Response to referee #1

We thank the reviewer for taking the time to perform a thorough and constructive review to our article. Below we provide detailed responses to each one of the points raised by the reviewer and explain the changes applied to the manuscript.

MAJOR COMMENTS

Reviewer comment 1 (RC1): Lines 220-235: The paper presents a consistency testing framework, but there are opportunities to employ more rigorous statistical validation methods. In statistical practice, methods exist to estimate out-of-sample performance and predictive power of models, such as cross-validation. While applying such methods to an entire PSHA model may be technically challenging, components of the PSHA (particularly the occurrence models) can be tested using these approaches. Since this paper focuses on physics-based synthetic catalogs, testing the catalog or the occurrence models based on the synthetic catalogs before integrating them into the full PSHA would strengthen the methodology significantly. If this is out of the scope of your work, or if that would lead to a lengthy article, please discuss this briefly and state these issues should be tested in a companion report or paper. If such tests have been conducted in other studies, they should be cited here. Additionally, the adopted testing procedure evaluates the joint model (occurrence rates combined with GMPE), which should be explicitly stated and briefly discussed.

Author's response 1 (AR1):

We agree with the reviewer that testing the different components of the PSHA model would be the adequate approach, especially for a national/regulatory hazard assessment where both the occurrences and the ground motion models need to be properly tested. In our approach, we wanted to focus only on the hazard outputs to assess whether physics-based models are able to return consistent forecasts without requiring any prior testing of the hazard components.

While performing these prior tests would imply a major re-structuring of our testing approach, in the revised version of the manuscript we have implemented the following changes to better clarify and discuss this point:

1. We included the EBSZ catalogue MFD in figure 1. Right now, the figure shows only the fitted MFD from the area source model, so a comparison with the catalogue is also pertinent.
2. In section 2.2.2 (lines 181-184), we have added a paragraph to discuss the agreement of each simulated model with the catalogue MFD.
3. In section 2.4 (lines 288-289), we have better clarified that our approach tests the joint model, and we discuss in section 4.3.2 (lines 654-658). how further consistency analyses might benefit from separate testing of the different components of the hazard model.

RC2: Lines 250-254: Some clarifications are needed to rule out methodological inconsistency. The synthetic catalogs are tested against the stationary Poisson process hypothesis, which is typically used in PSHA for declustered events (mainshocks only). The manuscript suggests that there may be some foreshocks and aftershocks in the synthetic catalogs, which do not significantly affect the PSHA results, which is acceptable. However, readers may be confused and think the paper performs PSHA for full non-declustered seismicity. This distinction must be clarified explicitly. Please discuss briefly whether the presence of clustering in the catalogs affects the interpretation of results and whether a sensitivity analysis excluding clustered events would alter conclusions.

AR2: Rate-and-state simulators like RSQSim simulate earthquake sequences, including foreshocks and aftershocks. Even though we demonstrate that, at the scale of the fault system, the catalogues follow a Poissonian process, these have not been declustered. This means that the earthquake ruptures we pipeline to the PSHA might reflect time-dependent behavior linked to the spontaneous earthquake interactions emerging from the simulator. To ensure consistency in the analysis, we performed a re-clustering to the occurrence rates of the ZESIS area source model (area 55), because the Gutenberg-Richter fit parameters from this model come from a de-clustered catalogue (lines 250-252 of the manuscript). The de-clustering follows the approach by Marzocchi and Taroni (2014), which defines multiplicative coefficients to calculate non-declustered (complete) frequencies for magnitude values starting from their declustered values. In the same way, the macroseismic and instrumental catalogues used for consistency tests have not been filtered to exclude foreshocks and aftershocks. Even though a proper full declustered analysis would imply a re-calculation of the GR parameters of area 55 directly from a non-declustered catalogue, our analysis performs an approximation to that.

In the revised manuscript (lines 290-295), we have better clarified this point (section 2.4).

RC3: Line 208: The authors use "Classical PSHA" in OpenQuake, but the OpenQuake Engine also allows event-based PSHA using synthetic catalogs directly. This approach would avoid fitting occurrence laws, a procedure that potentially loses information contained in the catalogs. Moreover, OpenQuake allows scenario calculations that could be run for every event in the synthetic catalogs to calculate exceedances directly. These alternatives should be discussed, including why the classical approach was chosen over these potentially more direct methods. This discussion would clarify whether methodological choices limit the ability to fully leverage the physics-based catalog information.

AR3: This is a very interesting point. There are two reasons why we did not choose the event-based formulation:

1) The event-based PSHA approach, as well as scenario-based calculations, are especially important for risk assessments. This is not our scope here. Our aim is to develop a PSHA approach that is appealing to regional/large-scale seismic hazard applications, which commonly employ the classical approach (e.g., the area source model of Spain).
2) In the current OpenQuake (OQ) version, event-based PSHA produces an earthquake catalogue by stochastically sampling ruptures from an earthquake rupture forecast (ERF). This catalogue is considered a realization of the seismicity that a source can produce, and it is an optimal solution for the classical parametric source approximation, where sources are defined by an occurrence law without specific information of the ruptures (seismicity). However, this approximation is unnecessary for physics-based synthetic catalogues, as each catalogue itself is both the ERF and the stochastic realization of the seismicity. In fact, applying the OQ event-based approach here would enable simulated ruptures to be sampled multiple times along the investigation time. This feature would imply an alteration of the actual spatial rupture characteristics of the simulated catalogues, as earthquake ruptures are always unique.

   With our formulation, the calculations do not lose or alter information from the catalogues. For one, complex rupture geometries are preserved in the hazard as each simulated rupture is treated as a unique grid source. For another, while each rupture has a single occurrence on the catalogue (1/duration), the stack of all rupture occurrences follows that emerging from the simulated catalogue.

The optimal solution to perform a physics-based event-based PSHA would be to bypass the simulated catalogues directly to the ground motion field generation, but this is not a supported feature in OQ at the moment. We developed the pipeline approach in Python because, eventually, we intend to implement it with other types of OQ calculations, including event-based, in collaboration with the engine developers.

In the revised manuscript (section 2.3, lines 209-214), we have better clarified how our classical PSHA approach allows also a full exploitation of the simulated catalogue information.

RC4: Line 137: While the authors state that Cat-21 and Cat-18 represent the best and worst performing catalogs based on benchmarking, the rationale for explicitly including the worst-performing catalogue requires explanation. What insights does Cat-18 provide that justify its inclusion in the hazard analysis? Is it meant to demonstrate the importance of proper model selection, or does it represent a plausible epistemic uncertainty branch? This clarification would help readers understand whether poor-performing models should ever be included in operational hazard assessments or logic trees.

AR4: The rationale behind including the two models is both to demonstrate the importance of model selection and epistemic uncertainty exploration. First, the fact that Cat-18 – the worst ranked catalogue – is also the model with poorer performance in our testing, highlights the relevance of a proper parametrization of the physics-based models. Second, the fact that both models pass the statistical test against observations (p-values always above 0.05) means that they both should be considered as plausible realizations in a PSHA. The optimal solution would be a logic tree exploration in which the relative performance of all models is used to define branch weights.

We have better clarified the rationale of including both models in section 2.2.2 of the revised manuscript (lines 160-162).

RC5: Line 23 (Abstract) and throughout: The manuscript advocates for combining physics-based and traditional approaches, describing this as "complementarity." This concept is well-established across many scientific fields as hybrid modeling, which systematically combines physics-based models with data-driven approaches. The authors should acknowledge this broader context and cite relevant literature on hybrid models in seismic hazard or related fields (e.g., hydrology, climate science). This would strengthen the theoretical foundation and help position the work within established methodological frameworks.

AR5: We thank the reviewer for this is a very interesting suggestion. We have mentioned this in the abstract (line 23) and in the discussion (section 4.4, lines 743-747 of the revised manuscript).

MINOR COMMENTS

ABSTRACT

RC6: Line 21: The phrase "both the lower-performing simulation and" could be omitted for brevity without loss of meaning.

AR6: We have removed the sentence.

RC7: Line 22: The term "reliable" should be verified to ensure it is fully supported by the results presented. Given the consistency testing (rather than validation) performed for some components, this wording may overstate the conclusions.

AR7: We agree with the reviewer. We have replaced “reliable” by “appropriate”.

1. INTRODUCTION

RC8: Line 28: The statement that PSHA was "formalized by Cornell (1968)" is not 100% fair. Please see McGuire (2008), Probabilistic seismic hazard analysis: Early history, Earthquake Engng Struct. Dyn. 2008; 37:329–338. DOI: 10.1002/eqe.765

AR8: We have added the relevant references by Esteva in the context (line 28 of the revised manuscript).

RC9: Line 86: The reference to Ellingwood and Wen (2005) does not support the statement about high-impact, low-probability events as written. This citation should be removed or replaced with more appropriate references.

AR9: Noted, we have removed this citation in the revised version.

RC10: End of Introduction: A paragraph should be added that clearly states the objectives of this paper and provides a roadmap of the manuscript structure. This would help readers understand the overall contribution and organization.

AR10: We have added explicit objectives in lines 99-102 of the revised manuscript.

2. DATA AND METHODS

RC11: Line 175: Please clarify whether each rupture in the catalogs has only a single occurrence, or whether repeated similar ruptures can occur.

AR11: Each rupture in the catalogue is unique in the sense that the specific rupture geometry for each event is different, even for a same magnitude and source. However, repeated ruptures with similar geometries and at similar locations can occur. Ultimately, the full set of ruptures is the one that defines the overall occurrence at the fault system level, even if each specific rupture is unique. We have clarified this point in the revised version (lines 212-214).

RC12: Line 262: The statement that macroseismic records at close distances are mostly related to faults requires explanation.

AR12: The macroseismic records that we selected for our analysis correspond to historical earthquakes whose epicentral area is located within the EBSZ faults modelled in the physics-based catalogues and/or that have been attributed to one of those faults. This is because, epicenters close to the EBSZ faults are more likely to be generated by such sources (e.g., Yazdi and Garcia-Mayordomo, 2025). Instead, macroseismic intensities recorded from earthquakes whose epicentral area is outside the EBSZ cannot be confidently related to the faults in the model and, therefore, cannot be used for testing the physics-based models. In the revised manuscript, we have better specified the meaning of “close distances” in lines 309-312

RC13: Line 290: The statement that PGV "is the most likely linked to damage" is too strong and not generally true. Please see the literature on seismic fragility curves, e.g. Luco N., Cornell C.A. (2007)

Structure-Specific Scalar Intensity Measures for Near-Source and Ordinary Earthquake Ground Motions, Earthquake Spectra, Volume 23, No. 2, pages 357–392. https://doi.org/10.1193/1.2723158

AR13: We acknowledge this point. In the revised manuscript, we have added justification of our selection in lines 339-341.

RC14: Figure 5c: The jump in the curve at the last point for intensity MI=12 requires explanation. Is this a computational artifact, a feature of the GMICE, or physically meaningful?

AR14: We thank the reviewer for pointing this out. The apparent jump in the curve at MI = 12 in Fig. 5c is not physically meaningful and arises from a combination of the calculation of incremental values of PGV in the hazard curve and the behaviour of the Ground Motion–Intensity Conversion Equation (GMICE) at the upper bound of the intensity scale. Because the macroseismic intensity scale is truncated at MI = 12, the probability mass associated with very high PGV values is concentrated in the last intensity bin. When the incremental (discrete) PGV levels from the hazard curve are combined with the GMICE probability distributions, this truncation can produce a visible increase in the occurrence rate at the highest intensity level. Therefore, the discontinuity at MI = 12 is a numerical effect related to the discretization of the PGV levels and the upper bound of the intensity scale in the GMICE, rather than a physically meaningful feature of the hazard model. We note, however, that this occurs at extremely low occurrence rates and does not influence the consistency tests or the interpretation of the results. To clarify this point, we have added a sentence in the footnote of figure 5 explaining the origin of this feature.

RC15: Line 324: Please provide the justification for taking the average p-value across the four cases and then computing its logarithm.

AR15: The four cases per MI class are four feasible forecasts, and the p-value reflects their agreement in front of observations in a Poisson process. Given that all four forecasts are plausible, we compute the average of the p-values as a representative measure of the sample for each MI class. The logarithm is computed because we choose a logarithmic scoring system to rank all testing components. Logarithmic scoring is a proper scoring metric widely recognized in literature (e.g., Gneiting and Raftery; 2007).

3. RESULTS

RC16: Line 340: The choice of 2% probability of exceedance in 50 years should be justified. While this is a valid choice, the most common selection for design purposes is 10% in 50 years. Was this choice made for specific reasons related to the EBSZ, or to match existing hazard maps? This should be stated explicitly.

AR16: The 2% probability of exceedance in 50 years is chosen to avoid relating our work to conventional design seismic hazard applications (e.g., building codes). Our aim is to develop a general methodological framework to compute and evaluate the consistency of hazard estimates from physics-based ERFs, not to associate the development to specific design PSHA purposes that would require further validation/testing (see comment AR1). We have added a sentence to justify the selection of this probability of exceedance in the revised manuscript (lines 395-397).

RC17: Figure 6: Please provide specific commentary on the differences between the hazard maps in the area of the city of Vera.

AR17: In the city of Vera, differences in the hazard maps across models are small because the nearby Palomares fault does not increase the hazard above the baseline level from the area source model. In fact, in some models (e.g., Cat-21), the hazard in Vera seems to be controlled by the influence of the Alhama de Murcia fault, rather than the Palomares fault. This is due to the low slip rate estimates on this fault. We have enhanced the explanation (lines 432-434 of the revised version).

RC18: Figure 7: The current presentation makes comparison difficult. It would be more effective to have one subplot per city/station showing all three hazard curves (Cat-21, Cat-18, and area source) overlaid. This would facilitate direct comparison of model performance.

AR18: We have changed figure 7 according to the reviewer suggestion and adapted the explanation of the hazard results accordingly in section 3.1 (lines 417-434).

RC19: Line 455: The summing of LogP values to rank models requires justification.

AR19: The idea is that the global score aggregates all testing metrics (MI and PGA thresholds per class and site) to reflect the overall best- and worst- performing models across all metrics. That is, the best/worst models are not necessarily those that out/under perform in all the metrics, but those that show better balance among them. This ensures independence across metrics and avoids overfitting the models to specific/selected observations that might bias the scoring.

We have better clarified this point in the revised manuscript (section 3.3, lines 520-522).

EDITORIAL COMMENTS

RC20: Line 234: The term "consistency check" is introduced but not formally defined until later in the text. Provide a brief definition at first use.

AR20: We did not find any instance of “consistency check” in line 234. We infer the reviewer is referring to line 243. We have provided a brief definition of “consistency” in the introduction and section 2.4 of the revised manuscript (lines 262-263).

RC21: Line 590: "areas model" should be "area source model" throughout.

AR21: We have replaced all “areas model” instances by “area source model” in the revised manuscript.

References mentioned

Gneiting, T., and Raftery, A.E.: Strictly proper scoring rules, prediction, and estimation, J. Am. Stat. Assoc. 102, no. 477, 359–378, https://doi.org/10.1198/016214506000001437, 2007.

Marzocchi W., and Taroni, M.: Some thoughts on declustering in probabilistic seismic hazard analysis, Bull. Seismol. Soc. Am., 104(4), 1838-1845, https://doi.org/10.1785/0120130300, 2014

Yazdi, P., & García-Mayordomo, J. (2024). Active fault interaction in the Eastern Betic Cordillera: A model of coseismic and postseismic stress transfer following historical earthquakes in SE Spain. Tectonics, 43, e2024TC008383, https://doi.org/10.1029/2024TC008383, 2024.

Response to referee #2

Overall, the manuscript is well written and technically sound, addressing an important and timely topic on the integration and evaluation of physics-based earthquake simulators within a PSHA framework. The methodology is rigorous, the workflow is clearly designed, and the consistency tests against both macroseismic and instrumental observations are carefully implemented, making the study suitable for publication after revision. That said, the introduction would benefit from a clearer hierarchical structure and stronger logical progression, particularly by more explicitly distinguishing the limitations of traditional PSHA, the advantages of physics-based approaches, and the specific scientific gap this work fills, as well as by summarizing the main objectives and novel contributions more concisely toward the end of the section. In addition, while the discussion is generally solid, its depth and breadth could be further enhanced by providing a more integrative synthesis of the results, clarifying the broader implications for seismic hazard practice in low-to-moderate strain regions, and more explicitly discussing the transferability of the proposed workflow to other tectonic settings and its role within future hybrid or logic-tree PSHA frameworks. Addressing these points would improve the clarity, balance, and overall impact of the manuscript without requiring major changes to the core analysis or results.

We thank the reviewer for the positive feedback given to our article, as well as for the suggestions made. In the revised version of the manuscript, we have:

1) Improve the introduction according to the suggestions, clearly distinguishing the limitations of classical PSHA, the potential solutions that physics-based models can offer, and the role of our study in this context. This has implied a considerable modification of the introduction (lines 26 to 113 of the revised manuscript, see tracked changes document).
2) Improved the discussion in section 4.4. to accommodate the suggestion made by the reviewer, explicitly stating the impact of our approach not only for Spain but for other low-to moderate regions worldwide. We have also mentioned how our workflow is transferrable to other tectonic settings (lines 754-755 of the revised manuscript).

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 and describe the changes made to the revised manuscript.

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 have better specified the characteristics of the initial model conditions (depth-gradient of the normal stresses), and the relationship with the shear stresses (lines 164-173).

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.

We have included the figures of this point in the supplements of the paper (figures S4-S6) and we have also added discussion on the impact of the nucleation in section 4.3.1 of the revised manuscript (Lines 611-630).

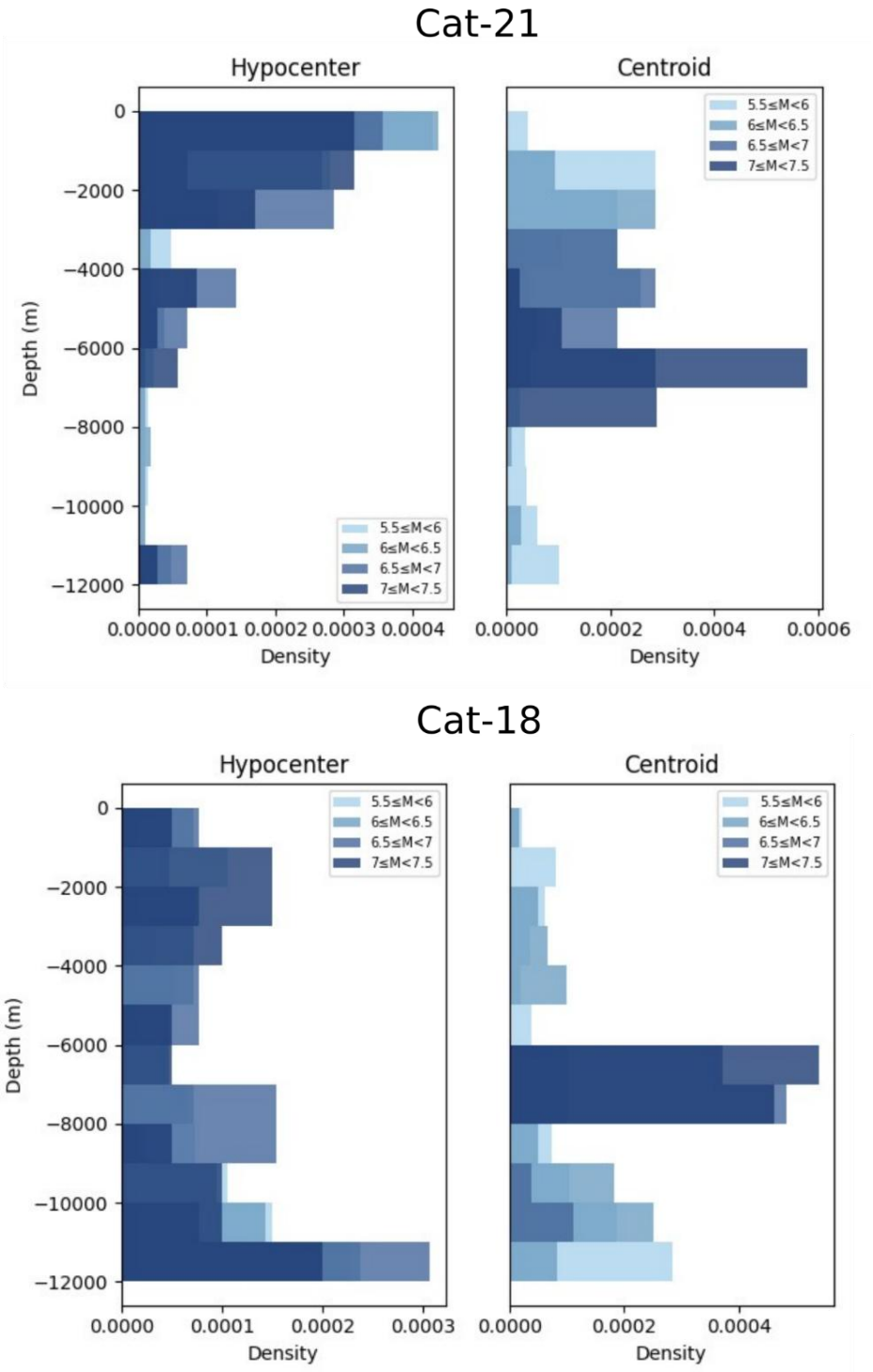

Figure 1. Depth distribution of hypocenters (left column) and rupture centroids (right column) for simulated catalogues Cat-21 and Cat-18.

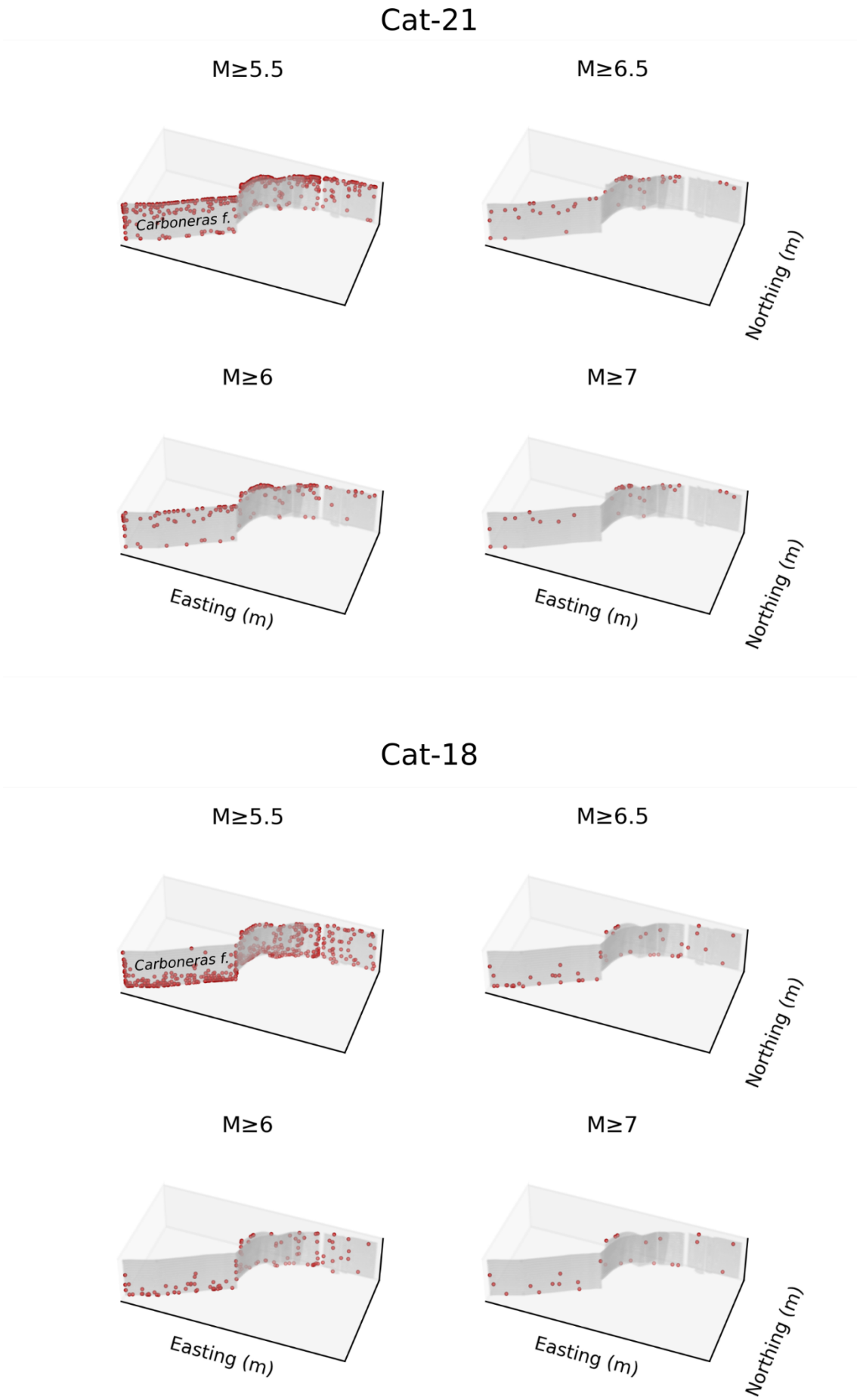

Figure 2. Distribution of hypocenters for simulated catalogues Cat-21 and Cat-18 along the EBSZ faults. The Carboneras fault is indicated for references in the text.

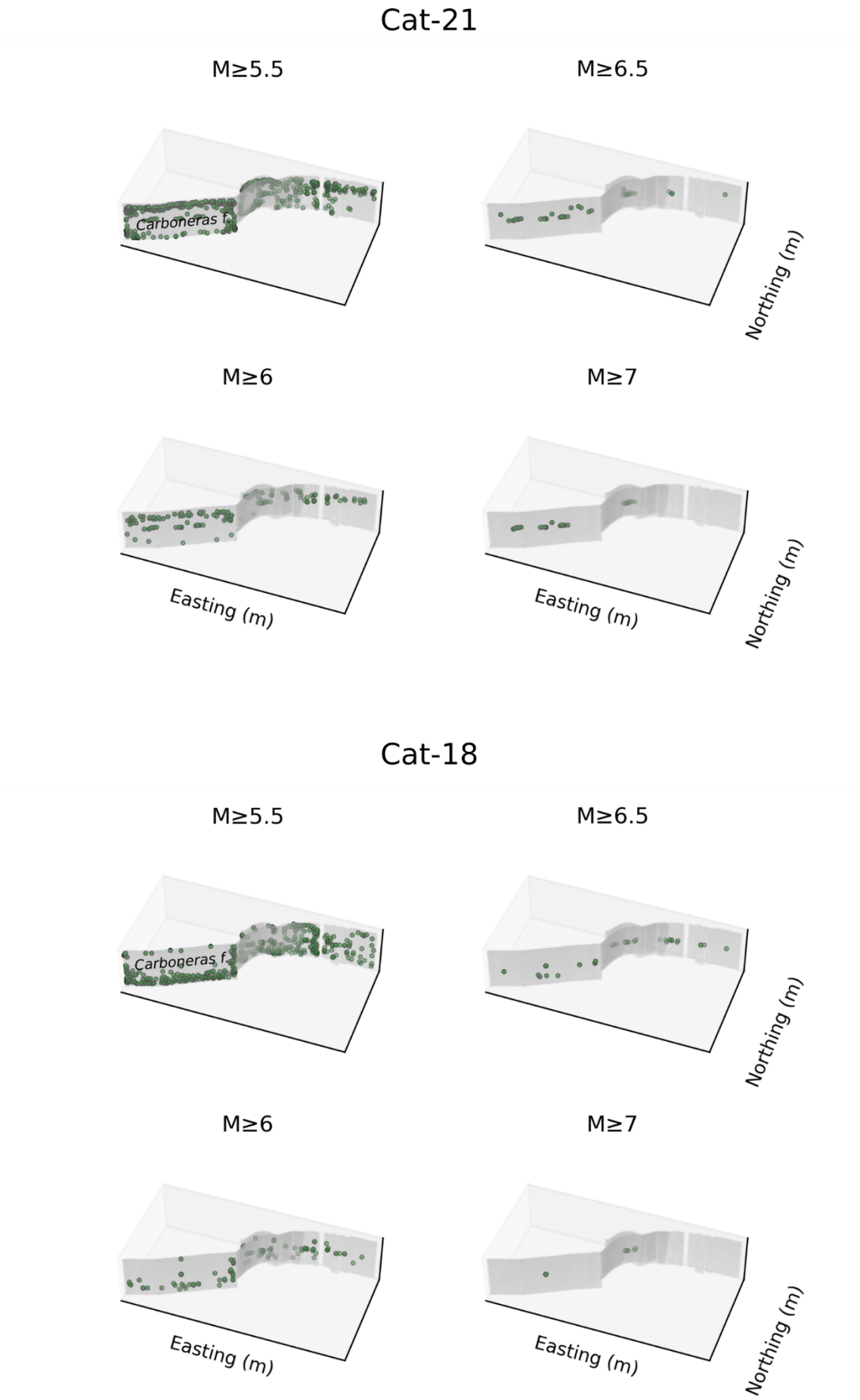

Figure 3. Distribution of rupture centroids for simulated catalogues Cat-21 and Cat-18 along the EBSZ faults. The Carboneras fault is indicated for references in the text.

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.

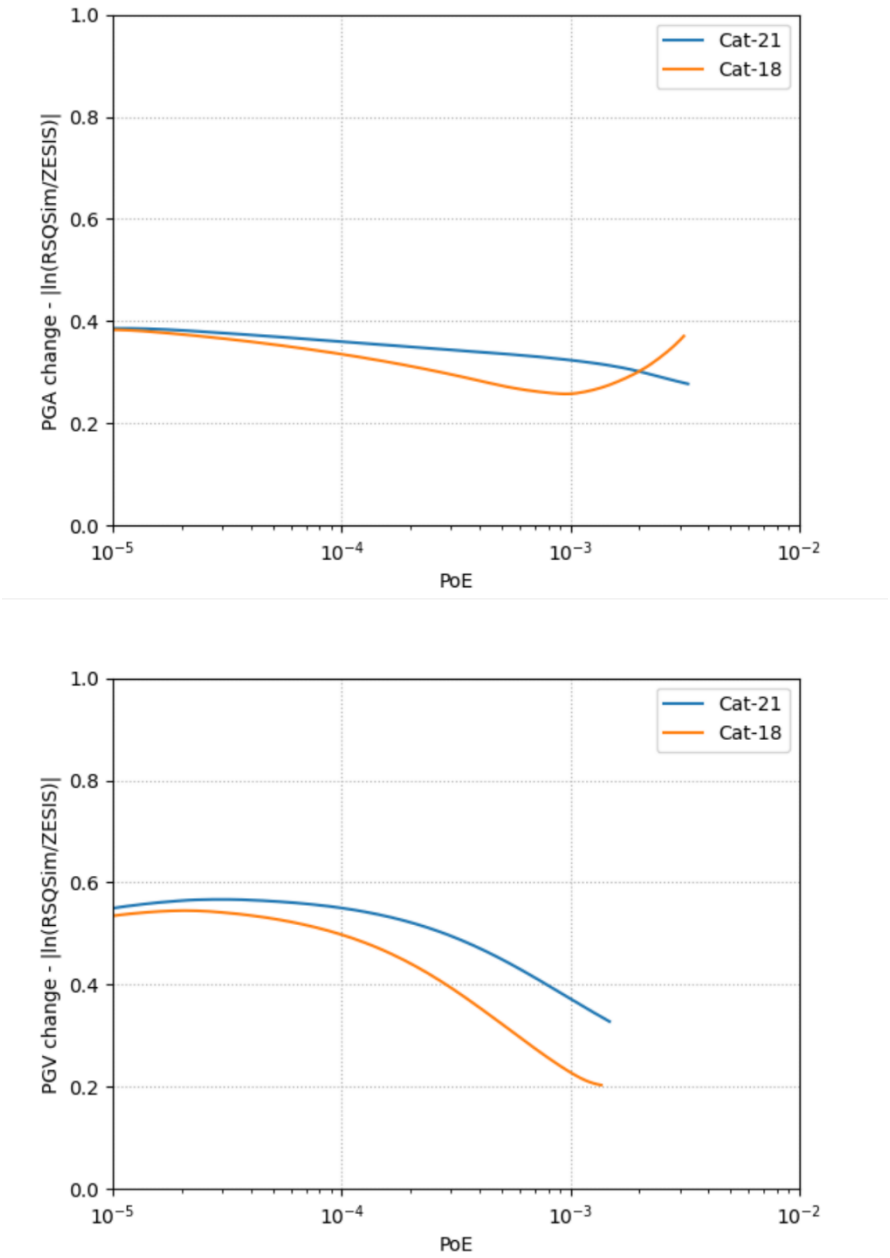

Figure 4. PGA (top) and PGV (bottom) variation of Cat-21 and Cat-18 with respect to the ZESIS area source model. Changes are computed using <|ln(RSQSim-ZESIS)|>, as in Shaw et al. (2018).

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 have discussed the points made in this comment. For one, we have added figure 4 in the supplements (figure S3) and we discussed the differences between

simulated and empirical models in the EBSZ for the short and long-term, and the implications for the region (lines 579-585 of the revised version).

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 have expanded this sentence by providing the main ideas described above (lines 556-558).

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.