Response to referee #3

The paper aims to test seismic hazard estimates derived from earthquake simulators, against historical intensity data and instrumental ground shaking recordings. This is an interesting attempt to connect with observational constraints and deserves publication. I offer a few questions and comments below to help improve the manuscript.

We thank the reviewer for the thorough and constructive review to our article. Below we provide detailed responses to each one of the points raised by the reviewer.

Reviewer comment 1 (RC1): The description of the models could be improved. In table 1 cat – 21 the initial normal stress is stated as 20 MPa per kilometer, but it is not mentioned that this is a vertical gradient, which only became clear from looking at the earlier paper references [Herrero-Barbero et al., 2021; Gomez-Novell et al., 2025]. Also, how can this low normal stress in the shallow layers be reconciled with the large initial shear stress of 60 MPa?

Author's response 1 (AR1): We agree with the reviewer that a homogeneous shear stress of 60 MPa is unrealistically high for low normal stress values at shallow depths. However, the key driver of stress evolution in the simulations is the normal stress. Previous studies like Liao et al. (2024), show that RSQSim simulations with an initial heterogeneous normal stress (vertical gradient) induce shear stresses to evolve into a depth-dependent profile once the simulations have stabilized, even if the initial shear-stress profile is homogeneous. This is because ruptures induce more changes in the shear stress compared to the normal, and because the shear stress is a fraction of the normal stress and thus, stress perturbations have proportionally more influence on the former (Liao et al., 2024).

Following this rationale, in our analysis, we conservatively discard the first 10-kyr from the simulation, considering them as the spin-up phase. This ensures that the models have diverged enough from the initial conditions and, thus, mitigates the potential effects that initial homogeneous shear stresses might have on the simulated catalogues.

In the revised manuscript (section 2.2.2), we will better specify the characteristics of the initial model conditions (depth-gradient of the normal stresses), and the relationship with the shear stresses.

RC2: I do not understand what the role of the initial shear stress plays in the simulation, once spin-up has been achieved, other than changing the specific sequences. Why is this a relevant parameter from a statistical point of view?

AR2: The initial shear stress conditions define the onset of the first ruptures in the system. From this point, the shear stress field is successively modified after each rupture, until it reaches internal consistency with the normal stress, friction, loading rate and fault geometry conditions of the model. This means that the statistical significance of the initial shear stress is mainly limited to this transient stress evolution phase (spin-up), where it controls the statistical properties of the catalogues until the stability phase, and the duration of the transient phase itself. After that, the long-term earthquake catalogue statistics respond to shear stress values that have evolved from the initial shear stress.

Having said that, the pre-existing shear stress conditions can have a large impact on specific short-term earthquake sequences, which makes it a relevant parameter for induced seismicity applications. In these cases, the short-term transient phase of seismicity after a specific initial stress condition (e.g., perturbation) is the target. Having said that, we reinforce that evaluating short-term sequences is not the objective of this study.

RC3: The stated loading suggest there would be uniform slip across the sections. Under backslip, this would be expected to nucleate a lot of events off the fault boundaries. Looking at the earlier references, I did not see any indications of where events are nucleating with depth and along-strike. Further, Figure 4 in [Gomez-Novello et al. 2025] suggest there are different slip patterns accumulating in different models. That would be unexpected in a steady state backslip. So these are either incomplete not long-term steady-state catalogs, or something else is going on. It would be good to clarify this all. One option is to do hybrid loading which uses regularized stressing rates to achieve self organizing slip rates, which can then be used in backslip mode. At a minimum, clarity in what is being used here, and plots of hypocenters would be useful.

AR3: First of all, we want to clarify that the figure 4 reference (Gómez-Novell et al., 2025) does not correspond to the RSQSim models of the EBSZ (Spain), but from a test model of a "single planar fault" used to exemplify the benchmarking method developed in that publication. In that figure, the slip patterns are used to showcase the differences in performance that different model parameter configurations can generate in output catalogues. The model with larger slip concentrations at the fault boundaries is used as a proxy to demonstrate the poor performance of that model, previously detected by the automatic benchmarking scoring with empirical relations.

In regards to the reviewer concern: we are aware that the current loading (uniform slip rate) under the back-slip assumption can generate nucleations at the fault boundaries and at shallow depths. In figure 1, we show hypocentral depth distributions and in figure 2, the hypocenter plots onto the fault system for both Cat-21 and Cat-18. As the figure shows, in Cat-21, hypocenter locations are mostly in the shallow parts of the seismogenic depth of the faults due to the low normal stress values in these regions. Conversely, Cat-18, shows more nucleations along the whole seismogenic depth, with higher concentration at the bottom due to the back-slip-related nucleations paired with a homogeneous normal stress profile.

Having said that, for a hazard application like ours, the rupture geometry on the fault is more important than the nucleation point. A few considerations:

1. The geometric centroid distribution of the ruptures (figure 1) shows that most ruptures are centered to the mid-seismogenic depth, a fact that is most consistent with real earthquake observations. This means that, even if earthquakes nucleate at shallow depths (e.g., Cat-21), the rupture develops towards deeper parts of the faults. Our hazard models take into account the rupture geometry, not the hypocenter. Hence, the centroid is a better proxy for determining the characteristics of the earthquakes than the hypocenter.
2. Centroid locations on the fault surfaces (figure 3) are less influenced by shallow and fault-tip nucleations, especially for larger magnitudes (M>6). However, such effects are not removed entirely for smaller magnitudes (M<6), meaning that a considerable number of earthquakes ruptures are still restricted to these fault regions. A couple of comments on this: i) Regarding the shallow depth of the nucleations, we expect this effect to have a limited impact in our hazard results because most of the faults affected are high-dip to subvertical (e.g., Carboneras fault; figure 3), and the distance metric used for the GMM is Joyner and Boore - JB. JB considers only the surface projection of the rupture, meaning that the source-to-site distances are insensitive to rupture depth in these faults. ii) Regarding nucleations at the fault tips for small magnitudes (M<6), we also expect a limited impact in the hazard, as the models nucleate earthquakes not only at these locations but all along the strike of the faults (see figure 3).

3. For site-specific purposes, the consideration of earthquake nucleation locations (e.g., in the vicinity of Carboneras fault) and detailed source-to-site distances would be more relevant for the hazard than in regional-scale applications like ours. This is currently out of our scope, but is worth mentioning.

In the revised version of the manuscript, we will include the figures provided here in the Supplements, and we will clarify the points discussed here.

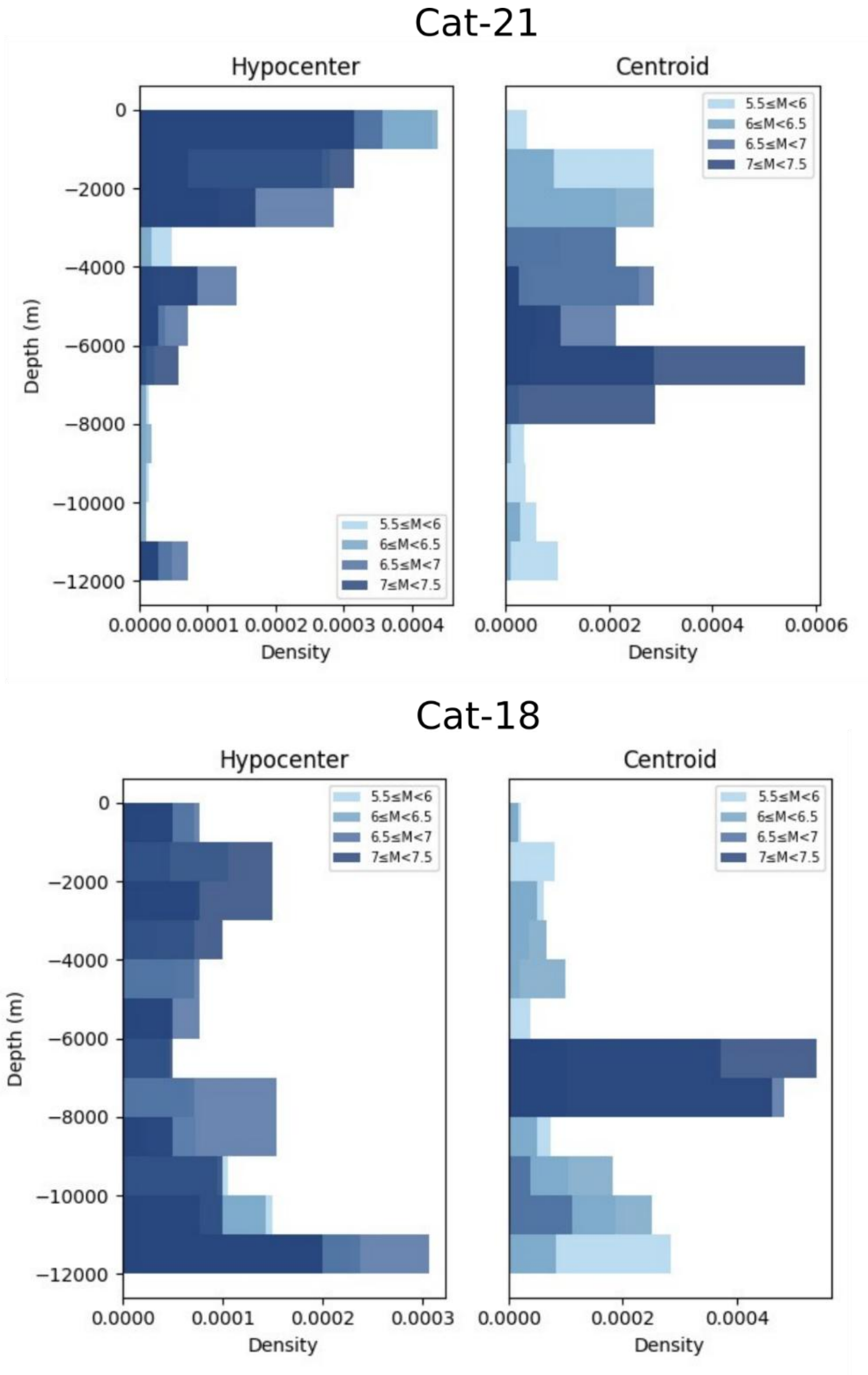

Figure 1. Depth distribution of hypocenters (left column) and rupture centroids (right column) for simulated catalogues Cat-21 and Cat-18.

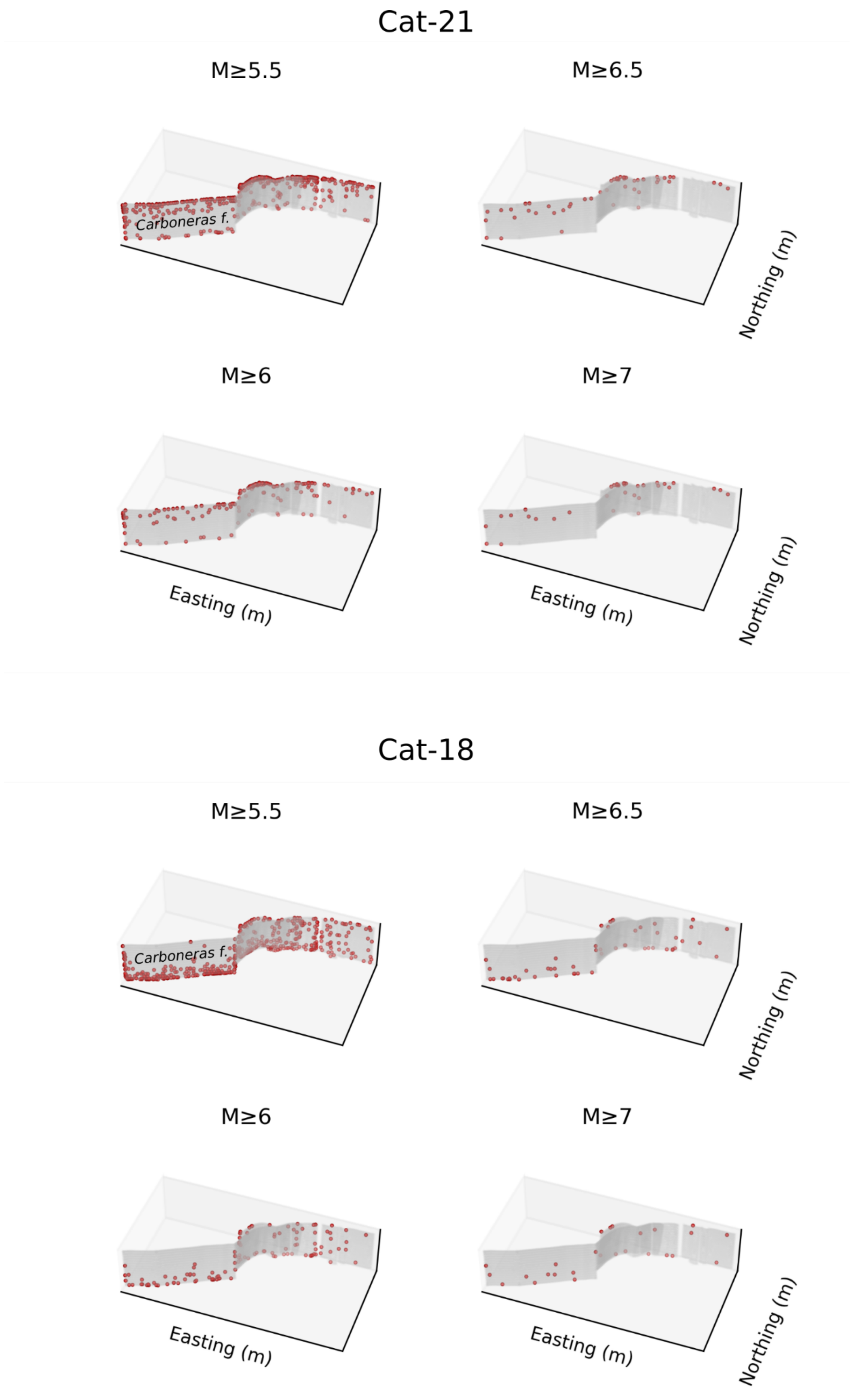

Figure 2. Distribution of hypocenters for simulated catalogues Cat-21 and Cat-18 along the EBSZ faults. The Carboneras fault is indicated for references in the text.

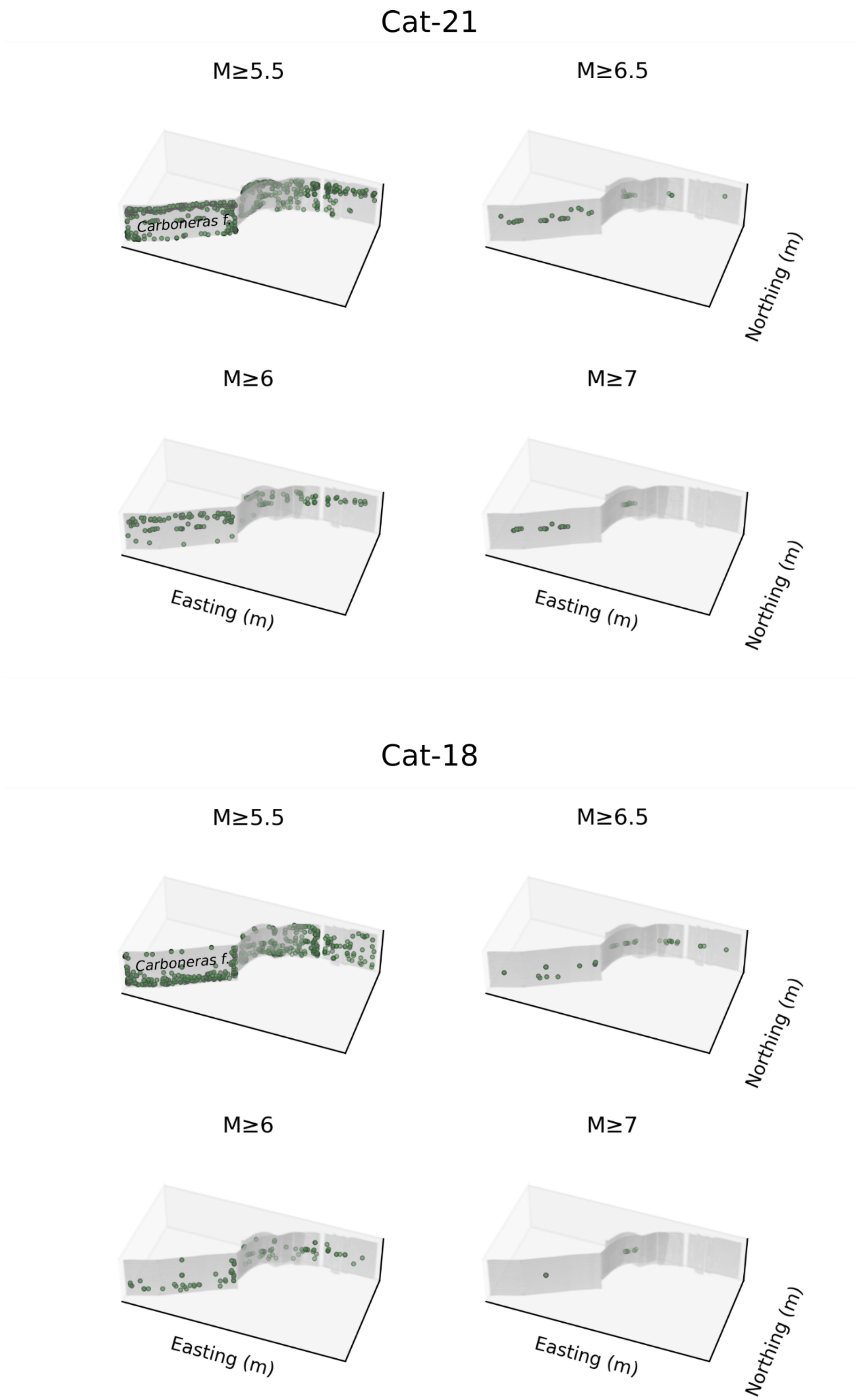

Figure 3. Distribution of rupture centroids for simulated catalogues Cat-21 and Cat-18 along the EBSZ faults. The Carboneras fault is indicated for references in the text.

RC4: There is a difficulty in trying to use the simulator models to compare against short time scale observations (less than large event repeat time), in that these shorter times scale features then become dominated by the small events. The problem there is that the small events tend to be dominated by smaller scale geometrical features which we know we don't know. In contrast, for longtime scales, the largest events dominate and there is more hope that large scale geometrical features play a more relevant role there. Note that in [Shaw et al. 2018] the hazard differences between the simulator and UCERF3 estimates in California were small at century time scales, but grew rapidly at times scales of a few decades and shorter where small events dominate the expected shaking. In low strain rate regions, this effect is only going to be exacerbated. The authors do speak to this in their comments regarding the utility of combining the simulators to go after longer term hazard, with the area sources to complement this. But a fuller discussion of what may be possible in these comparisons with this type of data over available and foreseeable time scales would be helpful. For measures which are going to be dominated by small events, hypocentral distributions are of additional importance due to ground motion model sensitivity to closest distances. This circles back to the previous comment, so some discussion of small event hypocenters along-strike, and with depth, is needed.

AR4: This is a very interesting point. We agree that comparing models against observations is a challenge. For the short term (smaller magnitudes), certainty of a causative relationship between seismicity and the modelled faults decreases, as small events might occur in unknown or unmapped structures. This issue is transversal to most fault-based hazard assessments – not only physics-based – and can explain discrepancies in short-term hazard estimates between model and observations.

In fault model-to-fault model comparisons, like the RSQSim-UCERF3 (Shaw et al., 2018), this is hardly the cause of the observed mismatches, because the distribution of magnitudes between background and faults is done prior to the fault source modelling. In this case, and as pointed by the reviewer, a contributor to the short-term discrepancies is the fact that smaller magnitudes might be related to smaller-scale physical features of the faults (e.g., geometries), which affect differently the RSQSim and UCERF-3 approximations. Interestingly, Shaw et al. (2018) attribute these discrepancies to hazard modelling features too: i) differences in the spatial extent of ruptures being smoothed in the shaking, ii) complementarity of large events with higher shaking and slightly smaller size events with lower shaking but higher ground motion probabilities, and iii) relative insensitivity of the GMM to larger magnitudes. This means that, while the rupture characteristics of earthquakes are important, there are other components controlling the hazard differences that do not strictly depend on the specific ERF.

In figure 4, we plot the changes (mean absolute error) in PGA and PGV as a function of PoE for each simulated catalogue vs. the ZESIS model. The plot shows how differences are, for the most part, larger for the long-term than for the short-term, especially for our preferred catalogue (Cat-21). At the EBSZ, the biggest challenge is characterizing the long-term, which is largely underrepresented in the historical and instrumental catalogues (ZESIS model). With this, our models suggest that the utility of physics-based models in low-strain regions is highly linked to their ability to increase observational resolution in the long-term seismicity than in reproducing short-term observations that are already captured in the seismicity records. Having said that, we remark that proper performance tests in this regard would ideally require model-to-observations tests rather than model-to-model comparisons.

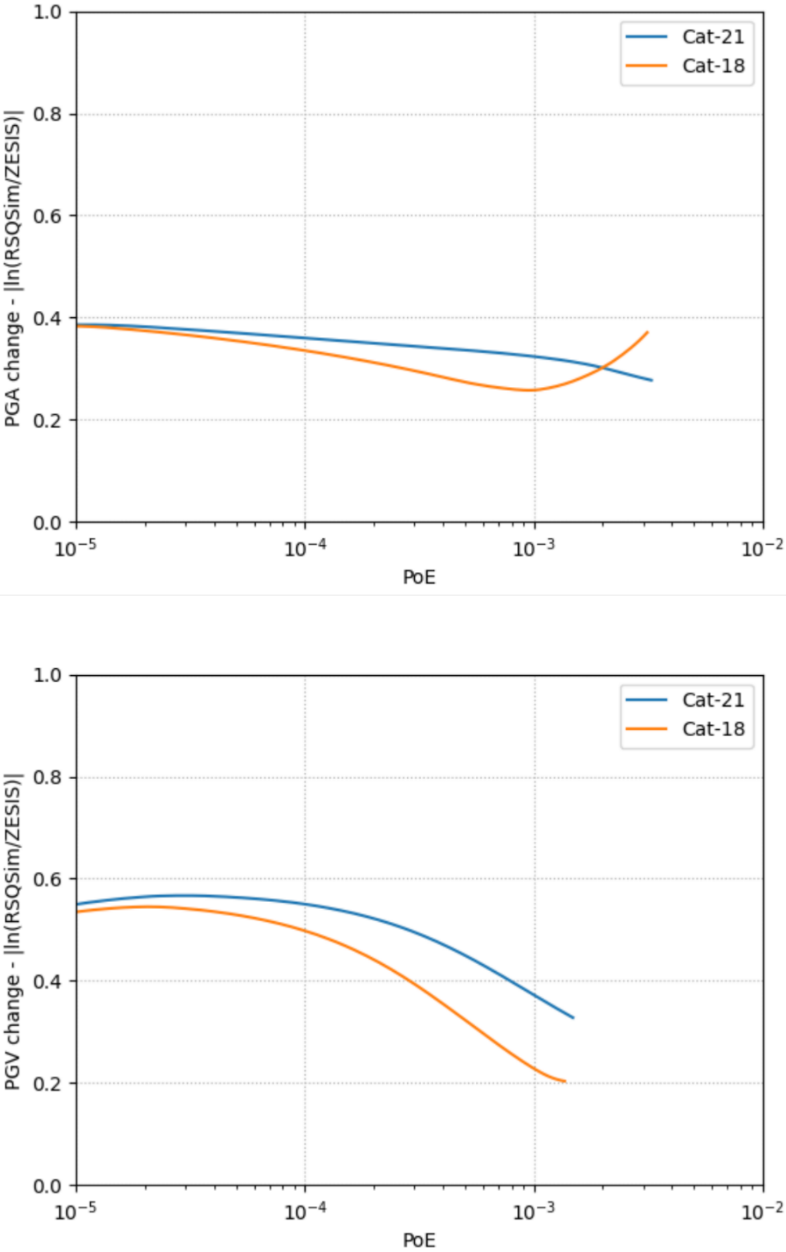

Figure 4. PGA (top) and PGV (bottom) variation of Cat-21 and Cat-18 with respect to the ZESIS area source model. Changes are computed using <|ln(RSQSim-ZESIS)|>, as in Shaw et al. (2018).

Regarding the nucleation of smaller earthquakes, our models still face difficulties in reproducing realistic along-strike and depth distributions (see comment AC3). We argued that the impact of these inconsistencies on the hazard estimates is likely limited, given our modeling choices (e.g., source-to-site distance metric, ground-motion model selection). However, we agree that this issue deserves careful consideration. In particular, developments of the presented testing approach for regulatory seismic hazard evaluations and site-specific studies, would require exhaustive evaluation of earthquake nucleation distributions to ensure consistency with hypocenter observations. That said, uncertainties in nucleation depth are not unique to physics-based ERFs. Similar challenges arise in traditional fault-based ERFs, where ruptures are floated along fault planes, including shallow sections and fault tips. Therefore, careful treatment of nucleation distributions is necessary in both approaches.

In the revised version of the manuscript, we will discuss the points made in this comment. For one, we will explain the differences in the long-term and short-term hazard between simulations and the area source model, including figure 4 as a supplement, and their comparison with other regions like

California. For another, we will also discuss the relevance of realistic earthquake nucleations for seismic hazard, given their sensitivity to certain measures of source-to-site distances.

RC5: Line 490 please be more specific on what is meant by the results support the simulators capturing not just large scale hazard patterns but also localized differences seen in empirical data. What is this referring to exactly? How robust are those results?

AR5: This sentence captures two main ideas:

1) Physics-based models are able to depict the influence of major faults in the territorial hazard. For instance, models show sharp hazard increases close to faults, as opposed to the area source model (see figure 6).
2) The agreement of the physics-based hazard estimates with macroseismic and station records at different sites across the EBSZ, indicates that the models are also able to produce forecasts consistent with – local – observations at those sites.

The robustness of the results is backed up by the statistical tests performed in the study, which demonstrate statistical consistency between models and observations.

In the revised version of the manuscript, we will expand this sentence by providing the main ideas described above.

References mentioned

Gómez-Novell, O., Visini, F., Pace, B., Álvarez-Gómez, J. A., and Herrero-Barbero, P.: A Benchmarking Method to Rank the Performance of Physics-Based Earthquake Simulations, Seismol. Res. Lett., 96, 231–243, https://doi.org/10.1785/022024002, 2025

Liao, Y.-W, Fry, B., Howell, A., Williams, C. A., Nicol, A., and Rollins, C.: The role of heterogeneous stress in earthquake cycle models of the Hikurangi–Kermadec subduction zone, Geophysical Journal International, 239, 574–590, https://doi.org/10.1093/gji/ggae266, 2024

Shaw, B. E., Milner, K. R., Field, E. H., Richards-Dinger, K., Gilchrist, J. J., Dieterich, J. H., and Jordan, T. H.: A physics-based earthquake simulator replicates seismic hazard statistics across California, Sci. Adv., 4, eaau0688, https://doi.org/10.1126/sciadv.aau0688, 2018.