Below in dark red the comments by the referee (John Clinton) and in black the reply.

Note: The replies are provided only for the important comments which required additional work and/or clarification.

This is an excellent article introducing a new dataset targeting machine learning applications using a near-complete set of earthquake records from Italy. It follows the example of STEAD described in Mousavi et al, 2019. The manuscript provides an overview of how the dataset was generated and organised, and then provides an overview of the general features of the dataset. The manuscript is well written, particularly the introduction that gives a strong motivation for why this type of product is sorely needed, and provides an overview of similar datasets. It is timely and important that high quality local earthquake records datasets outside the US W. Coast are highlighted and made easily available for researchers to use. I hope this sort of documentation of datasets and their preparation in research-ready format becomes the standard, and I expect that publication of this manuscript will lead to numerous publications on ML that use this dataset.

We are very pleased for the appreciation.

I know some information on the uniqueness of the dataset is dispersed through the article, but I suggest, in a single place, in the discussion, the authors extend and accentuate what is different with this collection compared to others, besides from the obvious that this is solely an Italian dataset based on the Italian earthquake catalogue. Are the metadata fields better? How do they differ? Are formats modified ? Is it unique to provide both raw and corrected waveform data? Data volumes similar to other datasets? Maybe a comparison table to STEAD and the Caltech datasets would be helpful. The authors could also provide stronger comments on the benefits of standardisation of formats / metadata for these datasets.

Yes, we will include a list of those components of the dataset that differ or are unique in the revised manuscript. Below some quick answers on the questions raised.

The metadata are many more than those included in previously published ML datasets. We will provide a summary table comparing our metadata with those presented in other similar datasets cited in the ms (i.e., STEAD and the Caltech datasets, LEN-DB and the dataset used by Lomax et al., 2019, when developing ConvNetQuake_INGV).

In summary,

- we are not aware of other datasets that include both raw and corrected waveforms. The only exception appears to be the article by Meier et al. (2019) that uses the SCSN data (https://scedc.caltech.edu/data/deeplearning.htm) and seems to include data in SI units but this is only stated in the README file linked above and not in the paper.

- The data volumes have been assembled following the indications of SeisBench - a novel initiative to standardize the seismic data for machine learning (https://meetingorganizer.copernicus.org/EGU21/EGU21-12218.html).

- we will reword and add stronger comments on the benefits of format standardization.

A comment on the metadata on earthquake parameters summarised in Table 2: numerous fields are provided, including location uncertainty, but in the text, there is no comment on what location algorithm or velocity model is used. Since the dataset spans 15 years and a very wide geographic range, its likely that despite efforts to ensure a continuous approach to manual catalogue review, the velocity model and location algorithms have changed across the catalogue. If they have not changed, this point should be made. If they have changed, this should be indicated in the document, and consideration should be given to add this information in future updates of the dataset.

The location of all the earthquakes was made using the same 1D standard model adopted by INGV for the locations reported in the 2008 Bollettino Sismico Italiano (Mele et al, 2010). We have added a small section in the appendix to provide the model. The earthquake locations are performed using the software IPOP developed by Alberto Basili (Bono, 2007). We can then assert that the source and the path/derived parameters are pretty much consistent through time.

The DOI of each of the network codes used in the dataset / publication must be made available in the references or in the data availability section

Yes, this is definitely important and they have been added in the data availability section.

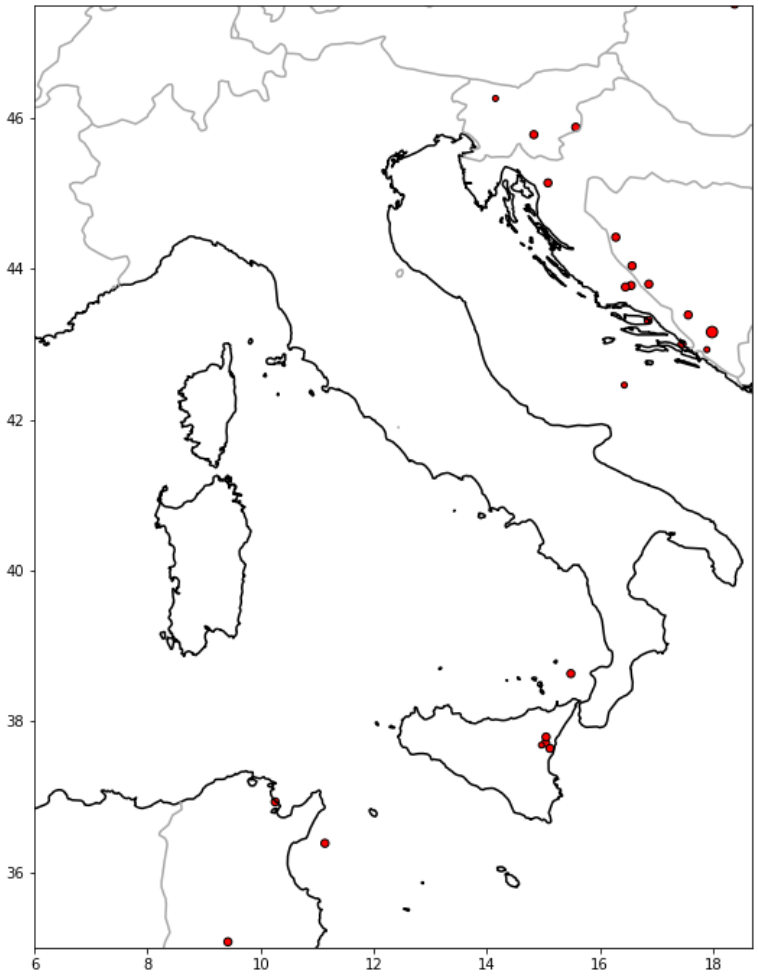
In general, the figures are not optimal, often use strange axis labelling (that may be a direct metadata field from Table 2 - if so mention it!), and often have captions that are too terse / insufficiently descriptive. I suggest the authors look through these carefully. Also font sizes on Figs 14,15,16 are too small- in particular the exponents are completely illegible.

We will add a note specifying that the labels of the figures follow from the naming of the metadata and we will make more exhaustive and self-contained descriptive captions for the figures. We are also fixing the font size issue in Figs 14, 15, 16 and 20.

...

P4 l25: some >M4.0 are rejected. A bit more info on significant events that have been removed is needed. Are these only those that include multiple events in the same time window in the catalogue? I hope no very significant events are rejected simply because a very small foreshock or aftershock is also catalogued…

We attempted to keep all the earthquakes with $M \geq 4$ but in some cases this was not possible because of missing data. As a consequence 30 earthquakes with $M \geq 4$ were removed and they are shown in the picture below. Almost all of them occur outside the Italian country borders. The four earthquakes in Italy (near Catania and in southern Tyrrhenian sea) were removed accidentally because of a download technical problem and will be reintroduced in updated versions of the dataset.



P4 station selection: mention in this section that only stations on Italian territory are used. Are the Civil Defence stations not added? If not, mention why this very significant dataset missing - is it technical or political?

Also stations belonging to the MedNet network outside Italy and some stations of the Albanian and the Greek networks have been used. Regarding the Italian Civil Protection (RAN network), these stations are not inserted in the dataset because they are not available in EIDA and they are not used for the compilation of the BSI upon which the data selection has been based. They may be included in future releases of the dataset although this would also involve the seismogram manual picking (P and S phases) which may be quite heavy to be completed with the available human resources.

P5 l1-5: Be more specific on what picks are made available. I assume the INGV catalogue makes first arriving P and S picks only. No additional phase type is indicated (Pg vs Pn), and secondary phases are also not identified (eg PmP)

In the BSI there is no distinction between Pg and Pn or secondary phases like PmP and in the dataset they are just referred to as P phases.

P5 l6-9 its should be accentuated either here or later that since 1/ not all stations used in the catalogue generation are included, eg foreign-operated stations;

Although very few, there are foreign stations (i.e., AC and HL neworks) in the dataset. There are no stations from the countries bordering northern Italy because they are not included in the EIDA Italian node that was used and our efforts were focused on Italian stations waveform data. Nevertheless, we consider that a dataset like INSTANCE can be easily integrated with additional data by following the same standards adopted here. In principle, it would be possible that other similar datasets be assembled for other regions following the same standards adopted here and then even merge them together in a single dataset.

2/ phases with large residuals or low weight are removed that it is not possible to use this dataset to relocate the catalogue.

The main point is that INSTANCE targets ML applications so that the results of these analyses could be then used to re-process the entire data and to rebuild the earthquake catalog. Since the average number of stations per earthquake is about 21, in many cases these would be more than sufficient to attain good and stable earthquake relocation.

P5 waveform data selection: in the case of multiple available sampling rates, I infer that the same sampling rate used to make the manual pick is selected for inclusion here. Or is it the highest available sampling rate for each channel?

The understanding is correct. We provide the trace data resampled at 100 Hz and the arrival times obtained from the original trace data used for the picking. In most cases there was no need for resampling and the trace data coincide with those used to pick the arrival times. P6l11: I don't understand what is 'arrival time samples'. Why not simply use time in seconds?

The 'arrival time samples' represents the sample number of the phase arrival time into the array available in the hdf5. The use of 'arrival time samples' serves to simplify the use of these quantities especially by non-seismologists. The arrival times are also provided.

P6 2.1.5: the authors should mention are all traces rotated into ZNE, or in the entire Italian catalogue in ZNE by default. If so, I am amazed!

They are by default all along ZNE. We do not have any waveform included that resulted from sensors oriented differently.

P7 l.29 2.2 Metadata: in source, the location method or velocity model are not included. They should be if either of these have changed over 15 years of the catalogue.

The velocity model used for location in the compilation of the BSI has not been changed in 15 years. Thus the earthquake locations in this sense are all consistent. Details on the BSI procedure are provided in Mele et al. (2010) and the velocity model is also provided in an additional appendix.

P8 l25 ‘missing data’ - please expand

In Jozinovic et al. (2020), the dataset used for ML consists of a fixed number of stations and when data from one or more stations are missing (either the whole trace or parts of it), the signal trace is set to be an array of zeros. The ML model used there was found to detect and learn the problematic values, and compensate for it, having a similar prediction accuracy on those stations as the accuracy on the stations which had the input data available .

P10 Table 2: location code is not part of the International Registry?

Thanks for noting it. Yes, the location code is not part of the International Registry. “Changed in table 2.

P13 l 5 I’m surprised to see selection criteria was for even number of traces for each channels? Seems in contradiction to 2.1.2, where all reasonable phases according to seismicity were selected. Was seismicity for smaller events actually selected according to numbers of station pick?

These are selection criteria for the *noise recordings* and there is no relation with the number of picks available for each station. We made an attempt to select all the station channels with a more or less even number of recordings.

P14 l4 ‘great majority’ seems an exaggeration.

The histogram in Figure 3d adopts a log scale. This issue was also reported by the other referee and the histogram scale will be changed to linear to better evidence the assertion.

P15 Fig 5 / l9 onwards: the number of up first motion polarities is double that of down. This is surprising, and possible concerning unless there is a reasonable explanation I do not see. The authors should explain this. Is it possible eventype=earthquake is not selected, and blasts are also included here?

We thank the referee for raising this issue which we did not address thoroughly in our manuscript.

As described in the manuscript, we have adopted the “event” FDSN web service implemented at INGV which adopts the standard FDSN parameters which at the moment do not include the “event_type” field for selection. This has been also noted recently by Gulia and Gasperini (2021). However, it is still possible to download the quakeML (a xml formatted file standard for seismology) for each event which includes the “event_type” parameter.

We have therefore proceeded to obtain the event_type value and we have included it as additional parameter (`source_type`) in the metadata file. Thus, the new metadata file now includes 115 parameters total. It appears, however, that the addition of the new parameter extracted from the INGV archive captures only a fraction of the non-earthquake sources. That is, many artificial sources are still catalogued as earthquakes in the INGV archive. The

table below provides a snapshot of the event_type included in the proposed dataset.

type_event	
anthropogenic event	1
controlled explosion	44
earthquake	53753
experimental explosion	8
explosion	5
landslide	1
quarry blast	194
volcanic eruption	2

In addition, the BSI distinguishes between earthquakes and other sources like quarry blasts only since 2012 (Gulia and Gasperini, 2021).

Given that the inclusion of the event_type above still misses several artificial sources, we have addressed the asymmetry noted by the referee between the number of positive and negative polarities by other means. To this end, we performed two different analyses to verify i.) how the inclusion of blasts can affect the reported asymmetry and ii.) how the region with its dominant tectonic style can condition the number of positive and negative polarities in INSTANCE.

Following Mele et al. (2010) who found that the 99.6% of the blasts have local magnitude ML ≤ 2.2 (Fig. 23 of their study), we have progressively increased the lower magnitude threshold to verify whether the nearly 2:1 ratio between positive and negative polarities persists as the magnitude is increased. The expectation is that as the magnitude increases, the ratio progressively levels out since the blasts (or other artificial sources) do not produce magnitudes greater than M=3 in Europe (Giardini et al., 2004).

Secondly, we have subdivided the Italian area into two zones: earthquakes inside the Apennines area [vertices (lat,lon) (41N,9E and 44N,15E)], and earthquakes elsewhere outside this area. This data selection is aimed to verify if the observed asymmetry of positive and negative polarities can result from the dominant extensional stress field characterizing the Apennines when compared to the other areas in Italy.

To address the variation of the proportion between positive and negative polarities with magnitude, the table below shows that the fraction (per cent values) of negative polarities increases progressively from 36% to ~41% when including earthquakes with M>0.25 and M>3, respectively. For larger minimum magnitudes (M>3), the percentage stabilizes around 42-43%. This would indicate that inclusion of the polarities of unrecognized blasts (i.e., with M<3) has a moderate but still significant impact on the observed asymmetry between the reported positive and negative polarities. This asymmetry, although somewhat surprising, seems to occur also elsewhere. For example, Ross et al. (2019) report in their analysis of the southern California earthquake dataset (before data augmentation) that their data contains 67% and 33% up and down polarities, respectively. We also note that the regional tectonic setting in Southern California is quite different from that in Italy.

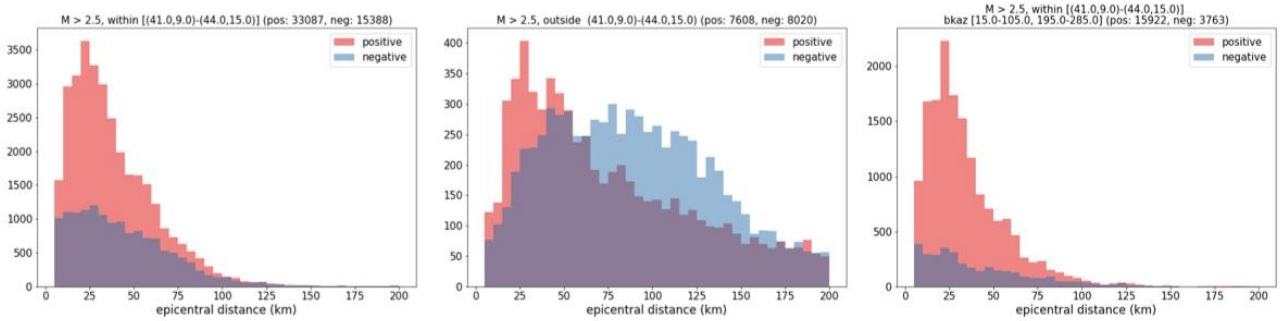
min_magnitude	total	positive	positive_percent	negative	negative_percent
0.25	236345	151544	64.12	84801	35.88
0.5	235806	151204	64.12	84602	35.88
0.75	234400	150335	64.14	84065	35.86
1	227810	146213	64.18	81597	35.82
1.25	219688	141096	64.23	78592	35.77
1.5	204277	131159	64.21	73118	35.79
1.75	194880	125020	64.15	69860	35.85
2	160464	102359	63.79	58105	36.21
2.25	118581	75072	63.31	43509	36.69
2.5	75907	46810	61.67	29097	38.33
2.75	57366	34740	60.56	22626	39.44
3	37183	21821	58.69	15362	41.31
3.25	27333	15748	57.62	11585	42.38
3.5	16749	9447	56.4	7302	43.6
3.75	12328	6979	56.61	5349	43.39
4	7200	4151	57.65	3049	42.35
4.25	4935	2810	56.94	2125	43.06
4.5	2232	1312	58.78	920	41.22
4.75	1369	833	60.85	536	39.15
5	814	468	57.49	346	42.51

For our second analysis (proportion between positive and negative polarities depending on the area), we have considered that in Europe the maximum magnitude of quarry blasts is usually assumed to be 2.5–3.0 (Giardini ed al., 2004) and, following the findings of Mele et al (2010), we focus only on earthquakes with M> 2.5. We have extracted the polarities for the target Apennine region and compared to those reported for earthquakes elsewhere in Italy. In the target area, the largest majority of the earthquakes are characterized by normal faulting mechanisms with the lobes of the seismic radiation pattern having negative polarities at short epicentral distances. Given these conditions, the observed asymmetry could result from the complex interplay between the source receiver geometry, the width of 200-300 km coast to coast from the Tirrhenian to the Adriatic seas of peninsular Italy and the dominant extensional faulting with faults striking NW-SE characterizing the Apennines and dominated by normal faulting.

In this setting, the radiation pattern predicts negative polarities in the near source and positive polarities farther away. Also, the negative polarity source radiation lobes map into a smaller extension region near the epicenter that, in general, will also have a smaller

number of stations when compared to the positive lobes of larger extension and, consequently, a larger number of stations.

The figure below shows the histograms of the distribution of the positive and negative polarities with distance. The panel to the left shows the distribution of the polarities for the chosen target area in peninsular Italy, in the rightmost the polarities in the same area but only along the approximate NE-SW direction of the backazimuth (i.e., the ranges 15-105 and 195-285 degrees) and, in the middle, the area outside this target area. We note that within the target area the polarities are overwhelmingly positive in gross agreement with what described above and, for further confirmation, we see that if we restrict to the NE-SW propagation direction perpendicular to the Italian peninsula (rightmost panel), the ratio between positive and negative polarities (%pos,%neg) increases from (68%,32%) to (81%,19%), respectively. Conversely, the number of polarities for the earthquakes outside the target area are pretty much well balanced (49%, 51%).



In conclusion, i.) the INSTANCE dataset does contain positive polarities resulting from the inclusion of quarry blasts misidentified as earthquakes for magnitudes less than ~2.5-3.0. This follows from what reported by Mele et al. (2010) (and very recently by Gulia and Gasperini, 2021) and the change in positive and negative polarities percentages reported in our table above confirms it; ii.) the current modalities of earthquake revision at INGV do not include identification of all the manmade sources and, the web service used does not include the *eventtype* identification but it was still possible to retrieve the event_type and, accordingly, add a new source parameter (`source_type`) to the dataset metadata; iii.) the target area in the selected Apennine region includes ~76% of the total number of polarities of the dataset; iv) In the Apennine region there is dominance of positive polarities which is likely the result of the dominant normal type of earthquake faulting in the area; v) the asymmetry observed in the target area disappears for M>2.5 elsewhere in Italy.

P17 l1 1/2: is it possible this can also be explained by systematically mis-identified first arrivals, rather than complications in the velocity structure?

It has been verified that these very long traveltimes belong to earthquakes that occurred during the 2012 Emilia earthquake sequence. The stations recording these events were located on the soft and thick alluvium characterizing the Po plain which features very low seismic velocities.

P27 l10: earthquake in INGV catalogue - so its very possible that noise traces include energy from regional and teleseismic events.

Yes the referee is correct. Will be pointed out in the manuscript

P27 l14: any effort to include the same spread of stations as found in the event dataset?

No, if for spread it is meant the same group of stations detecting earthquakes in a given area for the same time window. Anyhow, the stations are exactly the same as those of the event dataset as evidenced in Figure 2.

P39 FigA4 - over 100 records have PGA >2g, and many even over 4g. Which is rather unphysical. Is this understood?

The units are cm/s^2 and we do not see any value above 1g for PGA.

References

Basili A., 2005. Locator: Il manuale, Documentazione disponibile in forma digitale.

Bono, Andea. “SisPick! 2.0 Sistema interattivo per l’interpretazione di segnali sismici, Manuale utente,” 2008. https://istituto.ingv.it/images/collane-editoriali/rapporti%20tecnici/rapporti-tecnici-2008/rapporto59.pdf.

Gulia, Laura, and Paolo Gasperini. “Contamination of Frequency–Magnitude Slope (B-Value) by Quarry Blasts: An Example for Italy.” Seismological Research Letters, June 30, 2021. https://doi.org/10.1785/0220210080.

Meier, Men-Andrin, Zachary E. Ross, Anshul Ramachandran, Ashwin Balakrishna, Suraj Nair, Peter Kundzicz, Zefeng Li, Jennifer Andrews, Egill Hauksson, and Yisong Yue. “Reliable Real-Time Seismic Signal/Noise Discrimination With Machine Learning.” Journal of Geophysical Research: Solid Earth 124, no. 1 (2019): 788–800. https://doi.org/10.1029/2018JB016661.

Mele, F., Luca Arcoraci, Patrizia Battelli, Michele Berardi, Corrado Castellano, Giulio Lozzi, Alessandro Marchetti, Anna Nardi, Mario Pirro, and Antonio Rossi. “Bollettino Sismico Italiano 2008 (Italian Seismic Bulletin 2008).” *Quaderni di geofisica*, no. 85 (2010): 45.

Ross, Zachary E., Men-Andrin Meier, and Egill Hauksson. “P Wave Arrival Picking and First-Motion Polarity Determination With Deep Learning.” Journal of Geophysical Research: Solid Earth 123, no. 6 (2018): 5120–29. https://doi.org/10.1029/2017JB015251.