nature portfolio

Peer Review File

A machine learning estimator trained on synthetic data for real-time earthquake ground-shaking predictions in Southern California

Web links to the author's journal account have been redacted from the decision letters as indicated to maintain confidentiality.

13 Nov 2023

Dear Dr Monterrubio-Velasco,

Please allow us to apologise for the delay in sending a decision on your manuscript titled "Ground Motion Shaking Predictions Based on Machine Learning and Physics-based Simulations". It has now been seen by 3 reviewers, and we include their comments at the end of this message. They find your work of interest, but some important points are raised. We are interested in the possibility of publishing your study in Communications Earth & Environment, but would like to consider your responses to these concerns and assess a revised manuscript before we make a final decision on publication.

We therefore invite you to revise and resubmit your manuscript, along with a point-by-point response that takes into account the points raised. Please highlight all changes in the manuscript text file.

In particular, please ensure that the revised manuscript meets the following editorial thresholds:

* Clearly discuss and account for the limitations associated with your approach and it's relative utility to other regions lacking in such extensive data.

* Fully explain and justify the metrics and parameters used in your model and ensure that your method is entirely reproducible.

* Place your work and it's advance firmly in the context of other recent relevant studies.

We are committed to providing a fair and constructive peer-review process. Please don't hesitate to contact us if you wish to discuss the revision in more detail.

Please use the following link to submit your revised manuscript, point-by-point response to the referees' comments (which should be in a separate document to any cover letter), a tracked-changes version of the manuscript (as a PDF file) and the completed checklist:

[link redacted]

** This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first **

We hope to receive your revised paper within six weeks; please let us know if you aren't able to submit it within this time so that we can discuss how best to proceed. If we don't hear from you, and the revision process takes significantly longer, we may close your file. In this event, we will still be happy to reconsider your paper at a later date, as long as nothing similar has been accepted for publication at Communications Earth & Environment or published elsewhere in the meantime.

Please do not hesitate to contact us if you have any questions or would like to discuss these revisions further. We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Best regards,

Joe Aslin

Senior Editor,
Communications Earth & Environment
https://www.nature.com/commsenv/
Twitter: @CommsEarth

EDITORIAL POLICIES AND FORMATTING

We ask that you ensure your manuscript complies with our editorial policies. Please ensure that the following formatting requirements are met, and any checklist relevant to your research is completed and uploaded as a Related Manuscript file type with the revised article.

Editorial Policy: Policy requirements (Download the link to your computer as a PDF.)

Furthermore, please align your manuscript with our format requirements, which are summarized on the following checklist:
Communications Earth & Environment formatting checklist

and also in our style and formatting guide Communications Earth & Environment formatting guide .

*** DATA: Communications Earth & Environment endorses the principles of the Enabling FAIR data project (http://www.copdess.org/enabling-fair-data-project/). We ask authors to make the data that support their conclusions available in permanent, publically accessible data repositories. (Please contact the editor if you are unable to make your data available).

All Communications Earth & Environment manuscripts must include a section titled "Data Availability" at the end of the Methods section or main text (if no Methods). More information on this policy, is available at http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf.

In particular, the Data availability statement should include:
- Unique identifiers (such as DOIs and hyperlinks for datasets in public repositories)
- Accession codes where appropriate
- If applicable, a statement regarding data available with restrictions
- If a dataset has a Digital Object Identifier (DOI) as its unique identifier, we strongly encourage including this in the Reference list and citing the dataset in the Data Availability Statement.

DATA SOURCES: All new data associated with the paper should be placed in a persistent repository where they can be freely and enduringly accessed. We recommend submitting the data to discipline-specific, community-recognized repositories, where possible and a list of recommended repositories is provided at http://www.nature.com/sdata/policies/repositories.

If a community resource is unavailable, data can be submitted to generalist repositories such as figshare or Dryad Digital Repository. Please provide a unique identifier for the data (for example a DOI or a permanent URL) in the data availability statement, if possible. If the repository does not provide identifiers, we encourage authors to supply the search terms that will return the data. For data that have been obtained from publically available sources, please provide a URL and the specific data product name in the data availability statement. Data with a DOI should be further cited in the methods reference section.

Please refer to our data policies at http://www.nature.com/authors/policies/availability.html.

REVIEWER COMMENTS:

Reviewer #1 (Remarks to the Author):

The authors developed the machine-learning-based predictor of ground-motion intensity using many physics-based simulations of CyberShake. They demonstrated that the developed predictor has better prediction performance than the conventional regression model. Their approach of using many physics-based simulations as the training dataset is useful in overcoming the limitations of observations that limit predictor performance. This paper is interesting and siginificant for the field of ground motion prediction; however, there are several points for improvement. Therefore, I rate this paper as "major revision."

The machine-learning-based predictor of ground-motion intensity using many physics-based simulations was proposed by Withers et al. (2020). The authors should cite this paper and clarify the differences between their study and the previous study.

Withers KB, Moschetti MP, Thompson EM (2020) A machine learning approach to developing ground motion models from simulated ground motions. Geophys Res Lett 47:e2019GL086690. https://doi.org/10.1029/2019GL086690

The simulations of CyberShake are based on the finite fault model. The effect of finite fault is large in the simulation result of ground motion. However, the authors used the hypocentral location and the hypocentral distance as explanatory variables of the predictor. That is, the authors adopted the point-source assumption ignoring the finite-fault effect on ground motions. The authors should clarify this point in their paper and discuss the impact of using the point source assumption on the predictor.

The author used RMSE to evaluate the predictive performance of each predictor. However, RMSE does not provide any information on prediction bias (underestimation or overestimation). By looking at the residual between observations and predictions (e.g., log O/P), the prediction bias of each model can be evaluated. This should be addressed by the authors.

The authors demonstrated the prediction performance of their prediction in real records of five earthquakes. Although RMSE for each earthquake is shown in Fig. 5, there is no figure of map view figure like Fig. 3. The authors should add such figure.

The prediction of MLESmap has a similar feature to conventional regression model or not? The distance attenuation should be checked.

Although the authors mentioned the explanatory variables in “MLESmap: Deep Neural Network topology,” there is no mention of explanatory variables in “MLESmap: RF regression algorithm.”

What is the activation function?

Please clarify the input to ASK-14.

Reviewer #2 (Remarks to the Author):

The author presents Machine Learning (ML) strategies trained using one of the largest existing datasets (with more than 100M simulated seismograms) from CyberShake Study 15.4, developed by the Southern California Earthquake Center, covering the Los Angeles basin region. Moreover, the ML strategies were tested with a set of unseen samples from the synthetic database and real historical earthquakes. In my opinion, the idea behind the manuscript is very interesting, considering the growing application of machine learning approaches to support seismic risk mitigation. However, the manuscript does have some flaws. The main one is the underestimation of the limitations of the proposed approach by the authors, even though they state, "It should be noted that MLESmap models perform better than GMMs as long as the earthquakes are compatible with the CSS-15.4 dataset, that is, the locations, frequencies, and magnitudes interrogated are within the training dataset’s bounds." In my opinion, such limitations should be highlighted in both the abstract and the introduction. Additionally, the mentioned section should better emphasize other limitations of the proposed strategy: "The computation of the CSS-15.4 study required a total of 37.6M hours on tier-0 supercomputing facilities.". Hence, concerning the "trained area," ML outperforms the GMM thanks to the CSS-15.4 effort. In other words, the main potential of ML, to move towards a more accurate rapid response solution by combining the accuracy of physics-based simulations with the fast estimations given by GMMs, falters if one wants to apply the proposed ML strategies worldwide. Moreover, the proposed procedure cannot account for site conditions (topography, vertical and lateral heterogeneity) in other areas. To some extent, thanks to the huge amount of training data, the site effect can be evaluated, limiting it to the studied area. Hence, why should one adopt this strategy rather than what is provided by other researchers, including key parameters that allow accounting for site conditions? For instance, (Bindi et al., 2014; Kotha et al., 2020; Kubo et al., 2020; Mori et al., 2021; Zhu et al., 2022) provides ML and GMM including such parameters. What is the authors' opinion? In my opinion, the authors should include a discussion in the introduction regarding the points mentioned above.
Furthermore, a consideration about the selected periods for the prediction should be provided. In fact, predicting ground shaking for periods ranging from 0.1 to 2 s is crucial for seismic risk mitigation

strategies, especially concerning ordinary structures and infrastructures.
Some statements should be rephrased. For instance, "We propose a novel Machine Learning (ML) methodology that combines the best of both approaches, capturing physical information from simulations (directivity, topography, site effects) with times-to-solution analogous to empirical GMMs, helping to produce the next generation of shaking maps." In fact, directivity, topography, and site effects can only be accurately captured for the area of interest, thanks to the vast amount of available data. The ML strategy does not require input data such as slope, mean shear wave velocity in the upper 30 m, or key parameters for describing potentially buried morphology. Therefore, predicting directivity, topography, and site effects in other areas is expected to clearly fail. In this line, a brief discussion about site effects should be included ((Assimaki et al., 2005; Gazetas, 1982; Moczo et al., 2018)).
Furthermore, in my opinion, the title should be modified. The current title leads the reader to think of a combined approach, while, if I am correct, the results from ML and CSS-15.4 are not merged to provide a single Map. For instance, "Ground Motion Shaking Predictions Based on Machine Learning trained with data from Physics-based Simulations".

Finally, some valuable references should be added.

CONCLUDING RECOMMENDATION: Resubmit after major revisions.

Detailed comments are listed below.
English. Please check the English used and style. The use of present perfect not appropriate for actions concluded in the past (please, use the past tense when referring to actions concluded in the past).
Table 1. "NN" should be "DNN".
SI-Fig. 3 caption. "Mw=7.45" should be "Mw = 7.45". Leave a blank space before and after =. Please check the whole manuscript.
Lines 118-119. Please provide dataset bounds of the mentioned earthquake characteristics.
Line 249. ".RMSE ranges" should be ". RMSE ranges". Leave a blank space after the dot.
Tables. Please ensure that unit measure is included in all the tables.
R2 and r2. Please check the use or not of the capital letter.

REFERENCES
Assimaki, D., Gazetas, G., Kausel, E., 2005. Effects of Local Soil Conditions on the Topographic Aggravation of Seismic Motion: Parametric Investigation and Recorded Field Evidence from the 1999 Athens Earthquake. Bulletin of the Seismological Society of America 95, 1059–1089. https://doi.org/10.1785/0120040055

Bindi, D., Massa, M., Luzi, L., Ameri, G., Pacor, F., Puglia, R., Augliera, P., 2014. Pan-European ground-motion prediction equations for the average horizontal component of PGA, PGV, and 5%-damped PSA at spectral periods up to 3.0 s using the RESORCE dataset. Bulletin of Earthquake Engineering 12, 391–430. https://doi.org/10.1007/s10518-013-9525-5

Gazetas, G., 1982. Vibrational characteristics of soil deposits with variable wave velocity. Int J Numer Anal Methods Geomech 6, 1–20. https://doi.org/10.1002/nag.1610060103

Kotha, S.R., Weatherill, G., Bindi, D., Cotton, F., 2020. A regionally-adaptable ground-motion model for shallow crustal earthquakes in Europe. Bulletin of Earthquake Engineering 18, 4091–4125. https://doi.org/10.1007/s10518-020-00869-1

Kubo, H., Kunugi, T., Suzuki, W., Suzuki, S., Aoi, S., 2020. Hybrid predictor for ground-motion intensity with machine learning and conventional ground motion prediction equation. Scientific Reports 2020 10:1 10, 1–12. https://doi.org/10.1038/s41598-020-68630-x

Moczo, P., Kristek, J., Bard, P.Y., Stripajová, S., Hollender, F., Chovanová, Z., Kristeková, M., Sicilia, D., 2018. Key structural parameters affecting earthquake ground motion in 2D and 3D sedimentary structures. Bulletin of Earthquake Engineering 16, 2421–2450. https://doi.org/10.1007/s10518-018-0345-5

Mori, F., Mendicelli, A., Falcone, G., Acunzo, G., Spacagna, R.L., Naso, G., Moscatelli, M., 2021. Ground motion prediction maps using seismic microzonation data and machine learning. Natural Hazards and Earth System Sciences Discussions 22, 1–25. https://doi.org/10.5194/NHESS-22-947-2022

Zhu, C., Cotton, F., Kawase, H., Nakano, K., 2022. How well can we predict earthquake site response so far? Machine learning vs physics-based modeling. Earthquake Spectra. https://doi.org/10.1177/87552930221116399/ASSET/IMAGES/LARGE/10.1177_87552930221116399-FIG14.JPEG

Review of the manuscript titled "Ground Motion Shaking Predictions Based on Machine Learning and Physics-based Simulations" by Monterrubio-Velasco et al.

Ground motion prediction, in conventional practice, relies on statistical evaluation of worldwide recordings from past earthquakes, which is challenged by data scarcity, particularly for large and rare events, in addition to the impossibility of reasoning physical driving mechanisms. Physics-based modeling has been the main complementary approach to overcome the drawbacks of the statistical models, but to focus on specific investigations or reduce computational costs, their use has been mostly limited to few deterministic scenarios. The recent advances in machine learning stands therefore as an opportunity to empower physics-based models for handling computationally challenging tasks such as uncertainty quantification and data mining.

This manuscript presents one of the recent attempts to accelerate physics-based ground motion prediction by developing machine-learning based tools. Specifically, the authors use thousands of simulations of M6+ hypothetical earthquakes for southern California from CyberShake platform to train two different machine-learning algorithms and predict spectral accelerations at four different periods. In addition to testing set from the same database, the authors also verify their methodology on a select of real earthquakes in the region. They report a good performance for both tests and conclude validation of their method.

I think that the study would be a nice contribution to the literature, once certain points are addressed. I am especially confused by the issues below.

- The study is oriented towards shaking maps. Throughout the manuscript, shaking intensity and ground motion are interchangeably used, despite the two terms stand for different notions (see for example, USGS). I can understand that the utility of the developed tools includes near-real-time shaking map development, but the current version rather reads like a ground motion prediction study. If the main objective is indeed to accelerate shake map development, I would expect to see the application of the algorithms for shake maps and validate against real cases. Otherwise, as a ground motion prediction study, the writing needs reorganization which should clearly state the goal of predicting a specific ground motion metric, in a specific area as a function of magnitude and distance, faster than physics-based models. Such reorganization would also benefit from mentioning the existing machine-learning studies in the field, such as reduced order models, generative models, etc. (e.g., Khosravikia & Clayton, 2021 https://doi.org/10.1016/j.cageo.2021.104700, More et al., 2022 https://doi.org/10.5194/nhess-22-947-2022, Xu et al., 2022 https://doi.org/10.1080/13632469.2020.1826371, Florez et al., 2022 https://doi.org/10.1785/0120210264, and Mousavi and Beroza, 2023 https://doi.org/10.1146/annurev-earth-071822-100323), and what makes this study innovative compared to other studies.

- A detailed discussion on the physical aspects of the method and findings is missing. For example, the motivation often underlines that large events are more destructive, and the end product of the study would serve to post-disaster efforts. Yet, the most destructive part of an earthquake, or what governs peak acceleration, engineering design, and intensity distribution, is generally high-frequency ground motion (above ~1 Hz). The predictions here, on the other hand, are limited to 0.1-0.5 Hz, which is quite a low and limited frequency band, and the numerical resolution of the synthetic database is 1 Hz.

- Another issue is that the database simulations are likely to include event-specific source effects and site effects, since the study area is highly complex in terms of fault networks and crustal geological heterogeneity. The focus on very large events of M 7+, particularly, would contribute to event-specific spatial variability of ground motion. When looking at the training station grid (Fig. 6a) with respect to fault lines, stations are all in the near field, i.e., near fault. The choice of near field is unclear to me. In other words, wouldn't it be more straight forward to start with distant stations in the same area? The current version is not informative about whether the predictive power manifests repeated source or site effects.

- Lastly, I had difficulty to understand the reason of making comparisons to GMMs — empirical models. In general, such comparisons are made to verify far-field predictions of the physics-based models; for near-field, GMMs are less credible given their very large variabilities, and this is where physics-based models complement GMMs. Besides, such comparison is made for a synthetic event from the CyberShake database. It reads like authors verify the database, which is out of the objective of the study. Regarding real case applications, I again don't see the reason of comparisons with GMMs. The authors may be showing it to measure relative performance of different approaches, but it is unclear.

To clarify these issues, I provided my suggestions, together with minor comments, in the following.

Best regards,
Elif ORAL

<u>Major comments</u>

- Training dataset mostly includes very large and rare events (for the region), whereas ground motion station grid is quite small compared to fault sizes (red lines vs magenta points in Fig. 6a). Latest GMMs and physics-based models define the distance by shortest distance to fault, instead of hypocentral distance, to minimize directivity-related variabilities. Please justify your choice of using hypocentral distance for such large events and near field ground motion. Please also provide the range of hypocentral distances in the database.

- I would expect the first figure to inform about the big picture. Something like an assembly of Fig. S1 and Fig S6. Showing earlier the area that you train and test can make it easier to follow the text.

- Figures 1 and 2 show very similar patterns for different periods. You may consider only showing the representative cases and provide the remaining in Supplementary Material.

- Shaking intensity does not mean ground motion. I suggest using "ground motion metric," or "spectral acceleration" or the original definition "GMRotD50" throughout the manuscript.

- Please provide reference for GmRotd50 (Boore et al., 2006 https://doi.org/10.1785/0120050209)

- Since this study focuses on a single metric, please justify your choice of GmRotD50. It is orientation independent, which is obtained after a nonlinear change to acceleration data, i.e., you lose component-related information — potentially inherited data. You could have used for example peak ground motion. In addition, considering the low frequency resolution, peak ground displacement or velocity would be more informative, instead of acceleration or spectral acceleration.

- In general, please reorganize the writing to avoid confusion about intensity, ground motion, GMM use. For example, it should be clear to the reader why you work on this topic (shake map, or ground motion prediction?), why you choose such methodology, and why you focus on southern California (more data?).

- In general, current version highlights several times predictive power/good performance given small errors. Please also discuss why you get good/bad predictions with respect to physics-based knowledge.

- Fig 3: It is hard to understand why you evaluate the performance of an empirical method on a synthetic event. It is expected to be the other way around.

- Fig 6: In the same subplot (a) or another, you can also provide source area (rupture extents), that is included in synthetic database. This is important because, since you only use low-frequency ground motion metrics, ground motion will be dominated by either by basin resonance around LA (site effects), or since the source is very close, a combination of source and site effects. You can refer to previous studies on LA basin, to leave out or judge the significance of basin effects on the frequency band you focus on (e.g., Wald and Graves, 1998 https://doi.org/10.1785/BSSA0880020337; Kohler et al., 2020 https://doi.org/10.1785/0220200170)

- Looking at Figs. 4-5, the source is inside the station grid, for the cases (c) and (e). I suspect that this might be the reason you get largest errors for case (c). Are these real events covered by the database (at least by focal mechanism, rupture speed, slip distribution with respect to published studies on the events). Even if it is not, please report.

- Extrapolating for smaller and larger magnitude should be different. You see actually a good prediction for Mw 5.9 event, the case (e). This might relate to the fact that for smaller magnitude, rupture and near-field extent is smaller; you may be better capturing distance-dependent attenuation. Please comment.

- Fig. 5: It seems that overall you better predict the lowest frequency (T=10 s), which is expectedly the case in physics-based models (for example, Fig. 7c for distances > 10 km, in Oral et al., 2022 https://doi.org/10.1785/0120220064, and references therein). Please highlight the consistency.

- Please comment on why Random Forest seems to better perform? Or how do you explain that one method works better in some cases, and not so good in some other cases? Why DNN seems so bad for Fig. S5?

- How did you decide on DNN architecture? By trial and error, or after a reference study?

- Looking at Fig. 5, GMM and ML predictions look very close, which contradicts with your statement of “ML outperforms GMM.” Instead please quantify differences and comment on their significance.

- Discussion: Please also discuss potential directions that use larger sizes for other metrics and/or algorithms. You may consider citing other studies in the literature at this point.

- L131: I suggest to briefly mention data quality.

- L134: I again suggest to volume down this claim of ML better predicting than GMM, and to highlight the potential improvement of GMM when incorporating physics-based data through ML.

- L136: In addition to the computational aspect of the problem, highlighting the uncertainty quantification or enlarging datasets from a couple of deterministic cases to explore much larger parameter space deserves being mentioned. You say it only for focal mechanism, but especially for large events, it is beyond that (see kinematic vs dynamic rupture modeling differences, for example).

- L143: GMMs also are challenged by that. Overall, I suggest avoiding ML vs GMM, and rather highlight the existing challenge concerning data availability and physics-based knowledge. All approaches are at the end complementary.

- L153-L167: You can simplify this paragraph. Mainly you can highlight that simulations are based on kinematic source models, and wave propagation is calculated using the region-specific velocity profiles, and provide references for more details.

- Please comment on the limitation arising from kinematic source models, such as constant rupture speeds, preset slip distributions (even if they are stochastic). You can mention it as a potential improvement in future perspectives.

<u>Minor comments</u>

- Title: Ground motion prediction is the commonly used term. You may also consider to highlight the region of study.

- L13: All earthquakes are unpredictable. Please be specific. Do you mean that variabilities/uncertainties increase? If so why? Data scarcity? Uncertainties in physics-based model inputs?

- L19: Relief measure?

- L43: “just after the earthquake occurs.” Do you mean recordings?

- L33: times-to-solution?

- L58: Please avoid hype words, such as very good, excellent, and instead quantify.

- Fig. 3: Where is the hypocenter in this event? Pinky cross?

- Fig. 4 : You may consider adding a rectangle showing the boundary that separates inside/outside stations.

- Fig 6: Magenta markers are not star in my Pdf.

- L125: Stochastically attained freq.? Unclear. Do you mean stochastic methods of ground motion modeling? Even so, I suggest to generalize physics-based rupture and ground motion modeling.

- L130: “unrecorded” events: do you mean large and rare events?

- In general, you limit the benefit of this study to post disaster emergence response, but technically it is also applicable for hazard assessment as a part of preparedness phase, i.e., before and during hazard.
- L136: This reads contradictory. For 1 Hz, the problem arises from uncertainties associated with input parameters and idealizations in the physical processes. You can rather make your point for high-frequency simulations.
- L144: Any reference?
- In general, providing references along the discussions would enrich the manuscript.
- L149: Please avoid too generic phrases.
- L159: Seismic reciprocity?

Dear Reviewers,

We would like to thank you for your time and effort to provide such valuable comments that helped us to significantly improve the manuscript.

In this document, we addressed all comments and questions raised by the three reviewers. We highlight in yellow the specific changes that we have added in the manuscript. The corresponding changes in the manuscript itself are marked in red. In the supplementary material we also included Figures SS-Fig. 7, 8 and 9, as suggested by reviewer 1.

We hope that the implemented changes thoroughly address all questions and that you find them satisfactory.

Thank you for your consideration of this revised manuscript,

Best regards,

Marisol Monterrubio-Velasco and co-authors.

Reviewer #1:

The authors developed the machine-learning-based predictor of ground-motion intensity using many physics-based simulations of CyberShake. They demonstrated that the developed predictor has better prediction performance than the conventional regression model. Their approach of using many physics-based simulations as the training dataset is useful in overcoming the limitations of observations that limit predictor performance. This paper is interesting and significant for the field of ground motion prediction; however, there are several points for improvement. Therefore, I rate this paper as “major revision.”

1. **The machine-learning-based predictor of ground-motion intensity using many physics-based simulations was proposed by Withers et al. (2020). The authors should cite this paper and clarify the differences between their study and the previous study.**

Below we identify differences between Withers KB, Moschetti MP, Thompson EM (2020) A machine learning approach to developing ground motion models from simulated ground motions. Geophys Res Lett 47:e2019GL086690. https://doi.org/10.1029/2019GL086690 and our work. We also added the Withers et al. (2020) reference in the Introduction and discussed the key differences.

Objectives:

Similarities: Test ML models in order to build next-generation ML-based Ground Motion Models for estimating intensity measures.	
Differences:	
Withers et al. (2020)	MLESmap
This article outlines a machine learning approach trained on synthetic data to build a new GMM.The focus is on the comparison with the Next Generation Attenuation-West2 (NGA-West2) models, so they use the same predictor variables that were used in NGA-West2 models (see below for more detail).	MLESmap is focused on rapidly generating Ground Motion maps after the occurrence of an event of large magnitude. In order to fulfill this objective, we generate ML GMMs with predictor variables that are limited to the simple parameters that can be quickly obtained from the international agencies immediately after an earthquake (see below for more detail).

Database:

Similarities: The database used in both papers comes from the same CyberShake 15 study.	
Differences:	
Withers et al. (2020)	MLESmap
Withers et al. (2020) used CyberShake Study 15.12 (https://scec.usc.edu/scecpedia/CyberShake_Study_15.12) that takes the deterministic results from CyberShake Study 15.4 used in our study and adds broadband components. Cyber Shake	In MLESmap we used the CyberShake study 15.4 (https://strike.scec.org/scecpedia/CyberShake_Study_15.4), which is a computational study to calculate one physics-based probabilistic seismic

Study 15.12 is thus based on deterministic simulations that resolved ground motions up to 1 Hz and uses a hybrid approach (Graves & Pitarka, 2010) to predict ground motions at frequencies above this resolution (up to 10 Hz).CyberShake 15.12 includes 330 sites, and 10,000 variations of fault geometries.	hazard model for Southern California at 1 Hz (using CVM-S4.26-M01, the GPU implementation of AWP-ODC-SGT, the Graves and Pitarka (2014) rupture variations with uniform hypocenters, and the UCERF2 ERF). We thus consider only the deterministic simulations up to 1Hz, without the higher frequencies used in Withers et al. (2020).

Predictor Variables :

Differences:	
Withers et al. (2020)	MLESmap
Withers et al. (2020) used the same predictor variables as those used in NGA-West2 records, in order to replicate the same model dependencies that were used in the mixed-effects regressions, namely: From Cybershake → Mw, rake, dip, width, and others not specified Computed from fault geometry to station locations → Rx, Rrup,Rjb. Velocity and site effect → Vs30, Z1.0, Z2.5 were computed by extracting shear wave velocity profiles from the velocity model used in the CyberShake simulations (CVM-S4.26.M01) at 20 m depth intervals.	The predictor variables used in this work are much simpler and much more general than those of Withers et al,: Mw, hypocenter location, station location, euclidean distance and azimuth between the hypocenter and station. All our predictors can be rapidly estimated following a significant event and thus a new map can be proposed quickly. Although the target is the RotD50 for both Withers et al. (2020) and MLESmap, we used the logarithmic value of RotD50 since we observe a better accuracy in the predictions of the ML algorithms.

Methodology:

Differences:	
Withers et al. (2020)	MLESmap
They split the data set into a training and testing data set (80/20 split, respectively), based on station locations, to minimize cross contamination (the overlap between training and testing data sets) and overtraining. They normalize the input data into the approximate range [-1:1] to allow the activation function (sigmoidal) to find the best fitting weight coefficients in the stochastic gradient descent iterations.	We split the data set into a training and testing data set (90/10 split, respectively), based on the earthquake scenario: that is, we consider all stations for every event (scenario). We did a global normalization using the MaxMinScaler [0,1] for each predictor.

ML algorithms and NN topology:

Differences:	
Withers et al. (2020)	MLESmap
They use a shallow neural network (one hidden layer). They do not elaborate in the paper why they specifically chose a shallow neural network rather than other algorithms.	We used two different ML algorithms, RF and DNN with a very different topology than Withers et al. (2020). Our decision was made on the analysis of different regression models. For the case of the RF, before selecting this algorithm we also explored a Linear regressor, SVM, and RF. We chose RF since it shows better performance on our data. We found the hyperparameters by using a GridSearch algorithm as is described in Section Methods In the case of DNN, many factors were taken into account in determining the DNN architecture, including the nature of the application. Specifically, it is a regression problem with only 8 inputs and a single output. Tackling it with a "sophisticated" neural network model, such as a Convolutional Neural Network (CNN), a Recurrent Neural Network (RNN) or even a Transformer, was considered unnecessary (as confirmed by experiments with the simple topology adopted in the solution). For similar reasons, a significant increase in the number of parameters to be learned was considered unnecessary. In the end, the experience gained from working on problems of different types helped us to choose a Multilayer Perceptron (MLP) to tackle this regression problem. From there, as described in the article, we proceeded to optimise the hyperparameters via a search. We experimented with different data normalisation, number of layers, number of units per hidden layer, activation functions in hidden layers, learning rate schedulers, optimisation algorithms, etc.

2. **The simulations of CyberShake are based on the finite fault model. The effect of finite fault is large in the simulation result of ground motion. However, the authors used the hypocentral location and the hypocentral distance as explanatory variables of the predictor. That is, the authors adopted the point-source assumption ignoring the finite-fault effect on ground motions. The**

authors should clarify this point in their paper and discuss the impact of using the point source assumption on the predictor.

It is important to clarify the comment of the reviewer, as the methodology does not ignore the finite-fault effect. We are indeed using the hypocenter of the synthetic events as one of the inputs for the models generated by the MLESmap methodology. However, this does not imply that we are not considering the fact that CyberShake uses a finite fault model in the ground motion simulations: each finite fault model has a hypocenter in its specification. A central goal of our work is the rapid assessment of hazard given a specific earthquake in the study region. We chose hypocentral location and magnitude for that reason, these values can be quickly inferred from early data records. Of course our choice is not optimal if we focus only on accuracy. Having more specific parameters (ruptured fault, CMT, centroid location) should help in getting more accurate inferences. However we are willing to sacrifice accuracy for applicability, in our case, and we show that the patterns/relationships between the hypocentral location, the finite fault in the simulations, and the ground motions are learned implicitly by the models. We added an explanation in the MLESmap: Methodology section to make this clearer in the text.

3. **The author used RMSE to evaluate the predictive performance of each predictor. However, RMSE does not provide any information on prediction bias (underestimation or overestimation). By looking at the residual between observations and predictions (e.g., log O/P), the prediction bias of each model can be evaluated. This should be addressed by the authors.**

We include the metric suggested by the reviewer in Figure R1 here in the review, and also in the supplementary material of the paper (along with a relevant comment in the Results section in the main text). We also provide Figure R2 only here in the review for completeness.

According to a geometric mean ratio of Aida number K (*Aida, K. Reliability of a tsunami source model derived from fault parameters J. Phys. Earth, 26 (1978), pp. 57-73*) between the observed and simulated values

$$log\ K\ =\ 1/N \sum_{i=1}^{N} log(O_i / S_i)$$

Underestimation and overestimation correspond to K>1 and K<1, respectively (*Mulia, I.E., Ueda, N., Miyoshi, T. et al. Machine learning-based tsunami inundation prediction derived from offshore observations. Nat Commun 13, 5489 (2022). https://doi.org/10.1038/s41467-022-33253-5*).

In Fig. R1 we plot ***K*** with respect to magnitude for the synthetic validation. We also indicate the mean ***K*** value in the legend. No particular bias can be claimed for RF and DNN results here, whereas ASK-14 seems to tend to underpredict values for larger events (>6.5), and overpredict for smaller (<6.5)

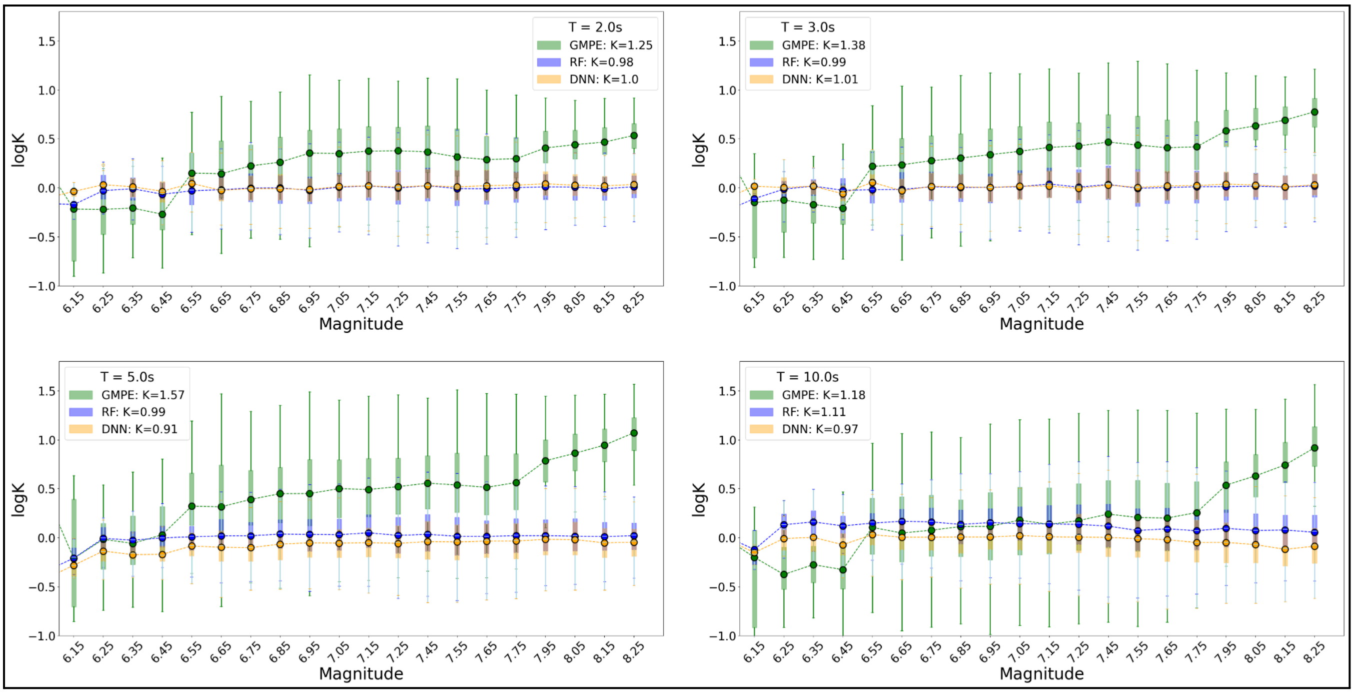


Fig. R1 related to the **log(*K*)** values for the validations (synthetic) dataset

Also we include this metric to analyze the values obtained from the five Real Earthquakes that are in the inside region (see Fig. R2). We cannot claim that RF or DNN models either consistently over- or underestimate the observations, as some cases result in K>1 and others in K<1. A similar conclusion can be drawn for ASK-14 predictions.

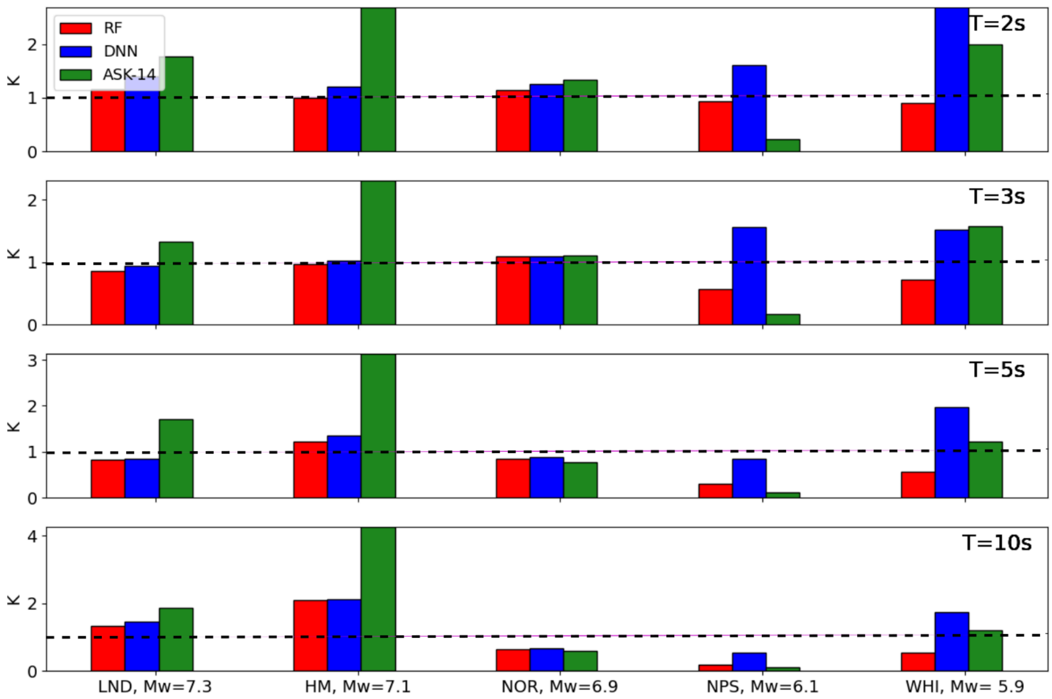
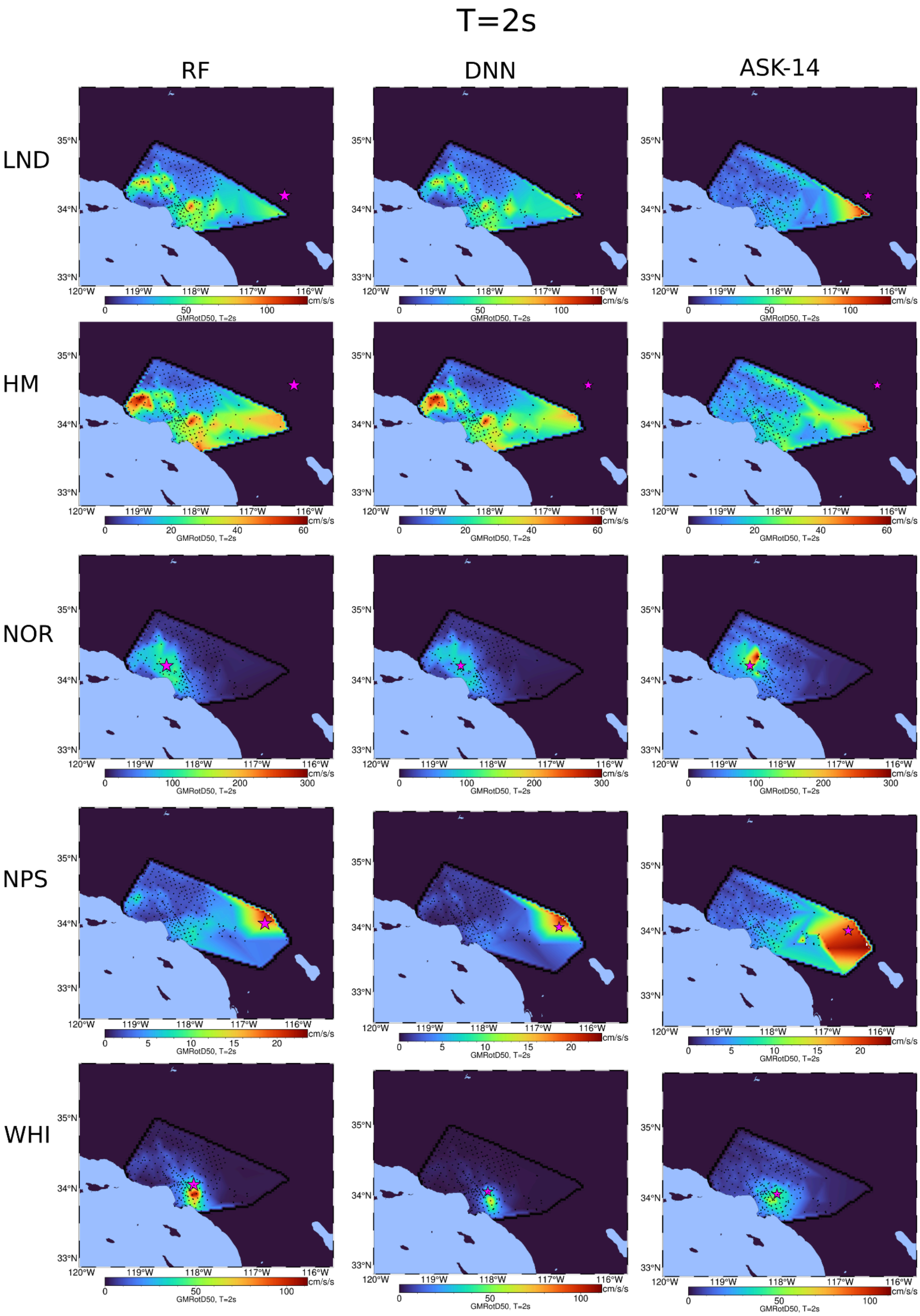


Fig. R2. *K*- metrics computed for the observation values on the stations in the defined 'INSIDE' region.

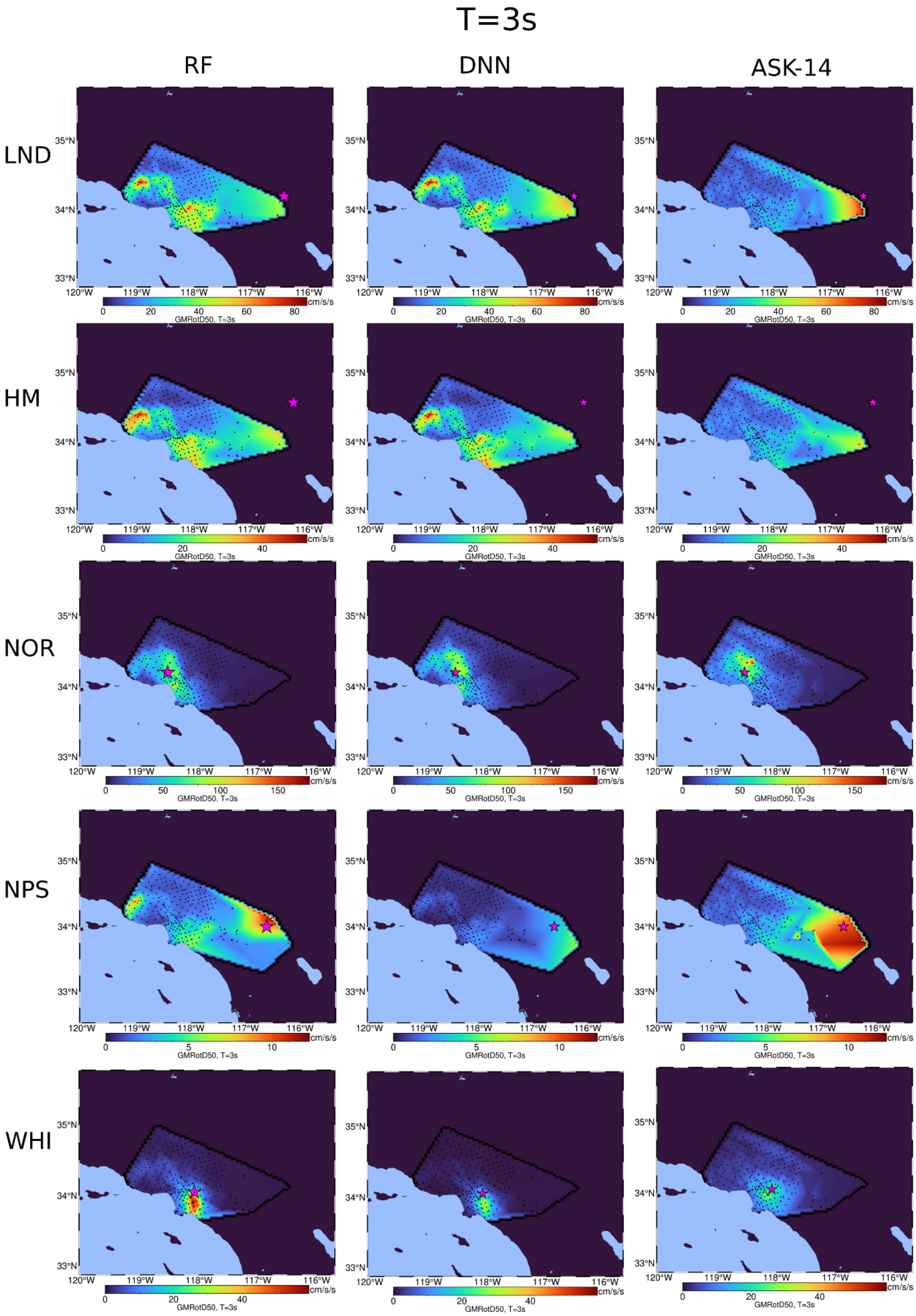
4. **The authors demonstrated the prediction performance of their prediction in real records of five earthquakes. Although RMSE for each earthquake is shown**

in Fig. 5, there is no figure of map view figure like Fig. 3. The authors should add such figure.

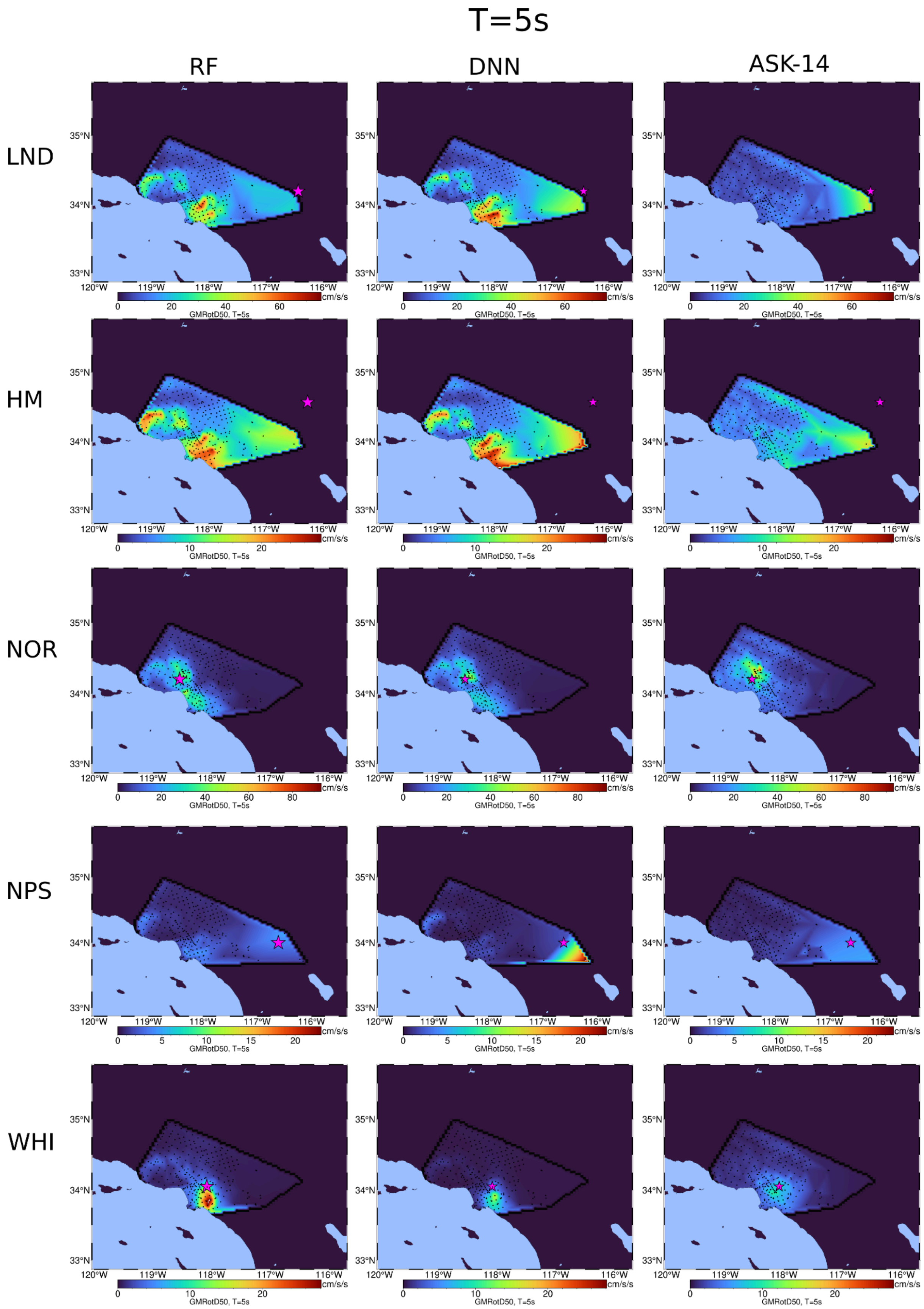
To plot a similar figure to Figure 3 for the real earthquakes, for the ASK-14 we have used the Thompson model (Thompson, E. M., Wald, D. J., & Worden, C. B. (2014). A VS30 map for California with geologic and topographic constraints. *Bulletin of the Seismological Society of America*, *104*(5), 2313-2321) for the Vs30 values and the finite surfaces for those earthquakes. The ASK-14 values were computed using the OpenSHA software. We include the T=2s figure in the manuscript, along with some additional comment to it in the text, and the other periods in the supplementary information.



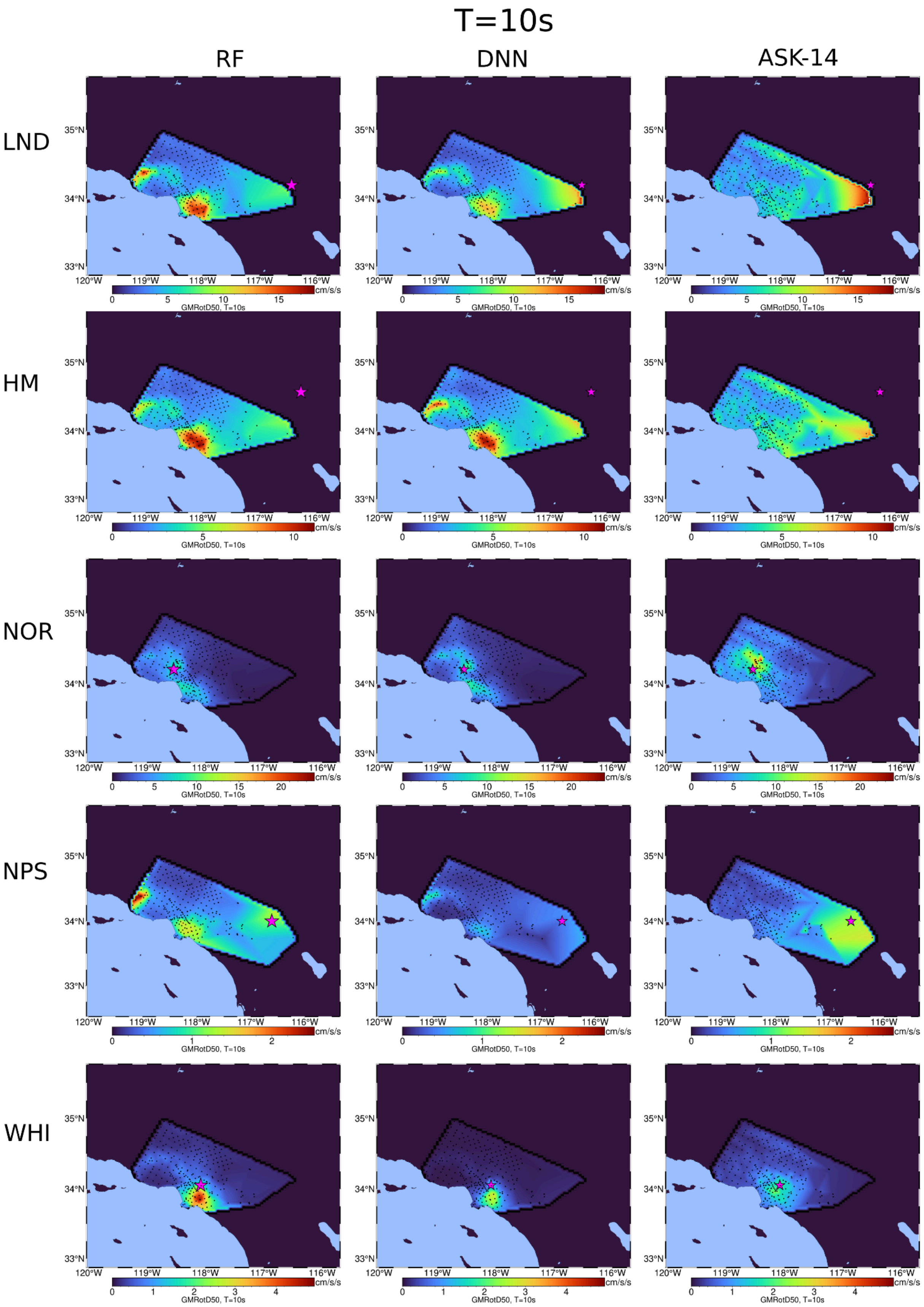
T=2s
RF
DNN
ASK-14
LND
HM
NOR
NPS
WHI
35°N
34°N
33°N
120°W
119°W
118°W
117°W
116°W
cm/s/s
GMRotD50, T=2s



T=3s
RF
DNN
ASK-14
LND
HM
NOR
NPS
WHI
35°N
34°N
33°N
120°W
119°W
118°W
117°W
116°W
cm/s/s
GMRotD50, T=3s



T=5s
RF
DNN
ASK-14
LND
HM
NOR
NPS
WHI
35°N
34°N
33°N
120°W
119°W
118°W
117°W
116°W
cm/s/s
GMRotD50, T=5s

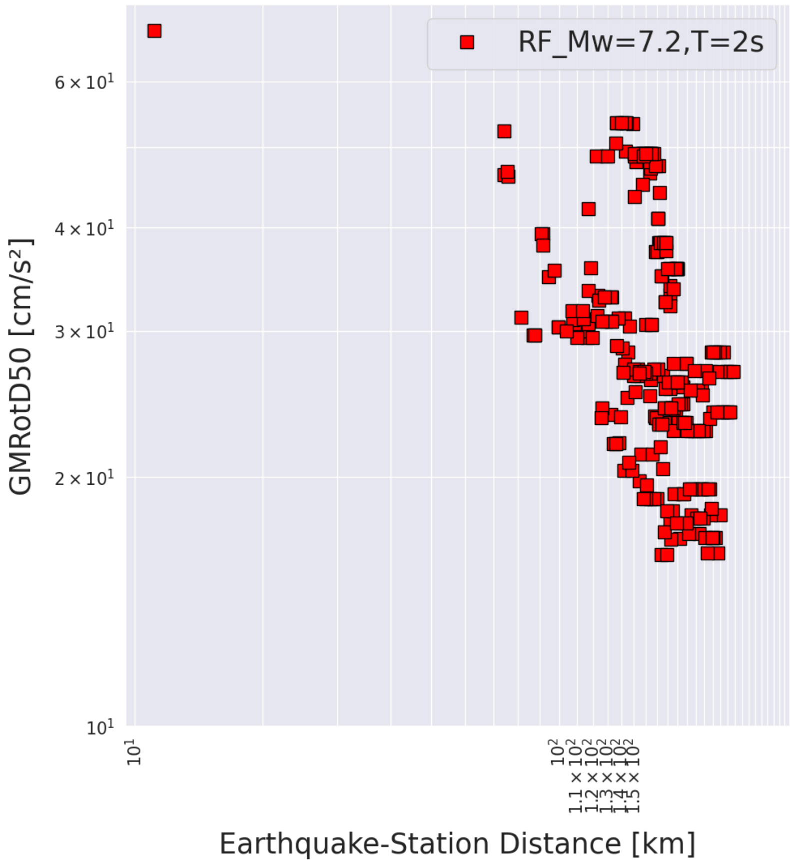

T=10s
RF
DNN
ASK-14
LND
HM
NOR
NPS
WHI
35°N
34°N
33°N
120°W
119°W
118°W
117°W
116°W
cm/s/s
GMRotD50, T=10s

5. **The prediction of MLESmap has a similar feature to conventional regression model or not? The distance attenuation should be checked.**

The distance attenuation is something that we inherit from CyberShake and the simulation results that our models are trained on. In the most general terms the attenuation relationship holds, but the point of physics-based simulations is to include effects that are omitted by conventional regression models, such as site effects, rupture directivity, etc. For example, citing the CyberShake authors:

“At low ground motion levels, there is little difference between the CyberShake and attenuation relationship curves. At higher ground motion (lower probability) levels, the curves are similar for some sites (downtown LA, I-5/SR-14 interchange) but different for others (Whittier Narrows, Santa Ana). We suggest this is due to rupture directivity and basin effects leading to increased ground motion in the CyberShake simulations. “ (*Gupta, N. & Callaghan, Scott & Graves, R. & Mehta, G. & Zhao, Li & Deelman, Ewa & Jordan, Thomas & Kesselman, C. & Okaya, D. & Cui, Yanghao & Field, E. & Gupta, Vasu. (2006). Calculating the Probability of Strong Ground Motions Using 3D Seismic Waveform Modeling - SCEC CyberShake. AGU Fall Meeting Abstracts.*)

This can be seen on the map plots for the real earthquakes in the answer above: the spatial distribution of the ground shaking indicates that the distance attenuation holds, except that we observe more localised effects. For completeness of the review we include a plot for LND below, but we do not include it in the manuscript:



6. **Although the authors mentioned the explanatory variables in “MLESmap: Deep Neural Network topology,” there is no mention of explanatory variables in “MLESmap: RF regression algorithm.”**

The input characteristics (also referred to as explanatory variables by the reviewer) for both algorithms are the same and are mentioned both in the MLESmap training and predictive capacity and MLESmap: methodology sections: magnitude, hypocentral location (latitude, longitude, and depth), site location (latitude and longitude) and the Euclidean distance and azimuth that define the spatial relationship between the hypocentre and the site. We have indeed mentioned this again under the DNN topology section in the methods and not in the RF section. We added this list of input characteristics in the RF section too to make it clear that these are the same for both.

7. **What is the activation function?**

On an artificial neural network the activation function is a mathematical expression used to determine the output of each neuron (Dubey, S. R., Singh, S. K., & Chaudhuri, B. B. (2022). Activation functions in deep learning: A comprehensive survey and benchmark. *Neurocomputing*.). It can be understood as a function that helps the network segregate information into “useful” and “not useful”, making it suppress information that is not relevant. The function decides whether a given neuron should be activated or not, that is, whether the given neuron’s input is important or not in the prediction process. As indicated in the MLESmap: Deep Neural Network topology section, in our DNN the Softplus was adopted as the activation function for the hidden layers and sigmoid for the output layer since the intensity values were previously normalized between 0 and 1.

8. **Please clarify the input to ASK-14.**

The inputs needed to compute the ASK-14 for each site/event are as follows:

<u>Rupture Parameters:</u>

* Magnitude
* Dip
* Down-Dip Width
* Rake
* Rupture Top Depth

<u>Propagation Effect Parameters:</u>

* Distance JB (horizontal distance from site to surface projection of rupture)
* Distance Rup (3D distance from site to rupture surface)
* Distance X (distance to rupture trace, extended infinitely in either directly. Sign indicates if the event is on the hanging or footwall: >=0 if site is on the hanging wall, < 0 if on the footwall. 0 is fine for strike-slip faults)

<u>Site Parameters:</u>

* Vs30
* Depth to Vs=1.0 km/sec

For this paper, the ASK-14 values for the CSS-15.4 events and the five real earthquakes were computed using the OpenSHA software.

We believe it is important to know how MLESmap-generated ML models fare against ASK-14 and validate the methodology against an existing solution. However, we believe that it is not relevant for our audience to know exactly how ASK-14 operates, so we choose not to include such detail in the paper. The manuscript includes the relevant reference: Abrahamson, N. A., Silva, W. J., & Kamai, R. (2014). Summary of the ASK14 ground motion relation for active crustal regions. Earthquake Spectra, 30(3), 1025-1055. We also included the following additional information: “The ASK-14 values were computed using OpenSHA software with the V_{S30} values from the Thompson model”

Reviewer #2

The author presents Machine Learning (ML) strategies trained using one of the largest existing datasets (with more than 100M simulated seismograms) from CyberShake Study 15.4, developed by the Southern California Earthquake Center, covering the Los Angeles basin region. Moreover, the ML strategies were tested with a set of unseen samples from the synthetic database and real historical earthquakes. In my opinion, the idea behind the manuscript is very interesting, considering the growing application of machine learning approaches to support seismic risk mitigation. However, the manuscript does have some flaws.

1. **The main one is the underestimation of the limitations of the proposed approach by the authors, even though they state, "It should be noted that MLESmap models perform better than GMMs as long as the earthquakes are compatible with the CSS-15.4 dataset, that is, the locations, frequencies, and magnitudes interrogated are within the training dataset's bounds." <u>In my opinion, such limitations should be highlighted in both the abstract and the introduction</u>.**

We included an extra comment regarding the limitations of our work both in the abstract and in the Introduction. However, it is important to mention that normally the Machine Learning models such as those used in this paper, applied to any problem, have similar limitations extrapolating values that are not in the training dataset. The mitigation of this issue is simply adding more training data with parameters that cover the entire range of interest.

2. **Additionally, the mentioned section should better emphasize other limitations of the proposed strategy: "The computation of the CSS-15.4 study required a total of 37.6M hours on tier-0 supercomputing facilities.".**

We are aware that our method uses a large amount of computational resources; however, this is the trade-off that must be paid to have more accurate inferences of ground motions. In this case, HPC machines are the infrastructures where these calculations can be carried out, putting them at the service of society. Furthermore, it should be stressed that we used a dataset that was not tailored specifically for our application. In fact, the dataset is suboptimal for statistical inference shown in the paper. If we would build a simulation dataset specifically for the purpose of training MLESmap, a significantly smaller amount of compute time would be sufficient in order to keep a similar prediction accuracy. We therefore choose to not mention specifically the node-hours for the CSS-15.4 database in the abstract and the introduction, as it is not indicative of the required resources for other purpose-specific databases. However, we added in the abstract and in the Introduction that the results are relevant "for a well-tailored synthetic database", making sure that it is clear that it is a key component/limitation. In the Outlooks for sytnethtic training databases of ground motions we also stress that database generation is "a process strongly constrained by computing power".

3. **Hence, concerning the "trained area," ML outperforms the GMM thanks to the CSS-15.4 effort. In other words, the main potential of ML, to move towards a more accurate rapid response solution by combining the accuracy of physics-based simulations with the fast estimations given by GMMs, falters if one wants to apply the proposed ML strategies worldwide. Moreover, the proposed procedure cannot account for site conditions (topography, vertical and lateral heterogeneity) in other areas. To some extent, thanks to the huge amount of training data, the site effect can be evaluated, limiting it to the studied area. Hence, why should one adopt this strategy rather than what is provided by other researchers, including key parameters that allow accounting for site conditions? For instance, (Bindi et al., 2014; Kotha et al., 2020; Kubo et al., 2020; Mori et al., 2021; Zhu et al., 2022) provides ML and GMM including such parameters. What is the authors' opinion? In my opinion, the authors should include a discussion in the introduction regarding the points mentioned above.**

We make no claims about the universality of our application. As the reviewer comments, MLESmap models are a region-dependent strategy, which implies that we cannot apply the same models from one region to another. In this paper, we present the first proof of concept of the methodology showing that we can have better performance than ASK-14 when we train ML algorithms with high-quality data from physics-based simulations. This implies that if we want to move towards other areas of interest we need to generate a set of data and repeat the process (training and validation). In fact, in areas with low to moderate seismic activity, simulation-based models may overcome several limitations of data-based models, which is an interesting outlook, in our opinion. To address this, we now make it more clear in the text that it is regional:

- In the Abstract we explicitly state that models are regional and that our results are in Southern California.
- In the Introduction we also state that we generate regional ML-based models, and that we present an application of MLESmap specifically in Southern California.

We are currently working in the Southern Iceland region, trying to replicate our success with California in a region with a different tectonic setting and geological characteristics, with encouraging results (*M. Monterrubio-Velasco, R. Blanco, D. Modesto, S. Callaghan, and J. de la Puente (2023). Machine Learning based Estimator for Ground Shaking maps. AGU Fall Meeting Abstracts.*). We discuss the challenges of generating a new database in the Discussion section Outlooks for synthetic training databases of ground motions.

Our goal for future versions of MLESmap is also to explore transfer learning (TL), which is a machine learning (ML) technique in which knowledge learned from one task is reused to improve performance on a related task (Torrey, L., & Shavlik, J. (2010). Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques* (pp. 242-264). IGI global.). We added a sentence in the Discussion in the Implications for ML in rapid ground motion IM estimates: “Another way to address the scope limitations would be to explore transfer learning, a technique where learning from one task can be reused to improve the performance of a related task (with tasks here understood as region-specific predictions).”

With respect to the proposed papers, Kubo et al (2020) was already in the Introduction of the manuscript. None of them use physics-based simulations. All of them are based on ground motion records in different areas, and we now cite them all in the Introduction of our paper in this context. In our methodology the ML models learn the effects of the physical phenomena of wave propagation. When the Earth model is well-calibrated the site-effects are considered, so the ML models can show those effects on the predictions. We are also including finite-fault effects into our predictions, because the dataset includes ruptures from all known main faults in the area.

Moreover, all but Kotha et al. (2020) use many complex predictor variables, but in this case uncertainty or unavailability of predictor values can become an issue at time of application. We use much simpler predictor variables that are quickly available at time of application with very good results. We require only the hypocentral location, the magnitude, and the site location. At time of application there is no need to access regional databases related to elevation, V_{S30}, surface roughness, and the like.

Below we summarise the algorithms and predictor variables in the proposed papers.

- Kotha et al. (2020) use robust mixed effects regression algorithm rather than the typical least-squares regression and the following predictor variables: event magnitude and distance metric. They propose a new regionally-adaptable GMM where the influence of outliers is down-weighted. This is a more traditional approach to GMMs than what we propose.
- Kubo et al (2020) use extremely randomized trees (ERT, a tree-based ensemble ML method derived from RF) augmented by conventional GMPEs and the following predictor variables: epicentral distance, moment magnitude, event depth, top depth to the layer whose S-wave velocity is 1400m/s at the site, and V_{S30}. The hybrid approach reduces the underestimation of strong motions.
- Mori et al. (2022) use a Gaussian process regression and the following predictor variables: hypocentral depth, moment magnitude, epicentral distance, V_{S30}, and five morphological variables (elevation, as well as its first and second order derivatives in the E-W and N-S slopes). The method improves both the precision and the accuracy in the estimation of the intensity measures with respect to the available near-real-time prediction methods (i.e., ground motion prediction equation and ShakeMaps)
- Zhu et al (2022) use Random Forest (RF) and the following predictor variables: surface roughness, peak frequency, V_{S30}, and depth $Z_{2.5}$. Their results suggest that ML can be a useful tool in site effect characterisation, especially in the context of model transferability across regions.

4. **Furthermore, a consideration about the selected periods for the prediction should be provided. In fact, predicting ground shaking for periods ranging from 0.1 to 2 s is crucial for seismic risk mitigation strategies, especially concerning ordinary structures and infrastructures.**

We are aware that periods ranging between 0.1 and 2 s are crucial for the seismic risk mitigation strategy. However, in this study we use long periods for the development of the MLESmap models because we are restricted by the periods available in the CS-15.4

database. What is important to highlight is that the algorithms used and the input characteristics selected are efficient in finding the target intensity values for a well-tailored synthetic database. We believe the MLESmap methodology will work for shorter periods. We will begin to analyze it through the CyberShake 2022 study that covers periods ranging from 0.1 to 10 s. The work we are publishing here is a proof of concept that is going to be further explored and developed, after our encouraging findings. We discuss this in the Outlooks for synthetic training databases of ground motions and we now added a special mention of the more relevant periods.

5. **Some statements should be rephrased. For instance, "We propose a novel Machine Learning (ML) methodology that combines the best of both approaches, capturing physical information from simulations (directivity, topography, site effects) with times-to-solution analogous to empirical GMMs, helping to produce the next generation of shaking maps." In fact, directivity, topography, and site effects can only be accurately captured for the area of interest, thanks to the vast amount of available data. The ML strategy does not require input data such as slope, mean shear wave velocity in the upper 30 m, or key parameters for describing potentially buried morphology. Therefore, predicting directivity, topography, and site effects in other areas is expected to clearly fail. In this line, a brief discussion about site effects should be included (Assimaki et al., 2005; Gazetas, 1982; Moczo et al., 2018).**

Our strategy is based on high-quality physics-based simulations. Those simulations consider directivity, topography, and site effects. As we mention in the paper, the site effects outside of the training domain fail (as shown in Figure 5 in Supplementary). MLESmap models can't extrapolate the inferences for areas outside of the training domain due to the heterogeneity in the site effects, not to mention extrapolating to other areas based on the same training dataset. As the reviewer mentions, directivity, topography, and site effects can only be accurate for the area of interest which we used for training. The point that we would like to make with this work is that, provided a good synthetic database, it is possible to use strategies such as MLESmap. We also tried to elaborate in our reply to the previous question above about further work towards training the method for other areas of interest, which is relevant to this.

In the text we are explicit about this and we make no claims about the generality of the application. The application works for well-curated synthetic databases, which we highlight multiple times in the manuscript, e.g.:
"Our methodology, **applied to a high-quality simulation dataset**, can predict RotD50 for real earthquakes more accurately than ASK-14, in a similar time, and using only primary information available shortly after an event."
"Other applications could use databases tailored specifically for the purposes of MLESmap, **allowing for applications in areas other than Southern California**. Populating ad-hoc databases for use in MLESmap, however, poses **problems in finding the optimal set of input parameters**."

We also adjusted the sentence mentioned by the reviewer in the Introduction which now reads "**well-curated** simulations" to highlight the need for a high quality database for this to work. We included the following sentence in the Introduction to be more specific about the results being conditional on the database and also applicable to the selected region: “Given the well-tailored synthetic database, the algorithms used and the input characteristics selected prove efficient in estimating RotD50 values in the region.” We also adjusted the Abstract so that it reflects better that the quoted results refer to models trained in Southern California, and added the region of study in the title of the paper.

6. **Furthermore, in my opinion, the title should be modified. The current title leads the reader to think of a combined approach, while, if I am correct, the results from ML and CSS-15.4 are not merged to provide a single Map. For instance, "Ground Motion Shaking Predictions Based on Machine Learning trained with data from Physics-based Simulations”.**

We thank the reviewer for this suggestion, as indeed it reflects the content better. We modified the title accordingly.

7. **Finally, some valuable references should be added.**

We added the following papers:

1. Graves, R. et al. Cybershake: A physics-based seismic hazard model for southern california. Pure Appl. Geophys. 168, 367–381, DOI: https://doi.org/10.1007/s00024-010-0161-6SCECContribution1354 (2010).
2. Li, Y. E., O’Malley, D., Beroza, G., Curtis, A. & Johnson, P. Machine learning developments and applications in solid-earth geosciences: Fad or future? J. Geophys. Res. Solid Earth 128, e2022JB026310, DOI: https://doi.org/10.1029/2022JB026310 (2023).
3. Mousavi, S. M. & Beroza, G. C. Machine learning in earthquake seismology. Annu. Rev. Earth Planet. Sci. 51, 105–129, DOI: 10.1146/annurev-earth-071822-100323 (2023).
4. Kotha, S. R., Weatherill, G., Bindi, D. & Cotton, F. A regionally-adaptable ground-motion model for shallow crustal earthquakes in europe. Bull. Earthq. Eng. 18, 4091–4125 (2020).
5. Mori, F. et al. Ground motion prediction maps using seismic-microzonation data and machine learning. Nat. Hazards Earth Syst. Sci. 22, 947–966 (2022).
6. Zhu, C. et al. How well can we predict earthquake site response so far? site-specific approaches. Earthq. Spectra 38, 1047–1075 (2022).
7. Florez, M. A. et al. Data-driven synthesis of broadband earthquake ground motions using artificial intelligence. Bull. Seismol. Soc. Am. 112, 1979–1996 (2022).
8. Yongjia Xu, Y. T., Xinzheng Lu & Huang, Y. Real-time seismic damage prediction and comparison of various ground motion intensity measures based on machine learning. J. Earthq. Eng. 26, 4259–4279, DOI: 10.1080/13632469.2020.1826371 (2022)
9. Withers, K. B., Moschetti, M. P. & Thompson, E. M. A machine learning approach to developing ground motion models from simulated ground motions. Geophys. Res. Lett. 47, e2019GL086690, DOI: https://doi.org/10.1029/2019GL086690328 (2020).
10. Petersen, M. D. et al. The 2014 united states national seismic hazard model. Earthq. Spectra 31, S1–S30, DOI:10.1193/120814EQS210M (2015).
11. Boore, D. M., Watson-Lamprey, J. & Abrahamson, N. A. Orientation-independent measures of ground motion. Bull. Seismol. Soc. Am. 96, 1502–1511, DOI: 10.1785/0120050209 (2006).
12. Aida, I. Reliability of a tsunami source model derived from fault parameters. J. Phys. Earth 26, 57–73, DOI: 10.4294/jpe1952.26.57 (1978).

13. Torrey, L. & Shavlik, J. Transfer learning. In Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques, 242–264 (IGI Global, Hershey, 2010).
14. Field, E. H., Jordan, T. H. & Cornell, C. A. Opensha: A developing community-modeling environment for seismic hazard analysis. Seismol. Res. Lett. 74, 406–419 (2003).
15. Thompson, E. M., Wald, D. J. & Worden, C. B. A VS30 Map for California with Geologic and Topographic Constraints. Bull. Seismol. Soc. Am. 104, 2313–2321, DOI: 10.1785/0120130312 (2014).
16. Poggi, V. et al. Rapid Damage Scenario Assessment for Earthquake Emergency Management. Seismol. Res. Lett. 92, 2513–2530, DOI: 10.1785/0220200245 (2021).
17. Rezaeian, S., Stewart, J. P., Luco, N. & Goulet, C. A. Findings from a decade of ground motion simulation validation research and a path forward. Earthq. Spectra 40, 346–378, DOI: 10.1177/87552930231212475 (2024).

Detailed comments are listed below.

- **English. Please check the English used and style. The use of present perfect not appropriate for actions concluded in the past (please, use the past tense when referring to actions concluded in the past).** Corrected.
- **Table 1. “NN” should be “DNN”.** Corrected.
- **SI-Fig. 3 caption. “Mw=7.45” should be “Mw = 7.45”. Leave a blank space before and after =. Please check the whole manuscript.** Corrected.
- **Lines 118-119. Please provide dataset bounds of the mentioned earthquake characteristics.** We refer now to current Figure 1 which describes the dataset.
- **Line 249. “.RMSE ranges” should be “. RMSE ranges”. Leave a blank space after the dot.** Corrected.
- **Tables. Please ensure that unit measure is included in all the tables.** Corrected.
- **R2 and r2. Please check the use or not of the capital letter.** Corrected.

Reviewer #3:

Review of the manuscript titled “Ground Motion Shaking Predictions Based on Machine Learning and Physics-based Simulations” by Monterrubio-Velasco et al.

Ground motion prediction, in conventional practice, relies on statistical evaluation of worldwide recordings from past earthquakes, which is challenged by data scarcity, particularly for large and rare events, in addition to the impossibility of reasoning physical driving mechanisms. Physics-based modeling has been the main complementary approach to overcome the drawbacks of the statistical models, but to focus on specific investigations or reduce computational costs, their use has been mostly limited to few deterministic scenarios. The recent advances in machine learning stands therefore as an opportunity to empower physics-based models for handling computationally challenging tasks such as uncertainty quantification and data mining.

This manuscript presents one of the recent attempts to accelerate physics-based ground motion prediction by developing machine-learning based tools. Specifically, the authors use thousands of simulations of M6+ hypothetical earthquakes for southern California from CyberShake platform to train two different machine-learning algorithms and predict spectral accelerations at four different periods. In addition to testing set from the same database, the authors also verify their methodology on a select of real earthquakes in the region. They report a good performance for both tests and conclude validation of their method.

I think that the study would be a nice contribution to the literature, once certain points are addressed. I am especially confused by the issues below.

A. The study is oriented towards shaking maps. Throughout the manuscript, shaking intensity and ground motion are interchangeably used, despite the two terms stand for different notions (see for example, USGS). I can understand that the utility of the developed tools includes near-real-time shaking map development, but the current version rather reads like a ground motion prediction study. If the main objective is indeed to accelerate shake map development, I would expect to see the application of the algorithms for shake maps and validate against real cases. Otherwise, as a ground motion prediction study, the writing needs reorganization which should clearly state the goal of predicting a specific ground motion metric, in a specific area as a function of magnitude and distance, faster than physics-based models. Such reorganization would also benefit from mentioning the existing machine-learning studies in the field, such as reduced order models, generative models, etc. (e.g., Khosravikia & Clayton, 2021 https://doi.org/10.1016/j.cageo.2021.104700, Mori et al., 2022 https://doi.org/10.5194/nhess-22-947-2022, Xu et al., 2022 https://doi.org/10.1080/13632469.2020.1826371, Florez et al., 2022 https://doi.org/10.1785/0120210264, and Mousavi and Beroza, 2023 https://doi.org/10.1146/annurev-earth-071822-100323), and what makes this study innovative compared to other studies.

We modified the manuscript to be clear that we propose a strategy that allows to generate regional-dependent ML-based GMMs (a suite of such models, for a range of periods and different IMs, and using different algorithms). We then specify that we apply that strategy specifically to generate models for log(RotD50) in Southern California. We develop the models for usage in a rapid manner: so this work regards fast and accurate IM prediction and, as an outcome of it, map generation. We also have more consistent language and refer either to the original definition RotD50 or to intensity measures (IMs) more generally. We hope that this clarifies it.

Below we summarise the proposed papers, highlight the main difference with our work, and indicate where we have included each of them in the manuscript.

- Khosravikia & Clayton (2021) use Artificial Neural Network, Random Forest and Support Vector Machine algorithms and the following predictor variables: event magnitude and hypocentral distance (defined as the site distance from the point of initiation of rupture on the fault plane) and V_{S30}. **Importantly, the methods are trained on recorded data, rather than synthetics. The paper was already included in the submitted manuscript in this context.** They study the advantages and disadvantages of the different techniques, with RF outperforming the others. They also find that, when sufficient data is available, all algorithms tend to provide more accurate estimates compared to the conventional linear regression. However, the conventional method is a better tool when limited data is available.
- Mori et al. (2022) use a Gaussian process regression and the following predictor variables: hypocentral depth, moment magnitude, epicentral distance, V_{S30}, and five morphological variables (elevation, as well as its first and second order derivatives in the E-W and N-S slopes). **Importantly, the method is trained on recorded data, rather than synthetics.** The method improves both the precision and the accuracy in the estimation of the intensity measures with respect to the available near-real-time prediction methods (i.e., ground motion prediction equation and ShakeMaps). We included the paper in the Introduction.
- Xu et al. (2022) do not use ML to predict intensity measures. They use three ML techniques (support vector machine, logistic regression, and the decision tree) to predict damage states *from intensity measures.* **In this case, the IMs are the input variables (48 IMs are selected), so the aim is very different.** We included the paper in the Introduction.
- Florez et al., (2022) use generative adversarial networks to synthesize earthquake acceleration time histories. **They therefore go well beyond intensity measures, as in our work, and generate realistic three-component accelerograms** that capture many relevant statistical features of acceleration spectra and waveform envelopes. They use magnitude, distance, and V_{S30} as predictor variables. We included the paper in the Introduction.
- Mousavi and Beroza (2023) provide a comprehensive overview of ML applications in earthquake seismology, discuss progress and challenges, and offer suggestions for future work**. They include a general discussion on the characterisation of ground motion and developing ML-based GMMs.** We included the paper in the Introduction.

B- A detailed discussion on the physical aspects of the method and findings is missing. For example, the motivation often underlines that large events are more destructive, and the end product of the study would serve to post-disaster efforts. Yet,

the most destructive part of an earthquake, or what governs peak acceleration, engineering design, and intensity distribution, is generally high-frequency ground motion (above ~1 Hz). The predictions here, on the other hand, are limited to 0.1-0.5 Hz, which is quite a low and limited frequency band, and the numerical resolution of the synthetic database is 1 Hz.

We are aware that periods ranging between 0.1 and 2 s are crucial for the seismic risk mitigation strategy. However, in this study we use long periods for the development of the MLESmap models because we are restricted by the periods available in the CS-15.4 database. What is important to highlight is that the algorithms used and the input characteristics selected are efficient in finding the target intensity values for a well-tailored synthetic database. We believe the MLESmap methodology will work for shorter periods. We will begin to analyze it through the CyberShake 2022 study that covers periods ranging from 0.1 to 10 s. The work we are publishing here is a proof of concept that is going to be further explored and developed, after our encouraging findings. We discuss this in Outlooks for synthetic training databases of ground motions and we have now added a special mention of the more relevant periods: “A direct follow-up of the present study is the application of MLESmap to such databases and the evaluation of their impact both on the training and on the predictive capabilities of the models, **in particular for shorter periods relevant to seismic risk mitigation strategies.**”

- Another issue is that the database simulations are likely to include event-specific source effects and site effects, since the study area is highly complex in terms of fault networks and crustal geological heterogeneity. The focus on very large events of M 7+, particularly, would contribute to event-specific spatial variability of ground motion. When looking at the training station grid (Fig. 6a) with respect to fault lines, stations are all in the near field, i.e., near fault. The choice of near field is unclear to me. In other words, wouldn’t it be more straight forward to start with distant stations in the same area? The current version is not informative about whether the predictive power manifests repeated source or site effects.

As we explain in the comment above, the database was not built ad-hoc for our purposes and therefore the training station grid is not something that we have explicitly chosen. It is the grid proposed by the CyberShake studies of the region. The further the stations are from the earthquake, the larger the computational grid needs to be, and the more expensive the calculations, so the area is chosen to manage the trade off between the computational resources and the areas where high intensity shaking is expected. CyberShake chooses this near-field area for its PSHA analyses to complement data scarcity, and we think of MLESmap models as the near-field complement of data-based approaches which perform better at larger distances, but tend not to perform well in the near field. We hope to close this gap with MLESmap-generated ML models.

In our opinion, studying the far field with MLESmap wouldn’t be too relevant because, in the far-field, similar-magnitude events become more alike and thus separating source vs site effects becomes feasible. This is the underlying principle of many empirical GMMs. The challenge is predicting the near-field, where source and site effects cannot be easily decoupled. In fact, the concept of site effect is a simplification of complex wave phenomena that often fails in the near-field, and thus when using underlying physical models needs careful reassessment. We remark that in MLESmap all sites are trained independently, hence capturing effects related to local rock properties can be understood as natural in this methodology.

- Lastly, I had difficulty to understand the reason of making comparisons to GMMs — empirical models. In general, such comparisons are made to verify far-field predictions of the physics-based models; for near-field, GMMs are less credible given their very large variabilities, and this is where physics-based models complement GMMs. Besides, such comparison is made for a synthetic event from the CyberShake database. It reads like authors verify the database, which is out of the objective of the study. Regarding real case applications, I again don't see the reason of comparisons with GMMs. The authors may be showing it to measure relative performance of different approaches, but it is unclear.

As the reviewer comments, the purpose of comparing the empirical GMMs with the MLESmap inferences is simply to measure relative performance of those approaches in the near-field. We want to show that where ASK-14 is less credible, we can try to complement the results with MLESmap, as we perform better in those areas. We do the comparisons both on the synthetic scenarios (with the map in Figure 3), and on real earthquakes for the few stations we have available (Figure 5). We do the map in Figure 3 for the synthetics not to validate the database, but to validate the predictions. For the real earthquakes we do not have a reference target map. The result on synthetics, in isolation, might not be so interesting. However, combined with the presented result on real events, it is a good indication that synthetics do a good job at predicting earthquake motion.

The main message is that 1) we can use as little information as empirical GMMs need to 2) produce predictions equally quickly but 3) obtain a better prediction. We consider this a novel and relevant contribution to the field of seismic hazard.

We added an explanation that: "GMMs, obtained by statistical regression from empirical observations in a given region, are the foundation of seismic hazard analysis, **so measuring the relative performance of the two approaches is important to showcase the operational relevance of MLESmap**." We also added: "It should be noted that the results on real earthquakes, combined with the results on the synthetic events, are a good indication that synthetics in the database can accurately predict earthquake motion."

We also added Figures in the main text and in the supplementary with maps for the real events, similar to Figure 3 for synthetics.

To clarify these issues, I provided my suggestions, together with minor comments, in the following.
Best regards,
Elif ORAL

Major comments

1.- Training dataset mostly includes very large and rare events (for the region), whereas ground motion station grid is quite small compared to fault sizes (red lines vs magenta points in Fig. 6a). Latest GMMs and physics-based models define the distance by shortest distance to fault, instead of hypocentral distance, to minimize directivity-related variabilities. Please justify your choice of using hypocentral distance for such large events and near field groundmotion.

Our objective is to generate information rapidly post-earthquake, therefore for our strategy we need to use information that is readily provided by the agencies immediately after the event, i.e., magnitude and hypocentral location. These are our key inputs. After our analysis, we suggest that this scarce information is enough for the ML algorithms to learn complex patterns from a high-quality dataset of physics-based simulations that already account for the mentioned directivity effects. Hence, in our case, the shortest distance to fault from the hypocenter is not needed, since those patterns are learned implicitly by the models and do not need to be accounted for explicitly. However, including said distances as predictors could be an interesting follow-up study. We elaborated in the MLESmap: methodology section on our choice of parameters to make it more clear.

2. Please also provide the range of hypocentral distances in the database.

We assume that the reviewer is asking for the distance between the hypocenter and the sites, which is one of the parameters of our method. We include the statistic for the requested parameter below:

Maximum: 637.6 km
Minimum: 0.20 km
Average: 135.4 km
Median: 129.4 km
Std: 68.6 km
Percentile 25% 80.2 km
Percentile 75% 184.7 km

- I would expect the first figure to inform about the big picture. Something like an assembly of Fig. S1 and Fig S6. Showing earlier the area that you train and test can make it easier to follow the text.

Indeed, Figure 6 (summarising the CyberShake dataset) sets the stage and we now moved that figure as Figure 1, and we also referenced it at the beginning of the first subsection in Results, namely in MLESmap training and predictive capacity, to show the area that we test and train. We keep SI-Fig 1 in the SI, as we do not feel like it is needed in the main text. (We believe that what the reviewer is referring to as Fig. S1 is the SI-Fig 1, while Fig S6 is Fig 6 from the main text).

- Figures 1 and 2 show very similar patterns for different periods. You may consider only showing the representative cases and provide the remaining in Supplementary Material.

We are grateful for the comment and we understand the rationale behind keeping a single frequency, but we find it valuable to show them. The main outcome is that multiple frequencies have been modelled by MLESmap. We rather keep the figures as they are. One of the challenges is that often modelling tends to work very well at long periods (e.g. 10s), and is much less accurate at shorter periods (e.g. 2s), and we show that here we remain accurate for all periods considered.

- Shaking intensity does not mean ground motion. I suggest using “ground motion metric,” or “spectral acceleration” or the original definition “GMRotD50” throughout the manuscript.

We thank the reviewer for pointing out this inconsistency. We modified the manuscript to have more consistent language and refer either to the original definition RotD50 or to intensity measures (IMs) more generally.

- Please provide reference for GmRotd50 (Boore et al., 2006 https://doi.org/10.1785/0120050209)

We added the reference.

- Since this study focuses on a single metric, please justify your choice of GmRotD50. It is orientation independent, which is obtained after a nonlinear change to cceleration data, i.e., you lose component-related information — potentially inherited data. You could have used for example peak ground motion. In addition, considering the low frequency resolution, peak ground displacement or velocity would be more informative, instead of acceleration or spectral acceleration.

We agree with the reviewer that other metrics, such as peak ground velocity or displacement, would be of interesting. However, for the present study we were limited by the dataset that we have obtained from SCEC where GmRotD50 is the only parameter made available to us. Future versions of our methodology may also include other metrics: as mentioned in the Discussion, we are planning to build further ad-hoc databases explicitly for MLESmap applications in other regions.

- In general, please reorganize the writing to avoid confusion about intensity, ground motion, GMM use. For example, it should be clear to the reader why you work on this topic (shake map, or ground motion prediction?), why you choose such methodology, and why you focus on southern California (more data?).

This work regards fast and accurate IM prediction and, as an outcome of it, map generation. We use an ML-based methodology and physics-based seismic simulations to obtain models that provide a quick and accurate response after a large-magnitude earthquake. We introduced numerous changes to the Introduction and to the Methods section to make this more clear.

We used the Californian dataset because, as mentioned in the Introduction, SCEC has one of the largest datasets of synthetics worldwide. Testing the methodology on this dataset helped us establish that ML models do learn a relationship between the selected rapidly available source characteristics and the chosen metric (RotD50). Testing our methodology with one of the largest and best-calibrated datasets of synthetics worldwide allowed us to validate the underlying idea and compute the accuracy of MLESmap. We also included an additional explanation of this in the Methods section: “The CSS-15.4 database is one of the largest and best-calibrated datasets of synthetics worldwide, thus serving as an ideal first implementation of MLESmap.”

We also made numerous changes to the text to avoid the intensity, ground motion, shake map and GMM confusion. We unified the nomenclature and use the IM and RotD50 more consistently. We also now make it clear that what we propose is a globally-applicable methodology which can be used to train *regional* ML-based GMMs that can be queried for an IM (a selected predictor, in the LA basin application it is RotD50), e.g.
Introduction: “The ML-based EStimator for ground-shaking maps (MLESmap) exploits supervised ML algorithms trained with data from well-curated physics-based simulations to generate regional ML-based models that accurately estimate an IM within a few seconds of an earthquake occurrence.”
Introduction: “We present an MLESmap application in Southern California, proposing regional ML-based GMMs trained on (...)”
Methods: “In particular, the MLESmap methodology provides a framework for training region-specific ML models of ground motions on databases of synthetic IMs with two algorithms: Random Forest (RF) and Deep Neural Networks (DNN). The resulting ML models, the RF-GMM and DNN-GMM, predict the distribution of a selected IM at a given period for given source characteristics of a new event in the region.”

- In general, current version highlights several times predictive power/good performance given small errors. Please also discuss why you get good/bad predictions with respect to physics-based knowledge.

Our ML algorithms learn from a physics-based data set. When we have well-calibrated data, such as in California from regional velocity models and the active fault system, this contributes to our good performance in ML predictions. However, it is important to note that uncertainty or poor calibration in the input parameters of the simulations will propagate errors to the ML predictions. This was mentioned in the Discussion with respect to generating ad-hoc databases for MLESmap: “Other applications could use databases tailored specifically for MLESmap, allowing for applications in areas other than Southern California. Populating ad-hoc databases for use in MLESmap, however, poses problems in finding the optimal set of input parameters. Accounting for large and rare earthquakes is particularly important, as they have the highest damage potential and are poorly represented in empirical GMMs. In a process strongly constrained by computing power, designing a new database requires a detailed analysis prior to implementation.”

Furthermore, as shown in SI-Fig. 5, spatial extrapolation of ML inferences for regions far from the training area is not accurate because these ML algorithms can't learn patterns where they have no examples to do so.

Regarding physics-based knowledge (e.g. amplitudes decrease with distance, basin amplification, directivity…) we do not perform any specific analysis for two main reasons: first, we trust in the work made by SCEC in calibrating and validating their synthetic dataset. Being a physics-based model, it is highly unlikely that such features would not be correctly modeled. And secondly, the fit of MLESmap vs Synthetic (Cybershake) data is very good. Hence, it is a reasonable hypothesis to assume that we inherit the physics-based knowledge embedded in the Cybershake runs.

- Fig 3: It is hard to understand why you evaluate the performance of an empirical method on a synthetic event. It is expected to be the other way around.

What we want to show in this figure is the predictive accuracy between the two strategies, MLESmap and one GMPE used in the region, mimicking a situation where we have a new earthquake and we want to rapidly estimate the RotD50 values. In this case, we can compare the inferences of the MLESmap and GMPE against the values coming from the simulations. We want to measure the relative performance of those different approaches in the near-field and show that where GMMs are less credible, we can try to complement the results with MLESmap, as we perform better in those areas. We do the comparisons both on the synthetic scenarios (with the map in Figure 3 referred to in this comment), and also on real earthquakes for the few stations we have available (Figure 5). In the 2nd case, however, we do not have a target or reference solution. The result on synthetics, in isolation, might not be so interesting. However, combined with the presented result on real events, it is a good indication that synthetics do a good job at predicting earthquake motion.

- Fig 6: In the same subplot (a) or another, you can also provide source area (rupture extents), that is included in synthetic database. This is important because, since you only use low-frequency ground motion metrics, ground motion will be dominated by either by basin resonance around LA (site effects), or since the source is very close, a combination of source and site effects. You can refer to previous studies on LA basin, to leave out or judge the significance of basin effects on the frequency band you focus on (e.g., Wald and Graves, 1998 https://doi.org/10.1785/BSSA0880020337; Kohler et al., 2020 https://doi.org/10.1785/0220200170)

As the reviewer mentions, source and site effects are highly coupled in the near field, which is the case in most parts of our data. We believe that a discrimination of site vs source effects is something that would be interesting, but our manuscript does not present or study the dataset. Hence, it would seem beyond the purpose of our work establishing a detailed analysis of the dataset itself. We are assessing the capability of ML methods to predict earthquake ground motion when fed with said synthetic data. For details on amplification for Cybershake experiments we refer to, e.g.,
https://www.jeff-bayless.com/wp-content/uploads/2021/09/SCEC-CScorr-Report.pdf and
https://www.wcee.nicee.org/wcee/article/16WCEE/WCEE2017-4423.pdf

In particular, we had the following information from the SCEC databases to perform our ML study: Site Name, Site Latitude, Site Longitude, Run ID, Source ID, Rupture ID, Rupture Variation ID, Magnitude, Hypocenter Latitude, Hypocenter Longitude, Hypocenter Depth, Intensity Measure Component, Period, Intensity Value

- Looking at Figs. 4-5, the source is inside the station grid, for the cases (c) and (e). I suspect that this might be the reason you get largest errors for case (c). Are these real events covered by the database (at least by focal mechanism, rupture speed, slip distribution with respect to published studies on the events). Even if it is not, please report.

No, the CyberShake earthquake rupture forecast does not include any real events. CyberShake includes events on the faults where actual events may have happened but with a range of hypocenters, magnitudes, and slips, such that the observed event should fall within the sampled distribution, but not the specific prescribed historic event. As stated in the Methods section, the earthquakes in CyberShake are all hypothetical: “In this work we have leveraged a large dataset generated via physics-based wave propagation simulations to generate ground motions from hundreds of thousands of hypothetical earthquakes to train ML algorithms.”

Regarding the source being inside the station grid in c), we believe that the largest RMSE in this case reflects the fact that it is a large event and RMSE scales with amplitude. As several stations are right above the source in this case, they likely recorded higher amplitudes. In the case of a) and b) the magnitude is higher, but the recording stations are further away from the source (and in particular the stations that we consider `inside`, that is, within our “training range”). For e) the RMSE errors are significantly lower, as the magnitude (and thus amplitudes) are lower, but our ML models do not do better than ASK-14, because the magnitude of that event is outside of the training range. We added a mention of the RMSE scaling in the Figure 5 caption (as well as in the corresponding figure for the outside stations in supplementary): Note that RMSE scales with amplitude, so the RMSE magnitude in each case depends on the event magnitude and on the location of the `inside' stations relative to the event. The relative performance of the ML-based predictions v.s. the empirical GMM are of interest rather than absolute values for each event.

- Extrapolating for smaller and larger magnitude should be different. You see actually a good prediction for Mw 5.9 event, the case (e). This might relate to the fact that for smaller magnitude, rupture and near-field extent is smaller; you may be better capturing distance-dependent attenuation. Please comment.

In Fig. 5 we actually show the *worst* predictions of our models as compared to the GMPE for the M5.9 quake, rather than the best. We also comment in the paper that our predictions are best when the parameters of the event fall within the parameters range of the training database, because otherwise the ML algorithm needs to extrapolate beyond what it was trained to do.

Indeed, the RMSE for Mw 5.9 and Mw 6.1 events is lower than for 7.3, 7.1 and 6.9 events, but we think that elaborating on the explicit reasons is beyond the scope of our paper as we have not studied this in detail. What is important, however, is that for the Mw 6.1 event, which is within our training range (as opposed to the Mw 5.9) and where the RMSE values are also low overall (likely due to the database capturing the physis effects for smaller events better, as the reviewer mentions), we very significantly improve the results with respect to GMM. Another factor is that RMSE scales with amplitude: for similar relative errors and broadly different amplitude ranges, the RMSE values will differ. We added a mention of the RMSE scaling in the section MLESmap v.s. empirical GMM for synthetic earthquakes, as well as in figure captions (see answer above).

- Fig. 5: It seems that overall you better predict the lowest frequency (T=10 s), which is expectedly the case in physics-based models (for example, Fig. 7c for distances > 10 km, in Oral et al., 2022 https://doi.org/10.1785/0120220064, and references therein). Please highlight the consistency.

RMSE indeed is lowest overall for 10s, and overall highest for 2s, for both MLESmap and ASK-14, a fact we did mention explicitly in the MLESmap v.s. empirical GMM for synthetic earthquakes section. We believe that this reflects the fact that RMSE scales with amplitudes, so lower period and lower amplitude results in lower overall RMSE. As mentioned above, we added a mention of the RMSE scaling in the section MLESmap v.s. empirical GMM for synthetic earthquakes.

- Please comment on why Random Forest seems to better perform? Or how do you explain that one method works better in some cases, and not so good in some other cases? Why DNN seems so bad for Fig. S5?

Our objective in this paper is not to show that an ML algorithm is better than the other, but to show that both algorithms - with a different foundational mathematical basis - can predict the target parameter that we are exploring with the predictor variables that we select, given that the respective models have a good representation of the predictor's parameter values (i.e. magnitude, spatial variables) in the synthetic training set.

However, it is correct that in Fig. S5 the DNN shows a very bad prediction for those stations in the 'outside' training region. Again, we do not expect the algorithms to extrapolate well to whatever is outside of the training region, so we expect both of them to perform poorly.

We can conclude from SI-Fig. 5 that - given their different mathematical foundations - the feature importance for each algorithm is different.The figure in question suggests that the DNN topology has a large influence of the spatial predictors, while the RF is less influenced by those parameters. Likely the non-spatial parameters play a larger role in the RF models (such as, for example, the event magnitude) and the station location falling outside of the training range has less of an effect than for DNN. We added this information in the figure caption.

- How did you decide on DNN architecture? By trial and error, or after a reference study?

Many factors were taken into account in determining the DNN architecture, including the nature of the application. Specifically, it is a regression problem with only 8 inputs and a single output. Tackling it with a "sophisticated" neural network model, such as a Convolutional Neural Network (CNN), a Recurrent Neural Network (RNN) or even a Transformer, was considered unnecessary (as confirmed by experiments with the simple topology adopted in the solution).

For similar reasons, a significant increase in the number of parameters to be learned was also considered unnecessary. In the end, the experience gained from working on problems of different types helped us to choose a Multilayer Perceptron (MLP) to tackle this regression problem. From there, as described in the article, we proceeded to optimise the hyperparameters via a search. We experimented with different data normalisation, number of layers, number of units per hidden layer, activation functions in hidden layers, learning rate schedulers, optimisation algorithms, etc.

We elaborated on the above in the Methods of the paper.

- Looking at Fig. 5, GMM and ML predictions look very close, which contradicts with your statement of "ML outperforms GMM." Instead please quantify differences and comment on their significance.

We do not agree with this observation, in Fig. 5 we show the improvement of ML inferences for those observations. For the stations located in the trained region and for events with a magnitude represented in the dataset we perform consistently better than the ASK-14 GMPE. The differences are quantified and indicated as percentage change relative to ASK-14 above the bars. In the figure we show that we do not outperform ASK-14 only for the WHI event, which has the magnitude range outside of the training data.

- Discussion: Please also discuss potential directions that use larger sizes for other metrics and/or algorithms. You may consider citing other studies in the literature at this point.

In the Methodology we mention that we do not expect other IMs to be an issue, we do not consider it the most prominent further direction of study with regards to MLESmap, but rather a relatively trivial one that we choose not to mention in the discussion. The methodology section states: "It should be noted that other IMs could serve targets without significant impact on the performance of the MLESmap models"

In the Discussion we mention different predictors (rather than outputs) of the models, other datasets (in particular generated specifically for MLESmap, and further CyberShake datasets with higher frequencies), integrating MLESmap with empirical GMMs to form an even more robust hybrid product, and explore transfer learning for integrating new regions. We consider those the most important future avenues for our work.

We added the following papers to the manuscript:

1. Graves, R. et al. Cybershake: A physics-based seismic hazard model for southern california. Pure Appl. Geophys. 168, 367–381, DOI: https://doi.org/10.1007/s00024-010-0161-6SCECContribution1354 (2010).
2. Li, Y. E., O'Malley, D., Beroza, G., Curtis, A. & Johnson, P. Machine learning developments and applications in solid-earth geosciences: Fad or future? J. Geophys. Res. Solid Earth 128, e2022JB026310, DOI: https://doi.org/10.1029/2022JB026310 (2023).
3. Mousavi, S. M. & Beroza, G. C. Machine learning in earthquake seismology. Annu. Rev. Earth Planet. Sci. 51, 105–129, DOI: 10.1146/annurev-earth-071822-100323 (2023).

4. Kotha, S. R., Weatherill, G., Bindi, D. & Cotton, F. A regionally-adaptable ground-motion model for shallow crustal earthquakes in europe. Bull. Earthq. Eng. 18, 4091–4125 (2020).
5. Mori, F. et al. Ground motion prediction maps using seismic-microzonation data and machine learning. Nat. Hazards Earth Syst. Sci. 22, 947–966 (2022).
6. Zhu, C. et al. How well can we predict earthquake site response so far? site-specific approaches. Earthq. Spectra 38, 1047–1075 (2022).
7. Florez, M. A. et al. Data-driven synthesis of broadband earthquake ground motions using artificial intelligence. Bull. Seismol. Soc. Am. 112, 1979–1996 (2022).
8. Yongjia Xu, Y. T., Xinzheng Lu & Huang, Y. Real-time seismic damage prediction and comparison of various ground motion intensity measures based on machine learning. J. Earthq. Eng. 26, 4259–4279, DOI: 10.1080/13632469.2020.1826371 (2022)
9. Withers, K. B., Moschetti, M. P. & Thompson, E. M. A machine learning approach to developing ground motion models from simulated ground motions. Geophys. Res. Lett. 47, e2019GL086690, DOI: https://doi.org/10.1029/2019GL086690328 (2020).
10. Petersen, M. D. et al. The 2014 united states national seismic hazard model. Earthq. Spectra 31, S1–S30, DOI:10.1193/120814EQS210M (2015).
11. Boore, D. M., Watson-Lamprey, J. & Abrahamson, N. A. Orientation-independent measures of ground motion. Bull. Seismol. Soc. Am. 96, 1502–1511, DOI: 10.1785/0120050209 (2006).
12. Aida, I. Reliability of a tsunami source model derived from fault parameters. J. Phys. Earth 26, 57–73, DOI: 10.4294/jpe1952.26.57 (1978).
13. Torrey, L. & Shavlik, J. Transfer learning. In Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques, 242–264 (IGI Global, Hershey, 2010).
14. Field, E. H., Jordan, T. H. & Cornell, C. A. Opensha: A developing community-modeling environment for seismic hazard analysis. Seismol. Res. Lett. 74, 406–419 (2003).
15. Thompson, E. M., Wald, D. J. & Worden, C. B. A VS30 Map for California with Geologic and Topographic Constraints. Bull. Seismol. Soc. Am. 104, 2313–2321, DOI: 10.1785/0120130312 (2014).
16. Poggi, V. et al. Rapid Damage Scenario Assessment for Earthquake Emergency Management. Seismol. Res. Lett. 92, 2513–2530, DOI: 10.1785/0220200245 (2021).
17. Rezaeian, S., Stewart, J. P., Luco, N. & Goulet, C. A. Findings from a decade of ground motion simulation validation research and a path forward. Earthq. Spectra 40, 346–378, DOI: 10.1177/87552930231212475 (2024).

- L131: I suggest to briefly mention data quality.

If we understand this brief comment correctly, the reviewer refers to the data quality/scarcity for large and rare events which affects the predictions of empirical GMMs for such events. To address this, the sentence that in the previous version was in L131 now reads “Accounting for large and rare earthquakes is particularly important, as they have the highest damage potential and are poorly represented in empirical GMMs **due to scarcity of high-quality data.**”

- L134: I again suggest to volume down this claim of ML better predicting than GMM, and to highlight the potential improvement of GMM when incorporating physics-based data through ML.

We rephrased the sentence that the reviewer referred to in order to highlight that complementary nature of the two approaches: “Since ML models trained on synthetics can predict the IMs of real earthquakes more accurately than empirical GMMs, and in a similar time, they are bound to become the next-generation tool for post-disaster analysis that **complements existing data-based approaches in guiding the relief efforts**". We also added in the introduction that the GMM limitations require research on "**complementary** strategies" to highlight that we do not talk about a competing strategy and modulate the message. Finally, the discussion already included a suggestion that a merge of the two strategies would likely render the best results.

- L136: In addition to the computational aspect of the problem, highlighting the uncertainty quantification or enlarging datasets from a couple of deterministic cases to explore much larger parameter space deserves being mentioned. You say it only for focal mechanism, but especially for large events, it is beyond that (see kinematic vs dynamic rupture modeling differences, for example).

The reviewer, if we understand the comment correctly, refers to:
L136 in the old version where we state that rapid hardware and software developments allow to generate databases that cover a large parameter space .
And also the beginning of a new paragraph in L138 in the old version: "Since decreasing uncertainty in predicted intensities is of paramount importance, further constraints on the events, such as the focal mechanism, could be included provided that such parameters can be assessed quickly. MLESmap models with that extra information could be developed in the hope of yielding even more accurate inferences."

The two belong to two paragraphs and refer to two different issues, which in the text may not be clear.
In L136 we mean that generating databases (be it with kinematic or dynamic rupture simulations) will be more routine, as computational costs decrease, so generating such large databases that allow exploring parameter space will be easier. We are not sure, however, why the reviewer mentions "a couple of deterministic cases", while the CyberShake dataset we use is a very large dataset that thoroughly explores the parameter space for PSHA.

In L138 we do not mean the constraints in the modelling of the events, i.e. "kinematic vs dynamic rupture" or other parameters for modelling, but we mean constraints for how the ML alrogithms learns to associate earthquake parameters with the IM in question. In the current implementation we use magnitude, and only the locations of the hypocentre and the sites, for the learning task. We could also use more elaborate parameters, such as the FM, or the rupture extent, etc, for the learning task. These parameters do not refer to the simulation parameters of the database. We clarified that and the sentences now read: As decreasing uncertainty in predicted IMs is of paramount importance, further event features, such as the focal mechanism or rupture extent, could be included to train MLESmap models provided that such parameters can be assessed rapidly for a new earthquake. Such additional input characteristics may result in ML models that yield even more accurate inferences. We have also added a mention of dynamic-rupture simulation databases in the Methods section on the Synthetic physics-based earthquake ground motions (see also one of the comments below).

- L143: GMMs also are challenged by that. Overall, I suggest avoiding ML vs GMM, and rather highlight the existing challenge concerning data availability and physics-based knowledge. All approaches are at the end complementary.

Indeed the phrasing that "GMMs excel at covering such gaps" that the reviewer refers to was unfortunate and did not relay exactly what we intended to say. The text now reads "our results suggest that GMMs could cover such gaps." - reflecting the fact that our work shows that empirical GMMs gave better predictions for events that are out of scope for MLESmap parameters.

We agree that the methods can be complementary and useful under different conditions. We state explicitly in the text that merging the two likely will give the best results: "Integrating the scope-limited MLESmap with the more general GMMs could also lead to significant improvements and render the approach more generally applicable." As mentioned in one of the comments above, we also added in the introduction that the GMM limitations require research on "**complementary** strategies" to highlight that we do not talk about a competing strategy. However, we also want to highlight that it is important to seriously consider the advantages that we can have using new technologies to fill the gap between the lack of data and the inclusion of physics-based assumptions that now are possible thanks to all the developments in this era of supercomputers, data, and simulations. We use the "ML vs GMM" for comparison of results, treating the GMM as the state-of-the-art and thus a benchmark.

- L153-L167: You can simplify this paragraph. Mainly you can highlight that simulations are based on kinematic source models, and wave propagation is calculated using the region-specific velocity profiles, and provide references for more details.

We shortened the paragraph by a third. However, we choose to keep the details about reciprocity, as this is what makes it computationally feasible to calculate such large databases and is an important feature. It is the backbone of the method that makes computing such databases possible, and thus makes MLESmap possible.

- Please comment on the limitation arising from kinematic source models, such as constant rupture speeds, preset slip distributions (even if they are stochastic). You can mention it as a potential improvement in future perspectives.

As reviewer suggests, having a richer set of rupture conditions could improve the applicability of the method. We added this in the Methods section: "Finally, it should be noted that the MLESmap methodology is not limited specifically to training on CS-like databases. In CS, although multiple kinematic rupture scenarios are derived for each rupture in the ERF by varying slip distributions and hypocenter locations across an input fault system, the rupture speeds and slip distributions in each simulation are pre-set. Having a richer set of rupture conditions by considering a synthetic database of dynamic rupture simulations could further improve the applicability of the method."

Minor comments

- Title: Ground motion prediction is the commonly used term. You may also consider to highlight the region of study.
The title now includes the region of study:
Ground Motion Shaking Predictions in Southern California Based on Machine Learning Trained with Data from Physics-based Simulations

- L13: All earthquakes are unpredictable. Please be specific. Do you mean that variabilities/ uncertainties increase? If so why? Data scarcity? Uncertainties in physics-based model inputs?
We do not compare large earthquakes to all earthquakes in this sentence, but to other phenomena, and do not want to make references to data scarcity or modelling. We believe that the sentence reads as follows: among natural phenomena, large earthquakes are among those most destructive and least predictable (v.s. a destructive hurricane or fire, where we can track them and evacuate in real time, for example).

- L19: Relief measure?
Added "emergency", now reads "emergency relief measures".

- L43: "just after the earthquake occurs." Do you mean recordings?
Text changed to: "which is rapidly estimated and made available by international agencies"

- L33: times-to-solution?
We leave this as is, time-to-solution of a computational science or engineering problem is a common term referring to how long it takes from starting calculations to obtaining a solution.

- L58: Please avoid hype words, such as very good, excellent, and instead quantify.
Modified the expressions.

- Fig. 3: Where is the hypocenter in this event? Pinky cross?
Yes, we added this in the figure caption, thank you for pointing this out!

- Fig. 4 : You may consider adding a rectangle showing the boundary that separates inside/outside stations.
Not all the stations inside/outside fit neatly into a rectangle. We tried this and the boundary does not render the figure more readable.

- Fig 6: Magenta markers are not star in my Pdf.
Changed caption to "magenta dots", thank you for pointing this out!

- L125: Stochastically attained freq.? Unclear. Do you mean stochastic methods of ground motion modeling? Even so, I suggest to generalize physics-based rupture and ground motion modeling.
CyberShake Study 15.12 includes stochastically simulated high-frequency (> 1 Hz) ground motion on top of the deterministic simulations up to 1Hz. The high-frequency simulation uses a physically motivated but simplified model for wave propagation and scattering to generate theoretically consistent ground motion amplitudes at these frequencies (based in part on the model of Boore (1983)). We changed "stochastically attained" to "stochastically simulated".

- L130: "unrecorded" events: do you mean large and rare events?
Yes, we corrected this accordingly.

- In general, you limit the benefit of this study to post disaster emergence response, but technically it is also applicable for hazard assessment as a part of preparedness phase, i.e., before and during hazard.
Yes, it could be used, as we can also generate what-if scenarios, but for preparedness more elaborate models with different predictors are likely more relevant. We show that we get good predictions with the simple predictors to get a prediction as quickly as possible with as little information as possible.

- L136: This reads contradictory. For 1 Hz, the problem arises from uncertainties associated with input parameters and idealizations in the physical processes. You can rather make your point for high-frequency simulations.
In this sentence we are not making a point about high-frequencies. The point is more general: the database used in this study has a high computational cost, and thus the approach might seem prohibitive for general applications (that is, for other regions than California), but this is changing rapidly.

- L144: Any reference?
This is the outlook we foresee for the application, no particular reference.

- In general, providing references along the discussions would enrich the manuscript.
We have included 17 extra references in the manuscript following the suggestions of all reviewers. We have listed those above in another comment by the reviewer. Relevant references have been included along the discussions.

- L149: Please avoid too generic phrases.
We removed the sentence.

- L159: Seismic reciprocity?
Yes, seismic reciprocity is used in CyberShake, rendering the computations feasible.

4th March 2024

Dear Dr Monterrubio-Velasco,

Your manuscript titled "Ground Motion Shaking Predictions in Southern California Based on Machine Learning Trained with Data from Physics-based Simulations" has now been seen by our reviewers, whose comments appear below. In light of their advice we are delighted to say that we are happy, in principle, to publish a suitably revised version in Communications Earth & Environment under the open access CC BY license (Creative Commons Attribution v4.0 International License).

We therefore invite you to edit your manuscript to comply with our format requirements and to maximise the accessibility and therefore the impact of your work.

EDITORIAL REQUESTS:

Please review our specific editorial comments and requests regarding your manuscript in the attached "Editorial Requests Table".

*****Please take care to match our formatting and policy requirements. We will check revised manuscript and return manuscripts that do not comply. Such requests will lead to delays. *****

Please outline your response to each request in the right hand column. Please upload the completed table with your manuscript files as a Related Manuscript file.

If you have any questions or concerns about any of our requests, please do not hesitate to contact me.

SUBMISSION INFORMATION:

In order to accept your paper, we require the files listed at the end of the Editorial Requests Table; the list of required files is also available at https://www.nature.com/documents/commsj-file-checklist.pdf .

OPEN ACCESS:

Communications Earth & Environment is a fully open access journal. Articles are made freely accessible on publication under a CC BY license (Creative Commons Attribution 4.0 International License). This license allows maximum dissemination and re-use of open access materials and is preferred by many research funding bodies.

For further information about article processing charges, open access funding, and advice and support from Nature Research, please visit https://www.nature.com/commsenv/article-processing-charges

At acceptance, you will be provided with instructions for completing this CC BY license on behalf of all authors. This grants us the necessary permissions to publish your paper. Additionally, you will be asked to declare that all required third party permissions have been obtained, and to provide billing information in order to pay the article-processing charge (APC).

Please use the following link to submit the above items:

[link redacted]
** This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete th e link to your homepage first **

We hope to hear from you within two weeks; please let us know if you need more time.

Best regards,

Joe Aslin

Deputy Editor,
Communications Earth & Environment
https://www.nature.com/commsenv/
Twitter: @CommsEarth

REVIEWERS' COMMENTS:

Reviewer #1 (Remarks to the Author):

Thank you for addressing my comments from the previous version of this manuscript. I find that you have addressed all my concerns and do not see any remaining issues.

Reviewer #2 (Remarks to the Author):

The paper has been improved. It is valuable for pubblication.