Review of: Synthetic ground motions in heterogeneous geologies: the HEMEW-3D dataset for scientific machine learning

General:

This is a well presented, if sometimes a little brief, manuscript describing a dataset of 3d physics-based simulations where the source remains the same (Le Teil earthquake, France 2019) and the deeper 3d geology is randomly varied within a given range of initial conditions, always for rock-to-hard-rock materials. Results (velocity traces) hold up to 5 Hz. It is envisioned that this set be used for enriching the limited empirical sets of seismic recordings. Some thoughts and concerns are expressed in what follows, which do not challenge the procedure or results per se, but rather its interface with earth science and engineering seismology and its usability at large.

Main/specific comments and concerns:

There are several references to recent work by the same authors (2022, 2023a, 2023b). It would be very helpful to state with even more detail and clarity the relation, differences and originality of the work at hand with respect to those. Two of these past works are also mentioned in the discussion section as 'applications' (presumably of the current work), which adds to the possible confusion of a reader who may not be already acquainted with this research group's output.

If the main purpose of this dataset is to enrich the existing (limited) recorded datasets for scientific machine learning, then a potential user may expect that its creation should follow strict scientific rigor in terms of simulating the physical phenomena involved, be as comparable as possible to recorded data and actual conditions, and not include any realizations that may be deemed unphysical. This in my view implies that:

- The stratification and overall selection of properties (mostly Vs) in the geological models needs to be plausible, and the randomization needs to be constrained by how formations are found in nature. Enough empirical knowledge exists in this field, which can serve to limit/guide the possible random cases based on credibility. For instance, not only the possibility of Vs reversals but also the impedance contrasts between formations need to be considered from the point of view of geological processes, etc. (contrasts in particular are important, as they determine the amplification levels). Lines 271-275 leave it to the user to perform the 'sanity checks' – on the contrary, if implausible data are left in (which they should not, in my view), they should at least bear clear labels/flags. Moreover, not only the variations/randomizations within the range, but also the range itself needs some justification: e.g., Vs ranges from 1070-4500 m/s for an assumed domain down to 8 km depth. Given that values of 3500 m/s are usually considered appropriate within the crust for even deeper seismogenic depths, is not 4500 m/s rather high?

- The geology is taken into account from the depth where Vs exceeds 1070 m/s and downwards. From the point of view of site response, this means that the top dozens or hundreds of meters of what is usually found as near-surface geological materials is ignored, or in engineering terms, eurocode-8 'A-class' rock sites are assumed. This opens some questions: 1. Rock sites often exhibit amplification at high frequencies (say, >8 Hz): these however would be invisible here, if the simulations only reach up to 5 Hz. Conversely, site response <5 Hz (which could actually be seen in the available bandwidth) is typically related to softer materials (say, Vs<600 m/s) which in turn are not included in the models. So either way, it seems like the site response/amplification is not captured fully, despite the effort to consider so many geology variations. 2. The 'spatial sampling' is of 300 m horizontally. However (and it is often the case also for rock sites with Vs>1000 m/s), the lateral variability can be much stronger, which again would mean that wave phenomena within distances <300 m would become invisible in these models, though likely important in nature. 3. Are there many regions where a surface Vs>1000 m/s is deemed probable, so that the synthetics here can represent surface motion? If not (as is my belief), is there any recommendation about how these simulations could be coupled with near-surface analyses that would include additional soft material effects (which are what really modifies ground motion in most observed cases) or even topography effects?

- The variability of the synthetic results should be somehow calibrated to that of empirical ground motion data. This means considering the components of what is known as sigma and its components in GMPEs, and comparing to the statistics of the group of simulations. This is mentioned in passing in line 185 (Convertito paper), but I feel it should be addressed more fully. E.g. fig. 3 shows the spread of the data in terms of PGV - how would it compare to observed data? Also, it would be nice to see more analyses and commentary such as that of fig. 3b and lines 186-187 (effect of varying geology on the variability of simulated ground motion) - this seems to be a rather central point of this exercise not stressed enough. Pages 12-14 focus on capturing variability of the model output from the point of view of number of realizations etc… but I am more concerned about the variability depending on the choice of initial constraints (on Vs, impedance contrast, thicknesses, etc.), which may not be sufficiently well planted into documented reality. This seems to me like a more urgent check to make.

- Making this huge effort only for one specific source (Le Teil seismic event, specific parameters in lines 100+) seems to me to subtract significantly from the usability of this dataset. This event is likely of great interest to France, and such a magnitude is likely of interest to some other stable continental regions, but an entire type of ground motion uncertainty (the between-event variability) is entirely left out of the dataset by keeping to a single magnitude/mechanism/location. It is mentioned as a future prospect/idea to investigate other events, but there also needs to be a reasoning why this study as it stands is self-sufficient and useful for users at large.

How much is PGV (a rather low-frequency parameter) expected to differ from a 'naturally recorded' PGV, given the band limitation of 5 Hz (line 182)? Also, because some disciplines are very strongly interested in acceleration (engineering), i.e. the derivative of the velocity results achieved here, which is much richer in high frequencies, please make some comment as to how viable it would be to derive acceleration time series and PGA values from these simulations, considering the implications of their upper bound of 5 Hz. This is also very important for the computation of Arias intensity (eq. 5), and the reliability of all computed durations hinges on it.
I am concerned about one more thing regarding the significant durations. According to page 11, most simulations have T< 2 or 3 s. However, in the example of page 9 (even if they are velocity traces), the duration by eye seems to be 5 s or more. How is that explained? Also, fig. 4a implies that many synthetics last significantly less than even 1 s, how is that explained? In recorded ground motion datasets, we rarely find such very short durations. Is it an artifact of he 5-Hz limit? If so, please propose a correlation to bring such 'compromised' T values closer to recorded ones.
0ne last thing that I do not understand: the commentary on fig. 5 says that 'significant ground motion happens between times 1.6-17 s'. But the synthetic only exceeds the P-arrival threshold from 2 s onwards. Even if accurate, I am not sure such a plot combining different time series is so meaningful or useful. Please explain its rationale/necessity.

The 'applications' section mentioned in the discussion could benefit from some more elaborating: 1. if basins are created in the mesh, then their fill material needs to be up to 4-5 times slower than the minimum current Vs - meaning 4-5 times longer run time for the analyses? 2. it is unclear how the 2023a and 2023b publications are applications of the current one; 3. as mentioned before, before synthetics can be used to actually infer conclusions about real ground motion, they should first be calibrated on the natural variability of observations; 4. exploring near-surface features is certainly needed, as discussed above, but more details as to the how would be welcome here. 0verall, it feels as if some of the issues 'left for later' could perhaps have been somehow included or at least better considered/discussed in this current effort, which –aside from the large number of realizations- seems somewhat limited in scope.

Other/technical/lesser comments:

In 2.1, it is unclear how these examples relate to 'data used in geophysics and seismology', except in a rather broad way. This reviewer, and likely the average reader, cannot see the relation in scope between CO2 leakage/flow databases and seismic ground motion simulations. On the other hand, the examples in 2.2, which are more closely related to the topic, could be detailed a little more. It is not fully understandable neither from this paragraph nor from the table exactly how they compare to the effort at hand. Please help by providing more context.

Lines 175+: P arrivals (and other phases) for the time being are most often identified by automated procedures e.g. comparing the short-term to the long-term average (STA, LTA), rather than by analysts or machine learning (though this may change soon). Why would this not be used here, and instead a velocity threshold is used?

The state of the art seems to focus a little more on the region of the authors. Though this is not objectionable, I wonder if it would be possible to make some additional references to the other synthetic works performed in other regions, including e.g. the SCEC broadband simulations.

Fig. 2 shows spatial variability of synthetics for 1 realization. It would be nice to add the 'input motion' (time series at the source) for comparison, and also to show the same figure layout for Fourier amplitude spectra (up t 5 Hz).

Please reconsider and possibly amend the number of decimals appearing in the various quantities described. For instance, page 11: duration of a seismogram cannot conceivably be given at an accuracy of 2 decimals of a second.

Line 160, how is the 100-Hz sampling frequency explained in view of the 5-Hz maximum threshold for wave propagation within the numerical grid?

Line 186: isn't scattering (loss of energy rather than its spreading in time) also a possibility?

A few phrases could be reconsidered in terms of use of language: e.g. words like 'incredibly' or 'tricky' can be avoided.