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.

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 will implement the following changes to better clarify and discuss this point:

1. We will include 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, we will (qualitatively) discuss the agreement of each simulated model with the catalogue MFD.
3. In section 2.4, we will better clarify that our approach tests the joint model, and we will discuss in section 4.3.2. 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, we will better clarify 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), we will better clarify 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 will better clarify the rationale of including both models in section 2.2.2 of the revised manuscript.

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 will make sure to mention it in both the abstract and the introduction of the article.

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 will remove 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 will replace “reliable” by a more appropriate term like “adequate” or “pertinent”.

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 will fix this in the revised version of the 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 will remove this citation in the revised version of the manuscript.

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 will add this in the revised version of the 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 will better clarify this point in the revised manuscript.

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 will better specify the meaning of “close distances” by providing context in the selection rules of macroseismic records mentioned above.

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 will better contextualize this statement by specifying that PGV can be used as a close estimate to damage, but that there are other ground shaking metrics that could be employed as well.

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 will add a sentence in the manuscript explaining the origin of this feature in the figure.

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 will better justify the selection of this probability of exceedance in the revised manuscript.

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 already comment this effect in lines 368-370 of the original manuscript, but we will enhance the explanation in 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: In the revised manuscript, we will implement this suggestion and make appropriate visualization changes to figure 7.

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 will better clarify this point in the revised manuscript (section 3.3).

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 will provide a brief definition of “consistency” in the introduction and section 2.4 of the revised manuscript. This will serve as a basis to introduce all references of consistency “tests” or “checks” throughout the manuscript.

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

AR21: We will replace 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.