Synthetic ground motions in heterogeneous geologies from various sources: the HEMEW^{S}-3D database

Fanny Lehmann[1,2], Filippo Gatti[2], Michaël Bertin[1], and Didier Clouteau[2]

[1]CEA, CEA/DAM/DIF, F-91297 Arpajon, France
[2]LMPS - Laboratoire de Mécanique Paris-Saclay, Université Paris-Saclay, CentraleSupélec, ENS Paris-Saclay, CNRS, Gif-sur-Yvette, France

Suggestions for revision

The authors have added some metadata labels to the database which were indeed needed, and have revised certain parts of the manuscript and its presentation. However, concerning the point-to-point rebuttal, although the authors have written many replies to the reviewer comments, the majority of them are not reflected in the manuscript. It would be a simple thing to add some of the explanations given to the reviewers into the main text, so that they are available to the general readership in a straightforward way, helping towards a better understanding –and most importantly, a better use of the data on offer. So my main and final recommendation for revision is to add explanations given in the rebuttal (and references) to help clarify/improve the main text. Some examples follow:

The authors thank the reviewer for their detailed comments and their accompanying explanation. Details have been added to the manuscript and a point-by-point response is given below.

1. Points 2.2.7, 1.2.1, 2.1.14: Both reviewers pose the question of how 5 Hz (maximum frequency that can numerically propagate through a grid) and 100 Hz (sampling rate) are really reconciled. The (identical) reply given to both reviewers is this: "100 Hz matches the usual temporal resolution of recorded time series available in public accessible earthquake engineering strong motion databases which is important for tasks such as seismic phase picking". Yet this is not explained in the revised manuscript, but only given as a personal reply in the rebuttal. But then the reader, who will likely ask him/herself the same, cannot benefit in the end. He/she should not have to read the commentary exchange in order to get the necessary clarifications for the article, so please explain your rationale in the paper.

A note regarding this specific reply: Please rephrase this explanation before adding it to the manuscript, because it is incorrect on a few accounts:

1. earthquake engineers do not access strong motion datasets to do phase picking, which is a purely seismological task/skill

2. the reason for the high sampling in strong-motion data is not for the sake of phase picking (wave windowing can be very rough in such applications, in stark contrast to seismic monitoring) – this investment is made in order to be sure to catch PGA correctly

3. in many important networks, the sampling rate of accelerometric data is actually not even 100 Hz but 200 Hz

Explanation for the choice of 100 Hz have been given l. 210-214, taking into account the reviewer's suggestions. It now reads "Although the sampling frequency is higher than the Nyquist frequency (i.e. $2 \times f_{max} = 10\,\text{Hz}$), the value of 100 Hz was chosen to match the temporal resolution of recorded time series in several publicly accessible datasets (e.g. STEAD [Mousavi et al., 2019]), INSTANCE [Michelini et al., 2021]). The sampling frequency is sufficient to allow an accurate computation of Peak Ground Velocity (PGV), derive the acceleration time series with finite differences, and compute the Peak Ground Acceleration (PGA). "

2. Points 2.1.2, 2.1.3, 1.1.1, 1.1.2, 2.1.5, 2.1.19: Again both reviewers pointed this out: the choice to include unphysical instances of various parameter values in the database. If the authors agree that some of the models are unrealistic from a geological/geophysical/seismological point of view, then please stress this in the text, and explain why you think there is this dire necessity to include them nevertheless for ML purposes. Also, and this is something I'd like to stress, please be very clear on what percentage of the data can be related to unrealistic, or in statistical terms, "extremely rare" or "coda" cases. Because it seems as if these rare cases may actually take up a lot of the database: from the numbers the authors give, it seems like a ratio of 1:3 between rare/unrealistic and normal (10,550 out of 30,000?), which seems too high, so is the coda being sampled or oversampled in the end? Please be clear on the statistics.

To say that the 'plausibility of data depends on the application' is, I think, a compromise detrimental to the earth science applicability of the work, in favor of ML. But natural occurrence does need to have a role here. (E.g. one may well sample 1,000,000 soil samples but will never get a density of, say, $5t/m^3$, and even so, it would certainly not happen 30% of the time.) So please add explicit commentary to the paper about all this. If 2/2 reviewers felt the need to bring it up, most earth scientists in the audience are likely to have similar questions.

Apart from agreeing that a density of $5\,\mathrm{t/m^3}$ is truly unrealistic and stating that no such case is included in the database, it is believed that neither the authors nor the reviewers could give an objective measure of what is a "realistic" model and thus, be able to compute the percentage of such cases in the HEMEW$^{\mathrm{S}}$-3D database. If we were to take impedance contrast as as criterium to discriminate between two types of geologies, the proposed database contains less than 1/3 of geologies with a minimum impedance contrast lower than 0.7 (l. 403-404). However, from a machine learning perspective, it is extremely important to provide *out-of-distribution* examples in order to demonstrate the generalization capabilities of proposed methods. Hence, our database offers users to select their own measure of *in-distribution* geologies on which their model will be trained.

3. Point 2.1.6: amplification. I find the arguments of the authors incomplete with respect to the existing knowledge on site response. Even so, please give your arguments in the paper explaining if, how and why amplification is or is not accounted for in your calculations, especially with respect to the frequency range <5 Hz (which may well be impressively high for such calculations but is lower than the range where hard sites amplify), and especially considering the large proportion of high-Vs cases. It is ok to say that it is not accounted for completely, but is at least dealt with better than it was in past papers, or is out of scope, etc. However, I think it is not ok to claim that there is absolutely nothing to talk about here and just ignore the issue: limitations exist and should be stated.

Ground motion amplification and site effects due to soft sediments are not the targeted applications of the HEMEW$^{\mathrm{S}}$-3D database. An improper formulation of the section 5.3. on "Other potential applications" led to confusion on this point. This section has been rephrased. In addition, the limitations section 6. now clearly mentions that the HEMEW$^{\mathrm{S}}$-3D database is not suited for site effects related to sedimentary basins (l. 384-385).

4. Limitations related to calculation speed/run time and memory/storage needs seem to underpin many of the decisions made and/or are the answer to many of the reviewer comments (duration of waveforms, spatial sampling, inability to fill basins with soft material, etc). It would be good to explain all these cases together in the end of the paper. So to speak, answer the question: when computing becomes faster/easier in future, what would be the top 5 things you'd like to do differently, without the need to worry about such issues?

As HEMEW$^{\mathrm{S}}$-3D is the first large-scale 3D database of seismological simulations for machine learning, it was crucial to ensure that the simulation objectives were reachable and that their outcomes remained manageable, both in terms of the memory constraints to store/download/reuse data and in terms of data variability that could make machine learning tasks too complex to be learnt. This is why feasibility concerns dictated several choices. Now that many predictive tasks have been proven possible, one can envision meaningful ways to extend the database. Following the reviewer's suggestion, perspectives are now listed l. 410-417.

5. points 2.1.8: I did not find this new explanation in the new text about surface velocity. Please add if missing, because it is very important the reader understands how you define 'surface velocity (30 or 300 m!). In the majority of locations in the world where seismic hazard is a concern, we would love to have 'near-surface' Vs of 1000m/s or more, but don't! Unless the dataset is more representative of certain regions (France? stable continental areas?), which it claims it is not. Also, please state in main text (as perspectives – we know it is not in the scope of this paper) if/how your methods and data could be combined with near-surface site effects calculations.

Explanation about the understanding of surface velocity was given at the beginning of the Limitations section (now l.385-389 in the revised manuscript). It is now referred to as upper velocity instead of surface velocity to avoid confusion. The definition of the upper velocity comes from the vertical resolution of **regional** geological models, which is 300 m. This means that Gauss points between 0 m and −300 m have the same value. As a consequence, the upper velocity should be understood as a 300-m average and cannot be compared with more common definitions, such as $V_{S,30}$.
Concerning the applications for near-surface site effects, the surface ground motion can be considered as an "outcropping bedrock" response which is classically used in 1D site-effect and Soil-Structure Interaction analyses, and which may require deconvolution (l. 376-377).

2.1.9: GMPEs. Although many GMPEs exist that are informed by simulations, even assuming they were all empirical, they still are a key tool in practice. And so if your database were to show a great divergence from what they predict, it would be extremely important to point it out. A comparison would be beneficial, and if there is disagreement then the various arguments the authors give can be proposed to explain why their work is better fitted than GMPEs for such and such a case. It is not a matter of believing in data more than in simulations, but there is an urgent need that the two communities finally start to acknowledge each other for science to move forward faster. Please help in this direction.

Comparisons with four GMPEs by [Atkinson, 2015], [Atkinson and Boore, 2006], [Chiou and Youngs, 2014], and [Shahjouei and Pezeshk, 2016] have been added in Fig. 8. They show a good agreement between the PSA computed from the HEMEW^S-3D database and the PSA from GMPEs. Figure 8 is reproduced below

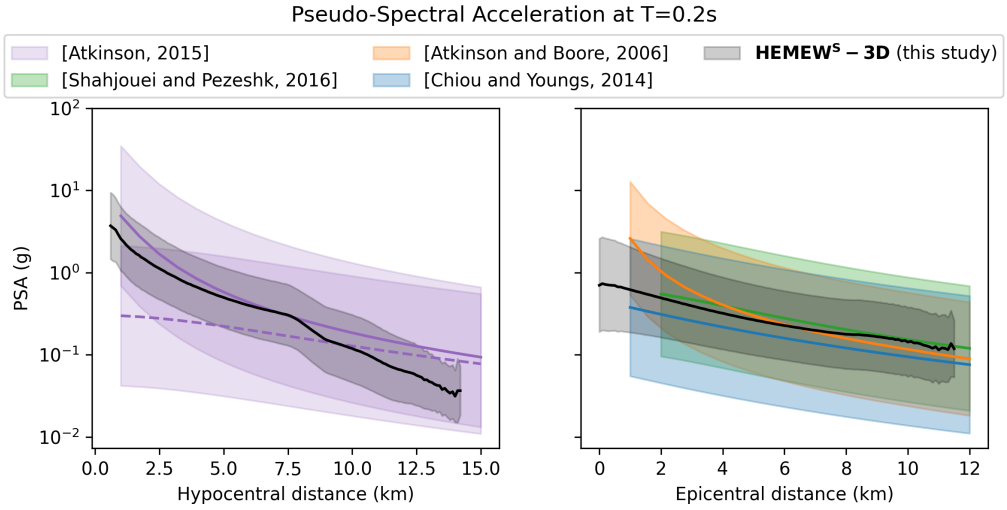

Figure 1: PSA at period T=0.2 s as a function of hypocentral distance (left) and epicentral distance (right) for GMMs by [Atkinson, 2015] (purple, left), [Atkinson and Boore, 2006] (orange, right), [Chiou and Youngs, 2014] (blue, right), [Shahjouei and Pezeshk, 2016] (green, right), and our HEMEW^S-3D database. Solid lines correspond to the mean PSA and shaded areas to one standard deviation.

2.1.15: please make this clarification in text about durations and lack of content

Details have been added l. 250-251 and now read "These short RSD values are related to the absence of high-frequency components in the coda and the dominance of high pulse-like time series in cases with shallow sources and low heterogeneity contrasts."

2.2.4: this was not answered

A dedicated paragraph 2.1. has been added in the Related work Section.

Outside reviewer bullet points: On lowering duration from 20 sec to 8 sec: Please include a phrase about why 8 sec only is sufficient duration (maybe based on distance and M combinations)

The choice of 8s has been justified l.251-252 from the consideration of P-wave arrival time and Relative Significant Duration that gives an estimate of the final time of significant ground motion.

References

[Atkinson, 2015] Atkinson, G. M. (2015). Ground-Motion Prediction Equation for Small-to-Moderate Events at Short Hypocentral Distances, with Application to Induced-Seismicity Hazards. *Bulletin of the Seismological Society of America*, 105(2A):981–992.

[Atkinson and Boore, 2006] Atkinson, G. M. and Boore, D. M. (2006). Earthquake Ground-Motion Prediction Equations for Eastern North America. *Bulletin of the Seismological Society of America*, 96(6):2181–2205.

[Chiou and Youngs, 2014] Chiou, B. S.-J. and Youngs, R. R. (2014). Update of the Chiou and Youngs NGA Model for the Average Horizontal Component of Peak Ground Motion and Response Spectra. *Earthquake Spectra*, 30(3):1117–1153.

[Michelini et al., 2021] Michelini, A., Cianetti, S., Gaviano, S., Giunchi, C., Jozinović, D., and Lauciani, V. (2021). INSTANCE – the Italian seismic dataset for machine learning. *Earth System Science Data*, 13(12):5509–5544.

[Mousavi et al., 2019] Mousavi, S. M., Sheng, Y., Zhu, W., and Beroza, G. C. (2019). STanford EArthquake Dataset (STEAD): A Global Data Set of Seismic Signals for AI. *IEEE Access*, 7.

[Shahjouei and Pezeshk, 2016] Shahjouei, A. and Pezeshk, S. (2016). Alternative Hybrid Empirical Ground-Motion Model for Central and Eastern North America Using Hybrid Simulations and NGA-West2 Models. *Bulletin of the Seismological Society of America*, 106(2):734–754.