The manuscript studies for the first time the conditions of reference stations across 60 rock stations belonging to the broadband and accelerometric networks of the National Observatory of Athens. The analysis is based on selecting stations situated on rock and whose data availability is sufficient for a meaningful collection of recordings. The study is relevant since no systematic previous site effects studies were conducted for the ensemble of stations under examination. Similar studies have been conducted in other contexts and at different scales by Lanzano et al. (2020, 2022) for Italy and by Pilz et al. (2020) for Europe. I believe that this work is a very useful contribution to the scientific community and to seismological studies in Greece. I suggest accepting the publication with minor revisions.
➔ We thank the reviewer for the overall appreciation of our rationale and outcomes, and suggested improvements.

Some comments below:

Station and data selection

1. Please explain further, on the basis of which criterion you consider the 60 stations to be installed on rock. Did you use geological and/or topographical proxies? Do any of the stations also have a geophysical survey?
   ➔ One of the main problems is that few strong-motion stations in HL have had the benefit of geophysical studies, and of those that have, most are either old (triggering mode) or lie on soil deposits, so they were of no interest to us at this point (the scope being modern continuous data on potential rock). No broadband stations have been characterized, as is the global standard practice.
   In selecting the stations, these were the criteria: 1. HH and collocated HH/HN stations: we selected all of them; 2. HN stations: we selected a HN station if it was thought to lie on rock according to any one of the following 3 criteria: by the network operator, as per existing information on accelnet.gein.noa.gr website; by ESM, based on proxies as rock; and according to the detailed investigation of surface geological conditions performed for all HN stations using the HSGME maps.
   The reviewer is correct in that we did not explain the selection criteria clearly enough and this can be amended in the revision.
2. Letters a and b are not present in Figure 1 and 2
   ➔ Yes indeed, thank you for catching that.

Creation of a new strong motion dataset

The analysis of the signals for the creation of the dataset is a very important step. I realised that the authors did a very thorough job, taking advantage of already available codes. However, I suggest that this section be reorganised schematically by indicating the data processing work in steps.
For example:
- Identification of clipped records (which criteria?)
- Waveform picking
- Calculation of signal-to-noise ratio
- Identification of corner frequencies - Etc.

I am not saying to repeat things that have been explained in other works, but to list the actions in a schematic and sequential manner. I think the work would benefit in terms of clarity. I think it would also be helpful to understand which signal analyses are done by the NGA-East code and which are not.
➔ We thank the reviewer for acknowledging the work load implied in creating this dataset. We also realize that we were not clear enough in our manuscript and created a misconception: we did not use any available off-the-shelf codes from PEER or elsewhere. We may have caused some of the confusion by mentioned PEER processing standards, but what we meant was that we were affected to a great extent by the logic followed in PEER NGA-West2 and East projects, where the first author was part of the data processing development team. But in this work, we created our own in-house processing code from scratch to analyse our data in the time and frequency domain for our specific purpose. We actually developed our code further as we processed more data, understood the needs arising and implemented improvements. We realize that the suggestions made by the reviewer will clarify the procedure we used and we can adopt them to better explain the flow and tools used.
Do you check for double events recordings? How do you treat?
➔ We are not sure we understand what the reviewer means by double recording: a recording of an event in whose tail another event takes place? In such cases, we try to salvage the recording if possible on a case-specific basis. If we understand the question better, we can address it in the manuscript.

I also suggest improving figure 3 to make it more self-explaining. A legend is missing.
➔ We can certainly do that.

Transfer function

1. For the calculation of transfer functions with the horizontal-to-vertical spectral ratio method, I would pay attention to the recordings of co-located stations.

From what I understand, all recordings for stations equipped with accelerometer and velocimeter were considered for the estimation of HVSR. My experience with INGV's Italian seismic network, which has a large number of collocated stations, is that this should be done carefully. I suggest conducting a preliminary analysis by keeping the instruments separate (consider them as two different stations) to verify that the transfer function is equal. First of all, the sensors could suffer from fixed scaling (a comparison of intensity measurements of recordings of the same event is also recommended) caused by incorrect conversion constants in the station xml. In addition, Hollander et al. (2020) and Castellaro et al. showed that the behaviour could be different at specific frequency ranges (especially in high frequency) due to different station installations.
➔ The reviewer makes an important comment here. We did not mix data from different channels (HH with HN) for any station analysed. When both were available, we selected the strong-motion one. Although a systematic comparison of the two sensor transfer functions is out of the scope of this particular article, we did make some cursory comparisons and these actually helped us identify a couple of mistaken sensor responses that were eventually corrected. In our case, if the gain is wrong on all three components in the same way, HVSR is unable to detect that, but the rotational sensitivity tests can indeed help identify cases where the N and E component are not consistent. Though an exhaustive account of all comparisons cannot fit in this paper, we can certainly address this question better in the manuscript, and we will take these suggestions on board for more detailed future work. The same holds regarding the useful comment of comparing high-frequency content from the two sensors in the light of installation differences.

1. I think it is also useful to explain how the groupings in Figure 7 were made. Was a clustering analysis conducted? Or is it based on a visual analysis of the curves? I have the impression that the transition from one group to another may not be clearly defined and some stations may be in one group rather than another arbitrarily. Wouldn't it be useful to set a quantitative criterion to isolate stations that have a flat response?

➔ The results shown in the submitted paper were grouped by visual inspection. This is not an optimal way to go about this. In the revision we will test some automated ways to yield the groups of stations, such as hierarchical clustering.

Discussion and conclusions

I noticed that the authors never considered the high-frequency near-site attenuation parameter k0 among the parameters for identifying reference sites. Considering the great experience of the authors, do you think this parameter could have any weight in the future? For example, in the work of Morasca et al. (2023) we started from the 36 stations of the study by Lanzano et al. (2020) and restricted to 6, as reference sites for a GIT in central Italy. The selection was made on the basis of k0, identifying those with k0<0.015s.
➔ The reviewer is correct - thank you for sharing your experiences with us. We appreciate the suggestion and vote of confidence, and we certainly wish to look into κ in future. However, as we responded to Reviewer 1 who had the same question -see point 2 above-, we need to reconsider the dataset and focus on an appropriate distribution that will include more near-source data to avoid results being overly dependent on path and distant events. When this is available, we can certainly try to use κ0 values as an additional piece of information when assessing reference stations.

Additional reference

Castellaro, S., Alessandrini, G., & Musinu, G. (2022). Seismic station installations and their impact on the recorded signals and derived quantities. Seismological Society of America, 93(6), 3348-3362.
Hollender, F., Roumelioti, Z., Maufroy, E., Traversa, P., & Mariscal, A. (2020). Can we trust high-frequency content in strong-motion database signals? Impact of housing, coupling, and installation depth of seismic sensors. Seismological Research Letters, 91(4), 2192-2205.
Morasca, P., D'Amico, M., Sgobba, S., Lanzano, G., Colavitti, L., Pacor, F., & Spallarossa, D. (2023). Empirical correlations between an FAS non-ergodic ground motion model and a GIT derived model for Central Italy. Geophysical Journal International, 233(1), 51-68.
➔ Thank you for sending these.