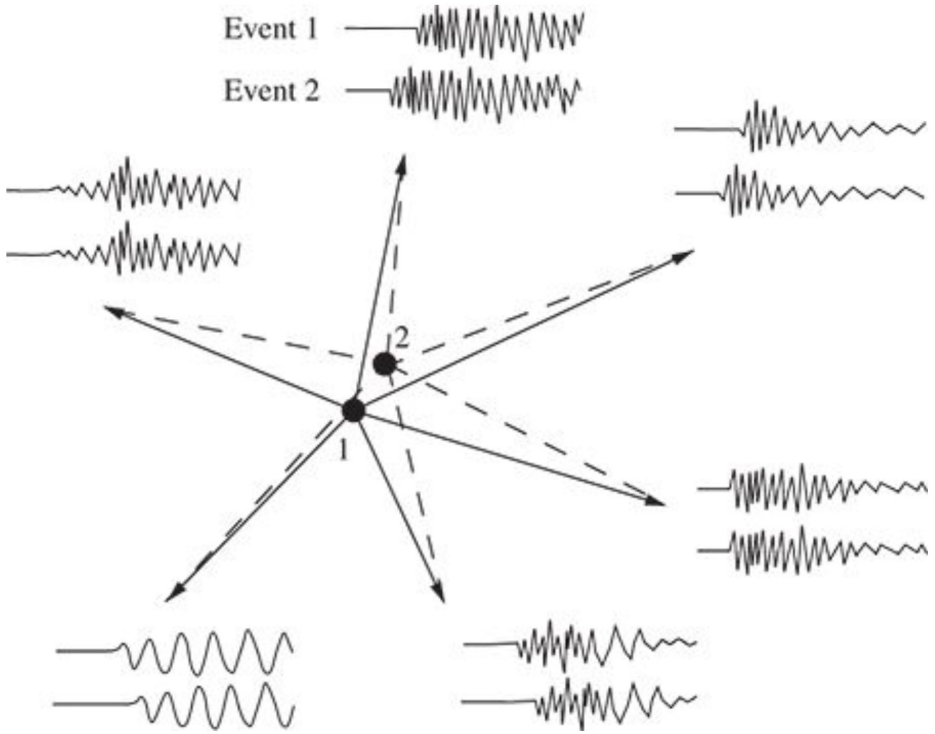


- Depth tradeoff

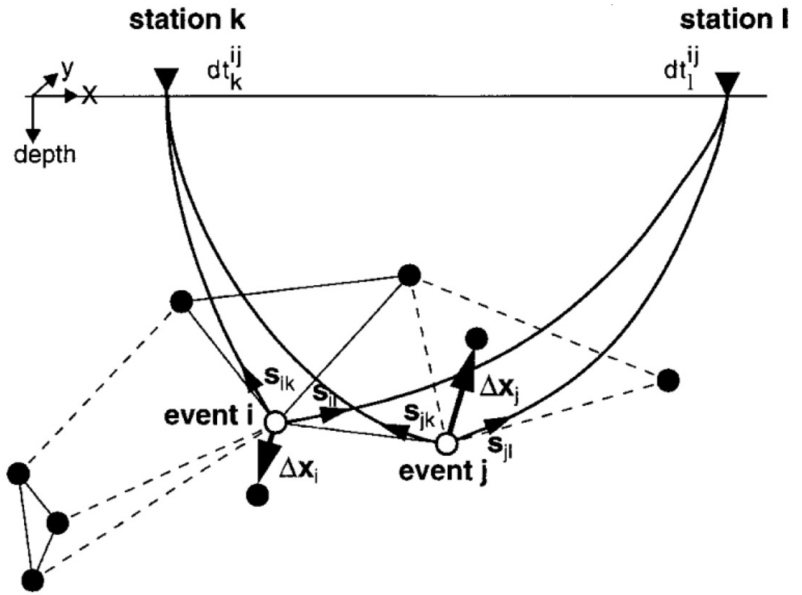


(Credit: Peter Shear)

Double-difference earthquake location



$$\Delta r_k^{ij} = \left(t_k^i - t_k^j\right) - \left(\hat{t}_k^i - \hat{t}_k^j\right)$$



HYPODD SOFTWARE FOR DOUBLE-DIFFERENCE EARTHQUAKE HYPOCENTER LOCATION

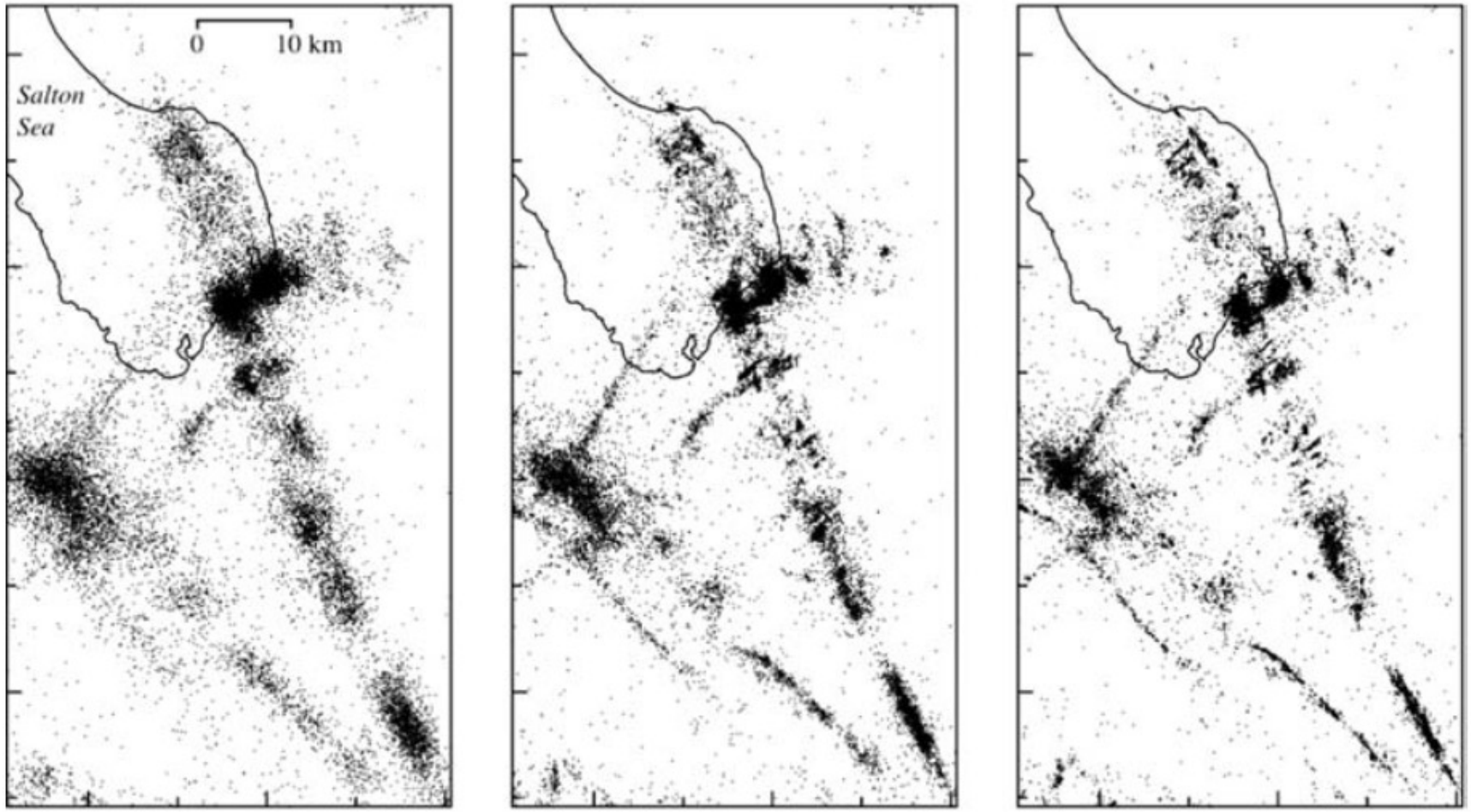
(Waldhauser and Ellsworth, 2000)

Comparison

single event location

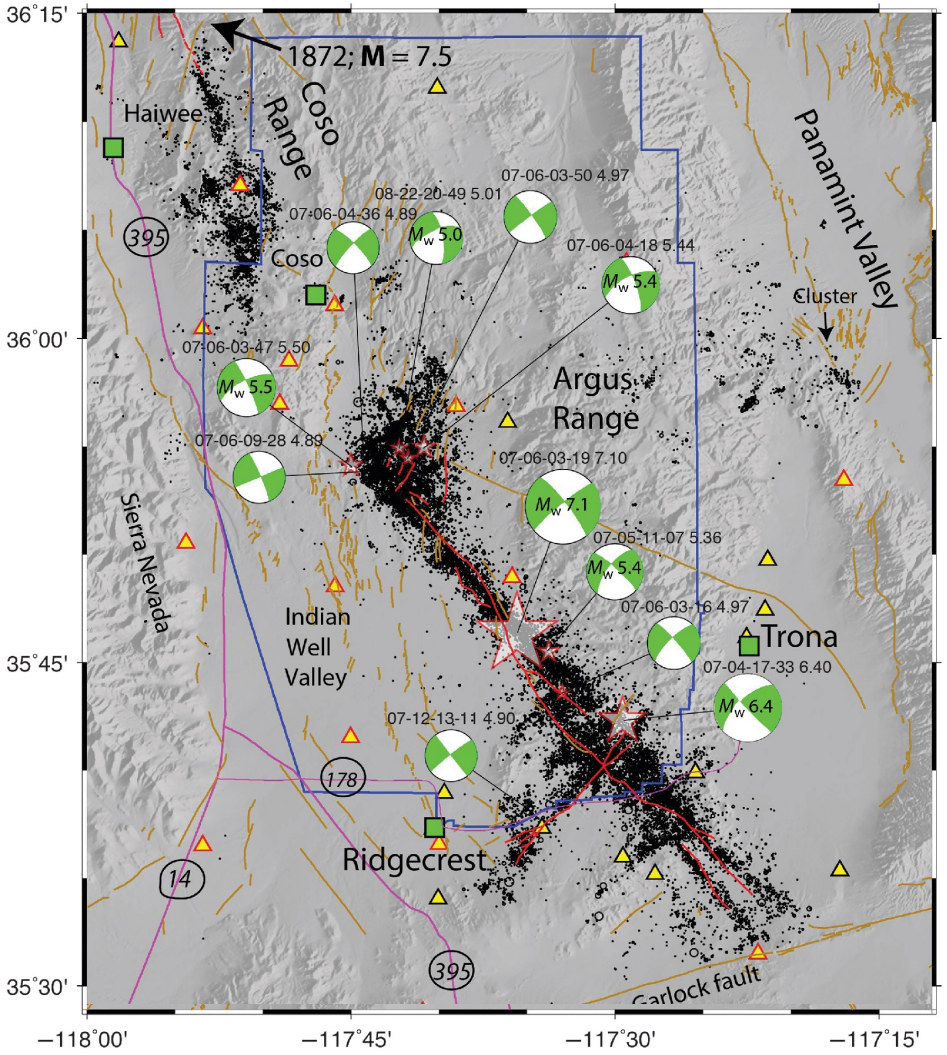
source-specific station term location (SSST)

waveform cross-correlation location



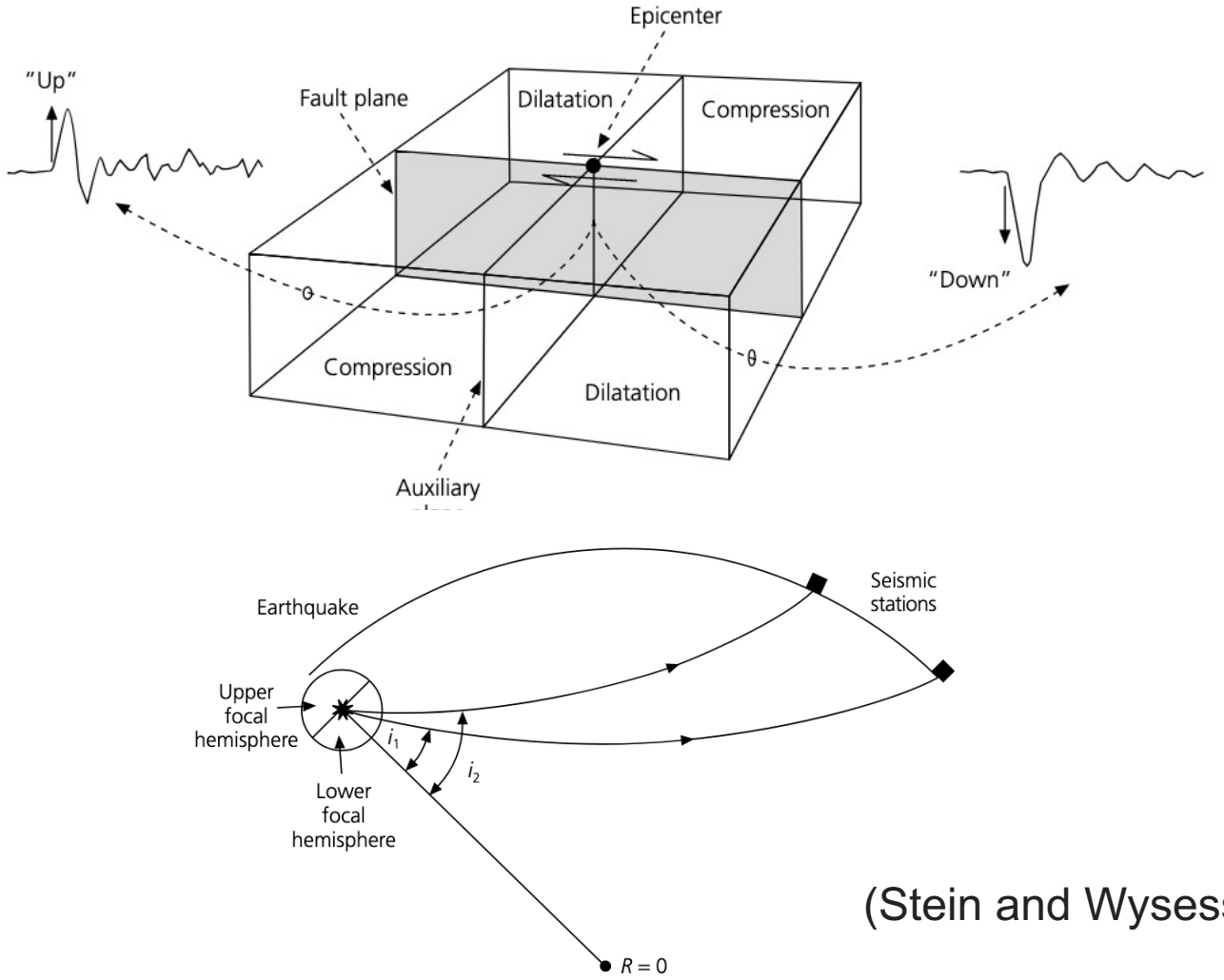
Lin et al. (2007)

Q6: How determine earthquake mechanism?

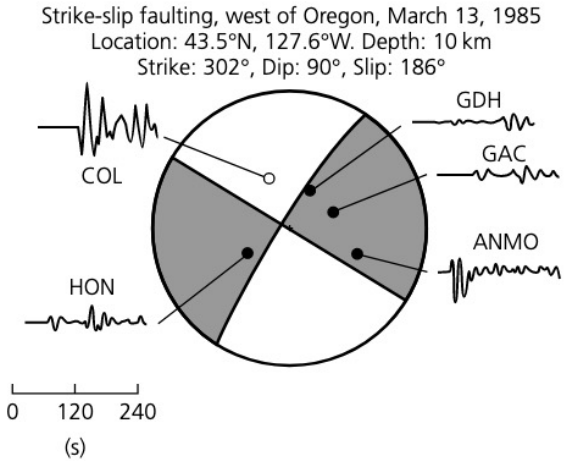


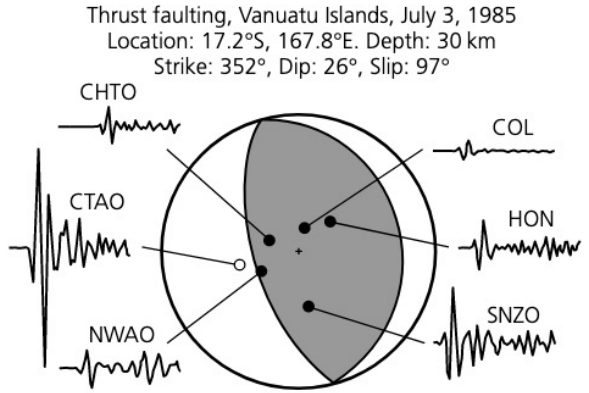
(Hauksson et al. 2020)

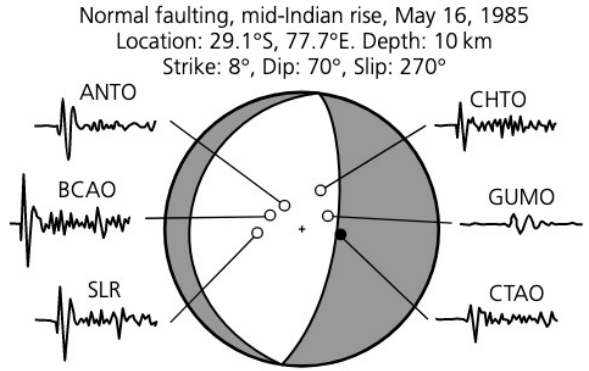
First-motion polarity



(Stein and Wysession 2009)







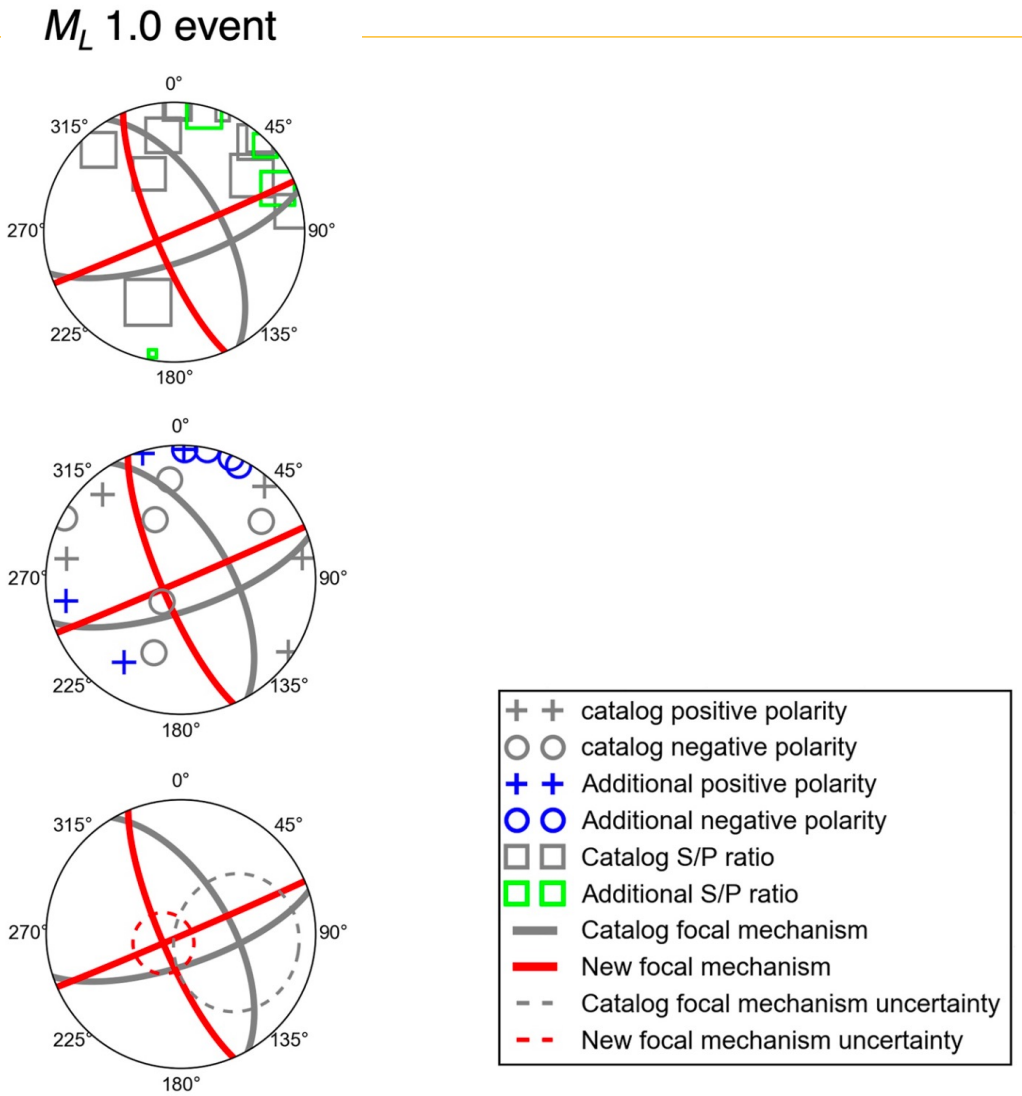
Constrain focal mechanisms

Input
S/P amplitude ratio

Input
P-wave polarity

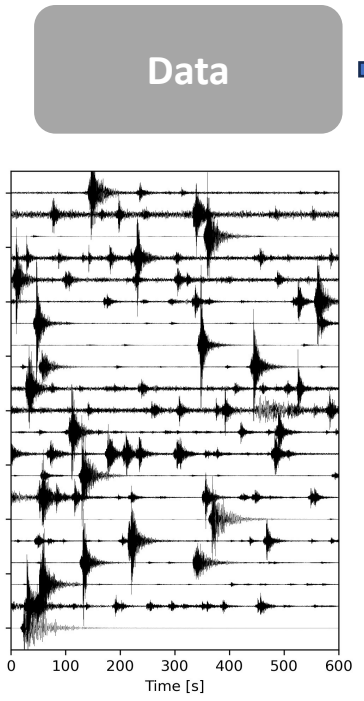
HASH

Output
Focal mechanism



(Hardebeck and Shear 2002)
(Cheng et al. 2023)
(Skoumal et al. 2024)

Multitask Learning: Arrival-time, Polarity, and Association



Data

Phase Picking

Phase Association

Polarity Picking

Earthquake Location

Earthquake Relocation

Catalog

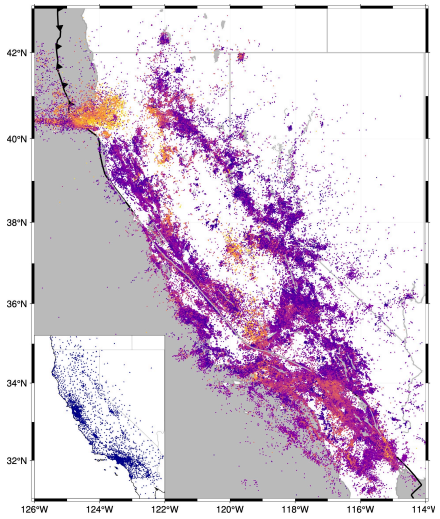
Focal Mechanism

Perception-based task

One Multi-task Neural Network
(e.g., PhaseNet+)

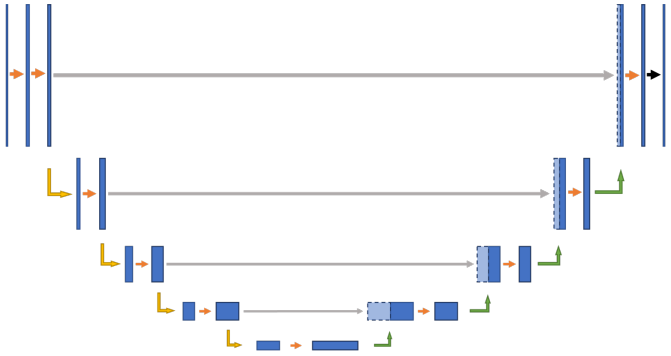
Physics-based task

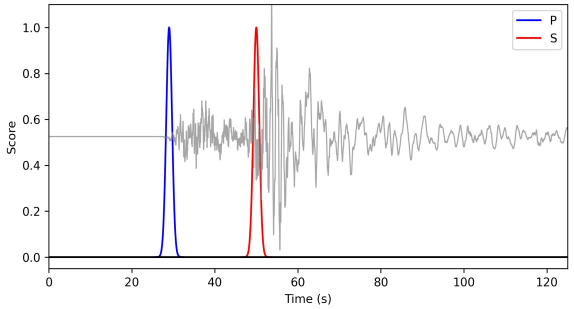
Geophysical Inversion



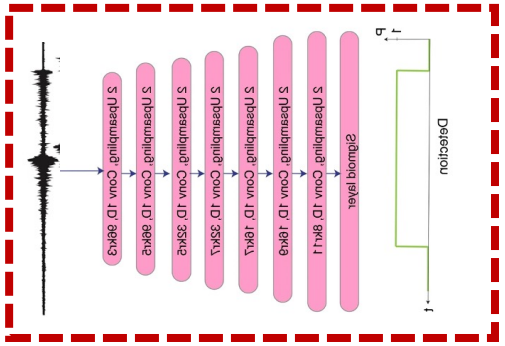
Neural Network Modules

Phase Picking

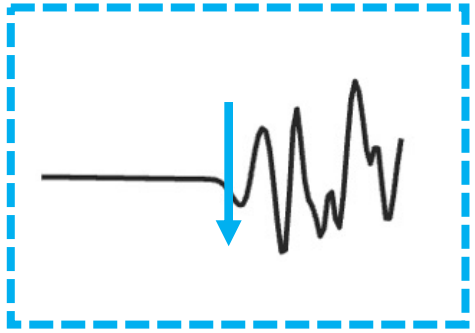




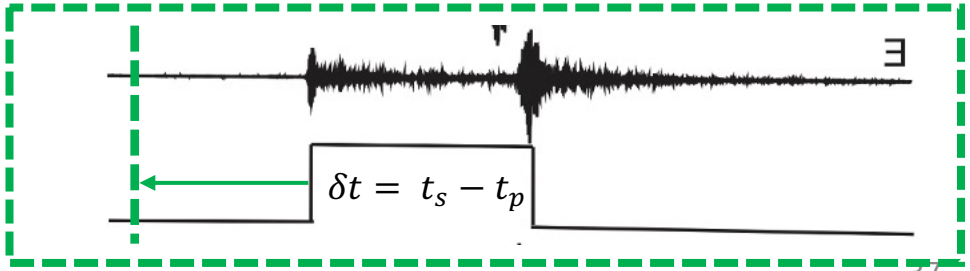
+ Event Detection



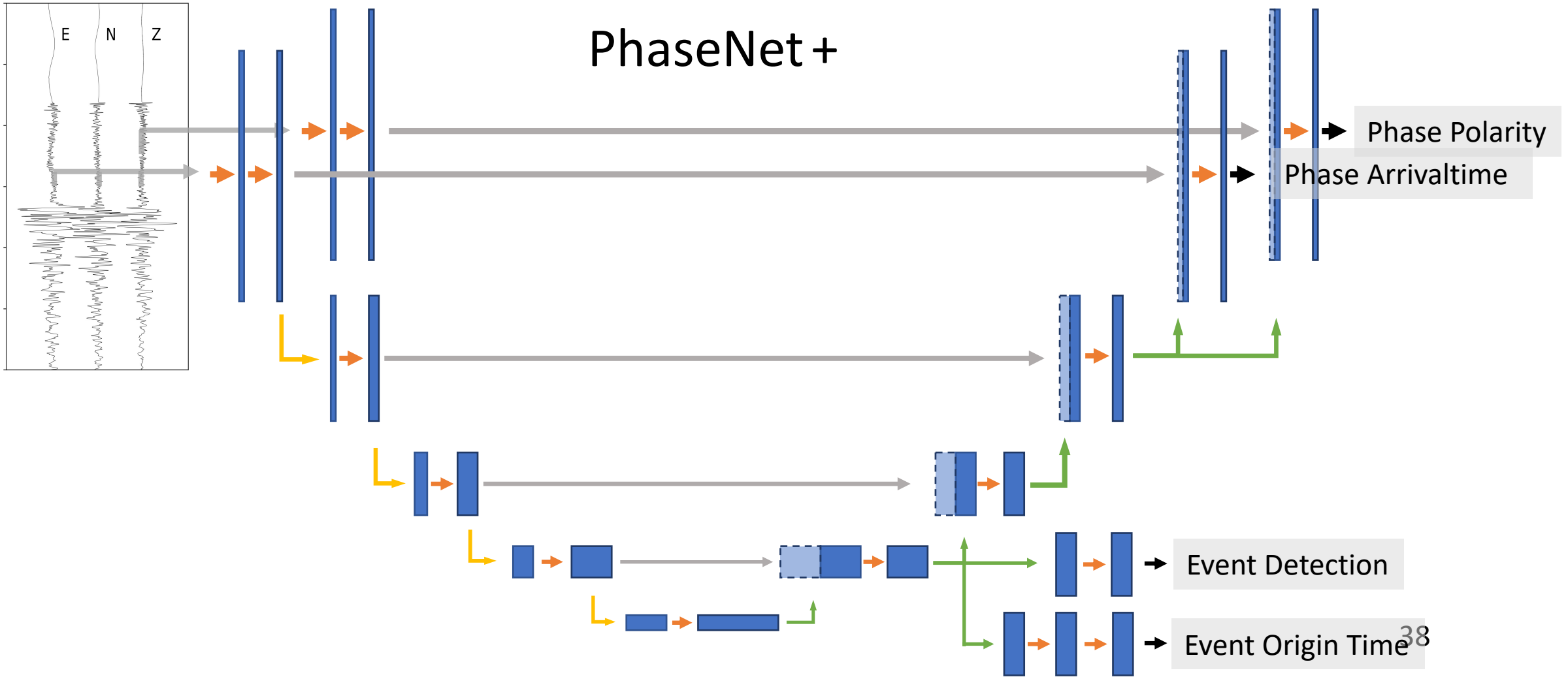
+ Polarity Picking



+ Phase Association based on origin time



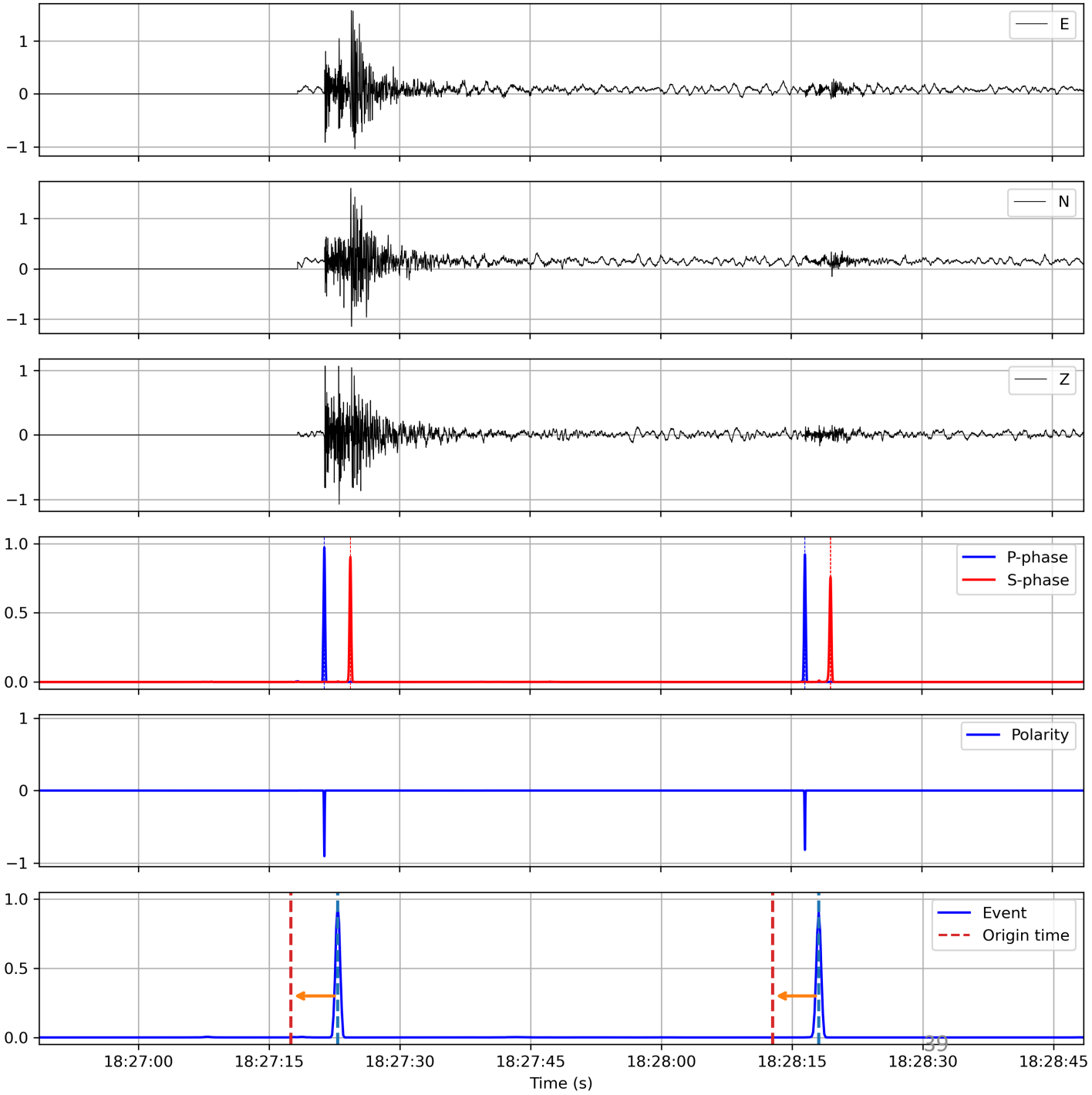
Multi-task Learning



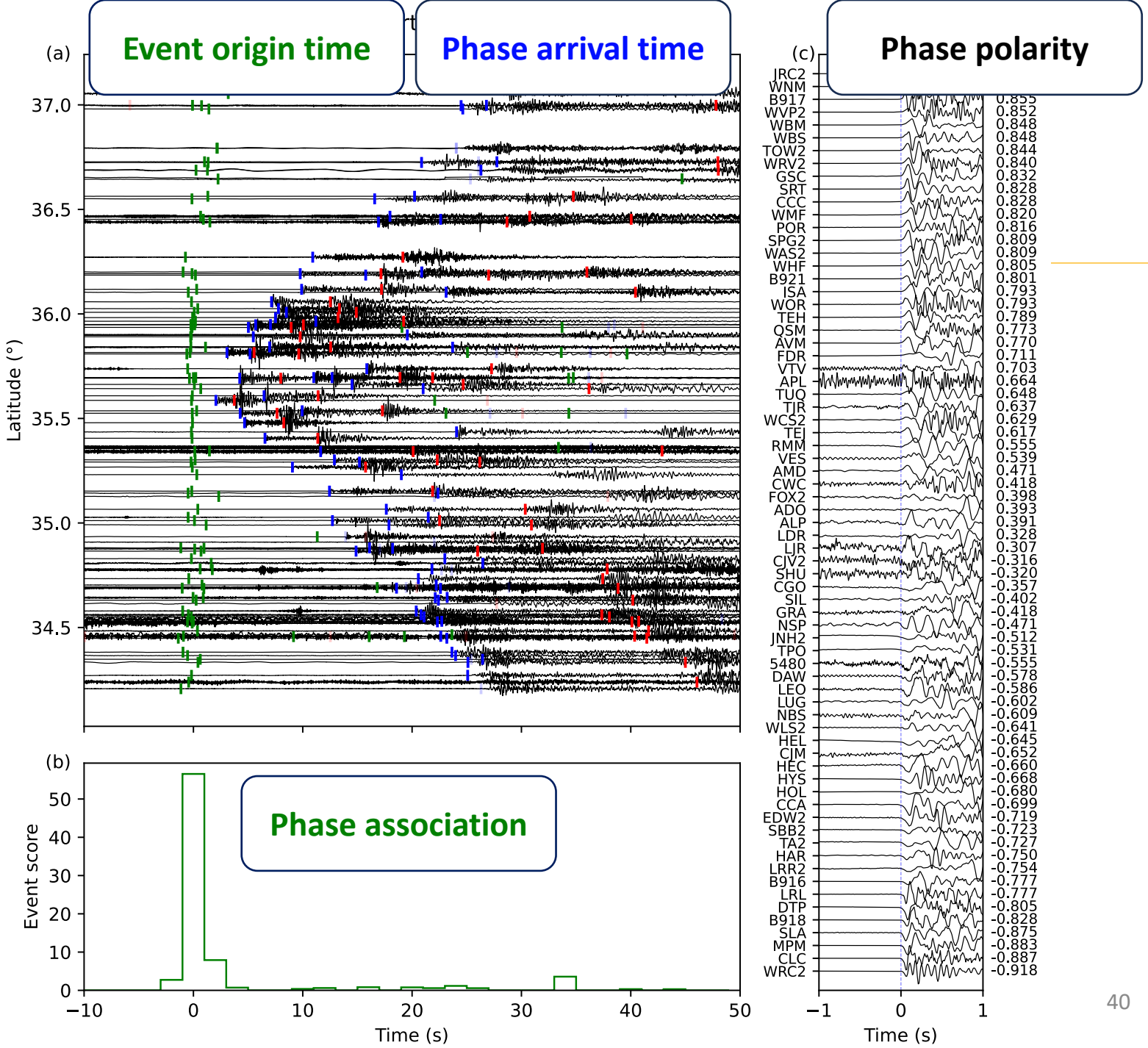
Examples:

PhaseNet+

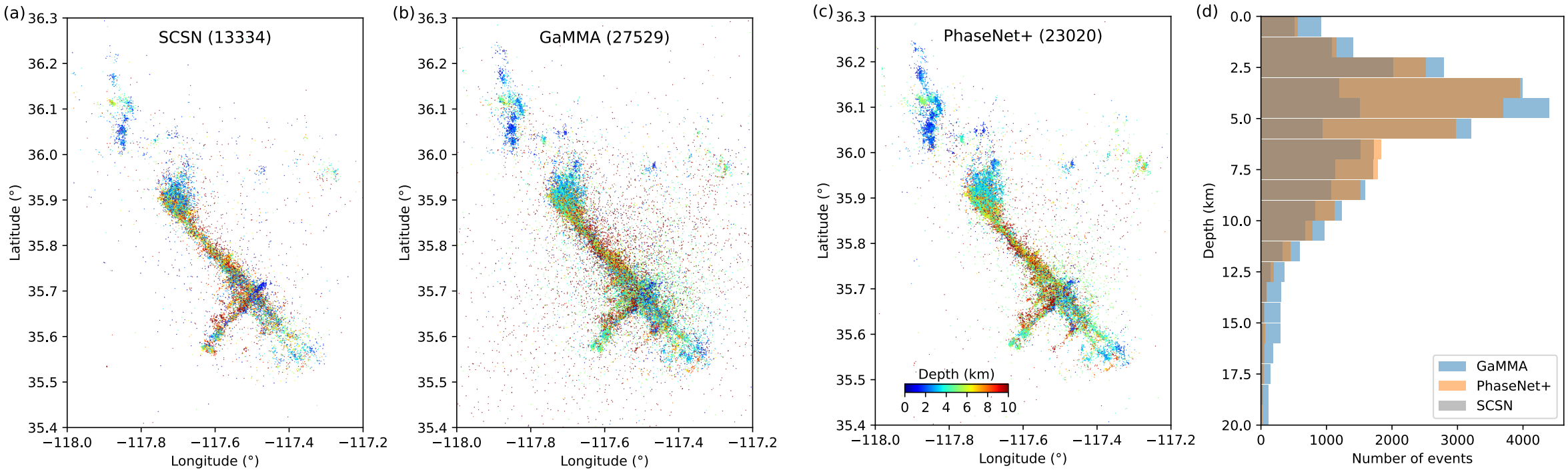
- Phase Arrival-time:
- Phase Polarity:
- Event Detection:
  Origin-time Prediction:



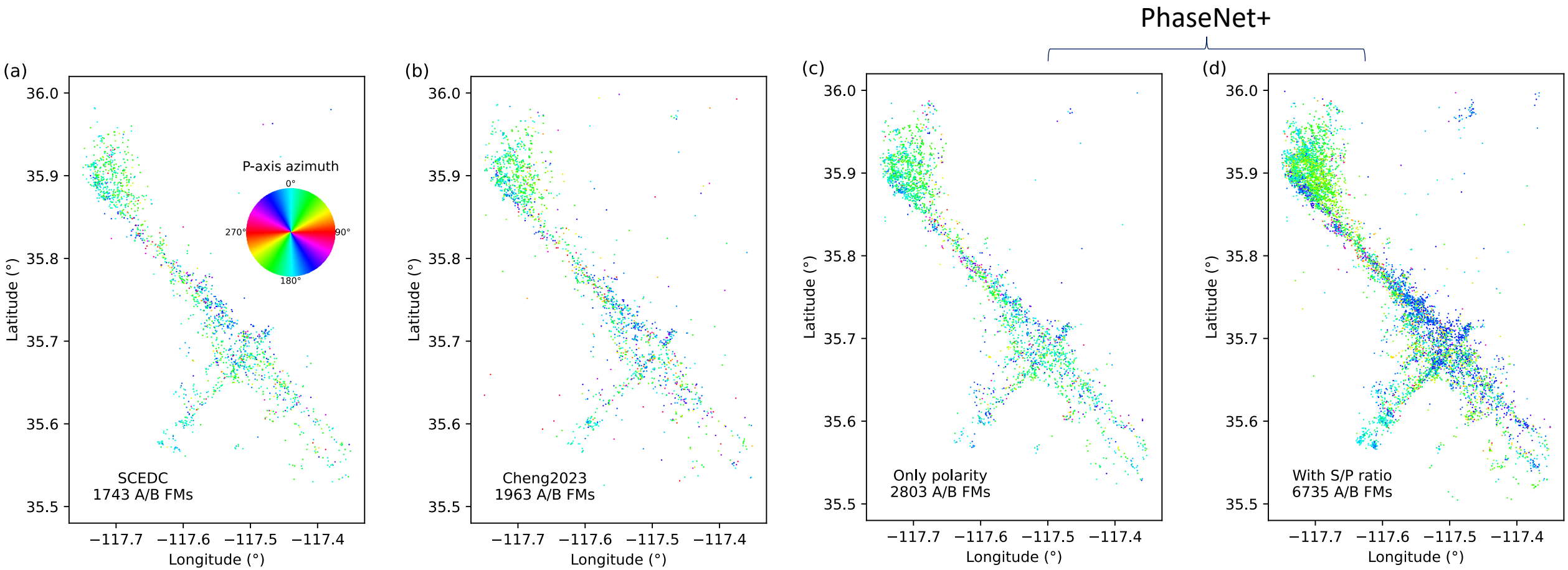
Examples:


Event origin time
Phase arrival time
Phase polarity
(a)
(b)
(c)
Latitude (°)
Event score
Phase association
Time (s)
Time (s)

Comparisons



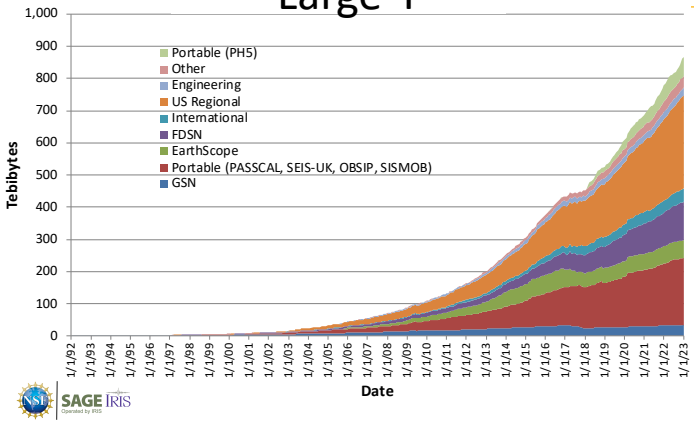
Comparisons



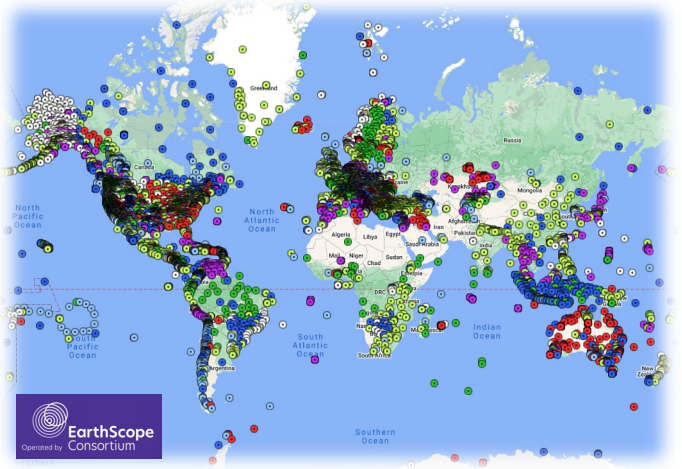
Applications of Earthquake Deep Catalogs

Seismic Datasets

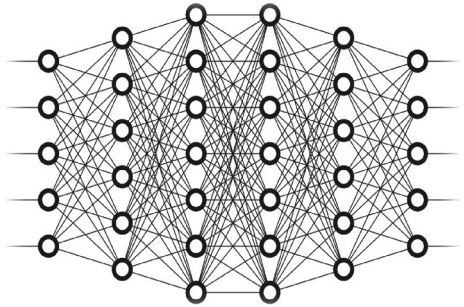
Large-T



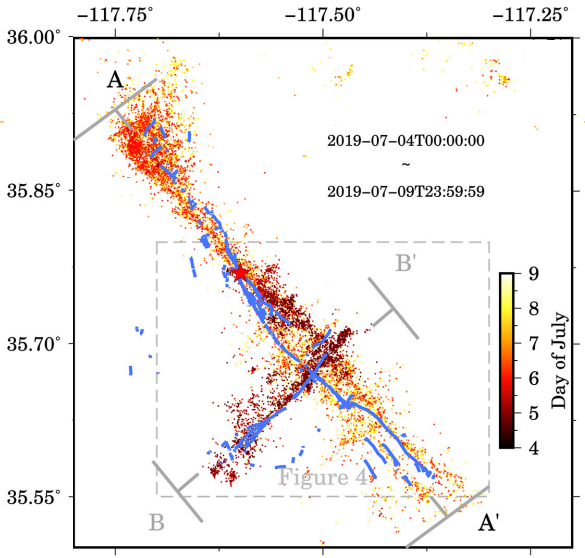
Large-N



Deep Neural Networks

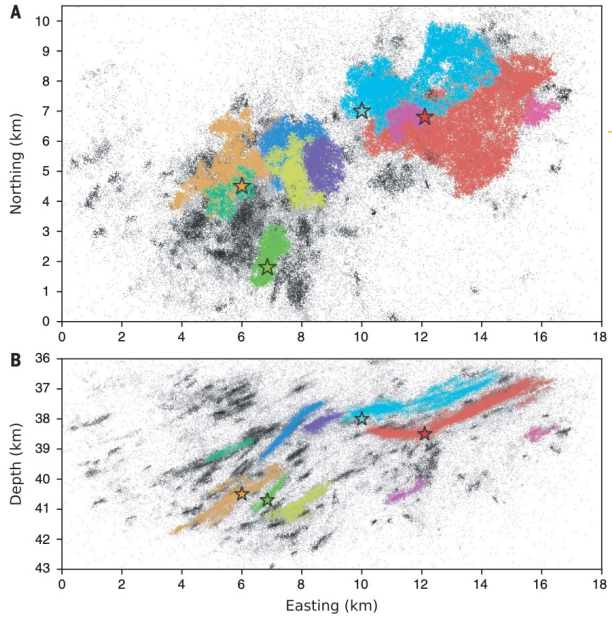


Tectonic earthquake



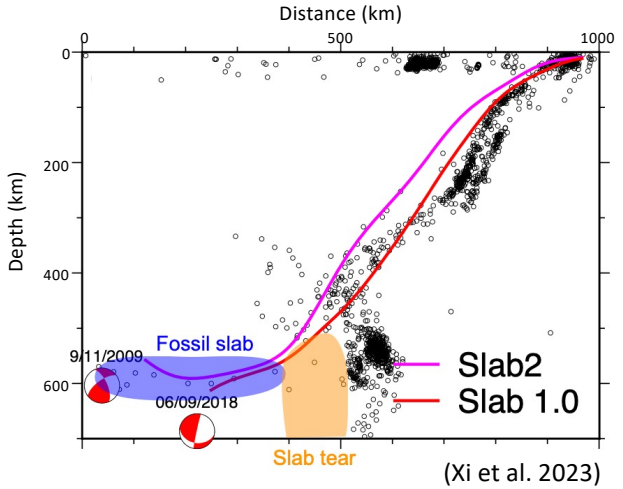
(Liu et al. 2020)

Volcanic earthquake



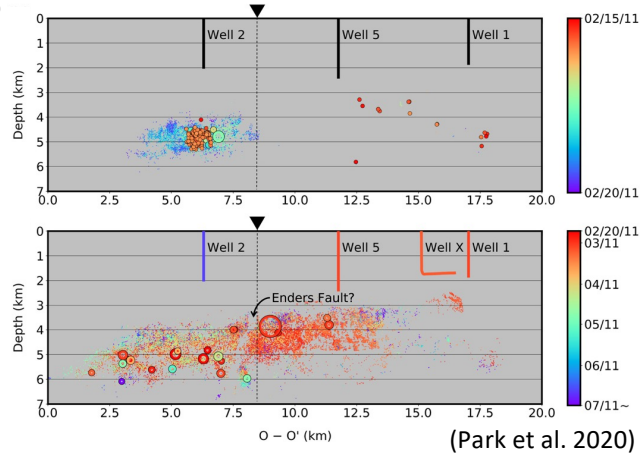
(Wilding et al. 2023)

Deep earthquake



(Xi et al. 2023)

Induced earthquake



(Park et al. 2020)