ISTITUTO NAZIONALE
DI GEOFISICA E VULCANOLOGIA

Benchmark datasets for ML in seismology

Alberto Michelini

Istituto Nazionale di Geofisica e Vulcanologia

g2net conference: WG2-3 workshop, June 8, 2022

Outline

- Motivations
- Benchmarking
  - Datasets
  - Platform
- Conclusions & Outlook

Background

- Machine & Deep Learning (ML, DL) applications are widespread in several fields of Science
- Several ML/DL *software* platforms *(e.g., TensorFlow/Keras, PyTorch, Scikit learn, Caffe, …)* to perform sophisticated analysis
- Well organized *datasets* are essential for exploiting the potential offered by the software platforms to perform the basic operations of learning, validation and testing
- Seismology is a data rich field: raw data (recorded waveforms) and databases with parametric measurements (phases, ground motion amplitudes, ...), and parameters obtained from analysis (location, magnitude(s), moment tensor, fault plane solutions, …)

Topics in seismology with ML & DL

- Event Detection, phase classification, picking of seismic phases (e.g., Ross et al., 2018; L. Zhu et al., 2019; Walter et al., 2020; Mousavi et al., 2020)
- Earthquake location, magnitude, fault mechanism (e.g. Perol et al., 2018; Trugman and Shearer, 2018; Kriegerowski et al., 2018; Zhang et al., 2020; Lomax et al., 2019; Mousavi and Beroza, 2020; Münchmeyer et al., 2021)
- Earthquake Early Warning (e.g., Li et al., 2018)
- Synthetic seismograms (e.g., Krischer & Fichtner, 2017)
- Inversion/tomography of seismic data for the Earth's interior (e.g., Bianco et al., 2019)
- Ground shaking estimation (e.g., Alavi, 2011; Derras et al., 2012, 2014; Jozinović et al., 2020; Münchmeyer et al., 2020)
- Analysis of massive seismic waveform data sets mining, clustering and dimensionality reduction
- Noise removal (e.g, Zhu et al., 2019)
- …

Qualified benchmark datasets

- Benchmark datasets and competitions are playing a crucial role in driving progress and innovation in ML research.

- High-quality benchmark datasets have two key benefits:
  - enabling rigorous performance comparisons and
  - producing better models.

- Competitions are common practice in ML practice to report performance of new algorithms on standard datasets
  - ~500 completed competitions on kaggle (kaggle.com)
  - In seismology, it was launched a ML competition for *laboratory earthquake forecasting* and Johnson et al. (2021, PNAS) and the *SeismOlympics* (Fang et al., 2017).

Examples of benchmark datasets

THE MNIST DATABASE

of handwritten digits

Yann LeCun, Courant Institute, NYU
Corinna Cortes, Google Labs, New York
Christopher J.C. Burges, Microsoft Research, Redmond

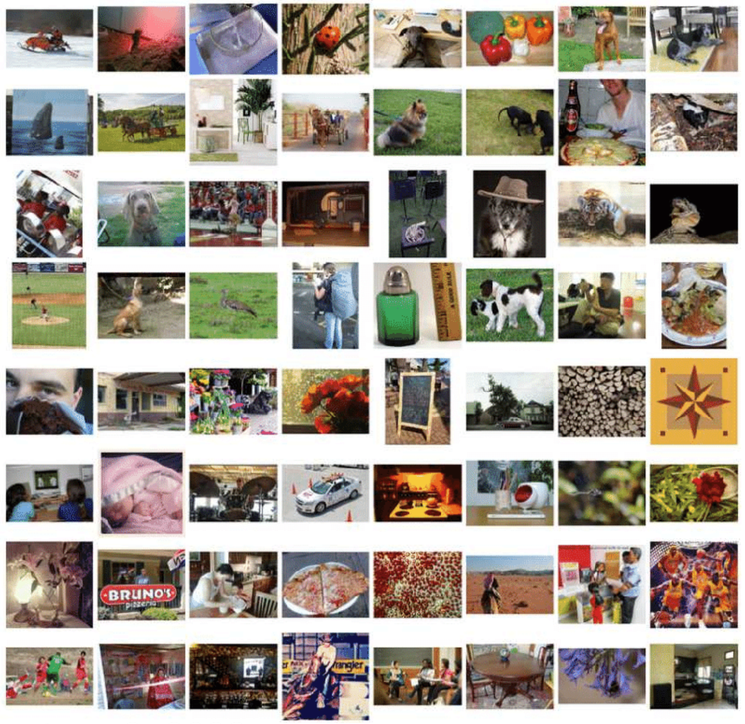
IMAGENET

14,197,122 images, 21841 synsets indexed

Home Download Challenges About

Not logged in. Login I Signup

ImageNet is an image database organized according to the WordNet hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images. The project has been instrumental in advancing computer vision and deep learning research. The data is available for free to researchers for non-commercial use.



Benchmark datasets in seismology

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2947848, IEEE Access

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

STanford EArthquake Dataset (STEAD): A Global Data Set of Seismic Signals for AI

S. MOSTAFA MOUSAVI[1], YIXIAO SHENG[1], WEIQIANG ZHU[1], and GREGORY C. BEROZA[1]

SCEDC Southern California Earthquake Data Center

Home | Cite the SCEDC | Recent Earthquakes | Earthquake Info | EQ Catalogs

Access Data

Special Data Sets

Training and Validation Data Sets for Deep Learning

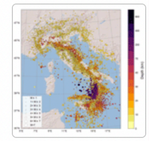
- P Wave Picking and First Motion Polarity
- Generalized Phase Detection
- Seismic Signal/Noise Discrimination

Earth Syst. Sci. Data, 13, 5509–5544, 2021
https://doi.org/10.5194/essd-13-5509-2021

Article | Assets | Peer review | Metrics | Related articles

Data description paper — 30 Nov 2021

INSTANCE – the Italian seismic dataset for machine learning



Alberto Michelini[1], Spina Cianetti[2], Sonja Gaviano[4,2], Carlo Giunchi[2], Dario Jozinović[1,3], and Valentino Lauciani[1]

[1]Istituto Nazionale di Geofisica e Vulcanologia, via di Vigna Murata, 605, 00143 Rome, Italy
[2]Istituto Nazionale di Geofisica e Vulcanologia, via Cesare Battisti, 53, Pisa, Italy
[3]Dipartimento di Scienze, Unversità degli Studi Roma Tre, Largo San Leonardo Murialdo 1, Rome, Italy
[4]Dipartimento di Scienze della Terra, Unversità degli Studi di Firenze, Via La Pira 4, Florence, Italy

Correspondence: Alberto Michelini (alberto.michelini@ingv.it)

Received: 11 May 2021 – Discussion started: 27 May 2021 – Revised: 08 Oct 2021 – Accepted: 17 Oct 2021 – Published: 30 Nov 2021

KeAi CHINESE ROOTS GLOBAL IMPACT

Contents lists available at ScienceDirect

Artificial Intelligence in Geosciences

journal homepage: www.keaipublishing.com/en/journals/artificial-intelligence-in-geosciences

Local earthquakes detection: A benchmark dataset of 3-component seismograms built on a global scale

Fabrizio Magrini [a,*], Dario Jozinović [a,b], Fabio Cammarano [a], Alberto Michelini [b], Lapo Boschi [c,d,e]

[a] Department of Science, Università Degli Studi Roma Tre, Italy
[b] Istituto Nazionale di Geofisica e Vulcanologia (INGV), Rome, Italy
[c] Dipartimento di Geoscienze, Università Degli Studi di Padova, Italy
[d] Sorbonne Université, CNRS, INSU, Institut des Sciences de La Terre de Paris, ISTeP UMR 7193, F-75005, Paris, France
[e] Istituto Nazionale di Geofisica e Vulcanologia, Bologna, Italy

USGS ScienceBase-Catalog | Communities | Help | Log in

ScienceBase Catalog → USGS Data Re… → Waveform Dat…

Waveform Data and Metadata used to National Earthquake Information Center Deep-Learning Models

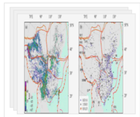
View

Dates

Publication Date : 2020-07-23

Citation: Ming Zhao, Zhuowei Xiao, Shi Chen and Lihua Fang. DiTing: A large-scale Chinese seismic benchmark dataset for artificial intelligence in seismology[J]. Earthquake Science

DiTing: A large-scale Chinese seismic benchmark dataset for artificial intelligence in seismology



Ming Zhao [1,2], Zhuowei Xiao [3], Shi Chen [1,2], Lihua Fang [1,4]

1. Institute of Geophysics, China Earthquake Administration, Beijing 100081, China
2. Beijing Baijiatuan Earth Sciences National Observation and Research Station, Beijing 100095, China
3. Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing 100029, China

Figures(7) / Tables(2)

OpenFWI: Benchmark Seismic Datasets for Machine Learning-Based Full Waveform Inversion

A Preprint

Chengyuan Deng
Los Alamos National Laboratory
charles.deng@lanl.gov

Yinan Feng
Los Alamos National Laboratory
ynf@lanl.gov

Shihang Feng
Los Alamos National Laboratory
shihang@lanl.gov

Peng Jin
Los Alamos National Laboratory
pjin@lanl.gov

Xitong Zhang
Los Alamos National Laboratory
xitongz@lanl.gov

Qili Zeng
Los Alamos National Laboratory
qzeng@lanl.gov

Youzuo Lin
Los Alamos National Laboratory
ylin@lanl.gov

February, 2022

KeAi CHINESE ROOTS GLOBAL IMPACT

Artificial Intelligence in Geosciences

Volume 1, December 2020, Pages 36-51

ShakeDaDO: A data collection combining earthquake building damage and ShakeMap parameters for Italy

Licia Faenza [a], Alberto Michelini [a], Helen Crowley [b], Barbara Borzi [b], Marta Faravelli [b]

Show more

+ Add to Mendeley | Share | Cite

Dataset compilation

In seismology and to the purpose of ML and DL, we can consider datasets consisting of either i.) *raw* or *instrument removed waveforms* (e.g., STEAD, INSTANCE, SCEDC,..) ; ii.) *synthetic waveforms* (e.g., OpenFWI) or iii.) big collections of *parametric data* (e.g., ShakeDado).

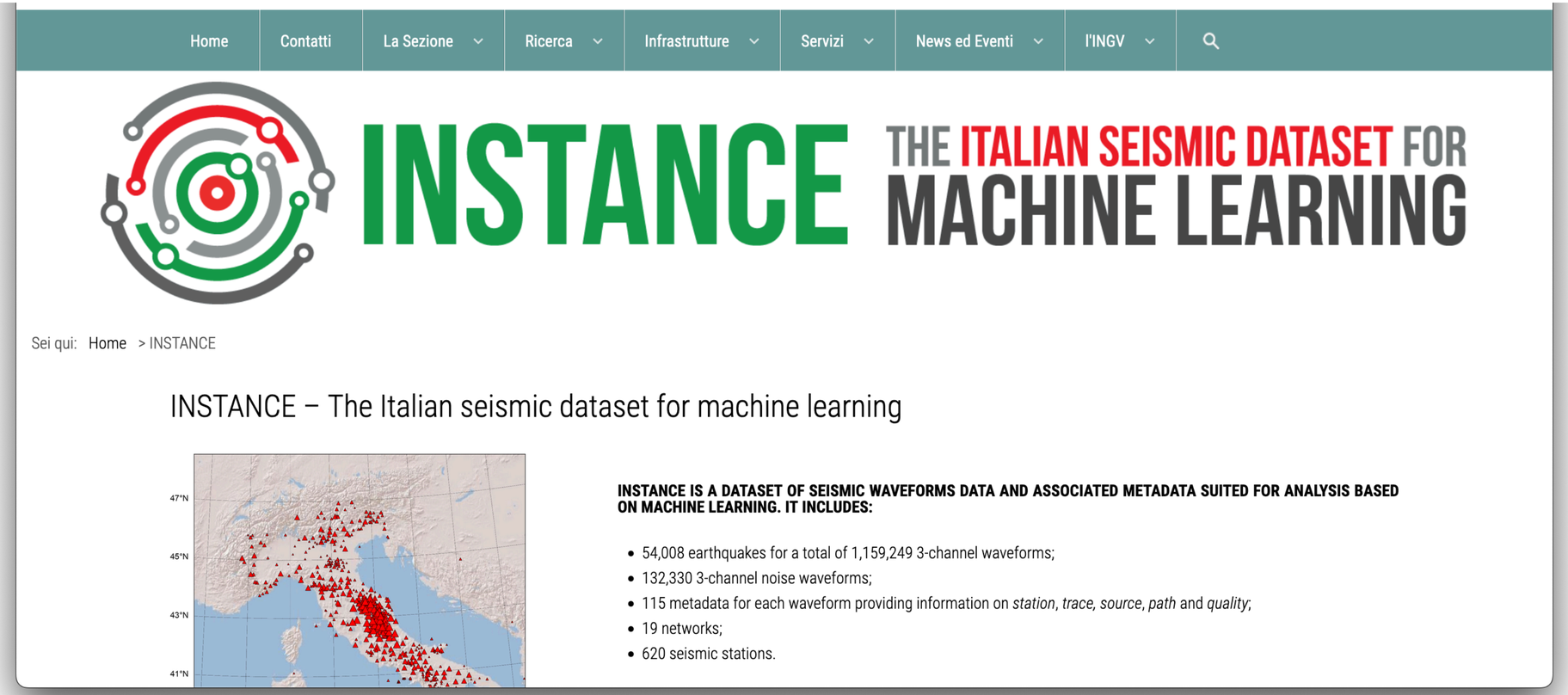
Metadata can serve as labels in supervised ML and for (sub)dataset selection.

Which data to include ? High-quality data only (i.e., faulty data are removed) *versus* dataset that include also “faulty” data but with a large number of trace diagnostic metadata (e.g., distributions of trace mean, median, quartiles,…).

Datasets consisting of $>10^6$ window traces at high sampling rate require weeks to download using standard web services and other technologies should be used to access to the data archives (e.g., Apache Spark)

Datasets should be formatted ready to be digested by ML/DL software platforms like Keras/TensorFlow, PyTorch, …

Dataset Compilation - INSTANCE



http://www.pi.ingv.it/instance/

Michelini et al. (2021). INSTANCE – the Italian seismic dataset for machine learning, Earth Syst. Sci. Data, 13, 5509–5544, https://doi.org/10.5194/essd-13-5509-2021, 2021.

Dataset Compilation - INSTANCE

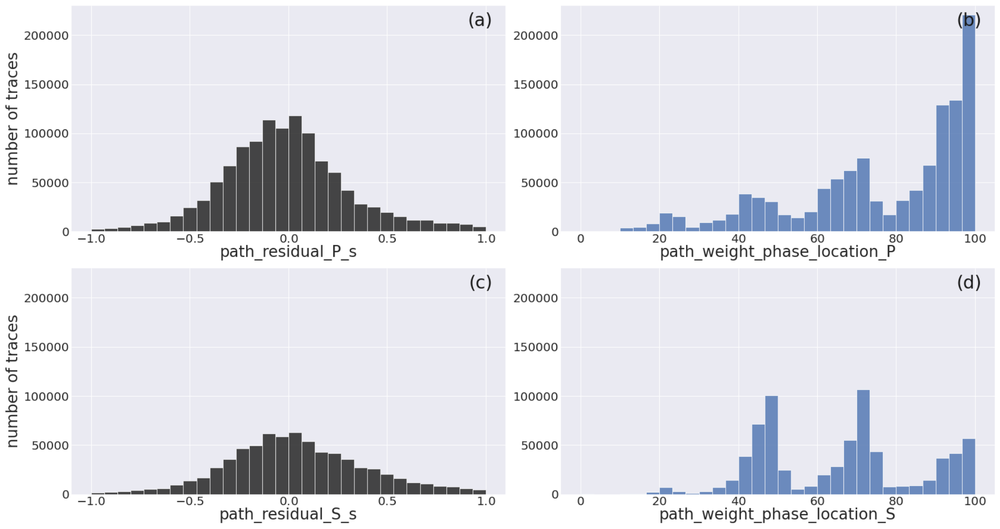
Data: 2005-2020 from EIDA INGV node

Event traces: 1,159,249 (90.0 %)
Noise traces: 132,288 (10.0%)

Total: 1,291,537

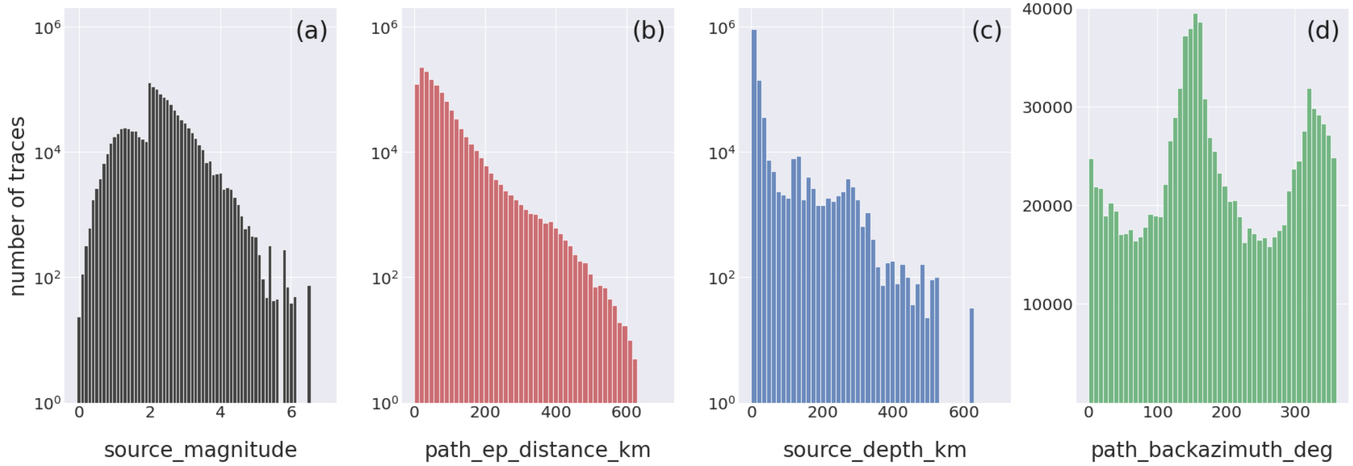
$\geq M_{min}$	$< M_{max}$	All	Selected	Percent kept	Nb. 3C records
0	1	57746	4462	7.73	39794
1	2	209652	15249	7.27	202572
2	3	43109	30845	71.55	757129
3	4	4342	3106	71.53	139338
4	5	342	315	92.11	18659
5	6	31	28	90.32	1593
6	7	3	3	100.0	164
0	7	315225	54008	17.13	1159249

P- and S-wave phase selection (residuals & weight)



P

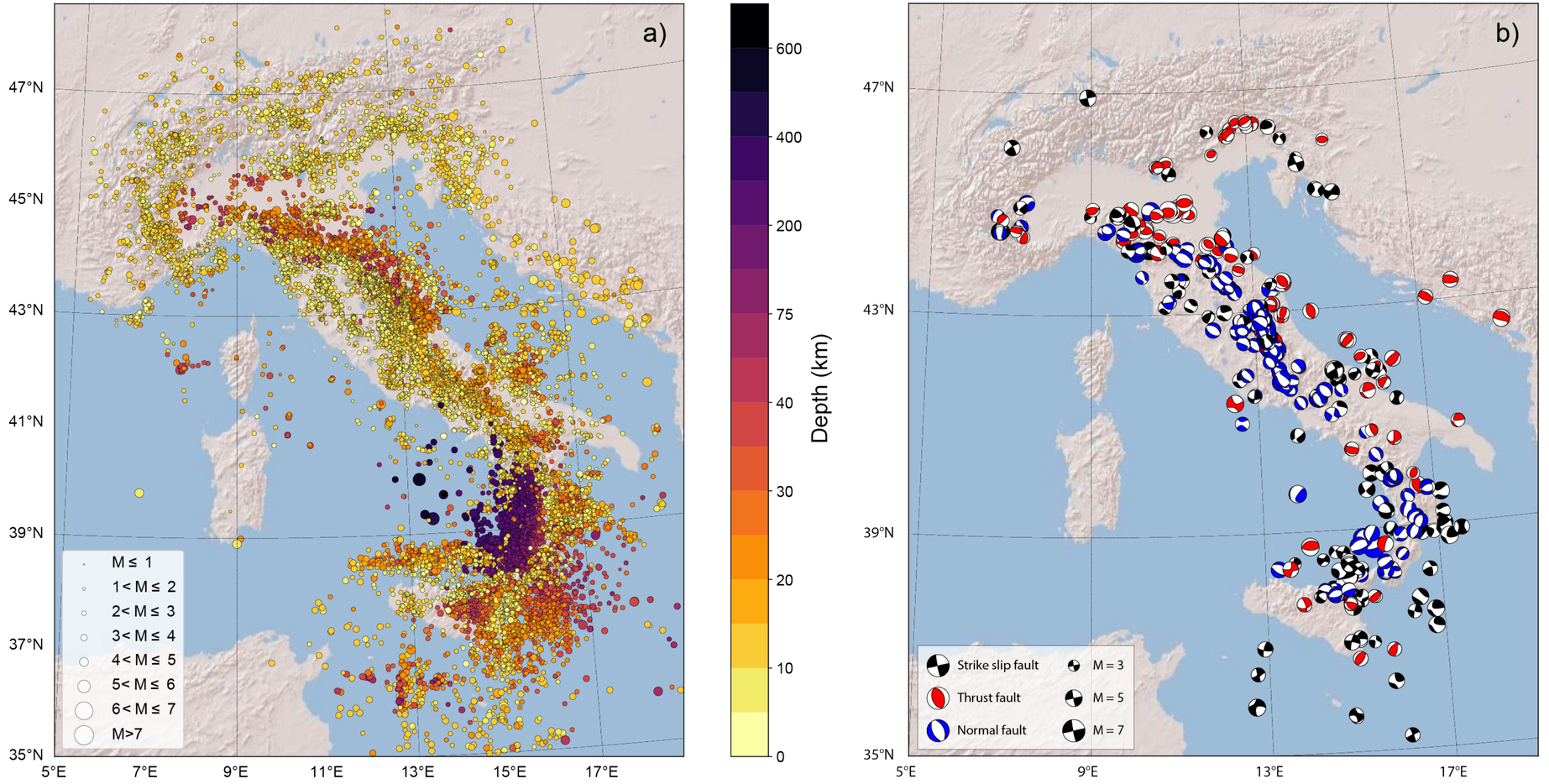
S



Dataset Compilation - INSTANCE

54,008 earthquakes -> 1,159,249 3C event traces

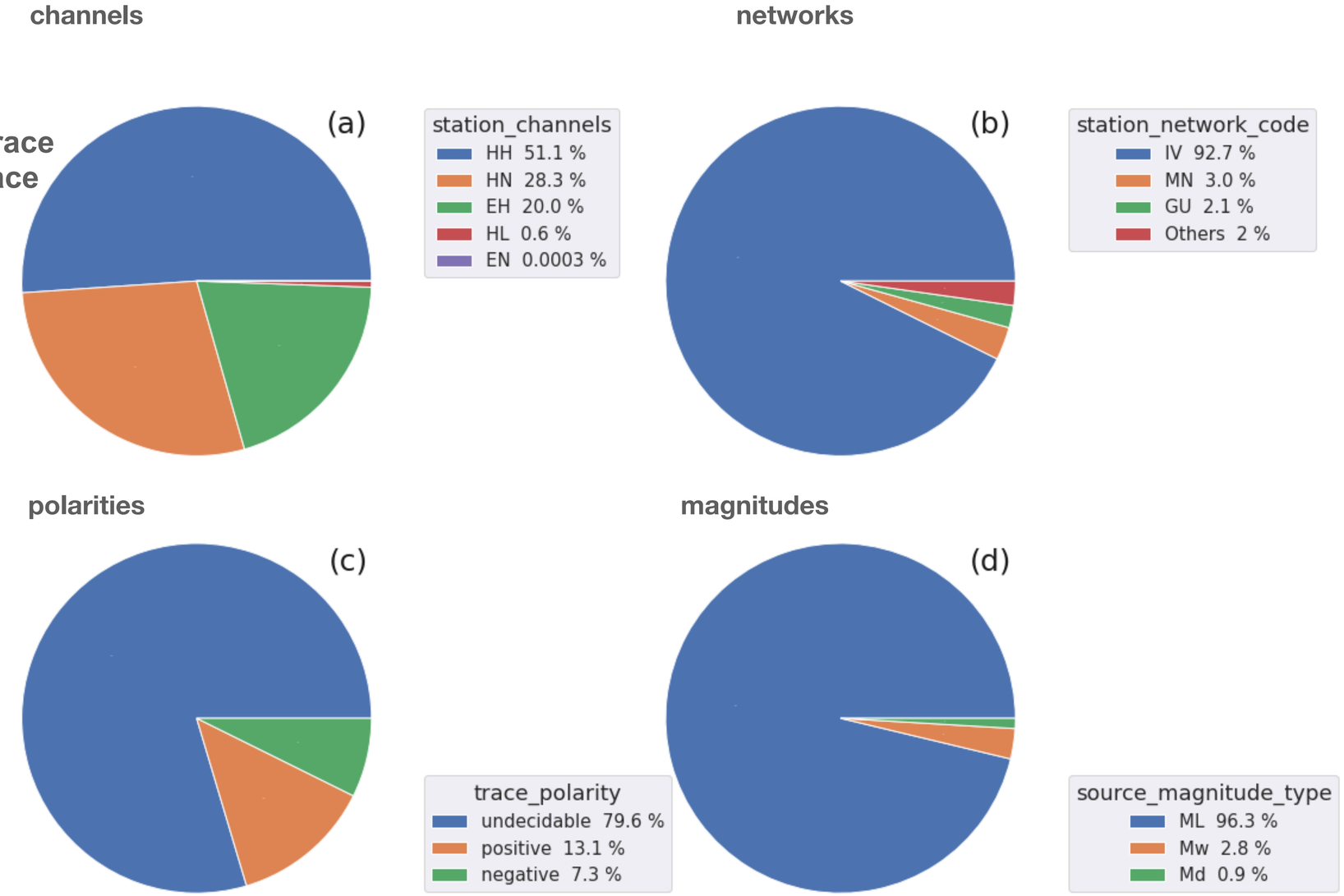
132,288 noise traces



Metadata

115 metadata associated to each event trace
46 metadata associated to each noise trace

4 types of metadata:
- **Source, station, trace, path (event)**
- **Station, trace (noise)**



Metadata

Metadata parameter name	Noise	Description
source_id	•	Earthquake and noise ID (INGV and UTC time, respectively)
source_origin_time		Location preferred origin time (YYYY-MM-DDTHH:MM:SS.SSZ)
source_latitude_deg		Location preferred latitude (°)
source_longitude_deg		Location preferred longitude (°)
source_depth_km		Location preferred depth (km)
source_origin_uncertainty_s		Location preferred origin time uncertainty (s)
source_latitude_uncertainty_deg		Location preferred latitude uncertainty (°)
source_longitude_uncertainty_deg		Location preferred longitude uncertainty (°)
source_depth_uncertainty_km		Location preferred depth uncertainty (km)
source_stderror_s		Preferred earthquake location standard deviation (s)
source_gap_deg		Location preferred location gap (°)
source_horizontal_uncertainty_km		Location preferred horizontal uncertainty (km)
source_magnitude		Preferred magnitude
source_magnitude_type		Preferred magnitude type
source_mt_eval_mode		Moment tensor evaluation mode (e.g., manual)
source_mt_status		Status of the evaluation ("reviewed" or "final")
source_mt_scalar_moment_Nm		Scalar moment (N m)
source_mechanism_strike_dip_rake		Strike, dip, rake of the two planes (two tuples)
source_mechanism_moment_tensor		Six components of the moment tensor (m_rr, m_tt, m_pp, m_rt, m_rp, m_tp)
source_type		Earthquake or other sources (quarry_blast, controlled explosion, experimental explosion, etc.)
station_network_code	•	Two characters FDSN network code (e.g., IV)
station_code	•	Station name (International Registry of Seismograph Stations, IR)
station_location_code	•	Location name identifier (Buland, 2006)
station_channels	•	Two characters identifying the sampling and the instrument gain (HN, HH, EH, etc.)
station_latitude_deg	•	Station latitude (°)
station_longitude_deg	•	Station longitude (°)
station_elevation_m	•	Station elevation (m)
station_vs30_mps	•	$V_{S,30}$ (m s^{-1})
station_vs30_detail	•	$V_{S,30}$ information
path_ep_distance_km		Epicentral distance
path_hyp_distance_km		Hypocentral distance
path_azimuth_deg		Direction from event location to station (°)
path_backazimuth_deg		Direction from station location to event epicenter (°)
path_residual_[P,S]_s		P- or S-arrival time residual between picked arrival time and traveltime using preferred location (s)
path_weight_phase_location_[P,S]		P- or S-phase location weight resulting from preferred location (range 0–100)
path_travel_time_[P,S]_s		P- or S-wave traveltime (s)

Metadata parameter name	Noise	Description
trace_name	•	Waveform name within the HDF5 file
trace_start_time	•	Waveform trace UTC start time (YYYY-MM-DDTHH:MM:SS.SSZ)
trace_dt_s	•	Sampling interval (s)
trace_npts	•	Number of samples in waveform trace (integer)
trace_[P,S]_uncertainty_s		Assigned P- or S-onset arrival time uncertainty (s)
trace_eval_[P,S]		P- or S-type of picking (currently only "manual")
trace_[P,S]_arrival_time		P- or S-arrival UTC start time (YYYY-MM-DDTHH:MM:SS.SSZ)
trace_polarity		P onset polarity ("negative", "positive", "undecidable")
trace_[P,S]_arrival_sample		P- and S-onset sample number on waveform trace (integer)
trace_[E,N,Z]_median_counts	•	*E*-, *N*-, or Z-component sample median (counts, integer)
trace_[E,N,Z]_mean_counts	•	*E*-, *N*-, or Z-component sample mean (counts, integer)
trace_[E,N,Z]_min_counts	•	*E*-, *N*-, or Z-component sample minimum (counts, integer)
trace_[E,N,Z]_max_counts	•	*E*-, *N*-, or Z-component sample maximum (counts, integer)
trace_[E,N,Z]_rms_counts	•	*E*-, *N*-, or Z-component sample root mean squared
trace_[E,N,Z]_lower_quartile_counts	•	*E*-, *N*-, or Z-component sample lower quartile (counts, integer)
trace_[E,N,Z]_upper_quartile_counts	•	*E*-, *N*-, or Z-component sample upper quartile (counts, integer)
trace_[E,N,Z]_snr_db		*E*-, *N*-, or Z-component signal-to-noise ratio
trace_[E,N,Z]_spikes	•	*E*-, *N*-, or Z-component number of spikes (integer)
trace_GPD_[P,S]_number	•	P and S number of picks retrieved with GPD
trace_EQT_[P,S]_number	•	P and S number of picks retrieved with EQT
trace_EQT_number_detections	•	Number of detections retrieved with EQT
trace_[E,N,Z]_pga_cmps2		*E*-, *N*-, or Z-component PGA (cm s^{-2})
trace_[E,N,Z]_pgv_cmps		*E*-, *N*-, or Z-component PGV (cm s^{-1})
trace_[E,N,Z]_pga_perc		*E*-, *N*-, or Z-component PGA (% g)
trace_[E,N,Z]_pga_time		*E*-, *N*-, or Z-component PGA UTC time (YYYY-MM-DDTHH:MM:SS.SSZ)
trace_[E,N,Z]_pgv_time		*E*-, *N*-, or Z-component PGV UTC time (YYYY-MM-DDTHH:MM:SS.SSZ)
trace_[E,N,Z]_sa03_cmps2		*E*-, *N*-, or Z-component spectral acceleration at $t = 0.3$ (cm s^{-2})
trace_[E,N,Z]_sa10_cmps2		*E*-, *N*-, or Z-component spectral acceleration at $t = 1.0$ (cm s^{-2})
trace_[E,N,Z]_sa30_cmps2		*E*-, *N*-, or Z-component spectral acceleration at $t = 3.0$ (cm s^{-2})
trace_pga_cmps2		Max. horizontal components PGA value (cm s^{-2})
trace_pgv_cmps		Max. horizontal components PGV value (cm s^{-2})
trace_pga_perc		Max. horizontal components PGA value (% g)
trace_sa03_cmps2		Max. horizontal components spectral acceleration ($t = 0.3$) (cm s^{-2})
trace_sa10_cmps2		Max. horizontal components spectral acceleration ($t = 1.0$) (cm s^{-2})
trace_sa30_cmps2		Max. horizontal components spectral acceleration ($t = 3.0$) (cm s^{-2})
trace_deconvolved_units		Ground motion units of the traces in the HDF5 volume (e.g., mps and mps2 for m s^{-1} and m s^{-2}, respectively)

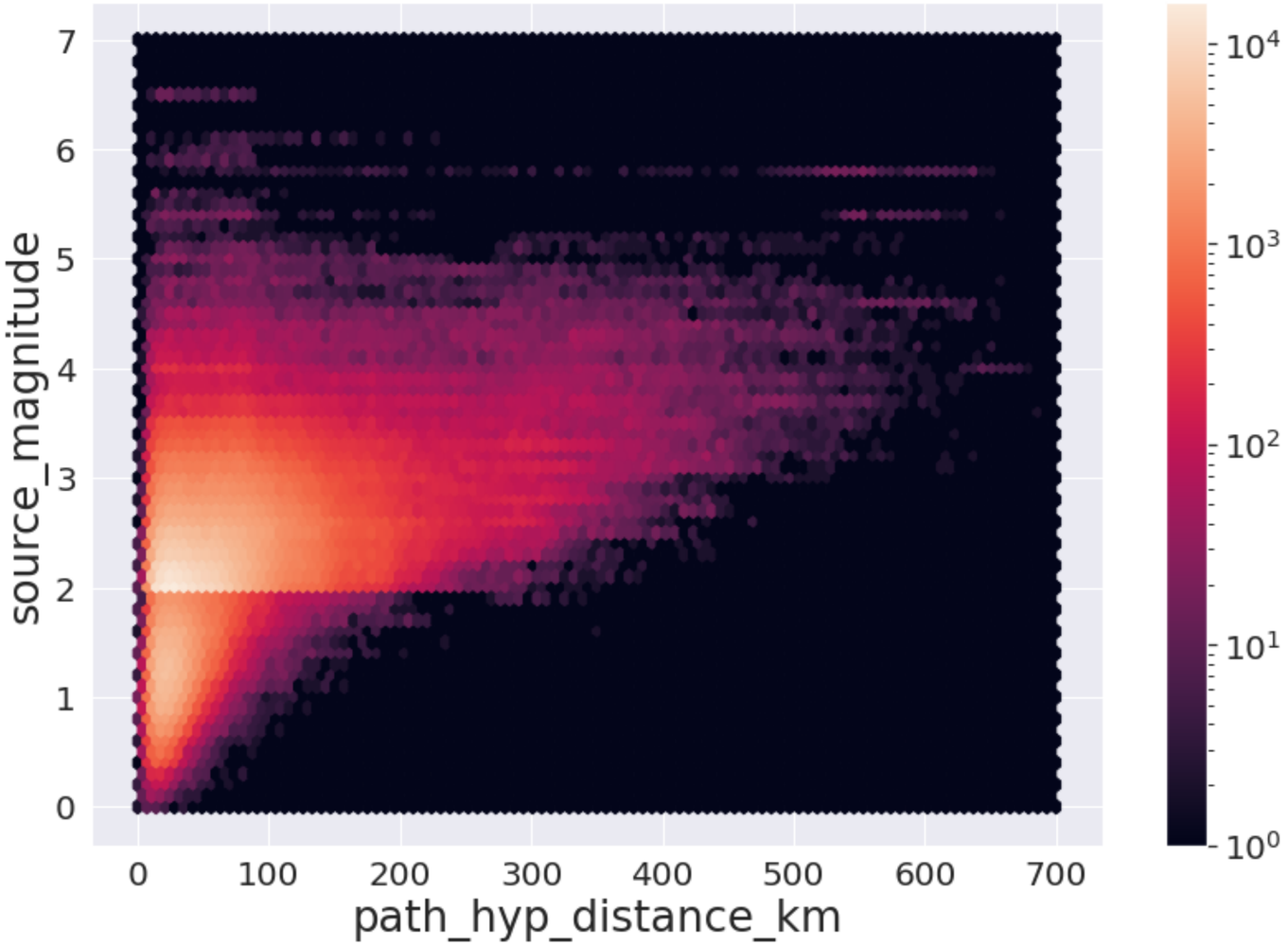
source

station

path

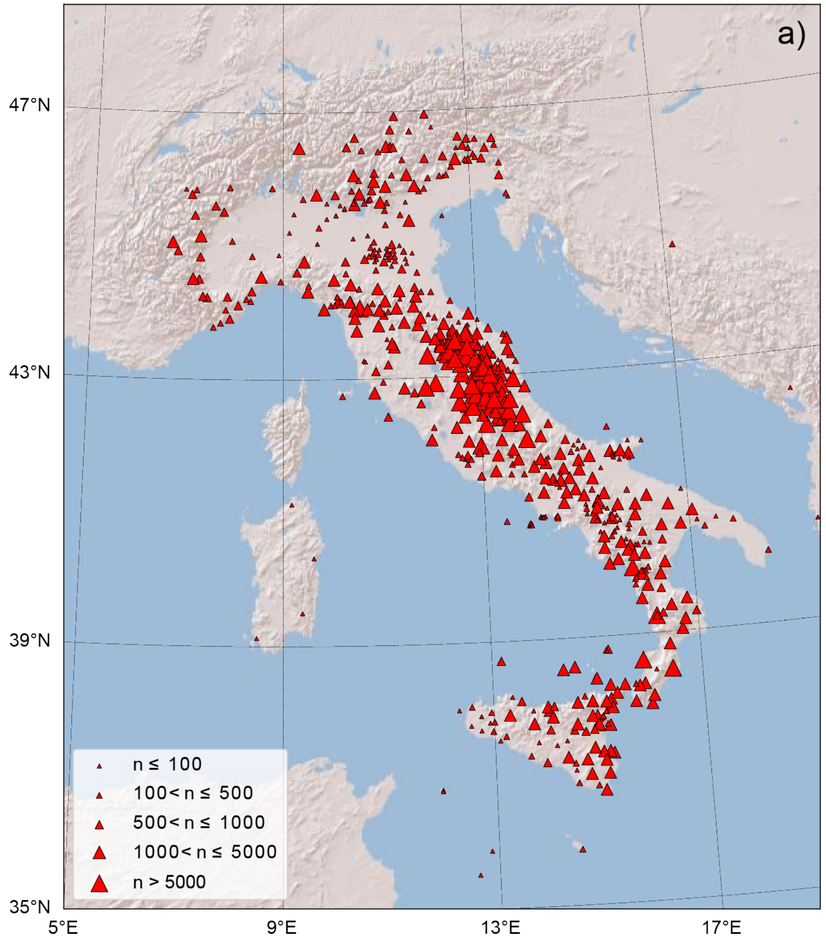
trace

Magnitude versus distance

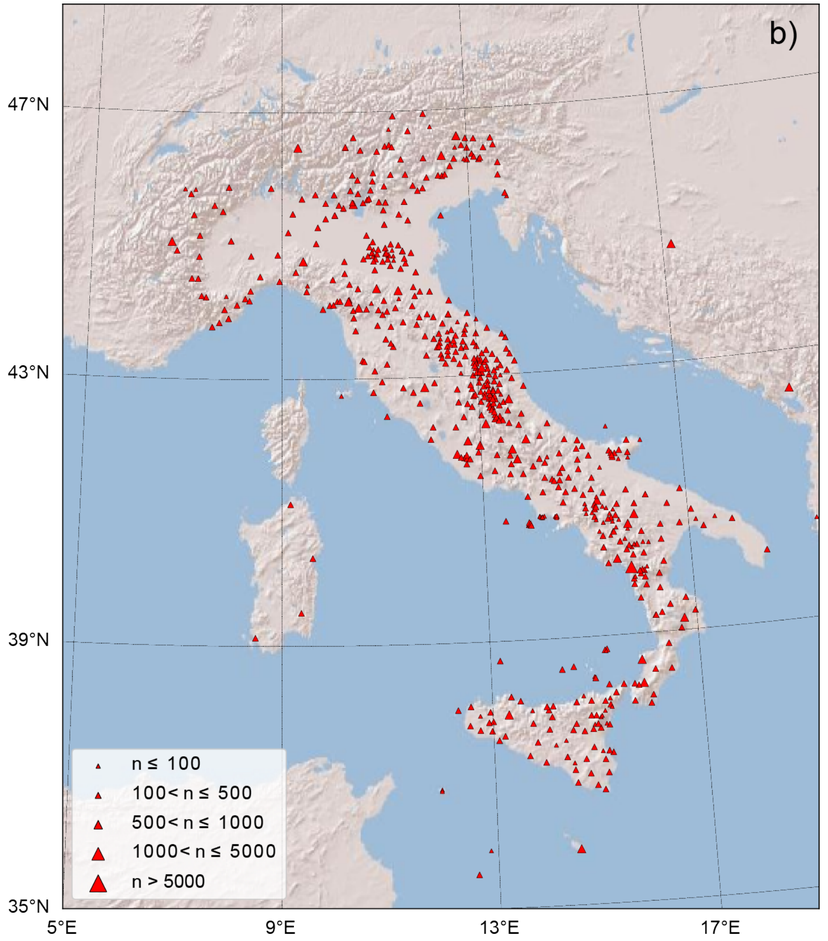


620 stations

Earthquakes

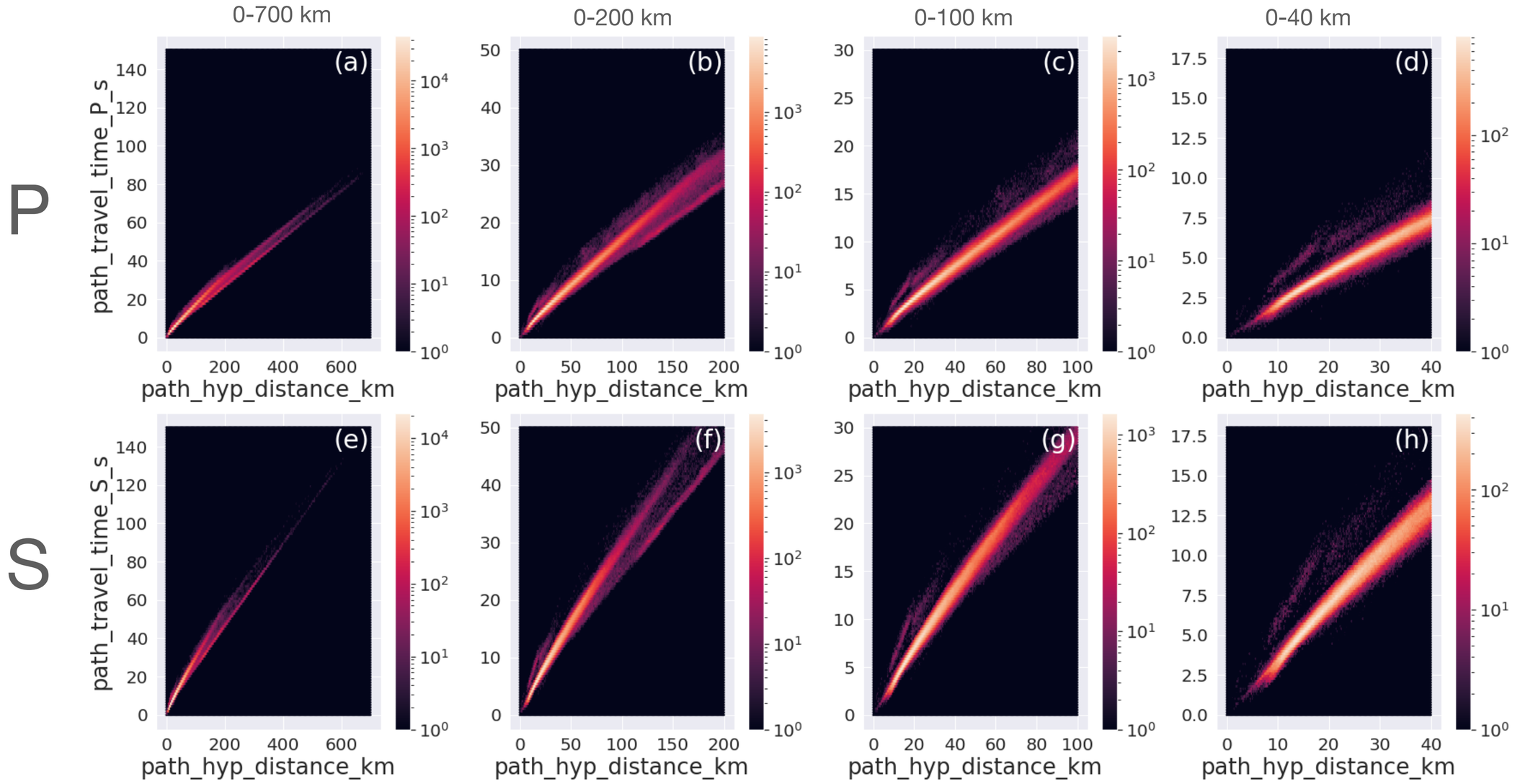


Noise



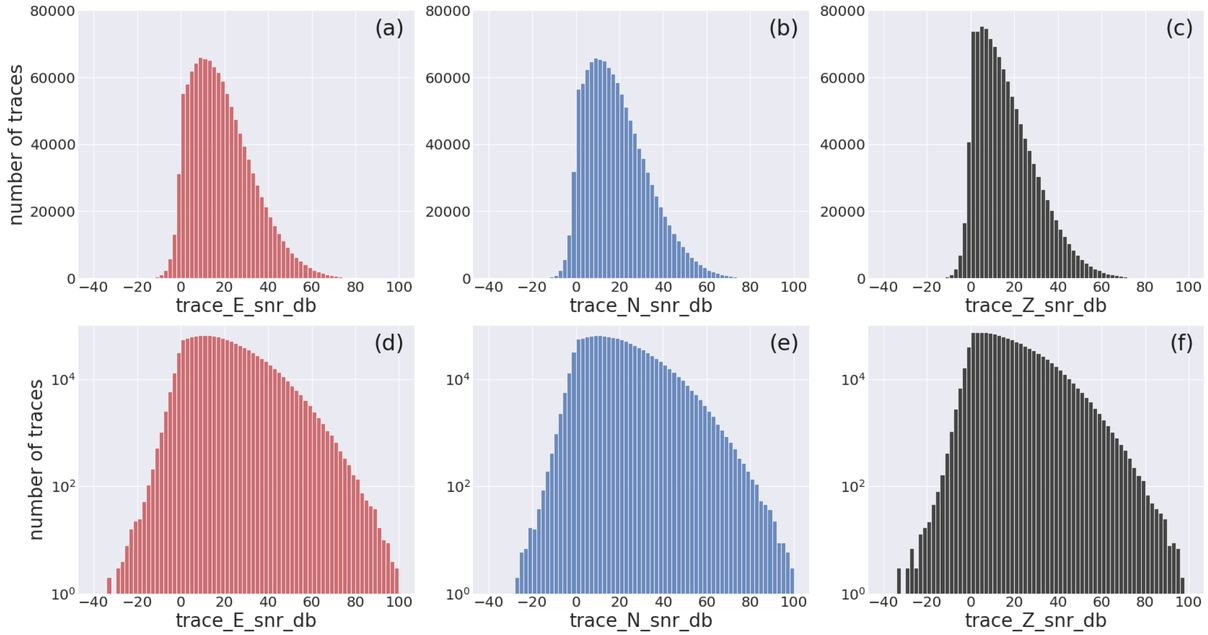
Symbol size proportional to no of. recordings

Path metadata: traveltimes

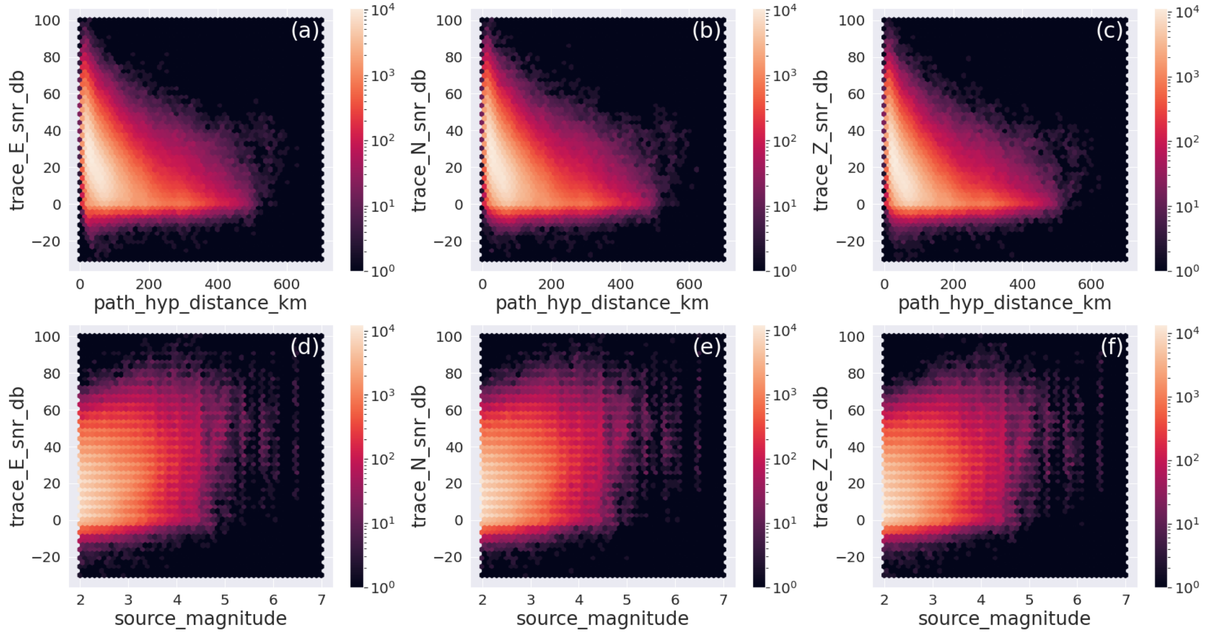


trace metadata: Signal to Noise Ratio (SNR)

SNR (top: linear; bottom: log10)



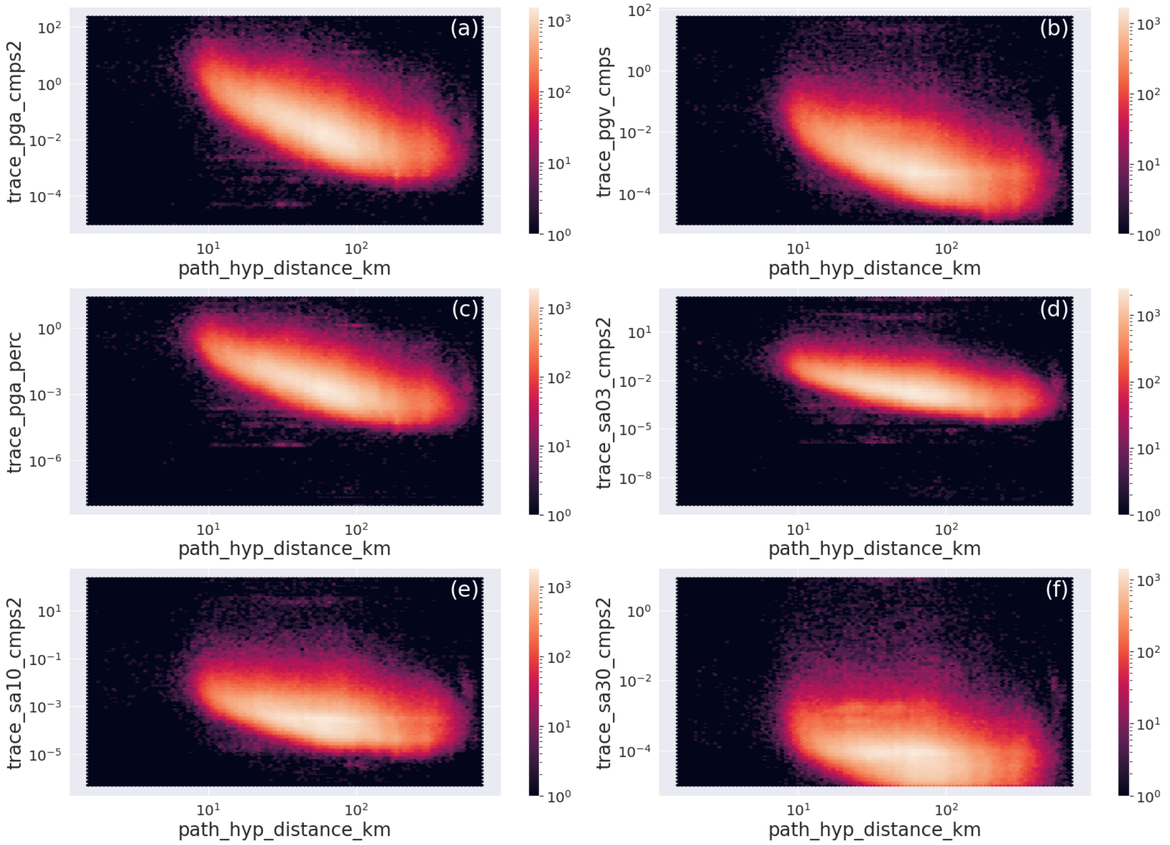
SNR vs distance (top) linear; SNR vs mag (bottom)



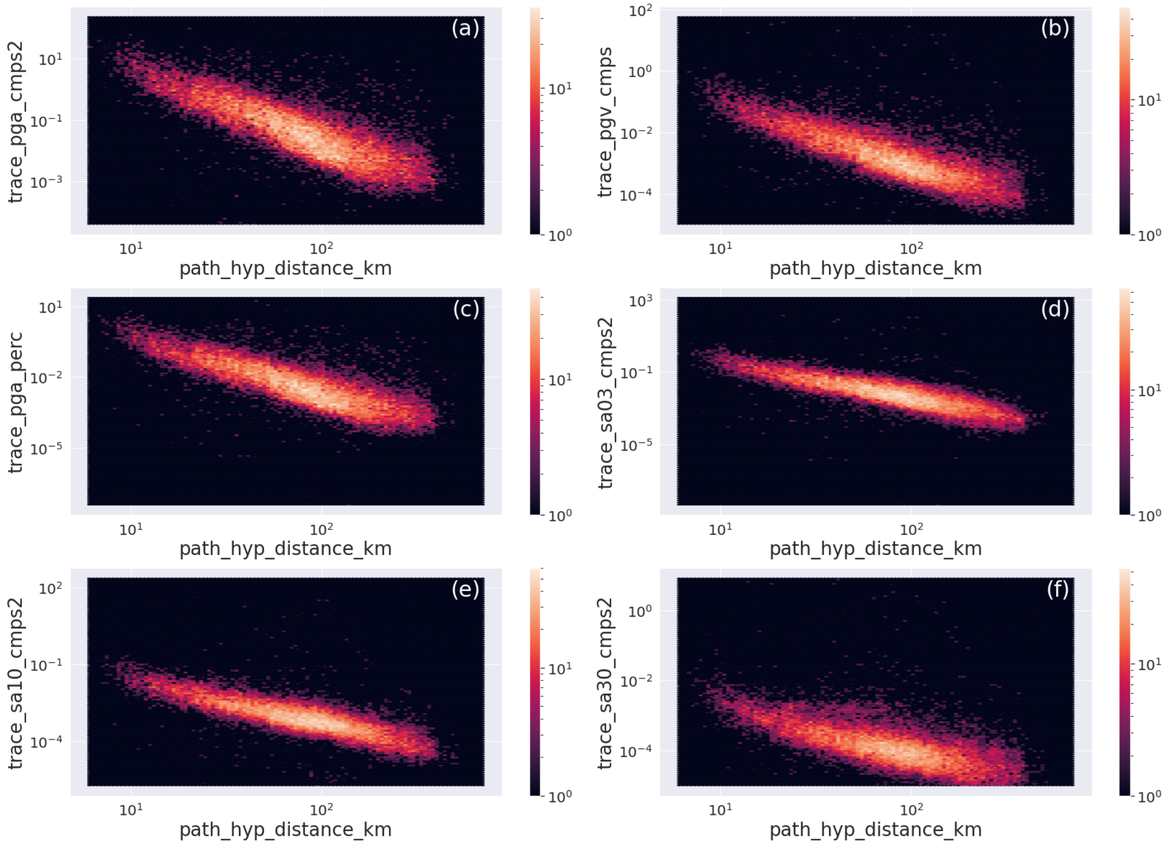
trace metadata: Intensity Measurements

PGA, PGV, PGA %g, SA [t=(0.3,1.0,3.0 s)] vs distance

IMs M ≥ 2



IMs M=3



trace metadata: Quality Control

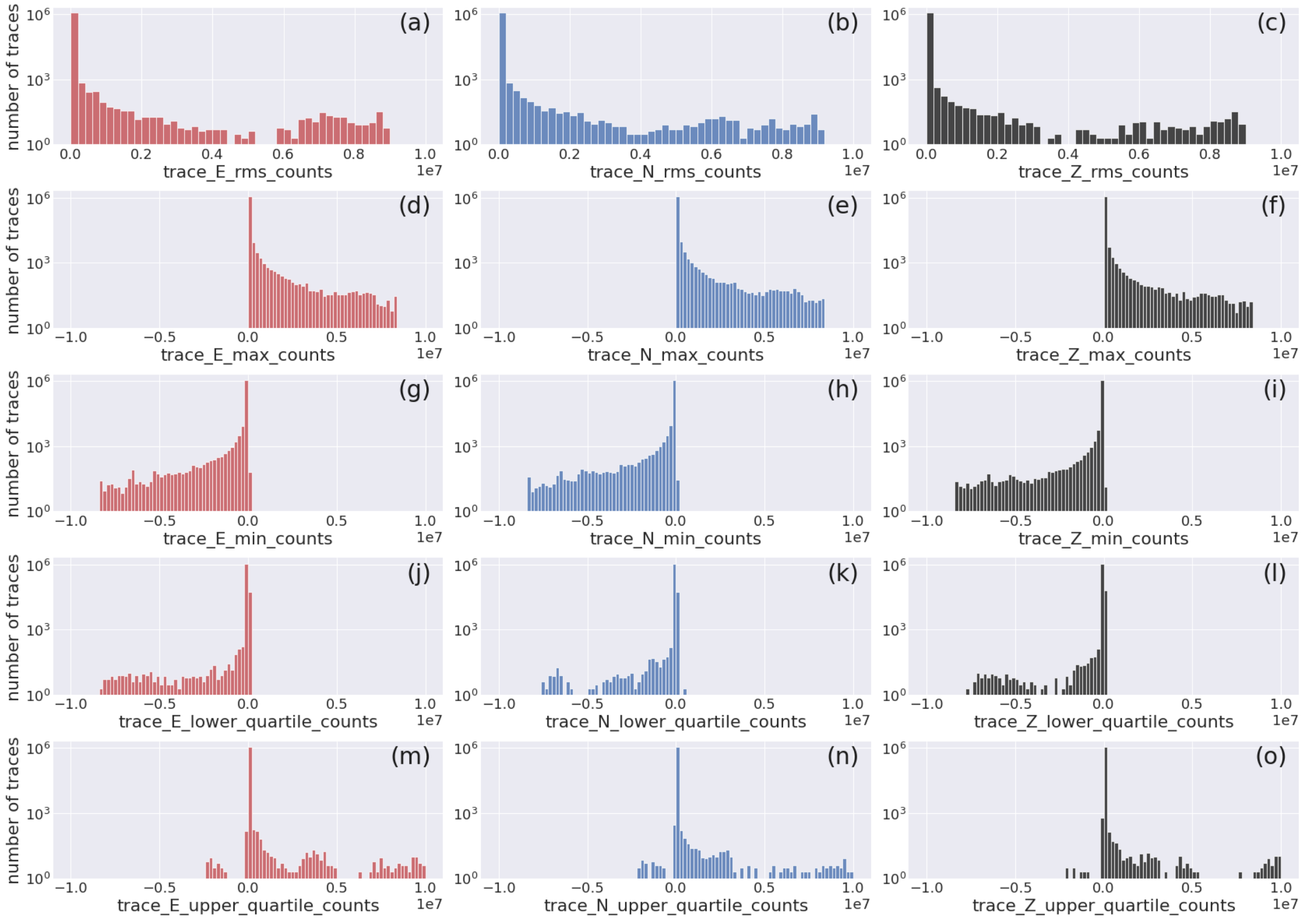
RMS

max

min

<25 %

>75 %



QC distributions (highlighted EQT and GPD)

Event

Table (3). Distribution according to different quantiles of selected metadata (cf. Table 2) for the HH channels of the event dataset.

Metadata parameter-name	min	10 %	25 %	50 %	75 %	90 %	max
trace_E_median_counts	-6.57e+06	-22	-6	0	6	22	3.03e+06
trace_N_median_counts	-6.51e+05	-22	-6	0	6	22.5	2.5e+06
trace_Z_median_counts	-7.63e+05	-11.5	-3	0	3	12	9.92e+05
trace_E_snr_db	-25.5	2.31	7.29	15	25	35.4	95.4
trace_N_snr_db	-26.9	2.3	7.27	15.1	25.1	35.5	95.8
trace_Z_snr_db	-23.3	1.21	5.51	12.5	22.2	32.5	95.4
trace_EQT_number_det.	0	1	1	1	1	1	7
trace_GPD_P_number	0	0	1	1	1	2	13
trace_GPD_S_number	0	0	1	1	2	3	22

Noise

Table (5). Distribution according to different quantiles of selected noise metadata (cf. Table 2) for the HH and EH channels.

Metadata parameter-name	10 %	25 %	50 %	75 %	90 %	max
trace_E_rms_counts (HH)	52.79	101.6	205	447.9	1013	1.919e+07
trace_N_rms_counts (HH)	53.47	102	207.3	465.8	1071	1.902e+07
trace_Z_rms_counts (HH)	44.68	85.42	166.3	364	793.1	9.986e+05
trace_EQT_number_det. (HH)	0	0	0	0	0	5
trace_GPD_P_number (HH)	0	0	0	0	1	31
trace_GPD_S_number(HH)	0	0	0	1	2	24
trace_E_rms_counts (EH)	7.53	22.92	58.29	141.8	327.1	7.54e+05
trace_N_rms_counts (EH)	7.864	22.88	57.65	140.9	332.6	2.913e+05
trace_Z_rms_counts (EH)	5.639	18.44	50.09	119.8	307.1	6.236e+05
trace_EQT_number_det. (EH)	0	0	0	0	0	5
trace_GPD_P_number (EH)	0	0	0	1	2	23
trace_GPD_S_number (EH)	0	0	0	2	4	26

Event Waveforms

Selection using different criteria based on the metadata

earthquakes 2≤M<3 (~67 % of the HH channels)

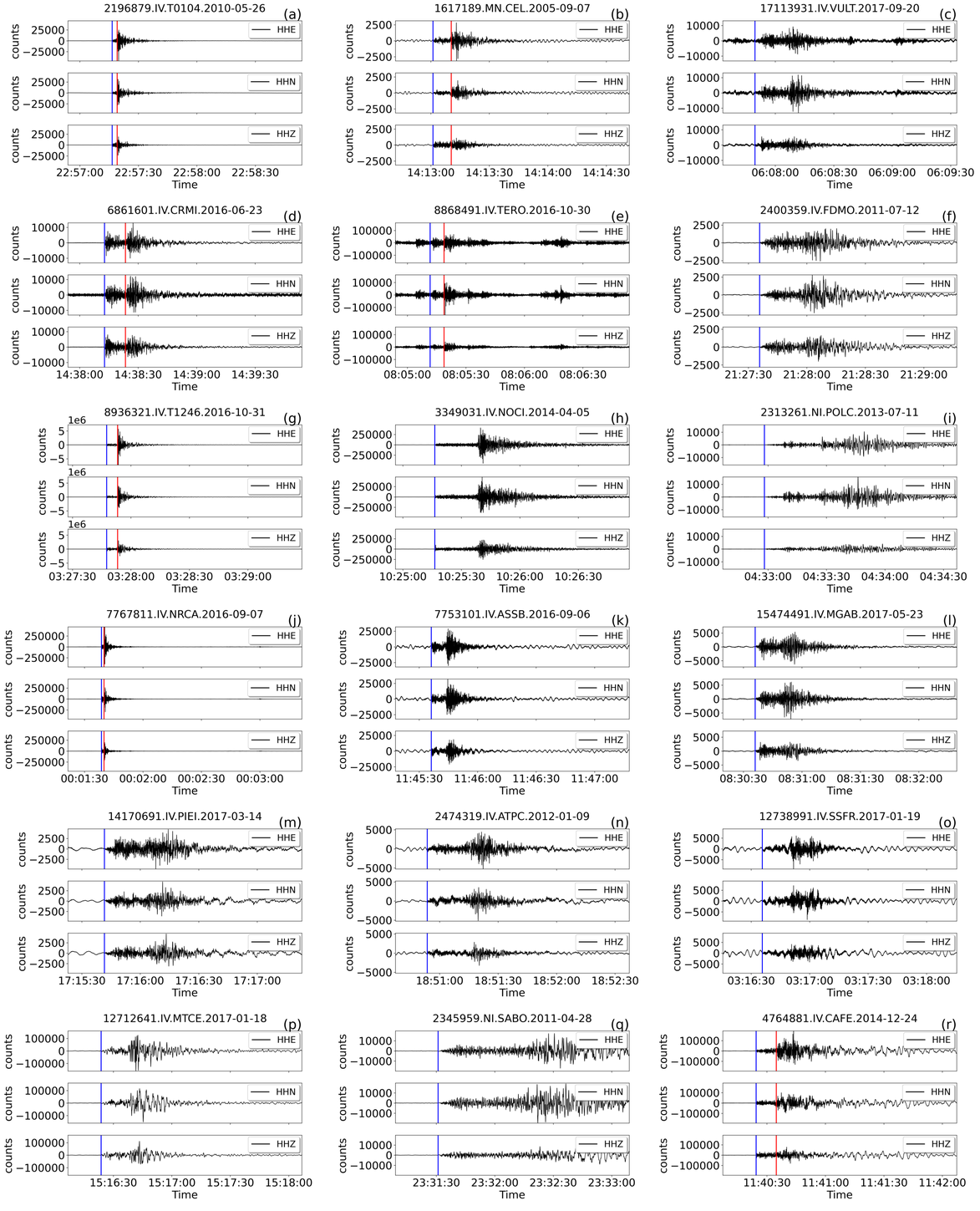
earthquakes 3≤M<4 (14 %);

earthquakes M≥4 (2 %)

earthquakes `trace_E_snr_db`≥10 and `path_ep_distance`<100 km (55 %)

earthquakes `trace_E_snr_db`≥10 and `path_ep_distance`≥100 km (11 %)

earthquakes M≥4 and `trace_E_snr_db`≥10 (2 %)



Comparison between available seismological ML datasets

	INSTANCE[1]	STEAD[2]	SCEDC[3]	LEN-DB[4]	CNQ_INGV[5]	NEIC[6]
Metadata (events)	115	35	–	14	6	5
Metadata (noise)	46	8	–	7	2	–
Trace length (s)	120	60	4,6	27	50	60
Units[7]	D, P	D	D	P	P	D
Events	54 008	~ 450 000	273 882	304 874	6213	136 716
Traces (events)	1 159 249	1 050 000	–	629 095	22 046	–
Traces (noise)	132 288	~ 100 000	–	615 847	12 543	–
Receivers	620	2613	–	1487	26	2361
Average receivers per event	21	2	–	2	4	–
Duration in hours (events)	38 641	~ 17 500	–	4718	306	–
Duration in hours (noise)	4409	~ 1700	–	4618	174	–
Epicentral distance range (km)	< 620	< 350	< 360	< 189	< 19 310	< 10 000
Magnitude range	0–6.5	0–7.9	−0.81–7.3	0.4–7.1	3–9.1	1–8.3
Sampling rate (Hz)	100	100	100	20	20	40
Storage size (GB)	331.2	91.4	–	18.4	0.9	~ 51
Focal mechanism	527	6200	–	–	–	–
Event type[8]	L, R	L	L, G	L	L, R, G	L, R, G
Data type[9]	BB, SM, SP	BB, SM, SP	BB, SM	–	BB	BB, SP?

[1] INSTANCE, https://doi.org/10.13127/instance. [2] STEAD, https://doi.org/10.1109/ACCESS.2019.2947848. [3] SCEDC, https://scedc.caltech.edu/data/deeplearning.html (last access: 19 November 2021). [4] LEN-DB, https://doi.org/10.5281/zenodo.3648232. [5] ConvNetQuake_INGV (CNQ_INGV), https://doi.org/10.5281/zenodo.5040865. [6] NEIC, https://doi.org/10.5066/P9OHF4WL. [7] D: digital; P: physical. [8] L: local; R: regional; G: global. [9] BB: broadband; SM: strong motion; SP: short period.

Benchmarking platform

- Accessing various benchmark datasets for training and implementing the standardization of models is a time-consuming process, hindering further advancement of ML techniques within seismology.

- The overall goal is to facilitate the analysis to data users through a standard analysis framework to allow:
  - *the analysis of different benchmark datasets using the same DL model*
  - *the use of different DL models on the same benchmark dataset,*
  - the inclusion of pre-processing tasks (e.g., *data augmentation)*

- SeisBench is a software package that tackles these issues.

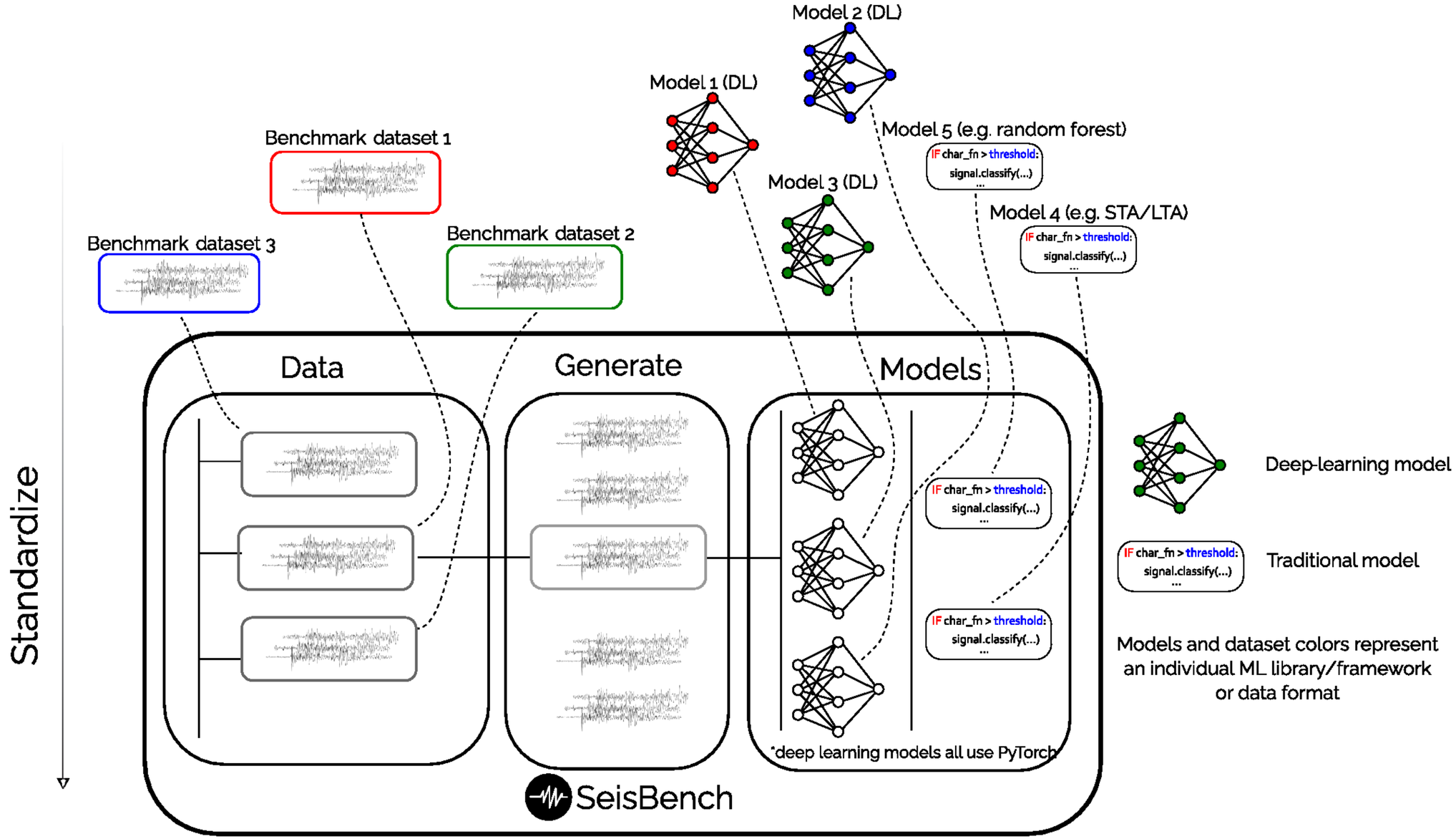
- SeisBench is an open-source framework for deploying ML in seismology—available via GitHub.

- SeisBench standardizes access to both models and datasets, while also providing a range of common processing and data augmentation operations through the API.

Woollam, J., Münchmeyer, J. et al. (2022). SeisBench—A Toolbox for Machine Learning in Seismology. *Seismological Research Letters*; 93 (3): 1695–1709. doi: https://doi.org/10.1785/0220210324

Münchmeyer, J., Woollam, J. et al. (2022). Which picker fits my data? A quantitative evaluation of deep learning based seismic pickers. *Journal of Geophysical Research: Solid Earth*, 127, e2021JB023499. https://doi.org/10.1029/2021JB023499

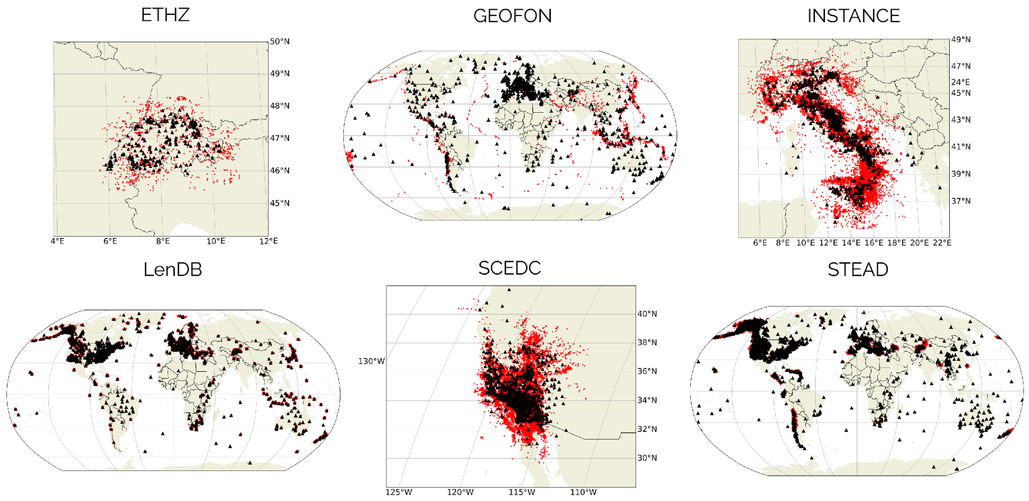
SeisBench as a unifying framework for developing models and applying them to seismic data

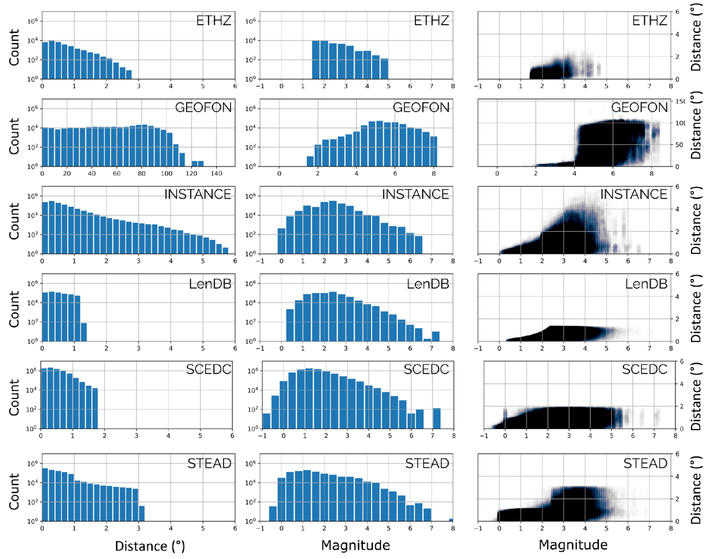


From Woollam et al. (2022)

Data, Generate, and Model tags highlight the different modules available within SeisBench

SeisBench example

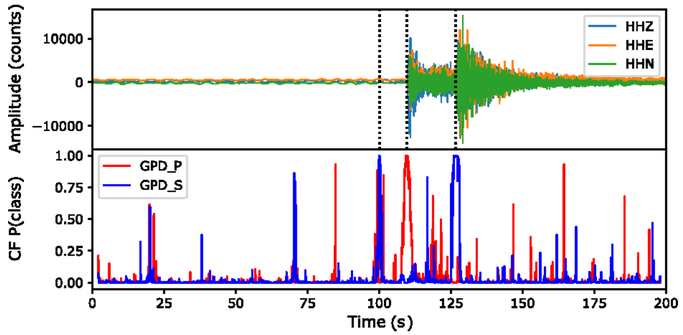




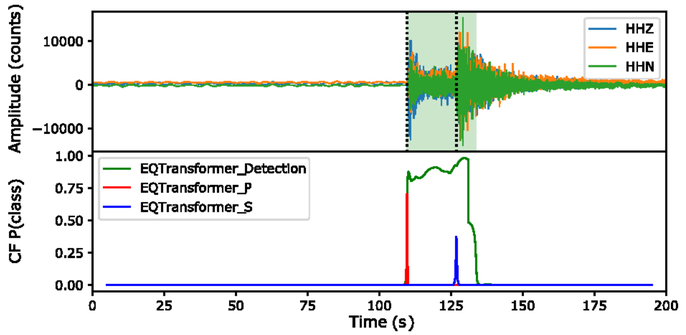
Example code blocks, which download a seismic waveform [1], then loads a pretrained deep-learning picking model and applies the model to predict on the seismic stream using either one of two ML architectures (GPD and EQTransformer) [2]. Resulting picks and characteristic functions from the output probabilities are displayed beneath the code blocks.

[1]:
```
from obspy.clients.fdsn import Client
from obspy import UTCDateTime
import matplotlib.pyplot as plt


# Get a seismic stream
client = Client("GFZ")

t = UTCDateTime("2007/01/02 05:48:50")
st = client.get_waveforms(network="CX", station="PB01", location="*",␣
↪channel="HH?", starttime=t-100, endtime=t+100)
```

[2]:
```
from seisbench.models import GPD

# Load pre-trained weights into model
gpd_model = GPD.from_pretrained("original")

# Apply GPD to stream
gpd_annotations = gpd_model.annotate(st, stride=5)
gpd_picks = gpd_model.classify(st, P_threshold=0.95, S_threshold=0.95)
```

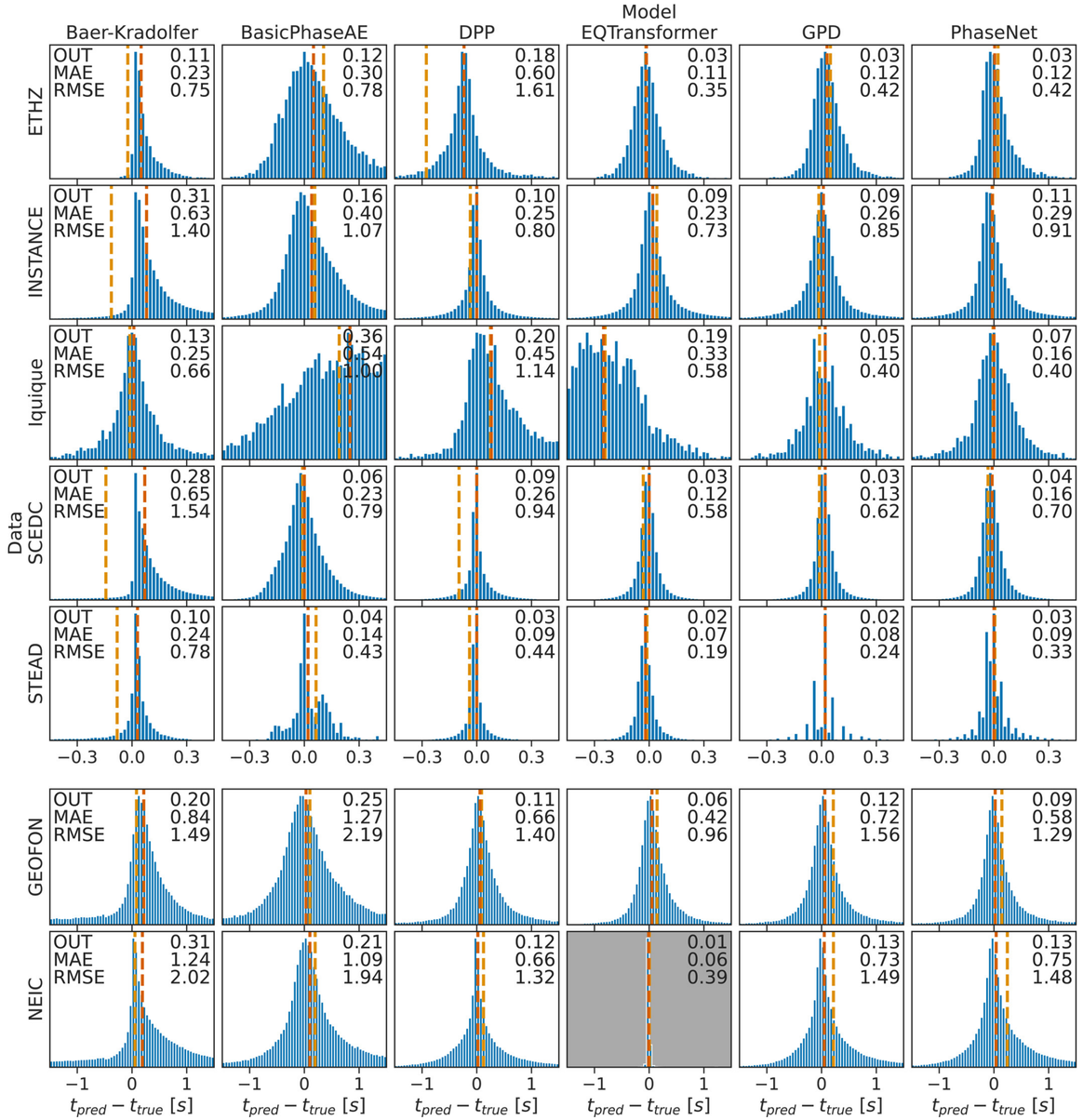


[2]:
```
from seisbench.models import EQTransformer

# Load pre-trained weights into model
eqt_model = EQTransformer.from_pretrained("original")

# Apply EQT to stream
eqt_annotations = eqt_model.annotate(st)
eqt_picks = eqt_model.classify(st)
```



From Woollam et al. (2022)

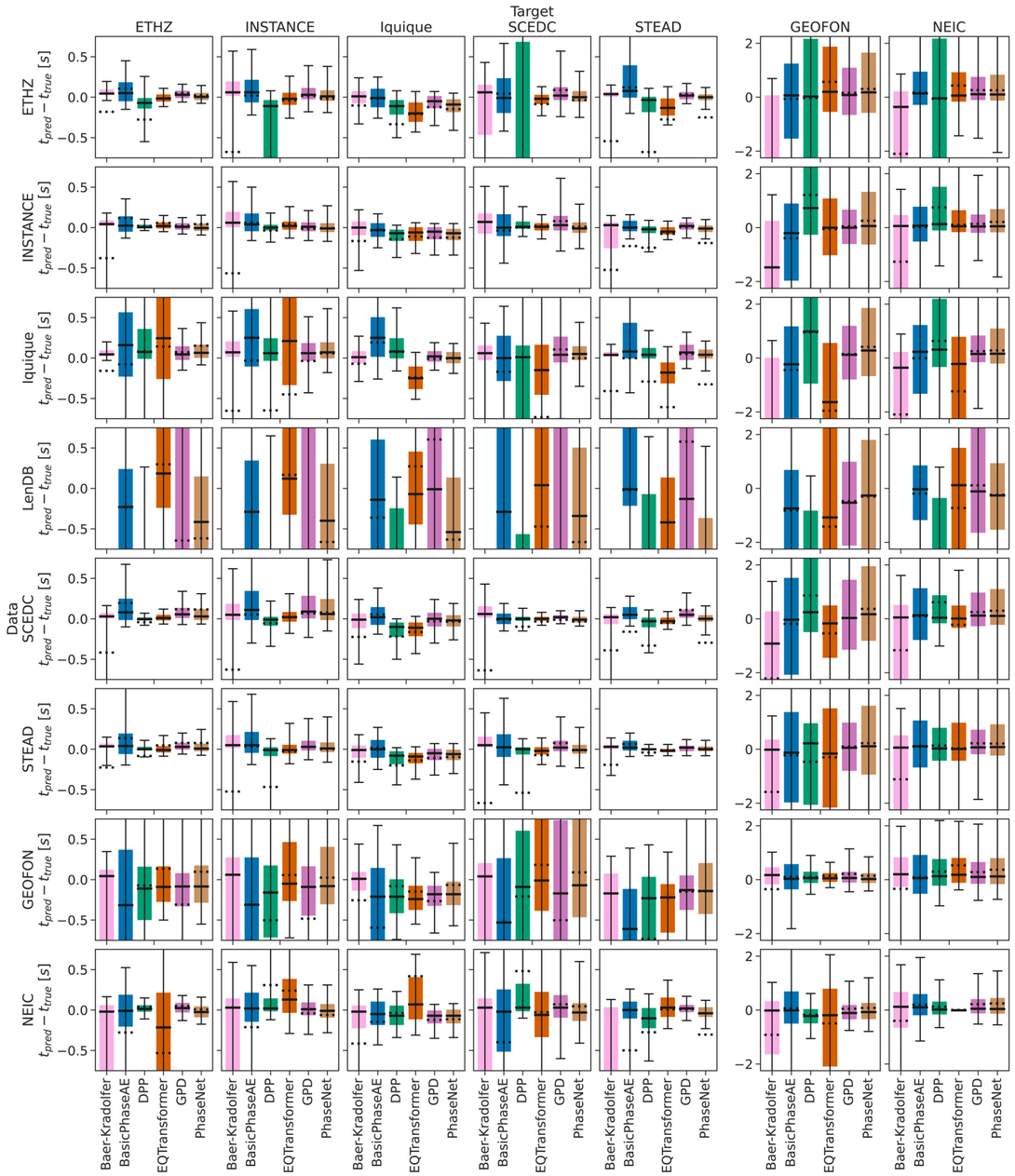
Example of benchmarking using SeisBench: P-picking



Histogram of P residuals from in-domain experiments. Vertical dashed lines show median (red) and mean (orange) of the residuals.

From Münchmeyer et al. (2022).

Example of benchmarking using SeisBench - Distribution of P pick residuals



Distribution of P pick residuals from cross-domain experiments. Each panel shows one combination of training (row) and evaluation (column) data set, each bar one model.

From Münchmeyer et al. (2022).

Conclusions & Outlook

- Benchmark *datasets* appear crucial for developing and testing ML/DL models
- Benchmark *platforms* are also highly desirable to facilitate the testing of *different models on the same dataset* or, viceversa, for testing the *same model on different datasets*
- Adoption of the same formats for data (e.g., HDF5) and the metadata (e.g., grouping according to *source, station, trace* and *path,* nomenclature standardization)
- Provision of a large number of metadata makes the dataset usable for many different analysis (e.g., earthquake detection, location, size estimation, denoising, ground motion estimations, …)
- SeisBench is open source available on GitHub and it has the potential to become the reference platform for performing ML in seismology

Acknowledgments

This work would not have been possible without the effort and dedication of the people that install and maintain the stations of the networks used here, and the skilled IT people that are in charge of archiving, curating and providing access to the data and the earthquake analysts that routinely perform the data analysis for the compilation of the earthquake bulletins.

The *INGV Friday Coffee* group and *SeisBench* group (Carlo Giunchi, Spina Cianetti, Sonja Gaviano, Dario Jozinović, Valentino Lauciani, Matteo Bagagli, Chris Zerafa, Anthony Lomax, Licia Faenza, Jannes Münchmeyer, Jack Woollam, Frederik Tilmann, Andreas Rietbrok, Dietrich Lange, Thomas Bornstein, Tobias Diehl, Florian Haslinger, Joachim Saul, Hugo Soto)

This work has been partially supported by the project INGV Pianeta Dinamico 2021 Tema 8 SOME (CUP D53J1900017001) funded by Italian Ministry of University and Research “Fondo finalizzato al rilancio degli investimenti delle amministrazioni centrali dello Stato e allo sviluppo del Paese, legge 15 145/2018” and by the European Union’s Horizon 2020 research and innovation program under Grant Agreement Number 821115, real-time earthquake risk reduction for a resilient Europe (RISE).

NumPy ObsPy A Python Framework for Seismology pandas matplotlib