Earth Syst. Sci. Data Discuss., author comment AC1
https://doi.org/10.5194/essd-2021-164-AC1, 2021

Reply on RC1

Alberto Michelini et al.

Author comment on "INSTANCE – the Italian seismic dataset for machine learning" by Alberto Michelini et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2021-164-AC1, 2021

Dear Martijn,

Thanks for the positive feedback.

- You are right about the sample dataset referenced in the GitHub repository. It would not have been possible to provide the complete dataset on GitHub due to size constraints. The purpose was to provide the python notebooks that we used. These can be easily modified to access the entire dataset and replicate the manuscript's figures or to make any other kind of data selection.
- Regarding the versioning of the dataset, we fully agree with you and we have already included this feature in our DOI scheme. However, we did not make it explicit in the manuscript. This will be clarified in the revised version.
- For what it concerns the NaN, it is used to indicate the missing values for the low gain channels where both EQTransformer and GPD have not been run. In case of high gain channels the lack of detection is indicated by '0'. High gain channels are (e.g. Figure 5a) ~71% of the data. This can be stated in the text. For the EQTransformer, we do agree that an "ad hoc" training could improve its performances, but this is beyond the aim of our manuscript.
- We agree with you that expliciting the disk space in the landing page is important and we will fix it.

On the minor comments

We do agree overall on your comments and improvements for the figures.

More specifically,

- For what concerns the velocity model used for the calculation, all locations come from the Italian Seismic Bulletin that adopt the velocity model provided in (https://istituto.ingv.it/images/collane-editoriali/quaderni-di-geofisica/quaderni-di-geofisica-2010/quaderno85.pdf). The traveltimes are calculated as the difference between the P- (or S-) arrival time and the origin time.
- For the time windows, yes, they do not overlap exactly for the same earthquake. It is explained earlier that the start time is chosen based on the P arrival time.
- Regarding the low SNR traces, please do keep in mind that still about 10% (~120,000 3C traces) have values less than about 2.3 db. This is already a significant number for

ML purposes.