https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

Open Access Earth System Science Data Discussions

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]Unversità degli Studi Roma Tre, Largo San Leonardo Murialdo 1, Rome, Italy
[4]Unversità degli Studi di Firenze, Via La Pira 4, Firenze, Italy

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

Abstract. The Italian earthquake waveform data are here collected in a dataset suited for machine learning analysis (ML) applications. The dataset consists of near 1.2 million three-component (3C) waveform traces from about 50,000 earthquakes and more than 130,000 noise 3C waveform traces, for a total of about 43,000 hours of data and an average of 21 3C traces are provided per event. The earthquake list is based on the Italian seismic bulletin (http://terremoti.ingv.it/bsi) of the "Istituto Nazionale di Geofisica e Vulcanologia" between January 2005 and January 2020 and it includes events in the magnitude range between 0.0 and 6.5. The waveform data have been recorded primarily by the Italian National Seismic Network (network code IV) and include both weak (HH, EH channels) and strong motion recordings (HN channels). All the waveform traces have a length of 120 s, are sampled at 100 Hz, and are provided both in counts and ground motion units after deconvolution of the instrument transfer functions. The waveform dataset is accompanied by metadata consisting of more than 100 parameters providing comprehensive information on the earthquake source, the recording stations, the trace features, and other derived quantities. This rich set of metadata allows the users to target the data selection for their own purposes. Many of these metadata can be used as labels in ML analysis or for other studies. The dataset, assembled in HDF5 format, is available at http://doi.org/10.13127/instance (Michelini et al., 2021).

1 Introduction

Important breakthroughs in the understanding of earthquake phenomena can be achieved through the analysis of the very large number of continuous waveform recordings stored in the existing seismic archives. To this end it can be important to dispose of well organized representative subsets of the archives together with their associated metadata information.

The recent developments of machine learning (ML) software platforms like https://www.tensorflow.org, https://pytorch.org, https://keras.io, https://caffe.berkeleyvision.org (Abadi et al., 2016; Paszke et al., 2019; Chollet and others, 2015; Jia et al., 2014, respectively), the availability of high performance computing hardware (i.e., GPUs) and the access to thoroughly selected

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

Open Access Earth System Science Data Discussions

benchmark datasets (e.g., STEAD https://github.com/smousavi05/STEAD, LEN-DB https://doi.org/10.5281/zenodo.3648231) offer new opportunities to apply ML methodologies to seismological and earthquake engineering problems. In particular, the use of sophisticated and optimized ML algorithms for the analysis of large amounts of seismic data can lead to remarkable improvements for automated tasks like seismic waveform onset picking, ground motion prediction, earthquake early warning or for the detection of hidden signals currently recognized as noise or for novel modeling and inversion strategies (see Kong et al., 2018; Bergen et al., 2019; Dramsch, 2020, for recent reviews). Specifically, the advent of ML into seismology has shown the importance of reference datasets for benchmarking the developed methodologies and it has fostered more thorough and statistically sound schemes for analyzing the data, like splitting all the available data into training, validation and test sets. Moreover, the introduction of competitions like those for predicting laboratory earthquakes launched on the Kaggle platform (https://www.kaggle.com/c/LANL-Earthquake-Prediction/data) or the SeismOlympics (Fang et al., 2017), that attracted several thousand teams, evidences even more the great potential of benchmark datasets (Johnson et al., 2021) and the general interest to tackle seismology problems with ML.

The application of ML techniques to seismological waveform data can be quite straightforward. Indeed, large amounts of labelled data are already available thanks to the analyses carried out since many decades by expert analysts that have compiled and reviewed earthquake catalogs (that include phase onset readings, earthquake location and size estimates) or that have assembled ground motion parameters in special flat files and maps of strong ground motion amongst the most common tasks. Their work provides effectively metadata that can be associated with the recorded waveforms and that can be used as labels when performing ML analysis. A main bottleneck in wide-scale implementation of ML is, however, the fast access to the waveforms and to the associated metadata. Open access waveforms archives available to the seismological community (e.g., EIDA, Strollo et al. (2021) http://eida.ingv.it/, or IRIS, Ingate (2008)) were mainly designed for preserving the continuous data and making them available to the scientific community. In practice, one of the main goals of seismological data centers has been the seamless acquisition of continuous data from the networks and the preservation, curation and archiving of the entire record of continuous waveforms. In this context, the users have complete flexibility in the selection of the data to download but accessing large data volumes can be very time consuming. Thus, despite the achievements attained in the last decades with the implementation of well tested and efficient web services (e.g., FDSN `dataselect`), the accessibility of remote servers still remains cumbersome (Quinteros et al., 2021).

Overcoming this difficulty and attracting a broader audience of developers can be achieved by providing publicly available benchmark datasets that can be readily used on the existing software platforms (Mousavi et al., 2019). Therefore, the matter consists of organizing the large amount of quality checked seismic data and metadata according to the format ready to be used in ML applications.

Recently, effort has been made to assemble and make publicly available datasets consisting of waveforms and associated metadata. In detail, the dataset used in the works by Ross et al. (2018a, b); Meier et al. (2019) is downloadable from the Southern California Earthquake Data Center at the web portal https://scedc.caltech.edu/data/deeplearning.html. This dataset includes 4.8 million time series recorded by nearly 700 receivers from more than 270,000 earthquakes in southern California. The STEAD dataset assembled by Mousavi et al. (2019) includes 1.2 million of 3C traces comprising 450,000 local earthquakes

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

Open Access Earth System Science Data Discussions

and 100,000 noise windows recorded by more than 2600 stations at global scale. The LEN-DB dataset (Magrini et al., 2020) is also a global dataset of local earthquakes and includes 1.2 million 3C waveform traces, half belonging to earthquakes and half to noise. The NEIC dataset (Yeck and Patton, 2020) includes global data and has been used by Yeck et al. (2020) to train the 1.3 million seismic-phase arrivals using three separate convolutional neural network models to predict arrival time onset, phase type, and distance.

Results attained by Ross et al. (2018b), Zhu et al. (2019a), Zhu et al. (2019b), Mousavi et al. (2020), and Mousavi and Beroza (2020) are excellent examples of successful applications of ML which can improve substantially the earthquake detection level with respect to most traditional methods, leading to the location of tiny and previously undetected earthquakes improving our knowledge on the heterogeneity of stress release on known and unknown faults. This enhanced information is crucial to make more thorough assessments of the ongoing seismotectonics and seismic hazard. The ML methods are likely to become an irreplaceable tool in seismology to extract as much information as possible from the large amount of data already stored in the archives. Among the indirect advantages, the enhanced detection can, to some extent, also govern network densification with sensible reductions in equipment investments and maintenance costs.

In general, the good performances of ML applications are strongly related to the availability of large amounts of data with associated properly labeled metadata. Large amounts of data are critical to perform proper training and avoid data overfitting. However, the preparation of a ML dataset is also tedious and very time consuming. These are the main reasons that motivated the work presented in this article. Our goal is to provide an open access dataset consisting of raw and instrument removed waveform data and associated metadata to study earthquake occurrence in Italy. The data collection, named INSTANCE, gathers seismic waveform data from weak and strong motion stations that have been extracted from the Italian EIDA node (Danecek et al., 2021). The metadata associated to the waveforms are extracted from the INGV earthquake catalogue and from the waveform traces themselves. We expect this reference dataset to be used for several different purposes spanning from improvements of the existing configurations of seismic monitoring in Italy to the development and testing of new techniques for earthquake detection and ground motion estimation.

2 Earthquakes

2.1 Data preparation

The data collection was assembled following the main stages listed below:

1. earthquake selection;
2. station selection;
3. waveform data selection and download;
4. cross-validation between phase-based station selection and downloaded waveform data;
5. processing of the data counts waveforms;

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

6. application of the instrument transfer function to the waveforms.

2.1.1 Earthquake selection

To compile the waveform dataset, we started from the INGV earthquake bulletin and seismic stations archives (http://terremoti.ingv.it/iside). These data are public and can be queried using the `fdsnws-event` (https://www.fdsn.org/webservices/fdsnws-event-1.2.pdf) and the `fdsnws-station` web services provided by INGV. The first step consisted of retrieving all the earthquakes with $M \geq 0$ from 1 January 2005 to 31 January 2020 in an enlarged area within the latitude and longitude corners (35.0,5.0) and (49.0,19.0). A total of 315,225 earthquakes were found. The beginning of the query corresponds approximately with the update, renovation and increase in the number of stations of the national seismic network (Michelini et al., 2016; Danecek et al., 2021; Margheriti et al., 2021). Around 2005, the INGV network (FDSN code IV) underwent a major upgrade with the existing, predominantly analog, instruments being replaced by high quality digital seismic data loggers and new, mostly broadband (and some extended short period), three-component (3C) sensors. Selected stations were also complemented with additional 3C strong motion sensors. The upgrade resulted in more than a two fold increase of the number of stations of IV network. Also and since 2005, there have been many temporary deployments of seismic stations coinciding with earthquake sequences and specific experiments which data are also available through the EIDA INGV node (Danecek et al., 2021). The total number of stations also increased thanks to the contribution of the networks belonging to other Italian institutions (e.g., the University of Genoa, National Institute of Oceanography and Experimental Geophysics-OGS, and the University of Naples amongst others). This increment resulted in a significant improvement of the detection of low magnitude earthquakes. At the regional scale of Italy, the magnitude of completeness of the INGV earthquake bulletin is around $\sim M1.7 - M1.8$ although significant differences occur depending on the area.

A relevant aspect when compiling a large dataset to be used for ML purposes consists of gathering a balanced distribution of data. In seismology, when taking earthquake magnitude for classification, balanced representation is impossible to achieve because small size earthquakes, following the Gutenberg-Richter magnitude versus the number of earthquakes power-law (Gutenberg and Richter, 1944), outnumber larger earthquakes. To address this issue (or at least to mitigate its influence), we choose to select in our target area:

- the great majority of the earthquakes with $M \geq 4.0$;
- earthquakes with origin times differing by more than 120 s in the range $2.0 \leq M < 4.0$;
- additional 20,000 earthquakes, randomly selected, with origin times differing by more than 120 s for $M < 2.0$.

The resulting distribution of the earthquakes according to their magnitude is detailed in Table 1 and they are mapped in Fig. 1a.

2.1.2 Station selection

In order to gather high quality earthquake signals, we based our choice on the most accurately picked P- and S-wave onset phases published in the INGV bulletin (http://terremoti.ingv.it/en/bsi). To this regard, the manual picking of the arrival phases is

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

Open Access Earth System Science Data Discussions

Table (1). Final data selection. “All” indicates the total number of earthquakes in then INGV bulletin in the time period between 01/01/2005 and 01/31/2020, “Selected” and “Percent kept” refer to the earthquakes, and “Nb. 3C records” to the waveform traces included in the dataset.

$\geq M_{min}$	$< M_{max}$	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

routinely performed by a group of about 20 INGV highly trained staff personnel who also review the hypocenter locations and magnitude determination before bulletin publication. These manually reviewed locations are indicated as *preferred* solutions in the INGV earthquake bulletin. In practice, we have selected only those stations that had P- and, if available, S-wave onset picks associated to the *preferred* location of the INGV bulletin.

In summary, we have adopted the following criteria to identify the waveform records to be included in the dataset after the earthquake selection above was applied:

- all stations that feature P-wave (and S-wave when available) onset phases used for the *preferred* earthquake location;
- P- and S-wave location residual times less than 1.0 s;
- P- and S-wave phases that contributed to the location with a weight larger than 10 %.

This selection procedure reduced the number of P- and S-wave phases from ~ 1.9 to ~ 1.2 and from ~ 1.1 to ~ 0.7 millions, respectively.

2.1.3 Waveform data selection and download

The selection procedure described in section (2.1.2) resulted in the compilation of a list of waveform data time windows to be downloaded from the EIDA continuous waveform archive. We choose a time window of 120 s in order to include both P- and S-waves from stations whose distance is up to ~600 km from the hypocenter. Indeed, in these cases, the $S-P$ time differences are approximately 75-80 s. Adding about 20 s of signal before the P-wave time and about 20 s after the S-wave, we end up with a 120 s window choice providing the most significant earthquake signals for either the most distant stations, in case of crustal depth earthquakes, or closer stations, in case of deep earthquakes of the Calabrian arc subduction. Traces that, according to the INGV earthquake catalog, would include more than one earthquake within the 120 s time window, were removed from the dataset.

More technically, the time windows were defined by inserting a randomly selected buffer time ranging between 15 and 20 s before the P-wave onset arrival phase and enlarging the time window to 125 s. This criterion ensured that the great majority of the waveform traces downloaded featured a pre-P-wave onset buffer time between 15 and 20 s. However, we found that, when dealing with such a large number of waveforms acquired by diversified instruments configured differently, some discrepancies may occur. In practice, since the data are archived in miniSEED compressed format that feature different sizes of the logical records, and since the web service extracts the full logical record containing the pre-defined trace start time, it can occur that the starting time of the trace is earlier than the predefined minimum time of 20 s (i.e., in this case, there is a longer time interval between the P-arrival and the actual trace start time). In contrast, when data are missing before the P-wave onset time (i.e., in the 15-20 pre-P-onset buffer time), start time of the extracted window can be delayed and a shorter time interval will separate the trace window start time from the P-wave arrival time (i.e., $<$ 15 s). See Fig. A1 in the appendix for the distribution of the P- and S-wave phase arrival time samples. The data (miniSEED format) were downloaded using the FDSN `dataselect` web services provided by INGV (http://terremoti.ingv.it/en/webservices_and_software). Using a set of 14 container based querying procedures running in parallel, this stage required about 7 days to complete the ~4 million waveform traces (i.e., ~1.3 million of 3C traces) download for a storage requirement of ~80 GB (miniSEED STEIM1 compression).

2.1.4 Cross-validation between phases-based metadata and downloaded waveform data

After the massive data download was concluded, a list of all the downloaded files was generated. This list was *intersected* with the originally selected metadata (section 2.1.2) to have a *one-to-one* correspondence between the miniSEED data and the metadata (i.e., each 3C waveform record — three miniSEED files —must correspond to a row of the metadata file).

2.1.5 Processing of the data counts waveforms

This part of our data assembling procedure targets the preparation of the digital counts waveform traces. It includes the following steps.

- removal of traces containing data gaps (i.e., missing data);
- trimming the waveform trace to the nearest sample to the start time;
- 120 s trace windowing;
- removal of mean and linear trends from the data;
- re-sampling at 100 Hz;
- calculation of the signal-to-noise ratio;
- extraction of the data quality metrics.

Open Access Earth System Science Data Discussions

For each waveform trace (i.e., each component), the maximum value of signal-to-noise ratio (SNR) was extracted and kept as metadata. The SNR was calculated as

$$SNR = 20\, log_{10} \frac{|S_{95}|}{|N_{95}|} \tag{1}$$

where $|S_{95}|$ and $|N_{95}|$ are the 95 % percentile of the data absolute values in a 5 s window immediately after the S-wave onset and right before the P-wave arrival time. If the S-wave onset were not available, the S-wave window was determined after calculation of the predicted S-wave arrival using an average velocity of 3.0 km s^{-1} and the hypocentral distance.

During this stage of the data preparation, we have also calculated some quality parameters extracted from the waveform traces to the purpose of a later inclusion in the metadata information. These additional parameters, providing the distribution of the trace values, have been computed using the `MSEEDMetadata` class of the `obspy` python software (Beyreuther et al., 2010; Megies et al., 2011; Krischer et al., 2015). To the same purpose, we have determined the number of spikes using a Hampel filter on a 161 samples sliding window to find outliers in the traces.

The final dataset consists of a total of $1,159,249$ 3C waveform data records from $54,008$ earthquakes in counts units assembled within an `hdf5` format file. Table 1 provides the number of traces within each magnitude interval of the final assembled dataset.

2.1.6 Application of the instrument transfer function to the waveforms

To make the dataset of more general use, we have also generated the associated ground motion units dataset after deconvolving the instrument response. To this end, we have downloaded the station response files for all the stations used and applied the transfer functions to the individual traces with frequency filtering corners 0.01, 0.04, 25, 40 Hz using a cosine flank frequency domain taper (see cosine_sac_taper in Obspy), and applying a 5 % cosine tapering at both ends of the trace signal. After removing the instrument response, we extracted the intensity measures (IMs, i.e., peak ground acceleration, PGA, peak ground velocity, PGV, and the spectral accelerations at 0.3, 1.0 and 3.0 seconds period) on each component so that they could be included amongst the metadata parameters. Peak ground displacements are not included since they result from single or double integration of velocity and acceleration records, respectively, and their determination can result inaccurate when performed automatically.

2.2 Metadata description

The 114 metadata associated to each 3C waveform trace of our collection are listed in Table 2. They provide different kind of information that can be subdivided into four main types — *source*, *station*, *trace* and *path* metadata. The unit of each metadata is provided in its denomination.

The *source* metadata provide information on the earthquake with description of the source origin time, location, size and, when available, the focal mechanism, the moment tensor, and the finite fault.

The *station* metadata provide information on the characteristics of the recording station which include the station, channel, network and location (SCNL) (cf. http://www.fdsn.org/seed_manual/SEEDManual_V2.4.pdf), the geographical coordinates

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

and the average shear-wave velocity of the top 30 m of the Earth, $V_{S,30}$, which is an important parameter for classifying sites in seismic engineering applications (e.g., Boore, 2004) and is extracted from the map used in the INGV implementation of the USGS-ShakeMap software in Italy (Michelini et al., 2019).

The *trace* metadata consists of parameters that are extracted from the waveform traces like maximum and minimum amplitudes, root mean squared values of the traces and, after application of the transfer function, intensity measures (IMs) of the ground motion. In this class of metadata, we include the P- (and S-wave) provided by the INGV bulletin and, in addition, the number of P and S picks obtained by processing the waveforms with two deep learning phase picking and event detection algorithms (GPD and EQTransformer; Ross et al., 2018a; Mousavi et al., 2020) to make the user aware that the waveform trace being used may include more than a single earthquake (see discussion further below).

The *path* metadata follow from the calculation of parameters that link the types of metadata above (e.g., traveltimes, hypocentral and epicentral distances).

The rationale of our metadata selection reflects our intention of providing the users with comprehensive information about the data. This appears an important issue since the data, being recorded automatically, can suffer of many diverse problems deriving from malfunctioning of the data loggers, of the sensors or from poor data transmission. Since we seek to assemble a data set that can be used also for analysing real time data streams using ML, we note that the automatic processing summarized above does not differ significantly from that routinely applied to the streamed data.

One alternative to our metadata "comprehensive" approach would have consisted of "cleaning" the dataset by removing the faulty traces from the dataset altogether. We do not think this approach appropriate since in this case the dataset would not be representative of the "true" data that are collected in real-time by the monitoring networks. Thus, the basic idea behind our criterion is that we would like to enable the users to make their own choices using opportune filters to exploit the data for their own purposes. For example, if a user looks for the cleanest data, this can be achieved by filtering the metadata accordingly (e.g., saturated velocimetric data acquired by broadband sensors equipped with 24 bit data loggers could be removed in a conservative fashion just by selecting only those traces with counts within $\pm 0.8 * 2^{23}$). In contrast, the user could also opt to leave the ML model to learn the "data problems" so that they can be detected when using real data. An approach of this kind has been used by (Jozinovic et al., 2020) for missing data. In practice, the provision of a rich set of waveform descriptive metadata is important not only to make use of an enlarged suite of labels that can be used for diverse purposes but also to identify problems with the waveform data and include or filter them out.

Our metadata includes P- and S-wave onset manually picked by INGV analysts as provided in the INGV bulletin. Recall that the traces were selected to include just one P-wave arrival time and possibly one S-wave arrival time since we sought to assemble one earthquake per window trace. This criterion was chosen to the purpose of facilitating the training of ML models upon traces containing just one earthquake (e.g., for phase picking, peak ground motions, ...). However, even though we have made considerable efforts to isolate only one earthquake per time window, more than one can be present effectively within the same time window (e.g., the analyst did not see or just disregarded other events with smaller amplitudes). Because the presence of additional, unidentified earthquakes adds complexities to the ML training phase, we followed the same approach taken by Mousavi et al. (2019) to run automatic picking algorithms upon the waveform dataset and include as metadata also the

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

number of P- and S-wave phases picked automatically by the generalized phase detection, GPD, technique proposed by Ross et al. (2018a) and the EQTransformer technique by Mousavi et al. (2020). In the analysis we have used as detection threshold: 0.99 for P- and S-phase detection for GPD, and 0.2, 0.1 and 0.1 for earthquakes, P- and S-phase detection, respectively, for EQTransformer.

As presented above, metadata are important constituents of data collections. They can be used for identifying the data to be analyzed and they can be used as labels in ML applications. Besides that not all the metadata information in INSTANCE is always available (e.g., moment tensors are available only for events with magnitudes $M \geq 3.5$ or the S-wave onset pick retrieved from the INGV bulletin may not be present), we have found that the automatically processed ground motion trace data may suffer from errors because the original traces contained already undetected malfunctioning problems (e.g., spikes, anomalous trends) which, after application of the instrument transfer function, are mapped into erroneous ground motion traces and IM values. Analogously, it may have also occurred that in isolated cases the coefficients of the instrument transfer functions were found incorrect producing also in this case incorrect traces and IM values. To address these problems, we have operated in two ways. First, we have chosen to detect the traces' maximum and minimum values lying outside the acceptable physical range and to replace them with *numpy* `nan` in the metadata file. This acceptable range was based on the IMs reported in the "flat" file of the ESM DB (https://esm-db.eu/, Lanzano et al., 2018) which includes all the IMs (obtained from analyst processing) of all the recordings available of earthquakes with $M \geq 4.0$ in Europe. Secondly, we have verified our instrument transfer function processing procedure by cross-validating all our IM values with those reported in the ESM DB "flat" file. To this regard, we found a very good correspondence between the IMs obtained using the two methodologies giving us confidence in the quality of the applied data processing and of the IM metadata being provided.

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

Open Access Earth System Science Data Discussions

Table (2). List of the metadata for the events and noise waveform traces. The units are given in parenthesis in the "Description" column. Only a subset of metadata can be associated to the noise traces (star in the "Noise" column).

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`		Scalar moment (N m)
`source_mechanism_strike_dip_rake`		Strike, dip, rake of the two planes (2 tuples)
`source_mechanism_moment_tensor`		6 components of the moment tensor (m_rr, m_tt, m_pp, m_rt, m_rp, m_tp)
`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 (International Registry of Seismograph Stations (IR))
`station_channels`	⋆	Two characters identifying the sampling and the intrument gain (e.g., HN, HH, EH, ...)
`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_azimut_deg`		Direction from event location to station (°)
`path_backazimut_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)

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

Table (2). List of the metadata, continued.

(a) The horizontal line separates the additional metadata obtained after application of the instrument response transfer function.

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-, 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 minumum (*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, S number of picks retrieved with GPD
trace_EQT_[P,S]_number	⋆	P, 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)

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

Open Access Earth System Science Data Discussions

2.3 Dataset description

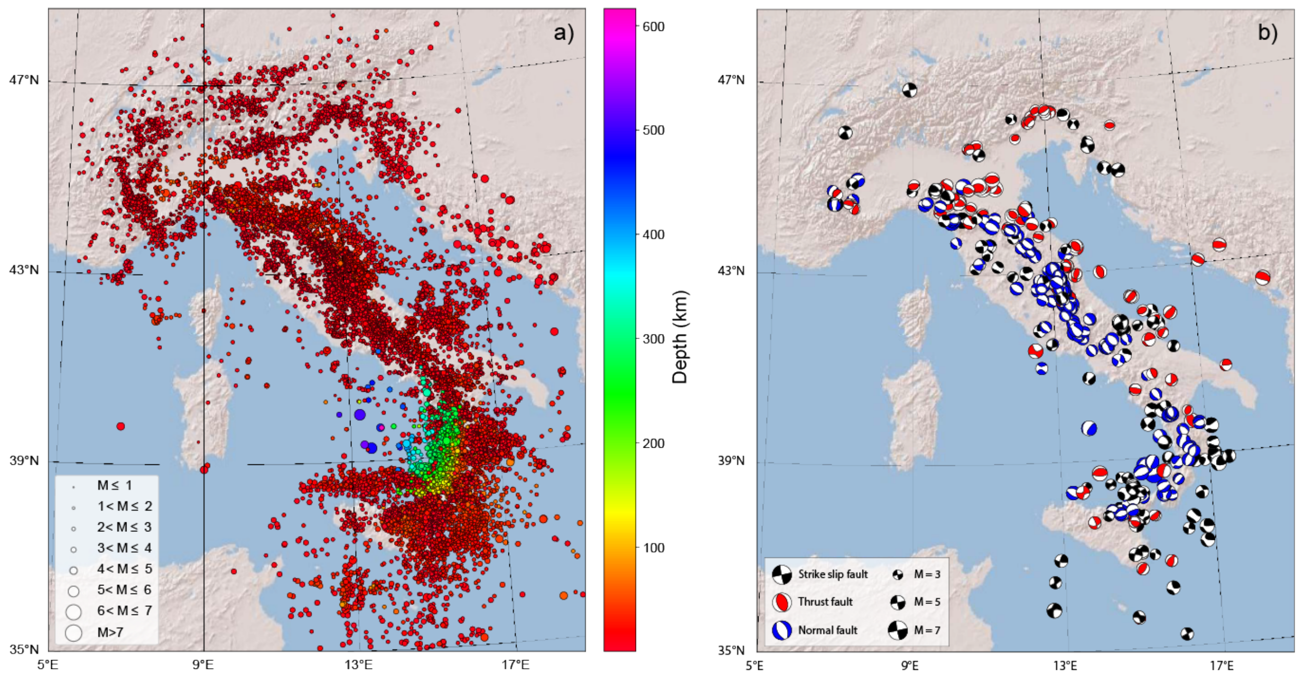

Figure (1). Earthquakes (a) and moment tensors (b) of the dataset. Symbol size, in both maps, is proportional to earthquake magnitude.

Figure 1a shows the earthquakes included in the dataset. The symbol size is proportional to the earthquake magnitude. We observe that the 54,008 selected earthquakes composing the dataset can be considered a representative subset of the entire seismicity in Italy and, for the larger events, also for those earthquakes occurring outside the Italian national borders. During the time span of our data selection three important sequences have occurred in Italy after the main shocks of the 2009 L'Aquila M 6.0, the 2012 Emilia M 5.9 and the 2016 central Italy extended sequence which featured three main earthquakes with magnitudes M 6.0, M 5.9, and M 6.5.

In Fig. 1b we plot the 527 moment tensors included in the metadata. The size of the moment tensors symbol is proportional to `source_magnitude` while the colors are defined according to the prevalent strain regime: black, red and blue for strike slip, normal and inverse faults, respectively. The prevalent strain regime is determined according to the fault's rake as derived form `source_mechanisms_strike_dip_rake`: strike slip for -45° $<$ rake $<$ 45° and 135° $<$ rake $<$ 225°; normal for 225° $\leq$ rake $\leq$ 315°; inverse for 45° $\leq$ rake $\leq$ 135°. In Fig. 2 we show the maps of the stations included in the events and noise datasets, respectively. The symbol size in panel a) is proportional to the number of reported phase arrivals by each station, while in panel b) is proportional to the number of waveforms included in the dataset for each station. Figure 2a evidences that quite different number of phases have been reported by the stations included in the event dataset. These differences depend on several factors like whether the stations are permanent or temporary, the time length of the acquisition, the noise level and the level of seismicity of the area where the stations have been deployed. For example, it is evident that many stations in

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

central Italy display many phases (and associated trace recordings) mainly because the area was struck by the 2009 and 2016 earthquake sequences. In contrast, stations that are located in the Po Plain generally feature small number of phases mainly because the noise level is high making it difficult the phase picking. The same diversification in the number of available traces is not observable for the noise dataset shown in Fig. 2b. This occurs because it was an intentional choice to select a more or less even number of traces for all the station channels.

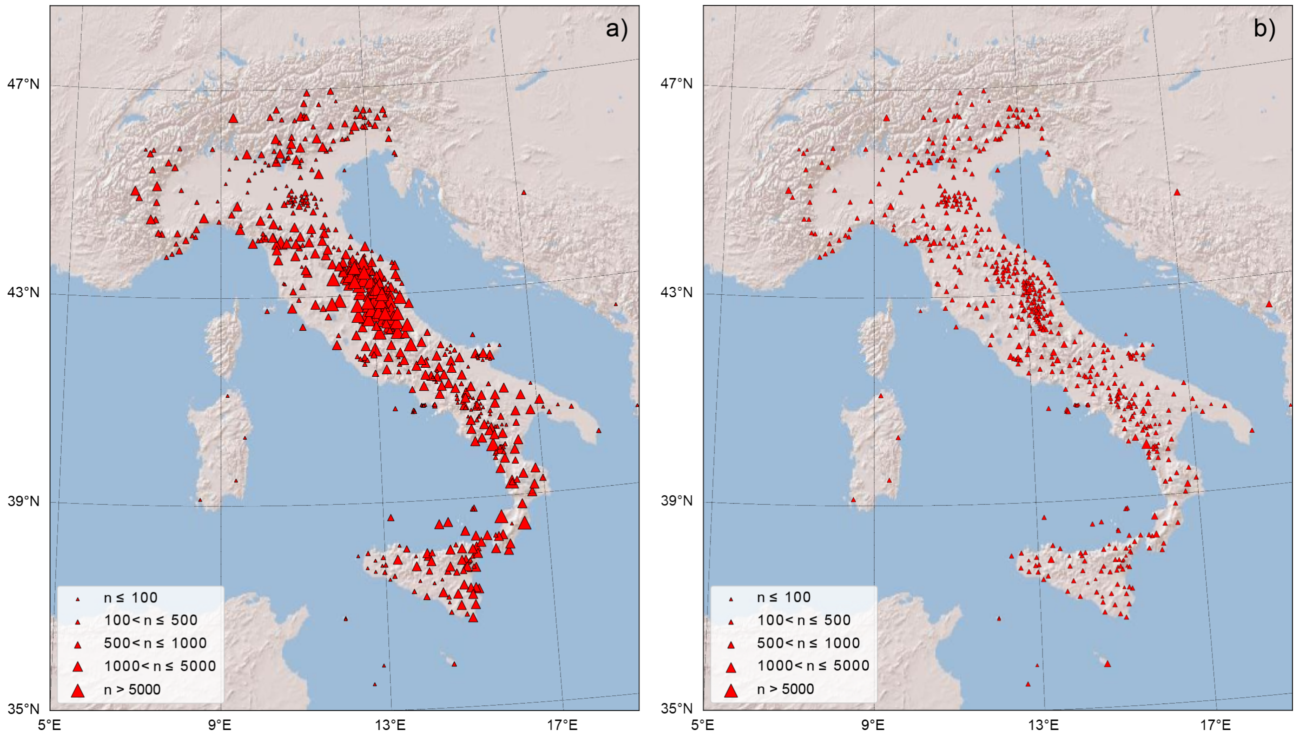

Figure (2). Stations included in the events (a) and noise (b) datasets. The symbol size in (a) is proportional to the number of the number of P-phases and corresponding waveform traces available for each station. In (b) the symbol size is proportional to the number of traces. A total of 620 stations are included.

In Fig. 3, we show the distribution according to magnitude, earthquake to station epicentral distance, earthquake depth and backazimuth of the 3C record traces composing the dataset. The panels show the histograms using the log_{10} scale to provide a complete representation of the distribution of the dataset. Despite the attempt to balance the distribution of earthquakes according to magnitude (Sect. 2.1.1), Fig. 3a shows that our selection still reflects (inevitably) the Gutenberg-Richter increase of the number of earthquakes at smaller magnitudes. The largest amount of trace records in the dataset belongs to earthquakes in the magnitude range $2 \leq M < 3$. The significant decrease of the number of traces for $M < 2$ follows from our choice to balance the dataset at small magnitudes by taking only about 7 % of the whole dataset. For what concerns the epicentral distances of the stations (Fig. 3b), the great majority of the traces have been recorded within 200 km. A better appreciation of the selected traces can be obtained from the observation of Fig. 4 where we show the magnitude versus hypocentral distance

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

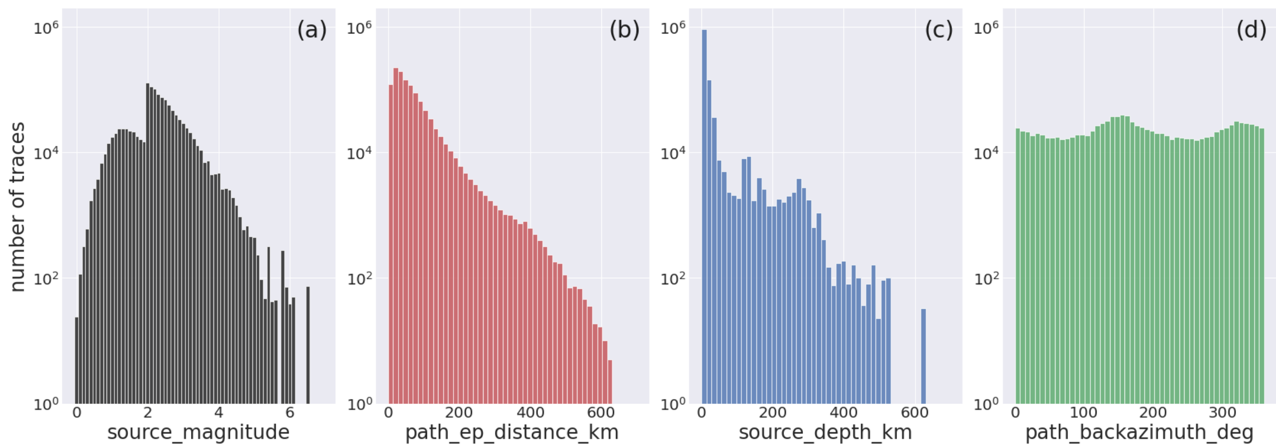

Figure (3). Histograms of the distribution of the trace records composing the dataset according to magnitude (a), epicentral distances (b), earthquake depth (c) and backazimuth (d).

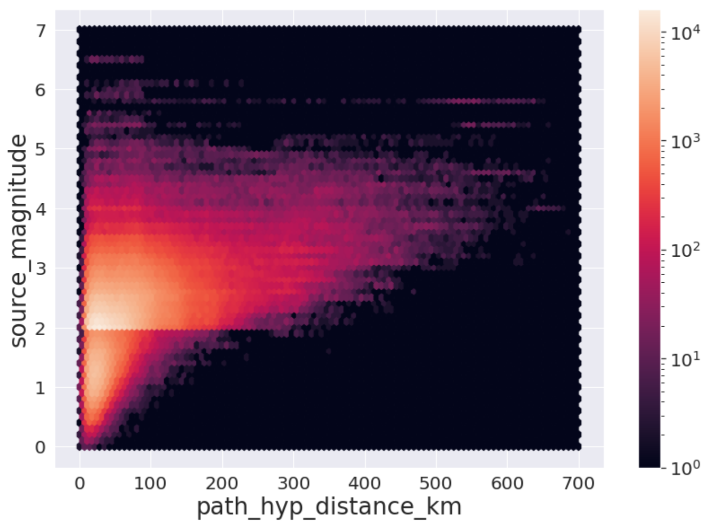

Figure (4). Diagrams of the magnitude distribution versus distance as hexbin plot.

distribution of the dataset traces represented as density plots using hexagon binning (hexbin, Hunter, 2007). The earthquake depth distribution (Fig. 3c) shows that the great majority of the traces belong to shallow crustal earthquakes although a few thousand included occur in the depth range 100 to 300 km. At greater depths, the number of traces decreases sharply and only a few hundred or less recordings are included in the depth range 400 to 550 km. Figure 3d shows that the great majority of the P- and S-wave onsets belong to paths more frequent along the NW-SE direction in agreement with the geographical trend of the Apennines and of peninsular Italy overall.

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

Figure 5a shows the distribution of the trace channels of the dataset (`station_channels`). The weak motion, high gain channels represent more than 70 % of the total number of traces. These are subdivided into HH channels associated to the broadband high gain velocimeters (51 %) of the total whereas the extended short period channels (EH) traces account for 20 %. The low gain accelerometric channels form the remaining part of the dataset. In Fig. 5b, we show the distribution of the

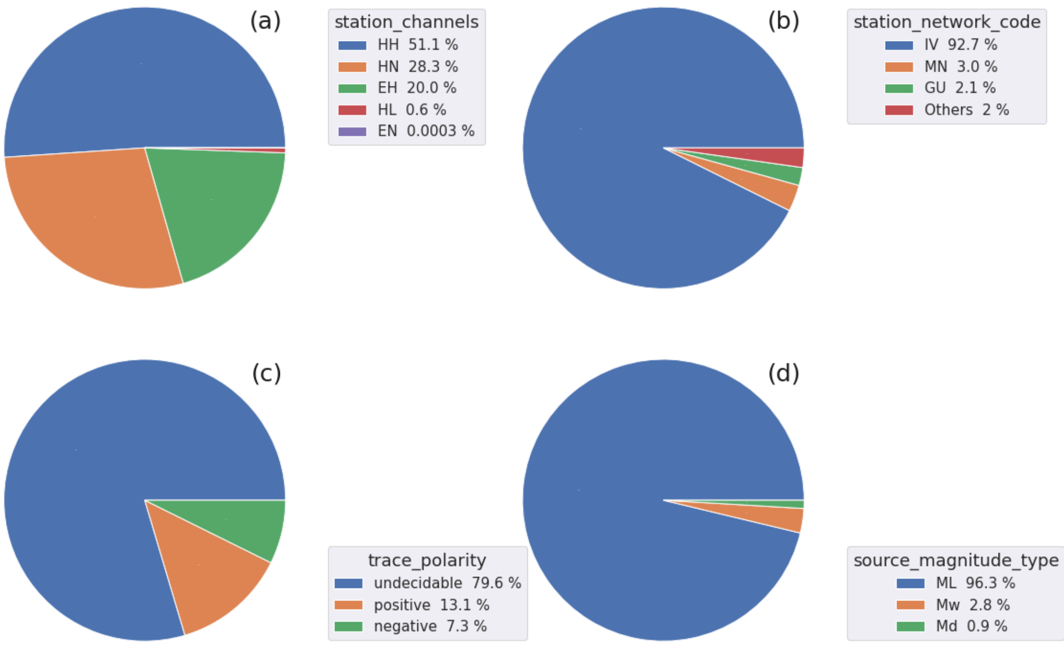

Figure (5). Pie diagrams summarizing the distribution of the channels (a), the data contributing networks (b), the P-wave polarities (c) and the magnitude types (d) of the dataset. The full list of `station_network_code` with % < 1 collected in Others in decreasing order is OX, ST, SI, XO, NI, IX, OT, RF, YD, TV, B1, AC, HL, ZM, 3A.

records subdivided according to the different networks (`station_network_code`) operating in Italy and in neighboring countries that have been included in the dataset. The dominant portion of the data (~96 %) have been acquired by the Italian National Seismic Network (IV code) and by the MedNet (MN code) both operated by INGV (Michelini et al., 2016; Danecek et al., 2021). The full list of the contributing networks is provided in the caption.

The polarities associated to the P-wave onsets (`trace_polarity`) are shown in Fig. 5c and have been reported in only 20 % of the total number of traces. Although this represents only a fraction of the dataset, we are confident that its number

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

(~235,000) is likely large enough to be used in a ML dedicated model (e.g., Ross et al., 2018b) for training and testing, and then used to recover the polarities of the remaining batch.

The magnitude type distribution (`source_magnitude_type`) is shown in Fig. 5d. The Wood-Anderson local magnitude M_L (Richter, 1935) is calculated predominantly (~96 %). The moment magnitude M_w is determined for earthquakes with ~$M_L \geq 3.3$ and when enough good quality station data are available (Scognamiglio et al., 2009). The M_d magnitude is used only when it is impossible to determine the M_L and it is provided mainly in the first years of the dataset when the IV network still included a considerable number of analog stations.

In Fig. 6a,c we present the histograms of the P- and S-wave residual times included in the dataset. Fig. 6b,d shows the phase arrival weights resulting from the earthquake locations for P- and S-phases, respectively.

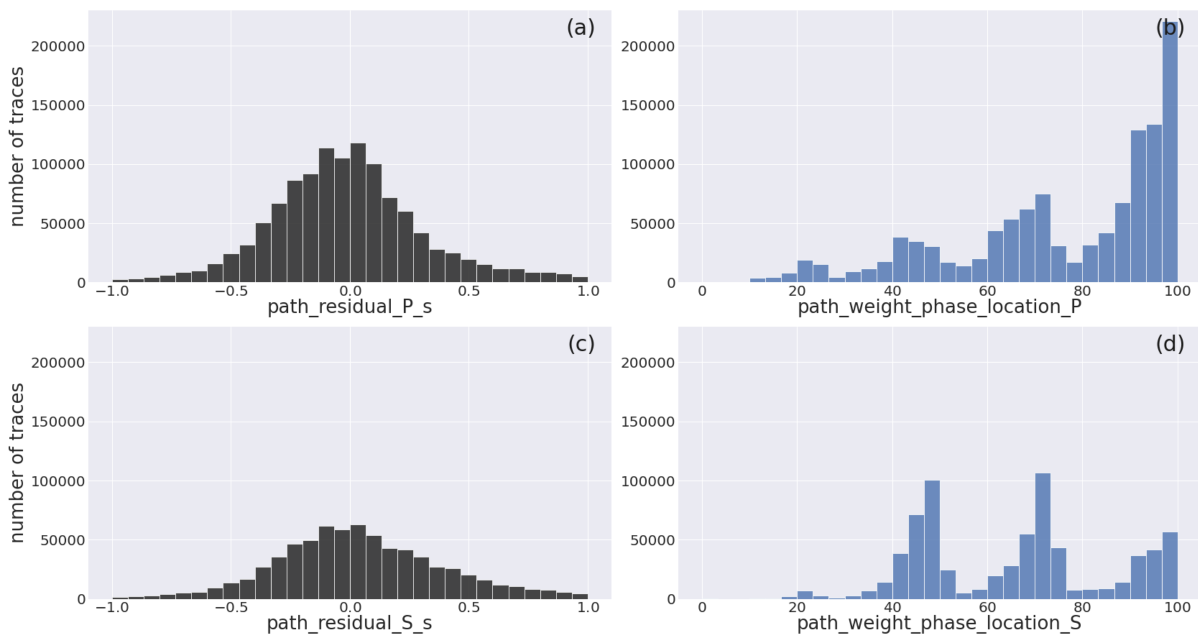

Figure (6). Histogram of the P- and S-wave residuals (a,c) and of the phase arrival weights, expressed as percent, resulting from the location (b,d).

To provide a broader perspective of the dataset and with the intent of showing the wide range waveform paths that have been included we present, in Fig. 7, the hexbin plots of the traveltime for both P- and S-wave arrival times used in the locations. These panels have been arranged using four different maximum distances and are useful for visualizing the dominant structure of the data selection given the large number of data. More specifically, it can be observed that the hexagon binning panels for the larger distance ranges (700 and 200 km max distance) and for both P- and S-wave traveltimes (Fig. 7a,b,e,f) highlight well both the direct and Moho refracted travel times. At smaller distance ranges (100 and 40 km, Fig. 7c,d,g,h), it is evident that our

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

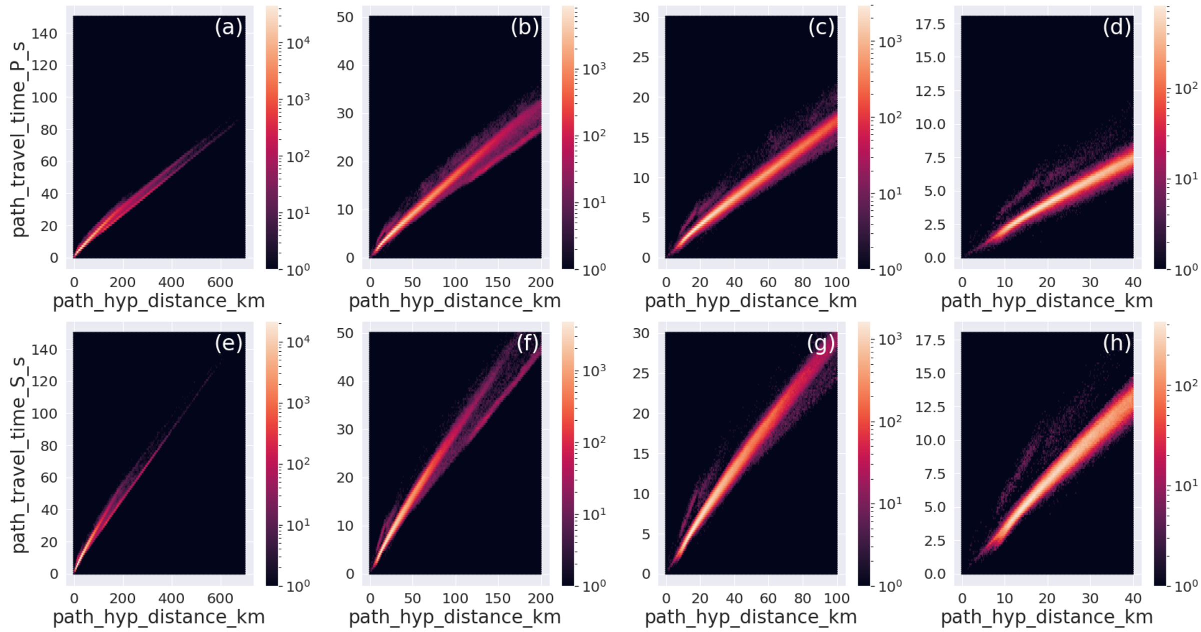

Figure (7). Hexbin plot of the traveltimes for different distance ranges for P- (top panels) and S-waves (bottom panels).(a,e) 0-700 km; (b,f) 0-200 km; (c,g) 0-100 km; (d,h) 0-40 km.

dataset includes waveforms that propagated across crustal structures with different velocities. This is very evident, for example, for both P- and S-wave in the hexbin plots where at small distance are observable very low Vp and Vs velocities.

In the following, we will focus on the *trace* amplitude metadata. These parameters, provided as metadata, are important for refined selection of the traces and are extracted from both the raw waveforms expressed in "counts" and from the traces in physical units after application of the instrument transfer function. Some of these parameters can be obtained without any knowledge on the earthquake source parameters whereas others, like the SNR, require knowledge on the arrival times of P- and S-wave onset times.

The panels of Fig. 8 display the median (`trace_[E,N,Z]_median_counts`) and the mean values (`trace_[E,N,Z]_mean_counts`) of the dataset traces. To evidence the whole range of values attained by these two metadata, we adopt the base-10 log scale. The histograms show, for all the three components, remarkable differences of the distributions. The mean values, which have been removed in the pre-processing preparation stage (cf Sect. 2.1.5), are tightly centered about the zero value whereas the median histograms, while being similarly centered about the zero value, do evidence a broader distribution of values around zero. This last behaviour occurs whenever the waveform trace values are unevenly distributed about the mean and it derives from the pre-processing of faulty traces that, for example, results into defective removal of the linear trend. The same figure for the full range of the parameters, is available in Fig. A2.

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

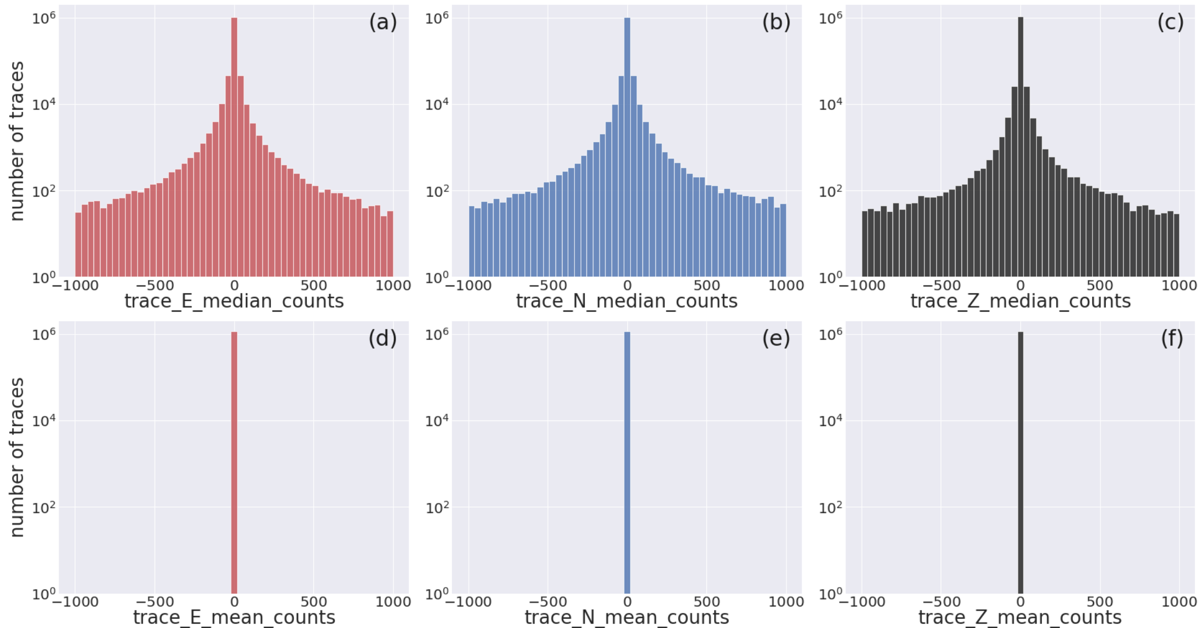

Figure (8). Close view of the histogram of the distribution of the median and mean values of the earthquake waveform traces.

In Fig. 9, we present the histograms of the distribution of the trace quality control parameters obtained from the application of the `MSEEDMetadata` module of the `Obspy` seismological software suite. The figure shows the distribution of the quality control parameters in a closer view (see the full range of values in Fig. A3). The histograms show that the largest majority of the traces feature root mean square values less than 2.5×10^4 with a minor contribution from traces featuring higher values. The minimum and maximum values follow a similar trend for negative and positive values, respectively. The lower and upper quartile values of the traces show that the peak of the distributions are at $\mp 2 \times 10^3$, respectively.

The SNR distributions are shown in Fig. 10 as histograms and versus distance and magnitude ($M \geq 2$) as hexbin plots in Fig. 11. The histograms show that the peak values for the whole dataset are at ~10 db for the two horizontal components and slightly less for the vertical (~6 db). This is expected because the S-wave motion is polarized perpendicular to the nearly vertical propagation direction at the surface, implying that the ground motion occur mainly along the horizontal components. In any event, the distribution of the SNR values of our dataset can be considered rather satisfying given that values larger than 2 already provide distinct earthquake signals. The hexbin plots of Fig. 11 provide a snapshot of the dominant levels of SNR with distance and magnitude. It is observed that higher values occur for nearby earthquakes and that the SNR progressively decreases at farther distances. Conversely and as expected, the SNR generally increases with larger magnitude earthquakes.

The hexbin plots of the distribution of the IMs with distance for earthquakes with $M \geq 2$ are shown in Fig. 12 whereas their associated distributions are shown in Fig. A4. The panels evidence a broad concentration of ground motion values deriving from the inclusion of earthquake recordings from different distances and magnitudes. The panels also evidence some horizontal

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

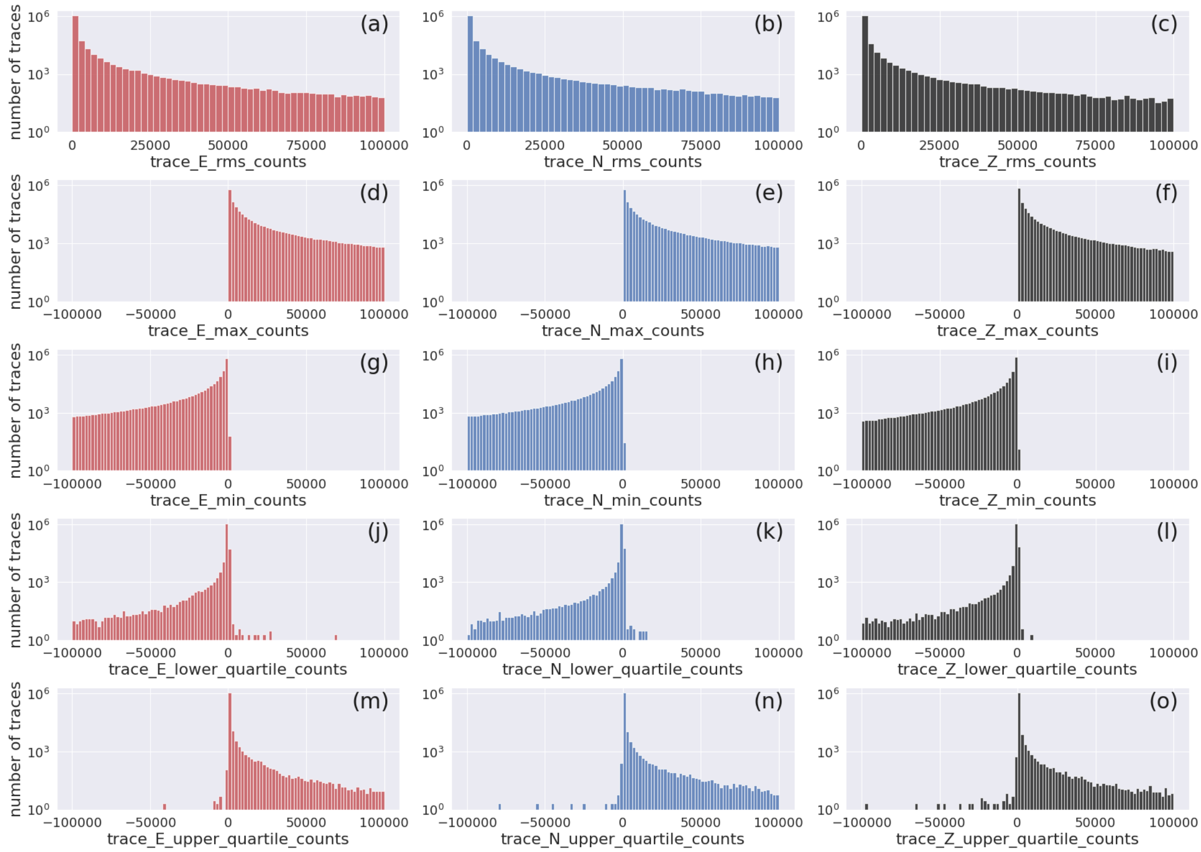

Figure (9). Histogram of the distribution of quality control metadata of the earthquake waveform traces: rms, min, max, first and third quartile. The width of the bins is 2×10^3. The full distribution of values is provided in Fig. A3.

stripes at higher values of ground motion resulting presumably from the acquisition and processing problems mentioned in Sect. (2.2).

To show the distance dependence of the IMs for a given magnitude, in Fig. 13 we plot the values for $M = 3$ earthquakes (i.e., IMs in the range $2.9 \leq M \leq 3.1$). The maximum concentration of IMs can be assimilated to an average ground motion model for for $M = 3$ earthquakes in Italy.

2.4 Examples of event data traces

Some examples of the data traces are shown in Fig. 14, 15 and 16. The traces have been selected randomly according to certainly non-exhaustive criteria described in the figure caption using, as a guideline, the metadata distribution illustrated in Sect. 2.3.

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

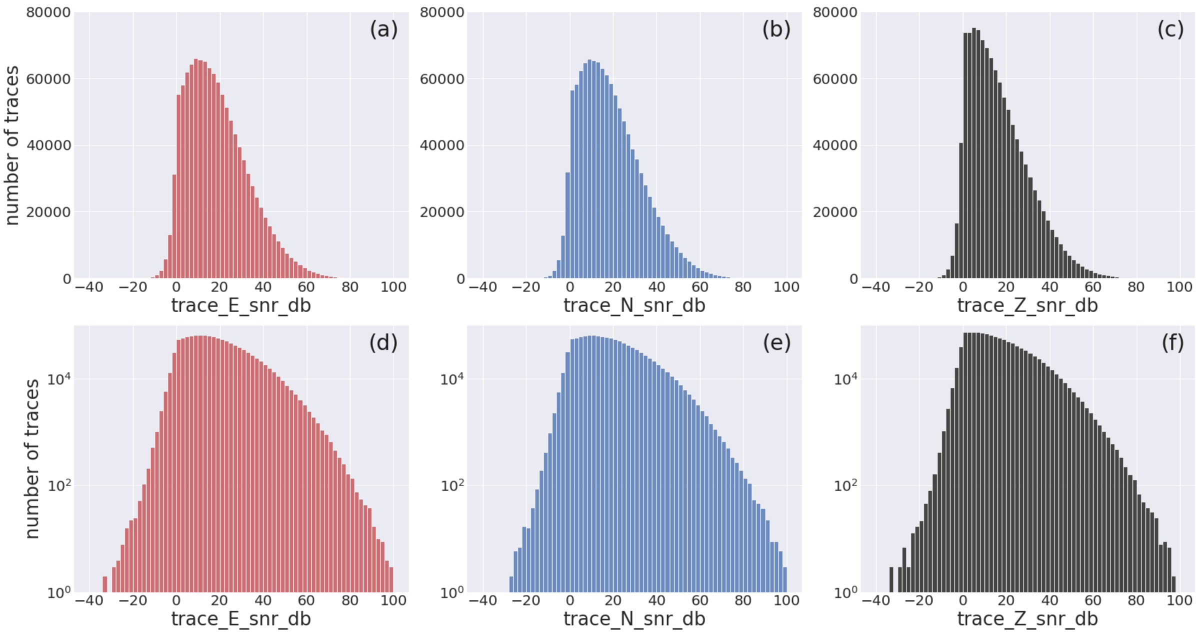

Figure (10). Distribution of the signal-to-noise ratio of the earthquake waveform traces.

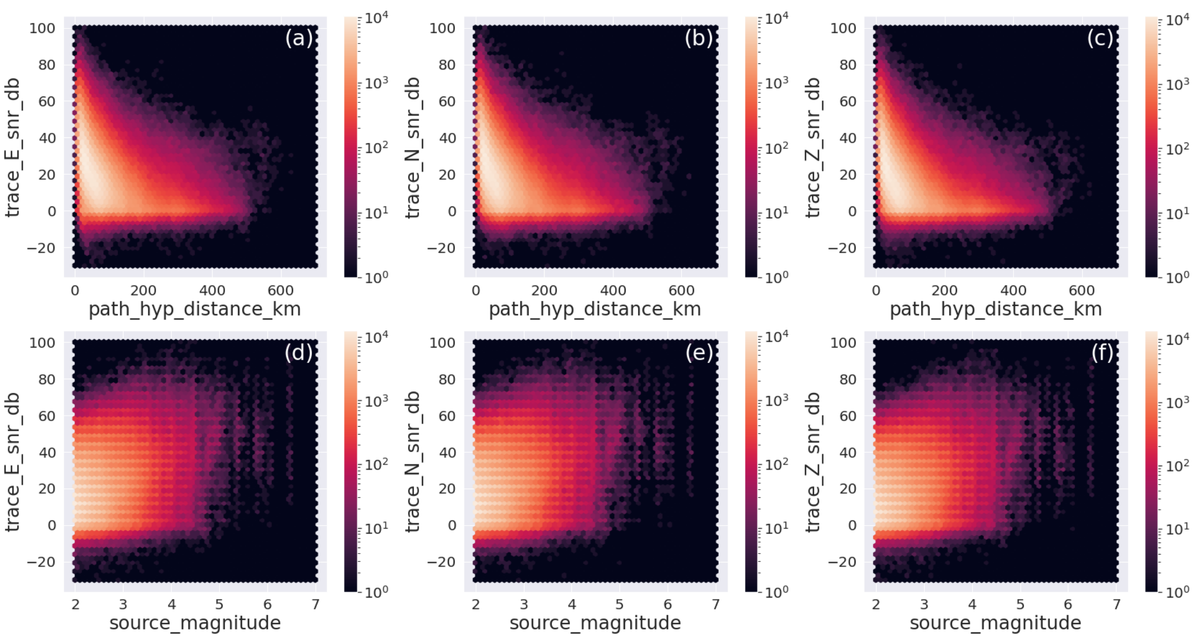

Figure (11). Hexbin representation of the distribution of the signal-to-noise ratio with distance and magnitude.

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

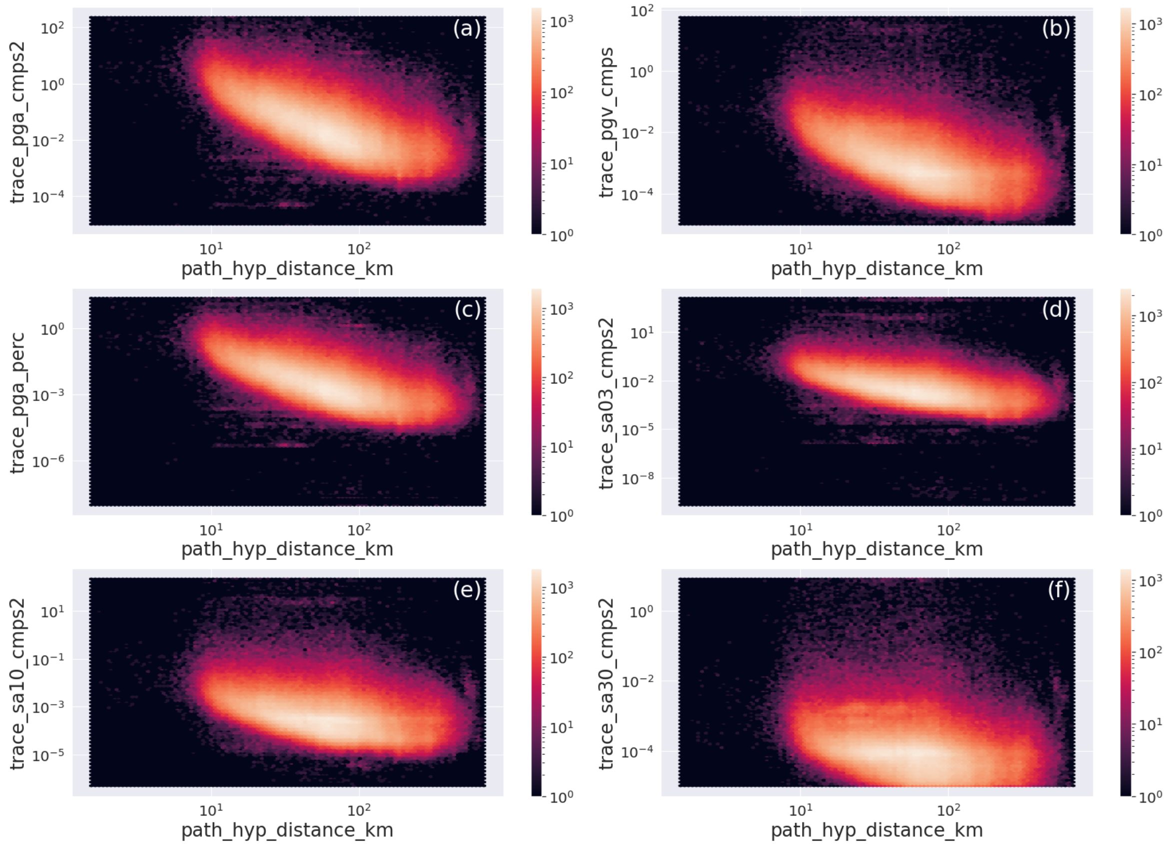

Figure (12). Hexbin plot of the distribution of the intensity measures (IMs) with hypocentral distance for the $M \geq 2$ earthquakes. The units are km along the horizontal axis in all panels, and, along the vertical axis, $cm\ s^{-2}$ in panels (a,d-f), $cm\ s^{-1}$ in panel (b), and % g in panel (c).

In Fig. 14 we show the traces in counts of events recorded by the broadband instruments (HH channels). Specifically, the first three rows (Fig. 14a-i) show traces for different ranges of magnitude which, taken together, represent more than 80 % of the total HH traces. In the following two panel rows (Fig. 14j-o) we show examples of traces selected according to SNR and distance criteria that evidence that more than 65 % of the traces feature relatively high SNR (i.e., ≥ 10) within the whole distance range covered by our data collection. The seismograms shown in the last panel row (Fig. 14p-r) provide some samples of recordings of the largest events ($M \geq 4$) where we found that ~ 87 % feature SNR ≥ 10.

To provide an exhaustive exploitation of the dataset, we focus next on examples of problematic traces. Although different criteria could have been used given the comprehensive set of metadata available, here for simplicity, we base our identification on: i.) the number of picks and detections resulting from application of the GPD and EQTransformer algorithms to isolate those traces likely containing more than a single event; ii.) the value of the SNR to identify poor quality noisy traces; iii.) the values of the trace median values which are expected to diverge from zero whenever the trace values are unevenly distributed

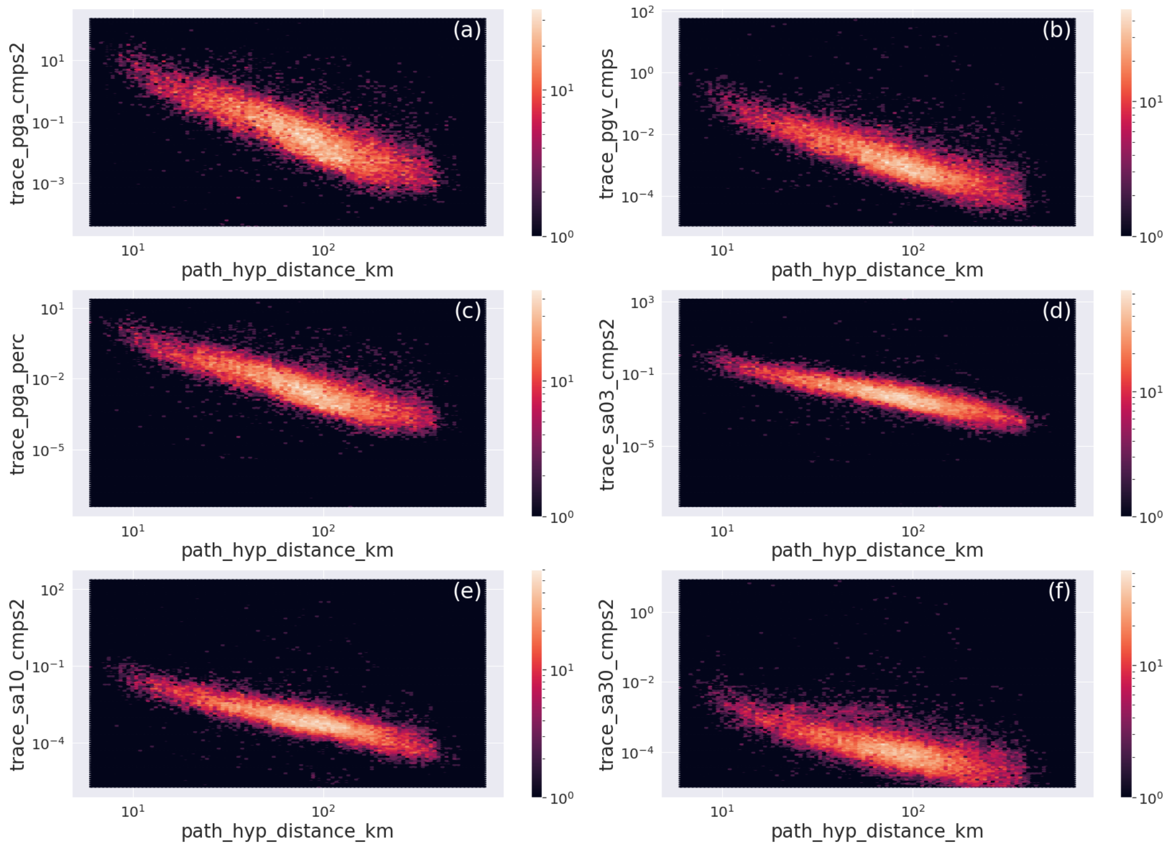

Figure (13). Hexbin plot of the distribution of the intensity measures (IMs) with hypocentral distance for $M = 3$ earthquakes. The units are km along the horizontal axis in all panels, and, along the vertical axis, $cm\ s^{-2}$ in panels (a,d-f), $cm\ s^{-1}$ in panel (b), and % g in panel (c).

about the mean value as result of acquisition or processing problems; and iv.) the values of peak acceleration and velocity ground motion parameters. The user, depending on desiderata, can customize the selection criteria. In Table 3 we provide a basic quantification of the distribution of the relevant metadata shown in Fig. 15 and, in Table 4, we present the distribution of the values of the maximum horizontal ground acceleration and velocity expressed as % g and $cm\ s^{-1}$, respectively. Some of the values reported in these two tables are used for our trace selections.

In Fig. 15 (a-c) and Fig. 15 (d-f) we show some traces that have been selected from the HH channels according to the number of P- and S-wave onset picks greater than 3 detected through the application of the GPD and EQTransformer techniques, respectively. Based on the values reported in Table 3, the presence of these multiple event traces in the dataset is less than the 10 %. In Fig. 15 (g-i), we focus attention on the traces that feature SNR values on at least one component within the lowest 10 % of the dataset. These traces are good examples of noisy traces and low amplitude event signals. In Fig. 15 (j-l), we plot three traces for which the median values of all the three components fall within the two 10 % extremes. They represent about 6 % of the entire HH channels dataset. In the bottom row of Fig. 15 (m-o), we show that excluding the very low 25 % of

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

Table (3). Distribution according to different quantiles of selected metadata 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

SNR values from the previous selection (Fig. 15 (j-l)) it is possible to select traces that do not suffer of particular problems. In particular, we see that just by selecting a higher threshold of SNR values, about 85 % of the first and last 10 % of the distribution of median values, the traces appear acceptable.

Table (4). Distribution according to different quantiles of IMs selected metadata for the HH, EH and HN channels

Metadata parameter-name	10 %	25 %	50 %	75 %	90 %	max
`trace_pga_perc` (HH)	0.000219	0.000506	0.00148	0.00532	0.0201	57.6
`trace_pgv_cmps` (HH)	7.81e-05	0.000158	0.000413	0.00139	0.00499	58.6
`trace_pga_perc` (EH)	0.000341	0.000927	0.00328	0.0139	0.0559	71
`trace_pgv_cmps` (EH)	0.000174	0.00035	0.000882	0.00302	0.0114	55.6
`trace_pga_perc` (HN)	0.000329	0.000875	0.00328	0.0141	0.0509	77.8
`trace_pgv_cmps` (HN)	0.000392	0.000582	0.00134	0.00501	0.0153	59.1

In Fig. 16, we show the instrument corrected traces randomly chosen in groups of 6 for each channel. The traces drawn from the entire dataset belong to the 75 % with the largest values of the maximum horizontal acceleration (i.e., second, third and fourth quartile of the value distribution, cf. Table 4). The total of traces satisfying this criterion amounts to more than 860,000 3C traces. Application of the instrument transfer function appears to be generally successful without introduction of particular side effects with the exception of some amplification of the very low frequencies for some very low amplitude traces of the EH channels (e.g., panels (h,k) in Fig. 16). This effect results from our choice to bandpass filter all the traces channels in the same frequency range: this has the negative effect of boosting the low frequencies of the narrower band EH channels although it can be promptly removed by high-pass filtering. Overall, the quality of the ground motion units dataset can be considered of satisfactory quality to perform analysis of ground motions.

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

Open Access Earth System Science Data Discussions

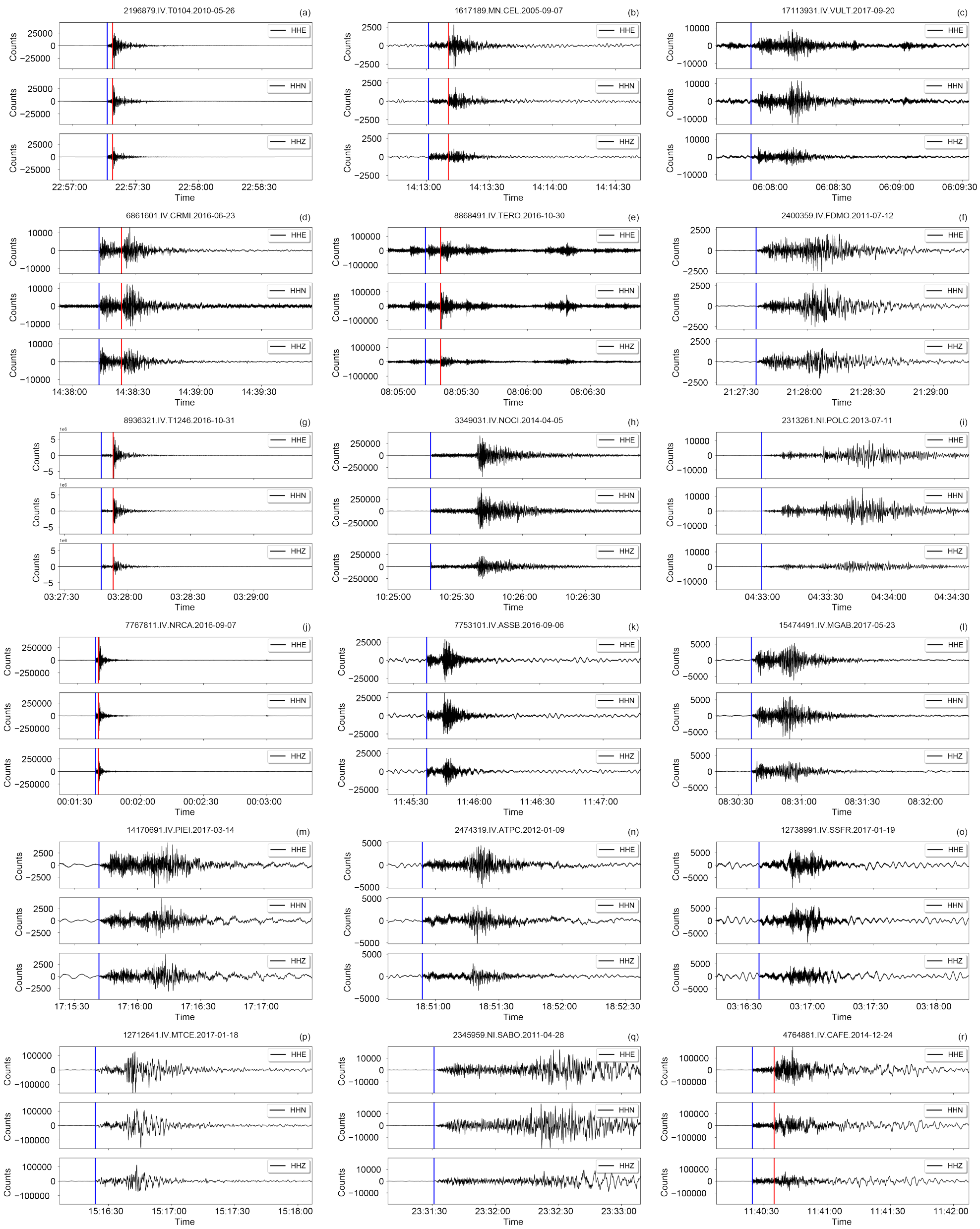

Figure (14). Example of randomly selected earthquake waveforms of the broadband HH channels contained in INSTANCE. Each row contains three randomly selected traces drawn according to the following criteria: (a-c) earthquakes $2 \leq M < 3$ (66.8 % of the total of the HH channels); (d-f) earthquakes $3 \leq M < 4$ (13.5 %); (g-i) earthquakes $M \geq 4$ (2.0 %); (j-l) earthquakes `trace_E_snr_db` ≥ 10 and `path_ep_distance` < 100 km (55.0 %); (m-o) earthquakes `trace_E_snr_db` ≥ 10 and `path_ep_distance` ≥ 100 km (10.8 %); (p-r) earthquakes $M \geq 4$ and `trace_E_snr_db` ≥ 10 (1.7 %). The arrival times of P- and S-wave onsets are shown by blue and red vertical lines, respectively.

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

Open Access Earth System Science Data Discussions

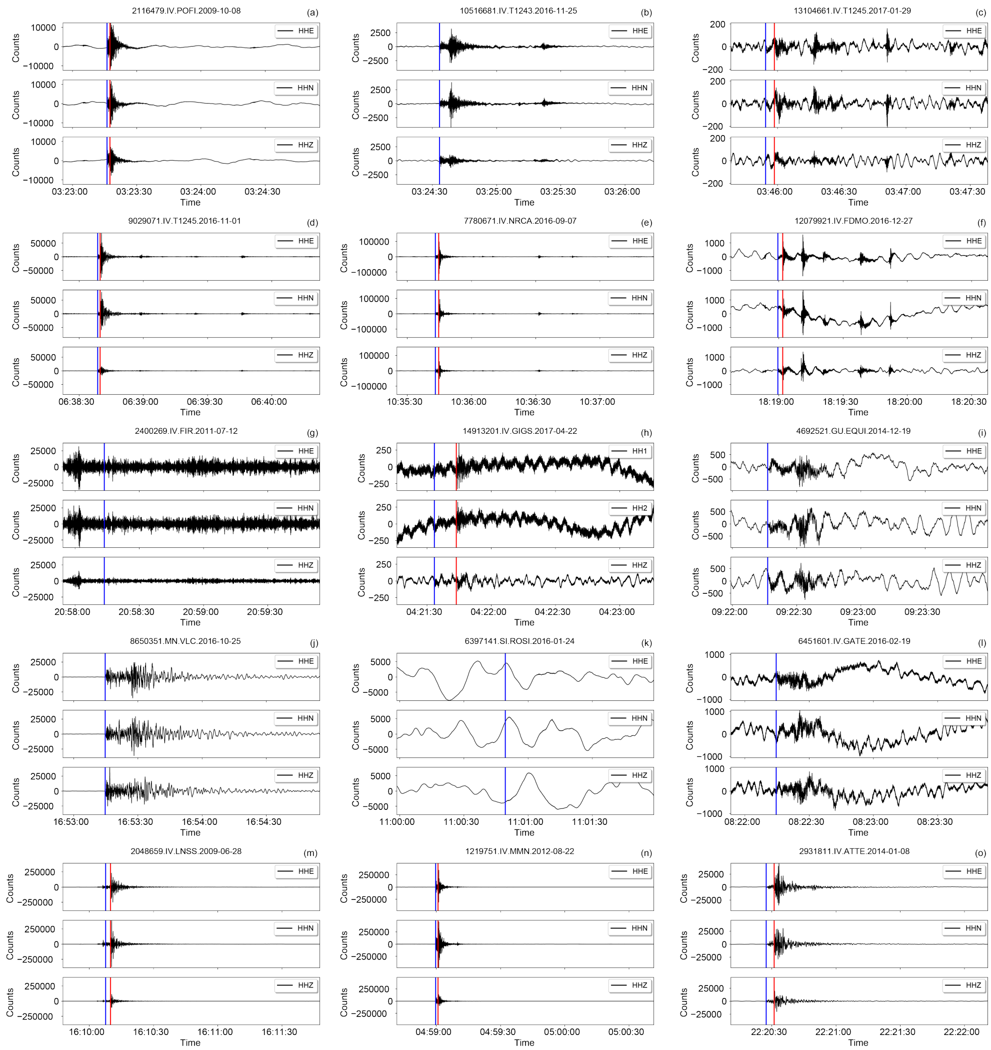

Figure (15). Example of randomly selected problematic earthquake waveforms of the broadband HH channels. Each row contains three randomly selected traces drawn according to the following criteria: (a-c) traces with `trace_GPD_[P,S]_number` > 3 (7.96 % of the total of the HH channels); (d-f) traces with `trace_EQT_number_detections` > 3 (0.38 % of the total of the HH channels); (g-i) traces `trace_[ENZ]_snr_db` with at least one component in the 10 % quantile (18.10 % of the total of the HH channels); (j-l) traces with all `trace_[ENZ]_median_counts` either in the first 10 % or the last 10 % quantiles (5.90 % of the total of the HH channels); (m-o) traces with `trace_[ENZ]_median_counts` either in the first 10 % or the last 10 % quantiles and corresponding `trace_[ENZ]_snr_db` excluded from the first quartile (5.06 % HH dataset). The arrival times of P- and S-wave onsets are shown by blue and red vertical lines, respectively.

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

Open Access Earth System Science Data Discussions

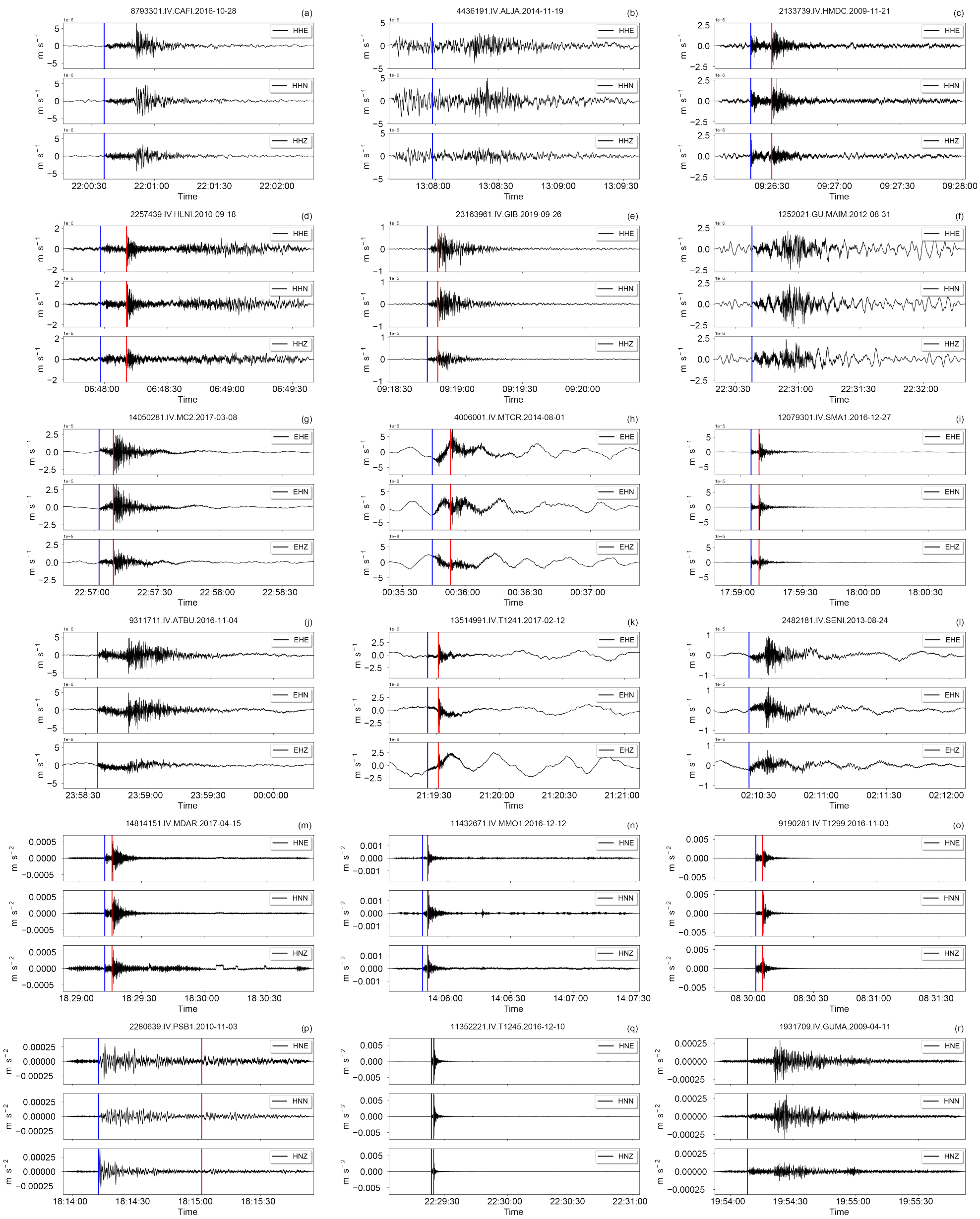

Figure (16). Example of randomly selected event waveforms in ground motion units of the HH, EH and HN channels in INSTANCE. The traces are representative of 75 % of the data and belonging to the second, third and fourth quartiles of each channel. Each row contains three randomly selected traces drawn as follows: (a-f) HH traces with `trace_pga_perc` > 5.1e-4 % g; (g-l) EH traces with `trace_pga_perc` > 9.3e-4 % g; (m-r) HN traces with `trace_pga_perc` > 8.7e-4 % g; The arrival times of P- and S-wave onsets are shown by blue and red vertical lines, respectively.

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

3 Noise

Noise is generated by many different sources such as ocean waves, wind, traffic, instrumental noise, electrical noise, etc. and its suppression in earthquake recordings represent a long standing objective (Zhu et al., 2019b, and references therein). The inclusion of noise data in a dataset like INSTANCE is thus important because it provides information on the noise characteristics of the individual stations in the absence of earthquake generated signal. ML models can reveal to be effective for noise removal or, in a classification analysis, for improving the detection of earthquakes. The noise data have been assembled starting from the stations gathered in the event selection stage described above.

3.1 Data Preparation

Starting from the entire catalogue consisting of more than 300,000 events (Table 1), we first identified 600 s long time windows free of any earthquake. Secondly, we obtained the operational times of acquisition of each station. The third step consisted of identifying the 120 s time windows to be included in the dataset for each station and channel. This was achieved by intersecting the time window series obtained in the previous two stages (i.e., the event free windows and the periods of station acquisition). For stations acquiring more than one channel type (e.g., HH and HN), noise windows for all the channels were identified and downloaded. The resulting total number of noise trace windows is 132,288 corresponding to about 10 % of the total number of traces of INSTANCE.

3.2 Metadata description

The 46 metadata (Table 2) used for the noise data selection include for each 3C waveform trace an identifier based on the start time, the *station* parameters, the *trace* quality control that include the automatic picks and event detection obtained using the GPD and EQTransformer procedures. These picks provide potential insights on whether any earthquake not catalogued in the INGV bulletin might be present in the selected time windows.

In Fig.17 we show the channel subdivision of the downloaded noise together with the networks the stations belong to. In Fig. 18 and 19, in analogy with what presented for the event data, we present the *trace* characteristics of the metadata. The `trace_[E,N,Z]_mean_counts` and `trace_[E,N,Z]_median_count` provide an outlook on the distribution of the mean and median values and, likewise the same parameters extracted from the event traces, could be used to identify high quality data. The histograms of the `trace_[E,N,Z]_rms_counts` noise values fall mainly in the range of values from 0 to 2000 counts with similar peak values for either `trace_[E,N,Z]_max_counts` or `trace_[E,N,Z]_min_counts`. This all would suggest that the gathered noise traces are of fairly good quality responding to the expectation of traces characterized by amplitudes with small number of counts.

3.3 Examples of noise data traces

Examples of the noise traces are shown in Figure 20. To perform the selection, we have used the distribution of the trace *rms* values that is provided in Table 5.

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

Open Access Earth System Science Data Discussions

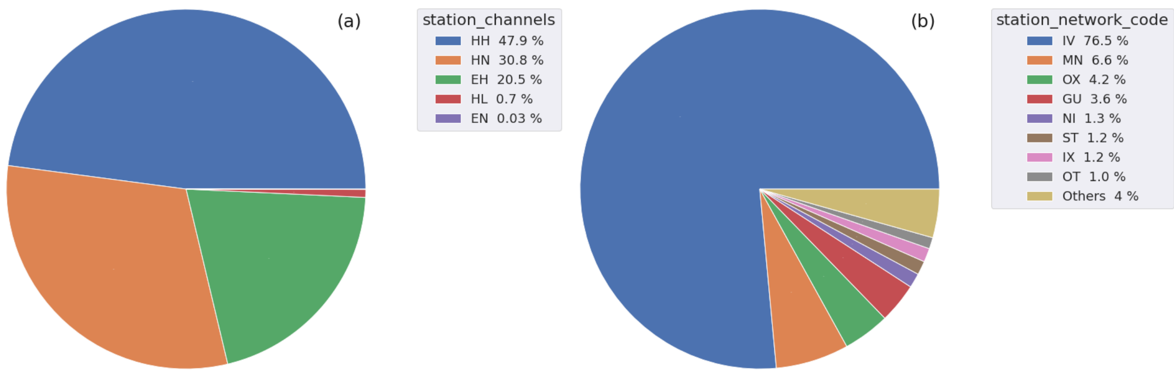

Figure (17). Pie diagrams summarizing the distribution of the channels (a) and the data contributing networks (b) of the noise dataset. The full list of `station_network_code` with % < 1 collected in Others in decreasing order is SI, YD, 3A, XO, ZM, BA, AC, HL, TV, RF.

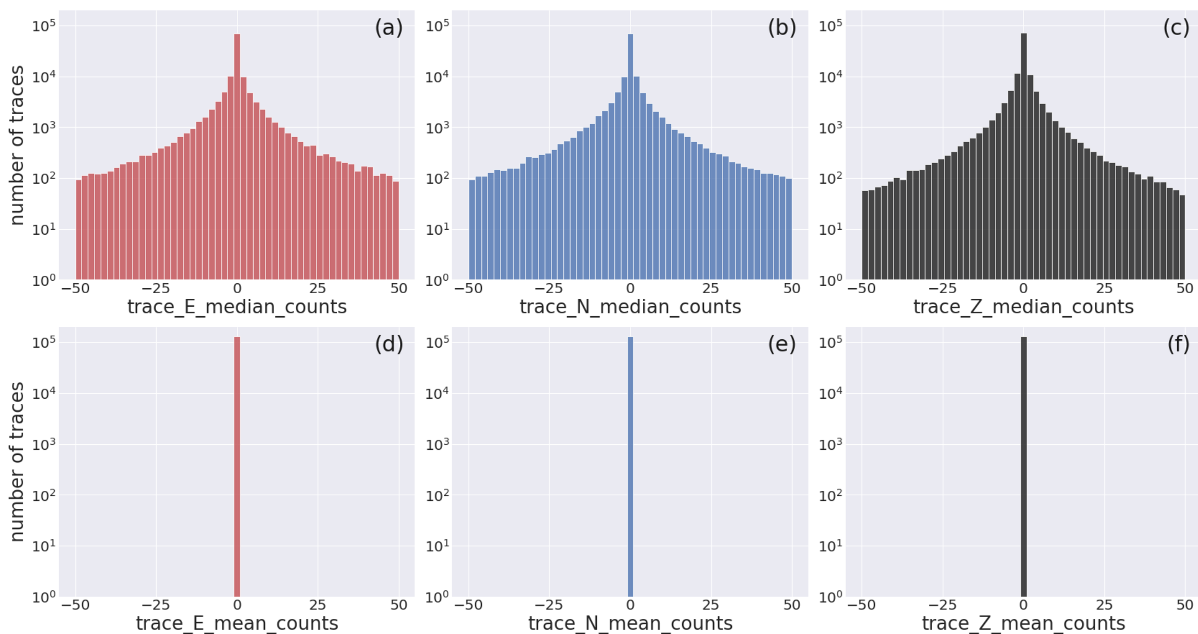

Figure (18). Histogram of the distribution of the noise quality control metadata: median (a-c) and mean (d-f). The width of the bins is 2×10^3. The complete distribution is provided in Fig. A5.

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

Earth System Science Data
Open Access
Discussions

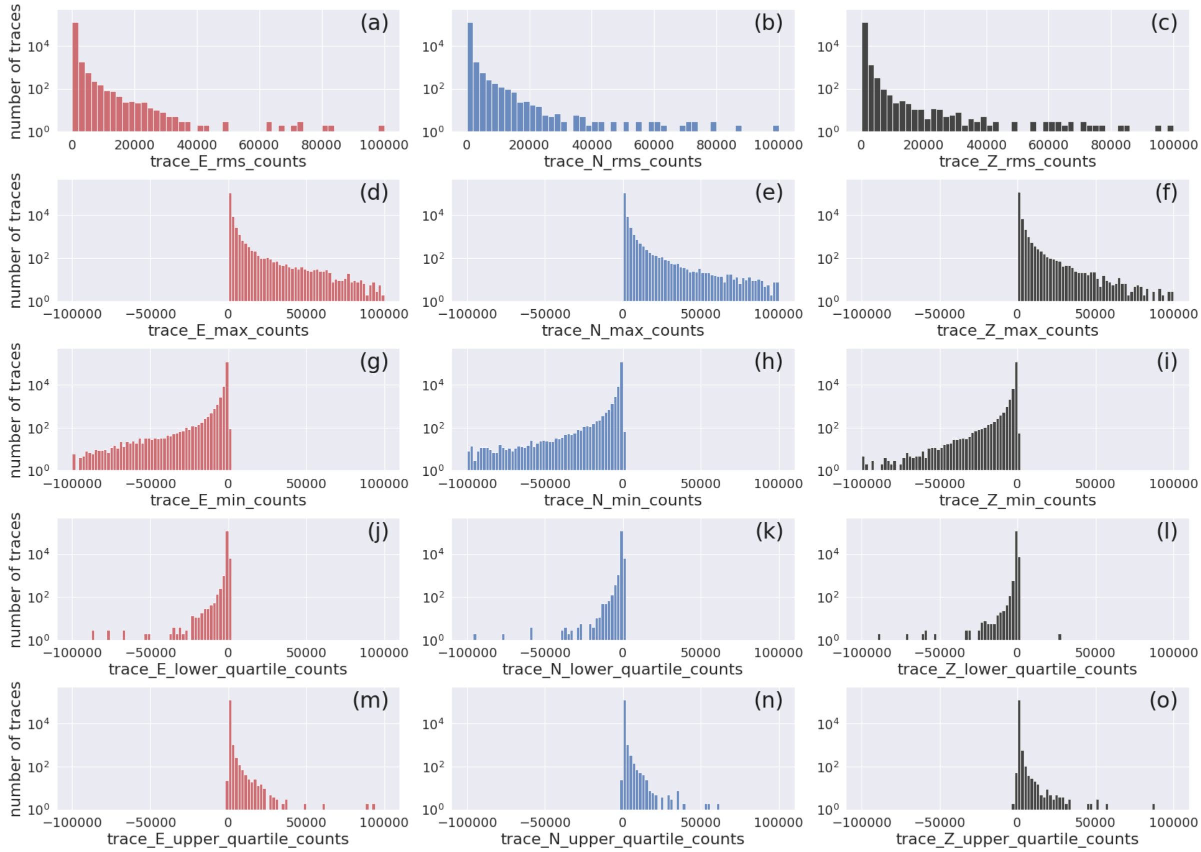

Figure (19). Histogram of the distribution of the noise quality control metadata: rms, min, max, first and third quartile. The width of the bins is 2×10^3. The complete distribution is provided in Fig. A6.

In the top two rows of Fig. 20, there are shown some examples of events detected using the GPD and EQTransformer algorithms on the EH channels. As it was the case with the event dataset, the noise traces also contain undetected events although their number according to our analysis seems rather small especially for the earthquakes detected by EQTransformer. This result gives us good confidence that the noise traces are for the great majority free of earthquake events. The following rows of Fig. 20 provide waveform samples drawn from the 90 % of the dataset (panels g-i and m-o) for the HH and EH channels, respectively. Both sets of panels exemplify some of the features of the great majority of the noise data. In contrast, the panels (j-l) and (p-r) have been chosen to show what could be considered traces exceeding noise values or that contain finite duration events of uncertain origin.

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

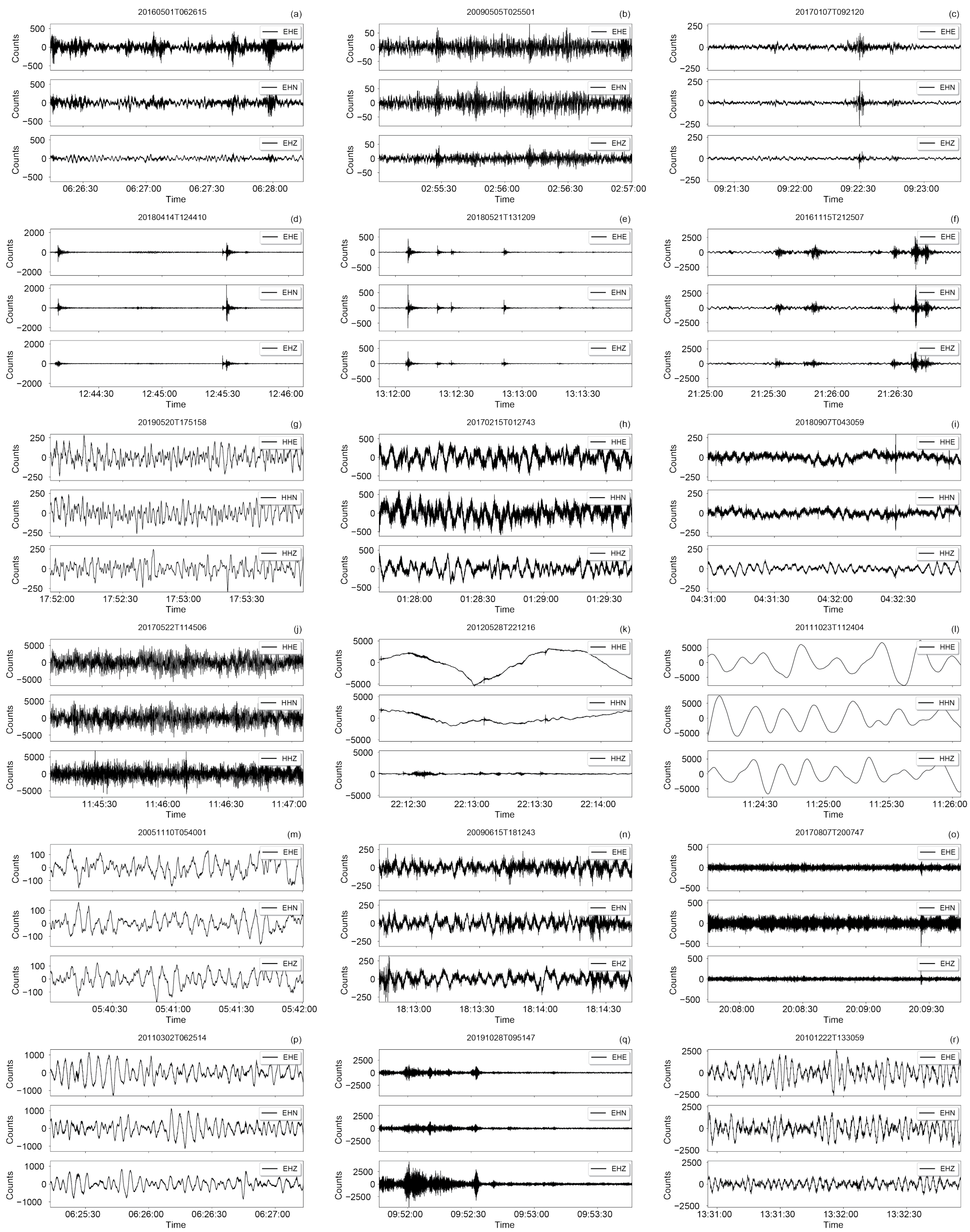

Figure (20). Example of randomly selected noise waveforms of the HH and EH channels contained in INSTANCE. Each row contains three randomly selected traces drawn according to the following criteria: (a-c) `trace_GPD_[P,S]_number` > 3 (11.6 % of the total of the EH channels); (d-f) `trace_EQT_number_detections` > 3 (0.13 % of the total of the EH channels); (g-i) all the `trace_[E,N,Z]_rms_counts` < [1013,1071,793] (86.31 % of the total of the HH channels); (j-l) any of the `trace_[E,N,Z]_rms_counts` > [1013,1071,793] (13.69 % of the total of the HH channels); (m-o) all the `trace_[E,N,Z]_rms_counts` < [327.1,332,307] (86.36 % of the total of the EH channels); (p-r) any of the `trace_[E,N,Z]_rms_counts` > [327.1,332,307] (13.64 % of the total of the EH channels);

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

Table (5). Distribution according to different quantiles of rms selected noise metadata 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

4 Discussion

The primary objective of this work has been to assemble a benchmark dataset consisting of seismic waveforms and associated metadata. It has been designed to be used for the analysis of earthquakes in Italy (and neighboring areas) using ML techniques and it could prove useful for ML analysis also elsewhere in other active tectonic regions by adopting transfer learning methodologies (Jozinović et al., 2021). The dataset consists of three HDF5 volumes — raw and instrument removed event traces, and raw noise traces — and of the associated metadata.

The selection of the waveform traces to be included was based on the availability of low ($\leq$ 1 s) P- and S-phase location residual times and large location weights taken from the *preferred* solutions listed in the Italian Seismic Bulletin (http://terremoti.ingv.it/en/help#BSI). To counteract the Gutenberg-Richter power law which affects the compilation of seismological datasets targeting ML analysis applications, attention was paid towards assembling a dataset that was not completely skewed by a large number of small magnitude earthquakes. To this end, we included all the traces available of the larger size earthquakes and then we decreased progressively the number of smaller size earthquakes and associated traces. Our effort, however, trades-off with the need of assembling a dataset that is sufficiently large for ML purposes. The distribution of the selected traces shown in Fig. 3 and Fig. 4 according to magnitude, distance and focal depth allows the users to make the appropriate choices for their purposes even though we recognize that the achievement of the sought balanced distribution remains difficult. Other data selection criteria could have been used (e.g., select all the data acquired within distance ranges depending on earthquake magnitude) but the (un)balanced magnitude and distance distribution would have persisted. Thus, given the criteria adopted it

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

Open Access Earth System Science Data Discussions

is pleonastic to remark that this dataset is not designed for studies addressing the earthquake magnitude power-law distribution (e.g., the b-value).

Our criterion, based on the available high-quality P- and S-phases with low location residual times, is expected to provide a large number of traces with distinct earthquake signal and high SNR ratios. The distribution of SNR values shown in Fig. 10 and 11 and in Table 3, and the example seismograms shown in Fig. 14(j-r) and Fig. 15(m-o), appear to confirm our choices.

The selection of 120 s trace length time window is longer than those made by other authors for analogous benchmark datasets (e.g., Mousavi et al., 2019; Magrini et al., 2020). This relatively long time window was required, however, because we sought to include the entire seismicity occurring in Italy that spans from very shallow to very deep (Fig. 3). Unfortunately, this long window trades-off with a higher probability of including earthquakes close in time that had not been reported in the INGV catalogue. For this reason, we carried out also a (preliminary) automatic picking and earthquake detection analysis using two well established recent ML techniques (GPD and EQTransformer; Ross et al., 2018a; Mousavi et al., 2020) to possibly isolate those traces that include multiple events. The results of this analysis summarized in Table 3 would indicate that about 90 % of the event data contain only one earthquake according to the EQTransformer analysis whereas the GPD analysis returned some slightly higher numbers of P- and S-phase detections.

Our metadata for the earthquake part of the dataset consist of more than 100 parameters. They are subdivided into three main classes plus one additional class derived from the previous ones. This is a rather rich set of parameters that can be used either to select subsets of the dataset, or as additional features to rely on when developing ML models, or as labels in supervised ML analysis, or for unsupervised ML applications. In addition, the metadata could be used by themselves for specific studies (e.g., seismic velocity model regionalization, travel time tomography, ground motion prediction models, local site corrections, ...).

Earthquake data gathered by seismic instruments and streamed in realtime to earthquake monitoring centers or preserved within archives can suffer from problems of different nature (e.g., sensor, data logger, equipment installation, data transmission, and processing among the most common). Thus, the compiled dataset could be useful for the development of robust techniques of analysis and this is one main reason for including several *trace* quality parameters as metadata since they can help the user to identify the possibly “faulty” records which can be then either removed or included to train the ML model just to “learn” them. This approach may seem to contradict one of the main purposes of compiling a high-quality dataset and it may also be an obstacle when attempting to reveal deep information therein but the expectation is that, by including all the data together, the rich set of metadata leaves the users enough freedom to identify the “good” data for their purposes. It is worthwhile to mention that Yeck et al. (2020) have found that inclusion of only “good”, high SNR trace data during training of various body waves resulted in lower performance when applied to real-time pick data.

The INSTANCE dataset includes intensity measures (i.e., PGA, PGV, spectral accelerations) obtained after deconvolution of the instrument response performed automatically and possibly affected by digital signal processing problems induced, for example, by the presence of abnormal drifts and spikes. Given the difficulty to verify the quality of all the individual processed traces, the availability of a rich set of trace metadata can be useful (again) to detect the faulty traces.

The example traces drawn randomly from the dataset that we have presented in Figs. 14, 15, 16 and 20 provide some evidence of the characteristics of the traces contained in the dataset and how they can be promptly selected through the provided

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

metadata. Although the great majority of the data appear to be of very good quality, we are also aware that low quality data are almost inevitable to occur. Inspection of the waveform traces by using other selection criteria than those shown here, and of the IM metadata (cf. Fig. 12) give us, however, good confidence on an overall good quality of the dataset.

The INSTANCE data collection assembles for the first time a very large amount of earthquake and noise data throughout Italy. If on one side this might seem a limitation when compared to other recent data collections like STEAD and LEN-DB that have gathered data globally, on the other hand, the dataset can be considered a representative subset of the seismicity in Italy and neighbouring areas. The dataset equals to more than 43,000 hours of continuous event and noise data and associated metadata. An average of 21 3C traces are provided per earthquake. In summary, the dataset features strengths such as the prompt availability of a large number of records assembled within a ready-to-use data volume that can be certainly considered representative of the whole waveform data archive of the INGV ORFEUS-EIDA node and that can be used for many diverse studies. In our opinion, the strengths of providing a diversified set of data outnumber the weaknesses and the latter ones can be isolated, and their negative contribution reduced through the exploitation of the very rich set of metadata.

5 Applications

To the purpose of describing the range of possible applications of INSTANCE we follow the basic exposition schema adopted by Mousavi et al. (2019) for the STEAD benchmark dataset. These authors addressed four main areas in which benchmark datasets can reveal very effective for improving seismological knowledge and seismic monitoring operational activities: earthquake trace denoising, earthquake detection and onset picking, classification/discrimination and direct earthquake characterization.

The seismic noise level at a station is frequency dependent and derives from many factors such as types of equipment, installation, meteorological conditions, anthropic generated noise, geography, season, and time of day (McNamara and Buland, 2004). Seismic trace denoising enhances the SNR that is crucial to lowering the magnitude detection level of earthquake catalogs and, by so doing, increase the number events detected. Analogously, denoising can be relevant to pre-process seismic traces when performing ambient noise cross-correlation analysis (e.g. Baig et al., 2009), or for detecting speed-of-light changes of the gravitational field (Vallée et al., 2017), or for the analysis of seismic data acquired in urban areas (e.g. Parolai, 2009) just to mention a few among many applications. ML techniques seem very promising to address the reduction of noise in seismic data. For example, Zhu et al. (2019b) (and references therein for a list of applications in applied geophysics and seismology) have proposed a denoising/decomposition method, DeepDenoiser, based on a deep neural network which is based on the adoption of signal and noise masks which are then used to effectively decompose the input data into a signal of interest and noise. The technique has been tested against a dataset composed of broadband recordings of the North California Seismic Network which is similar to the data of INSTANCE. The adoption of unsupervised machine-learning method has been instead advocated by Chen et al. (2019) who have proposed it in combination with an autoencoder algorithm that adaptively learns the features from the raw noisy seismological datasets and use the sparse constraint to suppress the learned trivial features that may be associated with partial noise component. They apply the technique to the waveform stacked data used in Shearer (1991) and similar stacks can be promptly prepared using INSTANCE at the local/regional scale and apply the denoising technique accordingly.

Earthquake detection (including phase picking), discrimination and rapid characterization represent main pillars of seismic monitoring and surveillance. During their lifetime, operational seismic centers alternate calm periods characterized by low levels of seismicity, in which it can become of relevance the detection of even the smallest possible events to delineate the activation of often hidden tectonic structures, to paroxysmal periods starting with significant earthquakes and followed by hundreds or thousands of aftershocks felt by people. To ameliorate the response of the centers in both these extreme cases, we find that the INSTANCE dataset can be of importance to calibrate and benchmark methodologies for i.) phase onset picking and earthquake detection methods to lower the magnitude detection level (e.g., Ross et al., 2018a; Zhu et al., 2019a; Walter et al., 2020; Mousavi et al., 2020, amongst others); ii.) discriminate between volcanic and tectonic earthquakes (e.g., Esposito et al., 2006) and, in the future, after updating INSTANCE with new data and metadata, discriminate earthquakes and other sources of seismic energy (e.g., sonic booms, quarry blasts, underwater explosions) often felt by the population (e.g., Del Pezzo et al., 2003; Linville et al., 2019); iii.) the rapid and accurate characterization of the earthquake source, distance, depth (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) and of the ground shaking (e.g., Alavi, 2011; Derras et al., 2012; Derras, 2014; Jozinovic et al., 2020; Münchmeyer et al., 2020).

Indeed, the field of application of the dataset is quite extensive and it can be used to address many diverse topics depending on how the data are grouped and it can also be useful for applications not relying on ML techniques. For example, the dataset features some stations with several thousand of traces recording earthquakes from different azimuths and distances that can be used to construct common-station gathers of seismograms for swaths of sources in almost any desired geometry (e.g., Korneev et al., 2003), or to study in detail the local site response. Analogously, the metadata alone provide a rich set of arrival times (cf. Fig. 7) that could be used as is for traveltime tomography at regional scale in Italy and, in addition, the availability of the associated waveforms makes it possible the application of methodologies that resolve the velocity structure jointly using arrival times and waveform data (e.g., Zhang* and Chen, 2014). For what concerns the ground motion amplitude data, the availability of these metadata can be of relevance in combination with the shakemaps (http://shakemap.ingv.it) to develop new tools for rapid earthquake ground motion estimation. Other applications of the data collection include the adoption of unsupervised ML algorithms to group the waveforms independently of the earthquake location and just on the waveform themselves (e.g., Seydoux et al., 2020). INSTANCE can also be used, as a dataset with a large number of data, for creating pre-trained models when using transfer learning techniques either for seismological or other applications which use time-series data (Otović et al., 2021).

Overall, we believe that the dataset will be useful for stepping up towards a new generation of earthquake monitoring tools that will profit of the ongoing very fast developments in machine learning. What is certain is that seismology is in great need of benchmark datasets (Mousavi et al., 2019) upon which test new and existing techniques. To this end, standardization of the input data and metadata formats is of great relevance and in constructing this dataset we have adopted the schema proposed by the Sesbench initiative (Wollam et al., 2021). Widespread adoption of the same metadata schema and data volume formats

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

Open Access Earth System Science Data Discussions

can foster the compilation of similar datasets also for other regions with the possibility to merge them all together. Perhaps more importantly, standardization of data and metadata formats will make it easier to test different datasets using the same ML model or, alternatively, benchmarking different models on the same dataset and in both cases the benefits appear clear.

6 Data availability

The dataset can be downloaded from http://doi.org/10.13127/instance (Michelini et al., 2021). A sample dataset is also provided on the same landing page.

7 Code and data availability

Routines and notebooks for analysis and display of the dataset (and the sample dataset) are available on https://github.com/cjunkk/instance.

The data used in this work were downloaded using the webservices provided by INGV (http://terremoti.ingv.it/en/webservices_and_software), the processing was performed using the *Obspy* python modules (Beyreuther et al., 2010; Megies et al., 2011; Krischer et al., 2015) and the graphics was prepared using the *Matplotlib* library (Hunter, 2007).

8 Conclusions

INSTANCE is the first dataset designed for the application of ML methodologies compiled using the seismic data archived on the ORFEUS-EIDA node of INGV for Italy (Danecek et al., 2021). One of the main scopes of the dataset is to provide a benchmark for developing improved techniques for earthquake and ground motion characterization. The dataset consists of about 1.3 M, 120 s long each, 100 Hz sampling, 3C traces subdivided into about 1.2 M containing seismic events and more than 100,000 that include noise. The traces are assembled within HDF5 formatted volumes to facilitate access and analysis. More than 100 metadata grouped according to *source*, *station*, *trace* (and derived quantities) are associated to each trace to give the user much flexibility and control for the selection of the most appropriate data for her/his scientific targets.

The event data include recordings of earthquakes in the magnitude range $0 \leq M \leq 6.5$ and in the distance range between 0 and more than 600 km although the great majority of the traces belong to earthquakes in the magnitude range $2 \leq M \leq 3$ and within 250 km. The depth of the earthquakes varies between very shallow crustal earthquakes and about 600 km depth in the Calabrian subduction slab. The data have been recorded by more than 600 stations operating in Italy in the time span January 2005 — January 2020. The dataset equals to more than 43,000 hours of continuous event and noise data and associated metadata. An average of 21 3C traces are provided per earthquake.

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

Open Access Earth System Science Data Discussions

Appendix A: Preparation of the HDF5 data containers

The waveform trace data are provided in binary HDF5 files volumes. The volumes have been prepared for event data in counts and ground motion units, and for noise data in counts. The HDF5 format allows for rapid and easy access to the individual traces without the need of loading the whole dataset into memory. The waveform trace datasets have been created using the HDF5-Group "data" structure https://portal.hdfgroup.org/display/HDF5/HDF5 which contains as many HDF5-Datasets as 3C waveforms. Every 3-component waveform is a separate HDF5-Dataset and is accessed by its trace name (`trace_name`) found in the metadata file.

Appendix B: Additional Quality Control figures

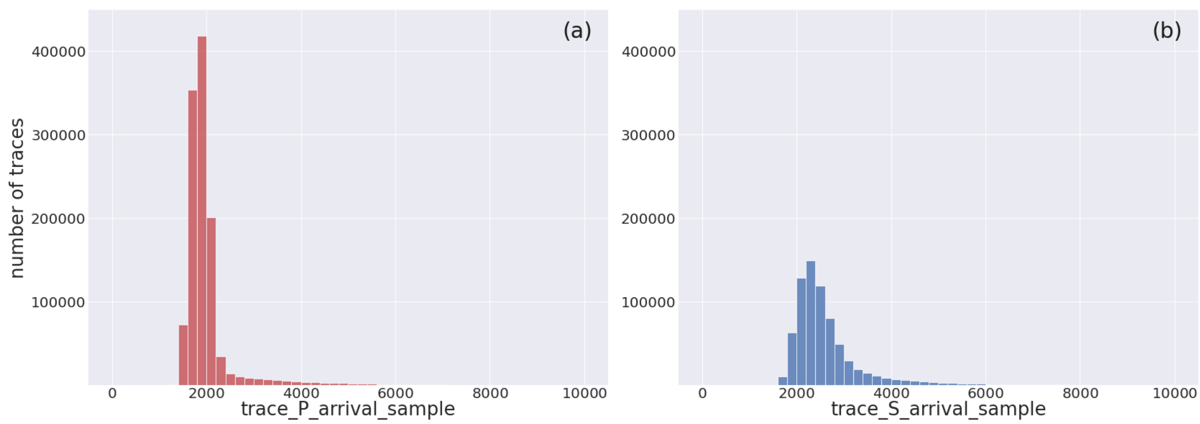

Figure (A1). Distribution of P- (left) and S-arrival (right) samples of the extracted waveform traces.

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

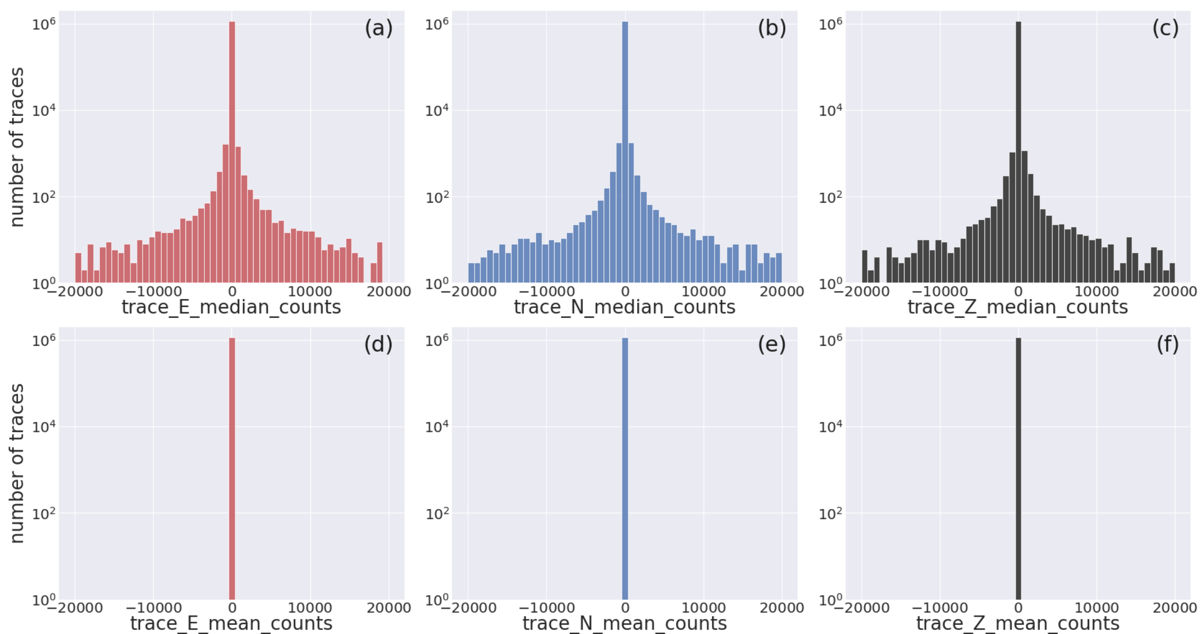

Figure (A2). Histogram of the distribution of the quality control metadata of the events: median (a-c) and mean (d-f). The width of the bins is 2×10^3 .

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

Open Access Earth System Science Data Discussions

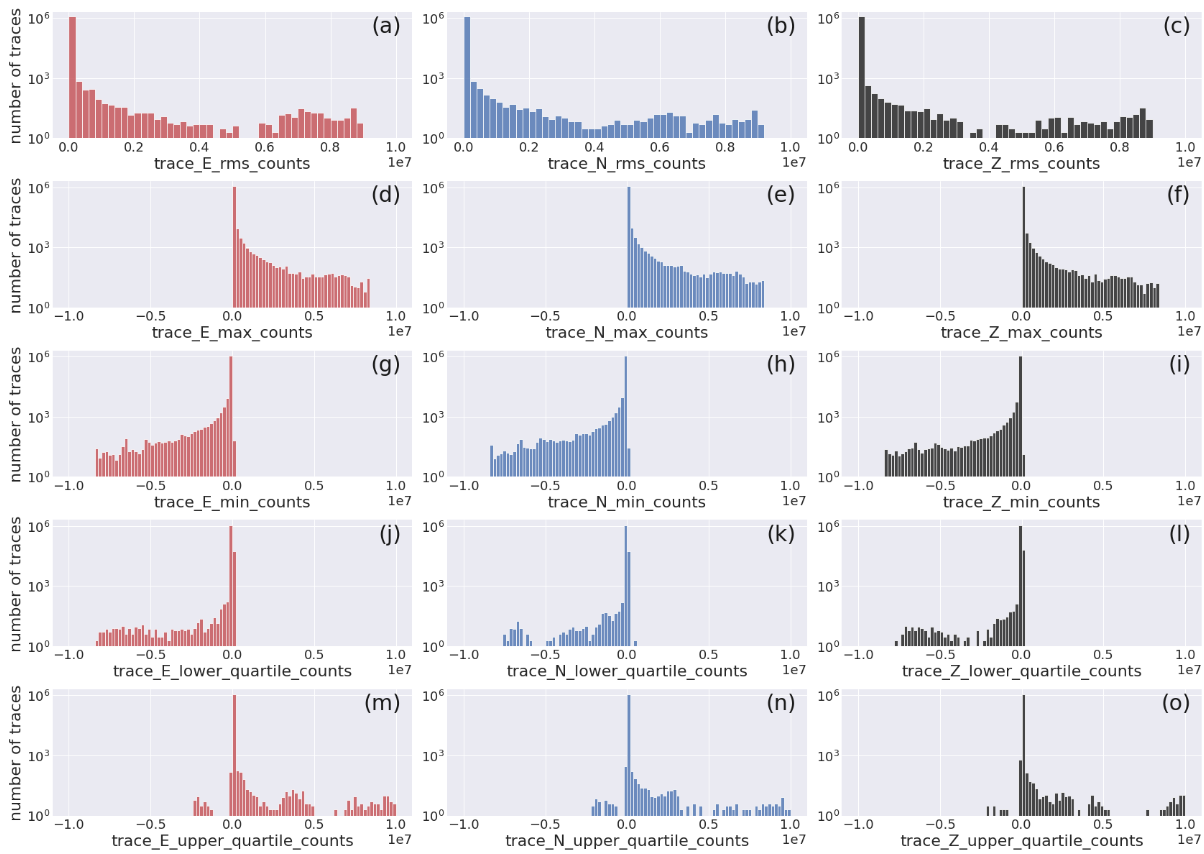

Figure (A3). Histogram of the full distribution of quality control metadata of the event waveforms: rms, min, max, first and third quartile. The width of the bins is 2×10^5 .

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

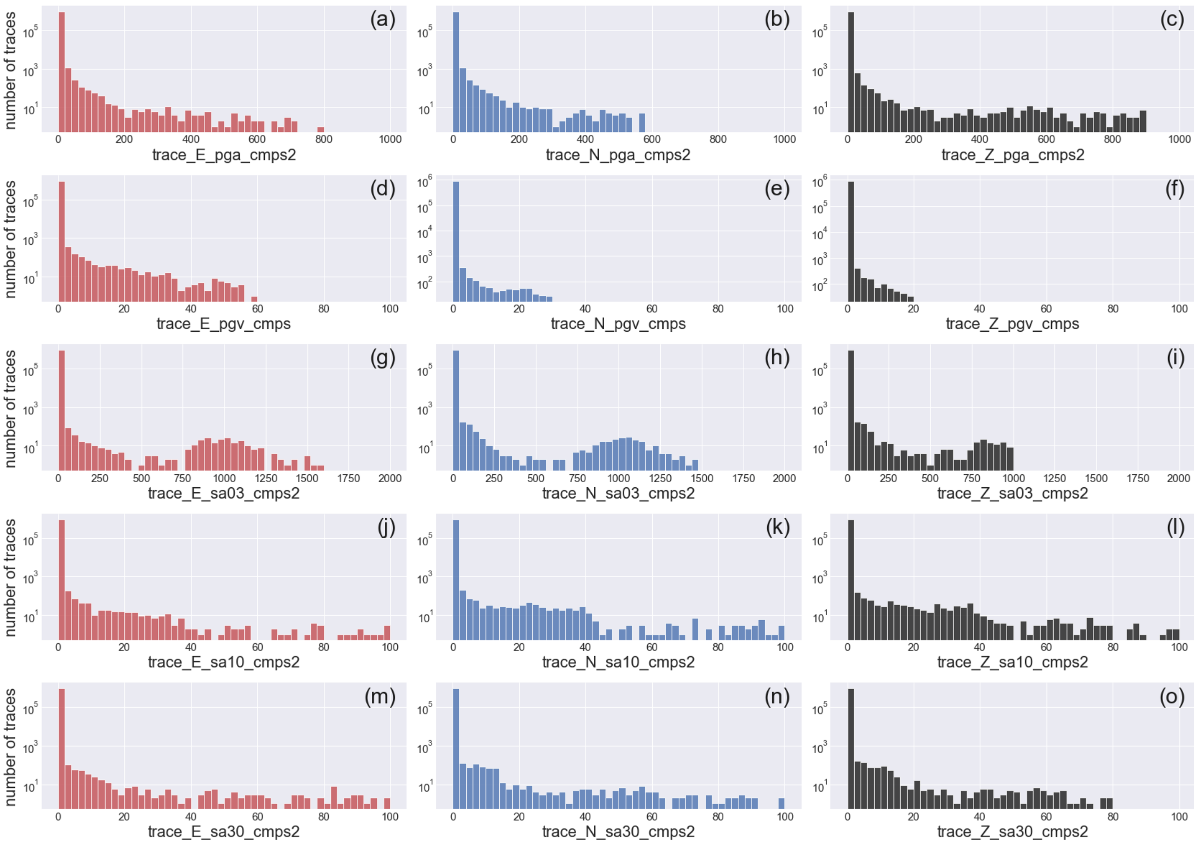

Figure (A4). Histograms of the distribution of the intensity measures (IMs) of the event waveforms for $M \geq 2$ earthquakes.

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

Open Access Earth System Science Data Discussions

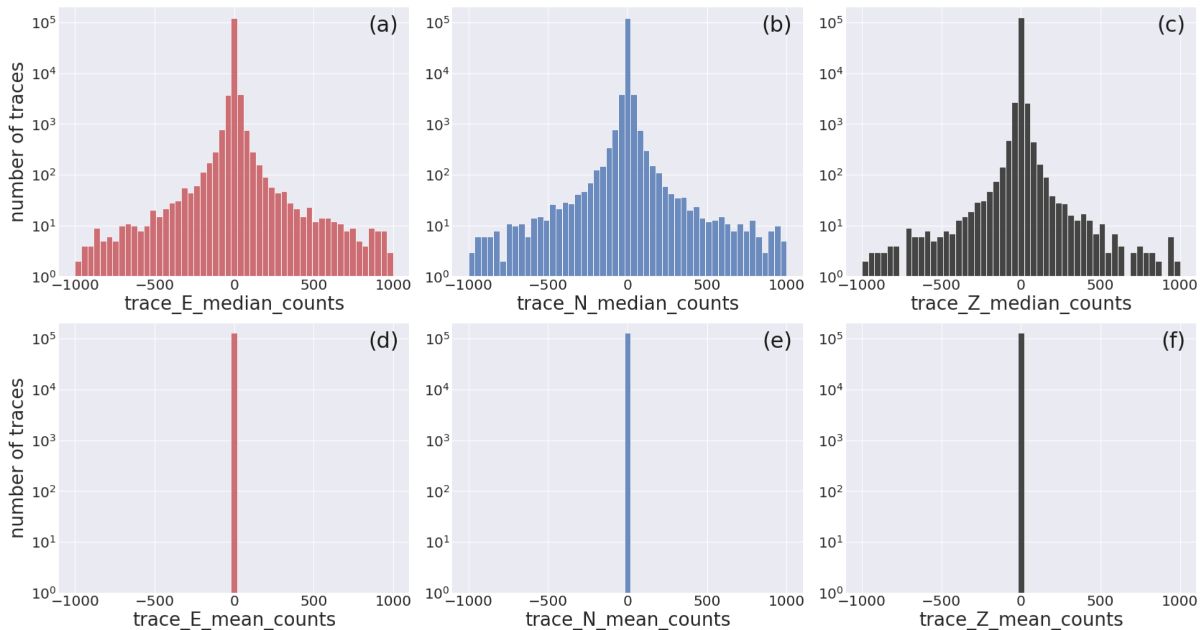

Figure (A5). Histogram of the distribution of the noise quality control metadata: median (a-c) and mean (d-f). The width of the bins is 2×10^3
.

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

Open Access Earth System Science Data Discussions

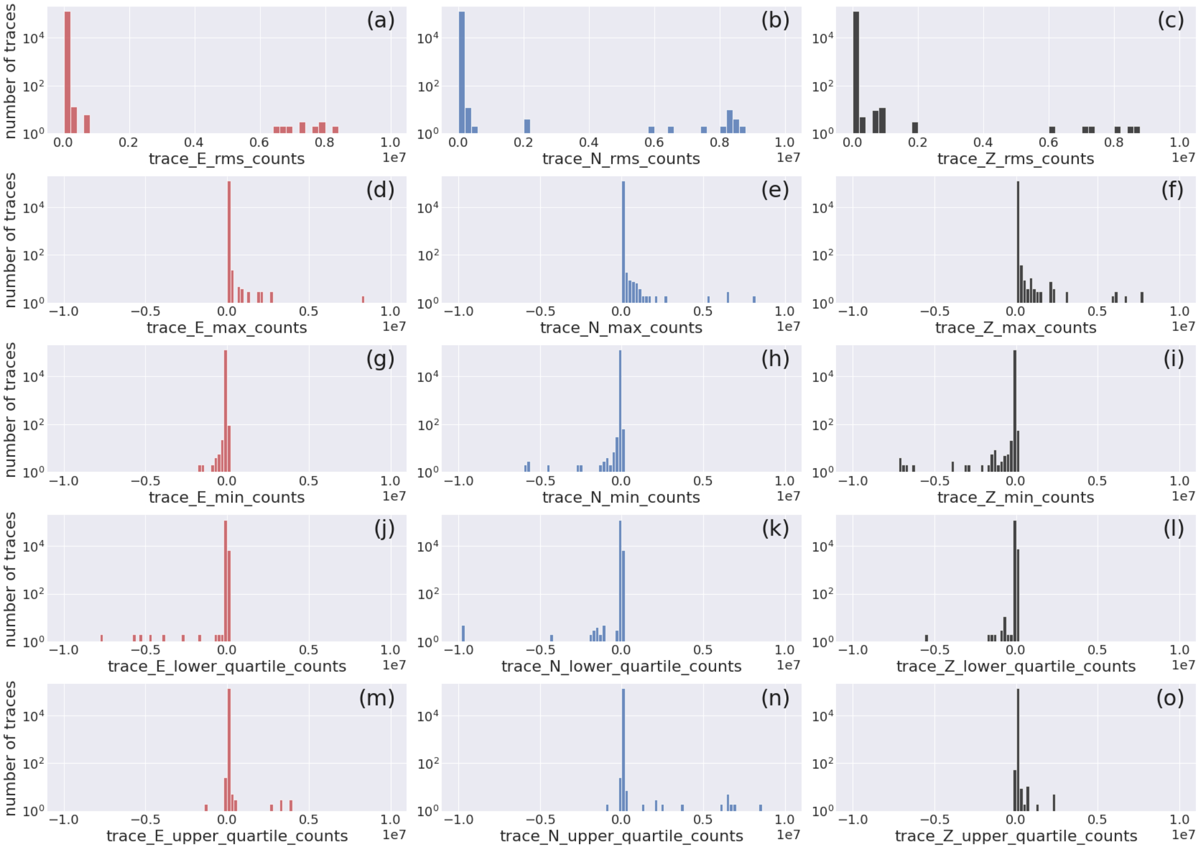

Figure (A6). Histogram of the distribution of noise quality control metadata: rms, min, max, first and third quartile. The width of the bins is 2×10^5 .

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

Open Access Earth System Science Data Discussions

Author contributions. AM prepared the initial data and metadata selection, drafted the manuscript and initial version of the figures. SC and SG provided ML estimation of the waveforms data, contributed to final versions of the figures and manuscript revision. CG contributed to data preparation and manuscript revision. DJ compiled the HDF5 formatted volumes and contributed to the data and manuscript revision. VL set up the virtual machines and the scripts to perform the massive data download.

Competing interests. No competing interests are present.

Disclaimer. The authors decline any responsibility for possible errors present in the INSTANCE dataset which can lead erroneous evaluations and to physical and economic damages.

Acknowledgements. 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. We also greatly acknowledge the earthquake analysts that routinely perform the data analysis for the compilation of the earthquake bulletins. The metadata nomenclature benefited from the interaction and discussions within the recently formed *SeisBench* group. We would like to thank Jannes Münchmeyer, Elisa Tinti and Anthony Lomax for reading the manuscript and providing suggestions. 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 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).

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X.: TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, arXiv:1603.04467 [cs], http://arxiv.org/abs/1603.04467, arXiv: 1603.04467, 2016.

Alavi, A. H.: Prediction of principal ground-motion parameters using a hybrid method coupling artificial neural networks and simulated annealing, Computers & Structures, 89, 2176–2194, https://doi.org/10.1016/j.compstruc.2011.08.019, 2011.

Baig, A. M., Campillo, M., and Brenguier, F.: Denoising seismic noise cross correlations, Journal of Geophysical Research, 114, 1478–12, https://doi.org/10.1029/2008JB006085, 2009.

Bergen, K. J., Johnson, P. A., de Hoop, M. V., and Beroza, G. C.: Machine learning for data-driven discovery in solid Earth geoscience, Science, 363, eaau0323–12, https://doi.org/10.1126/science.aau0323, 2019.

Beyreuther, M., Barsch, R., Krischer, L., Megies, T., Behr, Y., and Wassermann, J.: ObsPy: A Python Toolbox for Seismology, Seismological Research Letters, 81, 530–533, https://doi.org/10.1785/gssrl.81.3.530, 2010.

Boore, D. M.: Estimating Vs(30) (or NEHRP Site Classes) from Shallow Velocity Models (Depths < 30 m), Bulletin of the Seismological Society of America, 94, 591–597, https://doi.org/10.1785/0120030105, https://doi.org/10.1785/0120030105, 2004.

Chen, Y., Zhang, M., Bai, M., and Chen, W.: Improving the Signal-to-Noise Ratio of Seismological Datasets by Unsupervised Machine Learning, Seismological Research Letters, 90, 1552–1564, https://doi.org/10.1785/0220190028, 2019.

Chollet, F. and others: Keras, https://keras.io, 2015.

Danecek, P., Pintore, S., Mazza, S., Mandiello, A., Fares, M., Carluccio, I., Della Bina, E., Franceschi, D., Moretti, M., Lauciani, V., Quintiliani, M., and Michelini, A.: The Italian Node of the European Integrated Data Archive, Seismological Research Letters, 2021.

Del Pezzo, E., Esposito, A., Giudicepietro, F., Marinaro, M., Martini, M., and Scarpetta, S.: Discrimination of Earthquakes and Underwater Explosions Using Neural Networks, Bulletin of the Seismological Society of America, 93, 215–223, https://doi.org/10.1785/0120020005, https://doi.org/10.1785/0120020005, 2003.

Derras, B.: Peak Ground Acceleration Prediction Using Artificial Neural Networks Approach: Application to the Kik-Net Data, International Journal of Earthquake Engineering and Hazard Mitigation (IREHM); Vol 2, No 4 (2014), 2014.

Derras, B., Bard, P., Cotton, F., and Bekkouche, A.: Adapting the Neural Network Approach to PGA Prediction: An Example Based on the KiK-net Data, Bulletin of the Seismological Society of America, 102, 1446–1461, https://doi.org/10.1785/0120110088, https://doi.org/10.1785/0120110088, 2012.

Dramsch, J. S.: Chapter One - 70 years of machine learning in geoscience in review, in: Machine Learning in Geosciences, edited by Moseley, B. and Krischer, L., vol. 61 of *Advances in Geophysics*, pp. 1–55, Elsevier, https://doi.org/10.1016/bs.agph.2020.08.002, https://www.sciencedirect.com/science/article/pii/S0065268720300054, iSSN: 0065-2687, 2020.

Esposito, A. M., Giudicepietro, F., Scarpetta, S., D'Auria, L., Marinaro, M., and Martini, M.: Automatic Discrimination among Landslide, Explosion-Quake, and Microtremor Seismic Signals at Stromboli Volcano Using Neural Networks, Bulletin of the Seismological Society of America, 96, 1230–1240, https://doi.org/10.1785/0120050097, https://doi.org/10.1785/0120050097, 2006.

Fang, L., Wu, Z., and Song, K.: SeismOlympics, Seismological Research Letters, 88, 1429–1430, https://doi.org/10.1785/0220170134, https://doi.org/10.1785/0220170134, 2017.

Gutenberg, B. and Richter, C. F.: Frequency of earthquakes in California, Bulletin of the Seismological Society of America, 34, 185–188, 1944.

Hunter, J. D.: Matplotlib: A 2D graphics environment, Computing in Science & Engineering, 9, 90–95, https://doi.org/10.1109/MCSE.2007.55, publisher: IEEE COMPUTER SOC, 2007.

Ingate, S.: The IRIS Consortium: Community Based Facilities and Data Management for Seismology, in: Proceedings of the NATO Advanced Research Workshop on Earthquake Monitoring and Seismic Hazard Mitigation in Balkan Countries, edited by Husebye, E. S., pp. 121–132, 2008.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T.: Caffe: Convolutional Architecture for Fast Feature Embedding, arXiv:1408.5093 [cs], http://arxiv.org/abs/1408.5093, arXiv: 1408.5093, 2014.

Johnson, P. A., Rouet Leduc, B., Pyrak-Nolte, L. J., Beroza, G. C., Marone, C. J., Hulbert, C., Howard, A., Singer, P., Gordeev, D., Karaflos, D., Levinson, C. J., Pfeiffer, P., Puk, K. M., and Reade, W.: Laboratory earthquake forecasting: A machine learning competition, Proceedings of the National Academy of Sciences, 118, e2011362 118, https://doi.org/10.1073/pnas.2011362118, 2021.

Jozinovic, D., Lomax, A., Štajduhar, I., and Michelini, A.: Rapid prediction of earthquake ground shaking intensity using raw waveform data and a convolutional neural network, Geophysical Journal International, 222, 1379–1389, https://doi.org/10.1093/gji/ggaa233, 2020.

Jozinović, D., Lomax, A. J., Stajduhar, I., and Michelini, A.: SSA 2021 Annual Meeting, Seismological Research Letters, 92, 1213–1479, https://doi.org/10.1785/0220210025, https://doi.org/10.1785/0220210025, 2021.

Kong, Q., Trugman, D. T., Ross, Z. E., Bianco, M. J., Meade, B. J., and Gerstoft, P.: Machine Learning in Seismology: Turning Data into Insights, Seismological Research Letters, 90, 3–14, https://doi.org/10.1785/0220180259, https://doi.org/10.1785/0220180259, 2018.

Korneev, V. A., Nadeau, R. M., and McEvilly, T. V.: Seismological Studies at Parkfield IX: Fault-Zone Imaging Using Guided Wave Attenuation, Bulletin of the Seismological Society of America, 93, 1415–1426, https://doi.org/10.1785/0120020114, https://doi.org/10.1785/0120020114, 2003.

Kriegerowski, M., Petersen, G. M., Vasyura Bathke, H., and Ohrnberger, M.: A Deep Convolutional Neural Network for Localization of Clustered Earthquakes Based on Multistation Full Waveforms, Seismological Research Letters, pp. 1–7, https://doi.org/10.1785/0220180320, 2018.

Krischer, L., Megies, T., Barsch, R., Beyreuther, M., Lecocq, T., Caudron, C., and Wassermann, J.: ObsPy: a bridge for seismology into the scientific Python ecosystem, Computational Science & Discovery, 8, 014 003–18, https://doi.org/10.1088/1749-4699/8/1/014003, 2015.

Lanzano, G., Luzi, L., Russo, E., Felicetta, C., D'Amico, M., Sgobba, S., and Pacor, F.: Engineering Strong Motion Database (ESM) flatfile [Data set], Tech. rep., Istituto Nazionale di Geofisica e Vulcanologia (INGV), https://doi.org/10.13127/esm/flatfile.1.0, 2018.

Linville, L., Pankow, K., and Draelos, T.: Deep Learning Models Augment Analyst Decisions for Event Discrimination, Geophysical Research Letters, 46, 3643–3651, https://doi.org/https://doi.org/10.1029/2018GL081119, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018GL081119, _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2018GL081119, 2019.

Lomax, A., Michelini, A., and Jozinović, D.: An Investigation of Rapid Earthquake Characterization Using Single-Station Waveforms and a Convolutional Neural Network, Seismological Research Letters, 90, 517–529, https://doi.org/10.1785/0220180311, https://doi.org/10.1785/0220180311, 2019.

Magrini, F., Jozinovic, D., Cammarano, F., Michelini, A., and Boschi, L.: Local earthquakes detection: A benchmark dataset of 3-component seismograms built on a global scale, Artificial Intelligence in Geosciences, 1, 1–10, https://doi.org/10.1016/j.aiig.2020.04.001, 2020.

Margheriti, L., Nostro, C., Cocina, O., Castellano, M., Moretti, M., Lauciani, V., Quintiliani, M., Bono, A., Mele, F. M., Pintore, S., Montalto, P., Peluso, R., Scarpato, G., Rao, S., Alparone, S., Di Prima, S., Orazi, M., Piersanti, A., Cecere, G., Cattaneo, M., Vicari, A., Sepe, V., Bignami, C., Valoroso, L., Aliotta, M., Azzarone, A., Baccheschi, P., Benincasa, A., Bernardi, F., Carluccio, I., Casarotti, E., Cassisi, C., Castello, B., Cirilli, F., D'Agostino, M., D'Ambrosio, C., Danecek, P., Cesare, W. D., Bina, E. D., Di Filippo, A., Di Stefano, R., Faenza, L., Falco, L., Fares, M., Ficeli, P., Latorre, D., Lorenzino, M. C., Mandiello, A., Marchetti, A., Mazza, S., Michelini, A., Nardi, A., Pastori, M., Pignone, M., Prestifilippo, M., Ricciolino, P., Sensale, G., Scognamiglio, L., Selvaggi, G., Torrisi, O., Zanolin, F., Amato, A., Bianco, F., Branca, S., Privitera, E., and Stramondo, S.: Seismic Surveillance and Earthquake Monitoring in Italy, Seismological Research Letters, https://doi.org/10.1785/0220200380, https://doi.org/10.1785/0220200380, 2021.

McNamara, D. E. and Buland, R. P.: Ambient Noise Levels in the Continental United States, Bulletin of the Seismological Society of America, 94, 1517–1527, https://doi.org/10.1785/012003001, https://doi.org/10.1785/012003001, 2004.

Megies, T., Beyreuther, M., Barsch, R., Krischer, L., and Wassermann, J.: ObsPy – What can it do for data centers and observatories?, Annals of Geophysics, 54, 47–58, https://doi.org/10.4401/ag-4838, 2011.

Meier, M.-A., Ross, Z. E., Ramachandran, A., Balakrishna, A., Nair, S., Kundzicz, P., Li, Z., Andrews, J., Hauksson, E., and Yue, Y.: Reliable Real-Time Seismic Signal/Noise Discrimination With Machine Learning, Journal of Geophysical Research: Solid Earth, 124, 788–800, https://doi.org/https://doi.org/10.1029/2018JB016661, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018JB016661, _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2018JB016661, 2019.

Michelini, A., Margheriti, L., Cattaneo, M., Cecere, G., D'Anna, G., Delladio, A., Moretti, M., Pintore, S., Amato, A., Basili, A., Bono, A., Casale, P., Danecek, P., Demartin, M., Faenza, L., Lauciani, V., Mandiello, A. G., Marchetti, A., Marcocci, C., Mazza, S., Mele, F. M., Nardi, A., Nostro, C., Pignone, M., Quintiliani, M., Rao, S., Scognamiglio, L., and Selvaggi, G.: The Italian National Seismic Network and the earthquake and tsunami monitoring and surveillance systems, Advances in Geosciences, 43, 31–38, https://doi.org/10.5194/adgeo-43-31-2016, 2016.

Michelini, A., Faenza, L., Lanzano, G., Lauciani, V., Jozinović, D., Puglia, R., and Luzi, L.: The New ShakeMap in Italy: Progress and Advances in the Last 10 Yr, Seismological Research Letters, 91, 317–333, https://doi.org/10.1785/0220190130, https://doi.org/10.1785/0220190130, 2019.

Michelini, A., Cianetti, S., Gaviano, S., Giunchi, C., Jozinović, D., and Lauciani, V.: INSTANCE The Italian Seismic Dataset For Machine Learning, https://doi.org/10.13127/INSTANCE, http://www.pi.ingv.it/instance/, 2021.

Mousavi, S. M. and Beroza, G. C.: A Machine-Learning Approach for Earthquake Magnitude Estimation, Geophysical Research Letters, 47, 1691–7, https://doi.org/10.1029/2019GL085976, 2020.

Mousavi, S. M., Sheng, Y., Zhu, W., and Beroza, G. C.: STanford EArthquake Dataset (STEAD): A Global Data Set of Seismic Signals for AI, IEEE Access, PP, 1–1, https://doi.org/10.1109/ACCESS.2019.2947848, 2019.

Mousavi, S. M., Ellsworth, W. L., Zhu, W., Chuang, L. Y., and Beroza, G. C.: Earthquake transformer-an attentive deep-learning model for simultaneous earthquake detection and phase picking, Nature Communications, pp. 1–12, https://doi.org/10.1038/s41467-020-17591-w, 2020.

Münchmeyer, J., Bindi, D., Leser, U., and Tilmann, F.: The transformer earthquake alerting model: A new versatile approach to earthquake early warning, arXiv.org, physics.geo-ph, https://doi.org/10.1093/gji/ggaa609, 2020.

Münchmeyer, J., Bindi, D., Leser, U., and Tilmann, F.: Earthquake magnitude and location estimation from real time seismic waveforms with a transformer network, arXiv.org, physics.geo-ph, 2021.

Otović, E., Njirjak, M., Jozinović, D., Mauša, G., Michelini, A., and Štajduhar, I.: Intra-domain and cross-domain transfer learning for time series, Tech. Rep. EGU21-12142, Copernicus Meetings, https://doi.org/10.5194/egusphere-egu21-12142, https://meetingorganizer.copernicus.org/EGU21/EGU21-12142.html, conference Name: EGU21, 2021.

Parolai, S.: Denoising of Seismograms Using the S Transform, Bulletin of the Seismological Society of America, 99, 226–234, https://doi.org/10.1785/0120080001, https://doi.org/10.1785/0120080001, 2009.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library, in: Advances in Neural Information Processing Systems 32, edited by Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F. d., Fox, E., and Garnett, R., pp. 8024–8035, Curran Associates, Inc., http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf, 2019.

Perol, T., Gharbi, M., and Denolle, M.: Convolutional neural network for earthquake detection and location, Science Advances, 4, e1700 578, https://doi.org/10.1126/sciadv.1700578, https://advances.sciencemag.org/content/4/2/e1700578, publisher: American Association for the Advancement of Science Section: Research Article, 2018.

Quinteros, J., Carter, J. A., Schaeffer, J., Trabant, C., and Pedersen, H. A.: Exploring Approaches for Large Data in Seismology: User and Data Repository Perspectives, Seismological Research Letters, pp. 1–10, https://doi.org/10.1785/0220200390, 2021.

Richter, C. F.: An instrument earthquake magnitude scale, Bulletin of the Seismological Society of America, 25, 1–32, 1935.

Ross, Z. E., Meier, M., Hauksson, E., and Heaton, T. H.: Generalized Seismic Phase Detection with Deep Learning, Bulletin of the Seismological Society of America, 108, 2894–2901, https://doi.org/10.1785/0120180080, https://doi.org/10.1785/0120180080, 2018a.

Ross, Z. E., Meier, M.-A., and Hauksson, E.: P Wave Arrival Picking and First-Motion Polarity Determination With Deep Learning, Journal of Geophysical Research: Solid Earth, 123, 5120–5129, https://doi.org/https://doi.org/10.1029/2017JB015251, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2017JB015251, _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2017JB015251, 2018b.

Scognamiglio, L., Tinti, E., and Michelini, A.: Real-Time Determination of Seismic Moment Tensor for the Italian Region, Bulletin of the Seismological Society of America, 99, 2223–2242, https://doi.org/10.1785/0120080104, 2009.

Seydoux, L., Balestriero, R., Poli, P., Hoop, M. d., Campillo, M., and Baraniuk, R.: Clustering earthquake signals and background noises in continuous seismic data with unsupervised deep learning, Nature Communications, 11, 3972, https://doi.org/10.1038/s41467-020-17841-x, http://www.nature.com/articles/s41467-020-17841-x, 2020.

Shearer, P. M.: Imaging global body wave phases by stacking long-period seismograms, Journal of Geophysical Research: Solid Earth, 96, 20 353–20 364, https://doi.org/https://doi.org/10.1029/91JB00421, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/91JB00421, _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/91JB00421, 1991.

Strollo, A., Cambaz, D., Clinton, J., Danecek, P., Evangelidis, C. P., Marmureanu, A., Ottemöller, L., Pedersen, H., Sleeman, R., Stammler, K., Armbruster, D., Bienkowski, J., Boukouras, K., Evans, P. L., Fares, M., Neagoe, C., Heimers, S., Heinloo, A., Hoffmann, M., Kaestli, P., Lauciani, V., Michalek, J., Odon Muhire, E., Ozer, M., Palangeanu, L., Pardo, C., Quinteros, J., Quintiliani, M., Antonio Jara-Salvador, J., Schaeffer, J., Schloemer, A., and Triantafyllis, N.: EIDA: The European Integrated Data Archive and Service Infrastructure within ORFEUS, Seismological Research Letters, https://doi.org/10.1785/0220200413, https://pubs.geoscienceworld.org/ssa/srl/article/595802/EIDA-The-European-Integrated-Data-Archive-and, 2021.

Trugman, D. T. and Shearer, P. M.: Strong Correlation between Stress Drop and Peak Ground Acceleration for Recent M 1–4 Earthquakes in the San Francisco Bay Area, Bulletin of the Seismological Society of America, 108, 929–945, https://doi.org/10.1785/0120170245, https://doi.org/10.1785/0120170245, 2018.

https://doi.org/10.5194/essd-2021-164
Preprint. Discussion started: 27 May 2021

Vallée, M., Ampuero, J. P., Juhel, K., Bernard, P., Montagner, J.-P., and Barsuglia, M.: Observations and modeling of the elastogravity signals preceding direct seismic waves, Science, 358, 1164–1168, https://doi.org/10.1126/science.aao0746, https://science.sciencemag.org/content/358/6367/1164, publisher: American Association for the Advancement of Science Section: Report, 2017.

Walter, J. I., Ogwari, P., Thiel, A., Ferrer, F., and Woelfel, I.: easyQuake: Putting Machine Learning to Work for Your Regional Seismic Network or Local Earthquake Study, Seismological Research Letters, 92, 555–563, https://doi.org/10.1785/0220200226, 2020.

Wollam, J., Munchmeyer, J., Giunchi, C., Jozinović, D., Diehl, T., Saul, J., Michelini, A., Haslinger, F., Lange, D., Tilmann, F., and Rietbrock, A.: SeisBench: A toolbox for benchmarking and applying machine learning in seismology., 2021.

Yeck, W. L. and Patton, J.: Waveform Data and Metadata used to National Earthquake Information Center Deep-Learning Models, https://doi.org/10.5066/P9OHF4WL, https://www.sciencebase.gov/catalog/item/5ed528ff82ce2832f047eee6, type: dataset, 2020.

Yeck, W. L., Patton, J. M., Ross, Z. E., Hayes, G. P., Guy, M. R., Ambruz, N. B., Shelly, D. R., Benz, H. M., and Earle, P. S.: Leveraging Deep Learning in Global 24/7 Real-Time Earthquake Monitoring at the National Earthquake Information Center, Seismological Research Letters, 92, 469–480, https://doi.org/10.1785/0220200178, 2020.

Zhang*, J. and Chen, J.: Joint seismic traveltime and waveform inversion for near surface imaging, in: SEG Technical Program Expanded Abstracts 2014, pp. 934–937, Society of Exploration Geophysicists, Denver, Colorado, https://doi.org/10.1190/segam2014-1501.1, https://library.seg.org/doi/10.1190/segam2014-1501.1, 2014.

Zhang, X., Zhang, J., Yuan, C., Liu, S., Chen, Z., and Li, W.: Locating induced earthquakes with a network of seismic stations in Oklahoma via a deep learning method, Scientific Reports, 10, 1941, https://doi.org/10.1038/s41598-020-58908-5, https://www.nature.com/articles/s41598-020-58908-5, number: 1 Publisher: Nature Publishing Group, 2020.

Zhu, L., Peng, Z., McClellan, J., Li, C., Yao, D., Li, Z., and Fang, L.: Deep learning for seismic phase detection and picking in the aftershock zone of 2008 M7.9 Wenchuan Earthquake, Physics of the Earth and Planetary Interiors, 293, 106 261, https://doi.org/10.1016/j.pepi.2019.05.004, https://linkinghub.elsevier.com/retrieve/pii/S0031920118301407, 2019a.

Zhu, W., Mousavi, S. M., and Beroza, G. C.: Seismic Signal Denoising and Decomposition Using Deep Neural Networks, IEEE Transactions on Geoscience and Remote Sensing, 57, 9476–9488, https://doi.org/10.1109/TGRS.2019.2926772, conference Name: IEEE Transactions on Geoscience and Remote Sensing, 2019b.