Reviewer Comments

For author and editor

Technical and Moderate comments:

88: Specify the Worden et al., (2012) equation as there are multiple alternatives available (e.g., with or without M/dist corrections). Why was PGA preferred? PGA has been previously suggested to be not ideal for high MMIs (i.e., MMI > 7; see Wald et al.,). Would you expect any different results if you used a different parameter other than PGA?

138: It is difficult to determine the potential impact of these assumptions on the final results. For example, a deaggregation showing that the contributions of out-of-state earthquakes is insignificant would help in making this assessment.

138: Were “sparsely” populated regions where intensity data is not available thus also discarded from the later comparisons?

149: Equation 2 and its description is confusing without further explanation. I would suggest instead that the equation is removed and a simple explanation with a disaggregation plot would be a large improvement (i.e., highlighting the contributions that would be missing if a larger Mmin is used).

159-161: A consistent mean between hazard editions does not rule out a change in disaggregation shapes. The authors should provide further arguments for this assumption (e.g., no significant changes in sources that would significantly impact the shape of the deaggregation) and/or include this as a source of potential error in the discussion.

162-171: The authors should elaborate as to the degree of approximation this process results in. The better way of doing this would have been to rerun the PSHA map with a higher Mmin. Instead, they estimate the change by reaggregating for the probability of exceedance for a different Mmin for a specific return period. In effect, they are assuming that there is a simple scalar difference between the two hazard curves; this does not have to be the case.

General comment 1: No discussion on the possible impact of site effects. Is the CHIMP dataset and the hazard map on a consistent site condition?

General comment 2: No discussion on the reliability of the MMI conversion. The conversion from Intensity to ground motions represents a significant source of epistemic uncertainty. Is there any impact to the results using different CHIMP aggregation bins? Would there be any change on the results if the MMI conversion were reversed (i.e., CHIMP MMI à PGA)?

General comment 3: a greater discussion on potential sources of error (and bias) and/or the argued accuracy of a historical intensity database such as CHIMP would help a reader in drawing conclusions on the apparent overprediction. What is the difference in uncertainty between a PSHA ground motion and a historical intensity database? What if the uncertainty in both CHIMP, the hazard and the MMI conversions is included in the assessment?

General comment 4: the comparison aggregated results over the entire region. Are their regions which are more (or less) reliable?

General comment 5: Is there any Mmin that would make the results fit within the 95% CI?

Minor and editorial comments:

48: I would suggest using annual rate of exceedance (or something similar) as “annual hazard” may be interpreted ambiguously by some.

65 – 60: A reader of this section may interpret this as all or nearly all PSHA assessments that have been tested have been shown to overpredict historical shaking in these regions. Is this statement the intent of the authors? I suggest the section is slightly reworded to highlight that these are not exhaustive or to be more specific as to the study findings.

101: as defined

110: this list is not exhaustive and should not be framed as such

111: suggestion for clarity: median ground motions as predicted by the GMMs

112: “Alternatively” instead of “Or”

113: define Mmin

115: define aleatory variability

115-116: the use of standard deviation is confusing here when referring to hazard and “its distribution”. Reducing aleatory variability has an impact on each realization of the hazard.

123: removing “the hazard” is confusing and the text should be more specific.

126: Mmin of 5 for all sources?

146: As the document does not contain much of a description on deaggregation, references would be helpful.

150: in the case of eq 2, z is the predicted PGA.

150: ensure consistent use of disaggregation and deaggregation.

152: the use of a “/” for 6/6.5 is confusing, I suggest spelling it out.

170: I don’t understand the role of MMI here. Are you not just reducing the PGA map by the % suggested from the previous analysis (i.e., no MMI involved here as we are only dealing with the PSHA map and deag)?

189: Figure 4: The downwards error should be from the original target PGA as that is the point on the hazard curve that was used to scale the results.

189 Figure 4: some of the text on the figure (and the other figures) is very small (in particular the histogram and title for the maps).

189 Figure 4: the caption is difficult to read. I suggest breaking the sentences for each of the subplots rather than a continuous sentence.

189 Figure 4: Indicate where the Sierra foothills are.

260: overpredictions of the USGS hazard maps RELATIVE to the CHIMP dataset

Reviewer Comments

For author and editor

Review of Gallahue et al **Exploring the Effect of Minimum Magnitude on California Seismic Hazard Maps**

The manuscript by Gallahue et al uses the comprehensive CHIMP historical macroseismic intensity database for California to test the seismic hazard map. In particular, the paper explores the effects of the minimum magnitude used in the seismic hazard model as one source of apparent discrepancy between the historical intensity data and shaking forecast by the hazard model. The paper concludes that while differences in minimum magnitude between the model and observational dataset go part of the way towards explaining the differences, significant discrepancies still result.

Testing of seismic hazard model outputs is an important topic of research, and the use of historical intensity data provides an important avenue for such research considering that these typically span much longer time periods than instrumental observations. Furthermore, in many cases historical intensity data may be partially or fully independent of the data used to construct the PSHA model, further enhancing their value as test. However, historical intensity data has limitations, one of which is that the historical catalogue may be incomplete for smaller magnitude events, and it is this source of uncertainty that the paper focuses on. This focus on isolating a particular cause for (or contribution to) the discrepancy is a strength of the paper.

The authors compare fractional exceedances of shaking predicted by the PSHA model with that observed in the CHIMP dataset. They argue that shaking exceeds the hazard value predicted by the model at a significantly smaller fraction of sites compared with that expected from the hazard model. They argue that differences in minimum magnitude can explain some, but not all, of the discrepancy. While the large discrepancies based on the mean hazard suggest that this may well be the case, I do not believe that the statistical methods applied in the paper fully capture model uncertainties and randomness, and hence I am not yet convinced of the statistical significance of the results. Addressing these concerns will be critical to establishing the significance of the results. The main concerns I have are:

1. All observations in the CHIMP database are treated as independent, when this is clearly not the case. There are only 62 individual earthquake events in the full database; neighbouring sites in both model and observational datasets are seeing shaking due to the same events, and therefore are not independent. However, the statistical analysis (L239-258) appears to treat the observations at each of the ~2000 sites as an independent event. Furthermore, I think that the formula on L243 only applies to independent Poisson processes, which is not the case for the grids of observations and forecast hazard. This means that the 95% confidence ranges that are presented are narrower than I think they should be. I'm not sure how best to resolve this, but unless I've missed something, I believe the calculation of uncertainties is incorrect.
2. Related to the point above, it is not clear to me how the fractional exceedance of sites relates to the Poisson model used to constrain the rate of earthquake occurrence in the PSHA model. Clearly, for a given time frame many outcomes of the PSHA model may be realised. Would the results change if, for example, a large M8+ rupture of previously unruptured segments of the San Andreas Fault occurred within the observational period? Such an event would affect many sites at the same time, and possibly create 'new' observational points as well.
3. Uncertainties in the Worden et al (2012) PGA to intensity conversion equation are ignored (L205). Again, including these uncertainties might be expected to widen the 95% uncertainty bounds on the predicted fractional intensity exceedances.
4. The completeness of the observational dataset is based on magnitude. However surely there are also completeness issues based on intensity at any particular site, in terms of when a site as settled and which observations were recoded etc. I think this may be discussed in Salditch et al (2020), but should be mentioned here too.
5. PSHA models include a large range of epistemic and aleatory uncertainties. The approach here tests against the mean hazard (i.e. the hazard map), which may be considered valid as this is typically what is input into design criteria etc, and it is also what most other PSHA testing studies that I have seen do. However, full consideration of the uncertainty in the hazard may again lessen the significance of the results.

I'm not sure if this is possible with the hazard products that accompany the California model, but regardless, it should be stated explicitly that you are using the observations to test against the mean hazard, but not the full PSHA model.

All of the above issues make me question the statistical significance of the results. The large differences in the fractional area calculated from the mean hazard give overall support to the conclusions, and of course increasing the minimum magnitude in a hazard model with decrease the forecast hazard, but I think further work to establish the significance of the results will strengthen the conclusions of the paper.

Line by Line comments

L20-21: This statement seems to imply that shaking is always overpredicted in PSHA. I'm not sure that's uniformly the case – see further comments below.

L61-70. Some additional studies that use macroseismic intensities for testing include Rey et al 2018 and Griffin et al 2019. The Griffin et al (2019) study did not find a statistically significant discrepancy between observations and model when testing a PSHA model at several cities in Indonesia, while the Rey et al (2018) study noted that the PSHA model under-predicted hazard for roughly half of France. Note that some of the studies cited, included Stirling and Gerstenberger (2010) and Mark and Schorlemmer (2016) used instrumental data, not macroseismic intensity as implied in the text. Note also that Stirling and Gerstenberger (2010) suggest that, if anything, the hazard was under-predicted in New Zealand. Therefore the discrepancy is not always one of over-predicting hazard as suggested here.

L71: It would be good to mention some of the other potential sources of the discrepancy here. As I read the paper I kept thinking of other things that could be causing the discrepancy; when I went back to look at Salditch et al (2020) I found that many of these things were already discussed here. So better to briefly mention possible causes, and then explicitly state that here you are testing the sensitivity to one possible cause.

L94: Better to rephrase as 'the hazard value' rather than 'that predicted' as the PSHA model is not predicting a shaking level, it's giving a probability of that shaking being exceeded.

L141: Please state how many events are left in the catalogue for the two completeness end-members.

L149-150: Inconsistent use of Deagg and Disagg.

L155: Why are these 'approximate'

L170: I'm confused by the wording here. It reads as if a the hazard map is converted from MMI to PGA, but it should be the other way around.

L181: short period shaking. Is there a problem basing the analysis on PGA, when other shaking frequencies may correlate better with intensity in the Worden et al (2012) conversions?

Figure 4: parts f and h should be c and e.

L239-251 present methods and should be presented before the results.

L241: Shouldn't the uncertainty refer to the predicted rather than observed exceedances here? The observed exceedances is just what it is and doesn't have uncertainty; it is one realisation of the process and not necessarily the mean.

L257-258. As noted in my main comments, I'm not convinced that the uncertainties have been calculated correctly and therefore am not yet convinced by this conclusion.

References:

Griffin, J., Nguyen, N., Cummins, P. and Cipta, A., 2019. Historical Earthquakes of the Eastern Sunda Arc: Source Mechanisms and Intensity-Based Testing of Indonesia's National Seismic Hazard Assessment. *Bulletin of the Seismological Society of America*, *109*(1), pp.43-65.

Rey, J., Beauval, C. and Douglas, J., 2018. Do French macroseismic intensity observations agree with expectations from the European Seismic Hazard Model 2013?. *Journal of Seismology*, *22*(3), pp.589-604.

Reviewer Comments

For author and editor

This is my second review of the work of Gallahue et al.,

The work seeks to address an apparent disconnect between a historical intensity database (CHIMP) and the USGS seismic hazard maps. The authors' previous work has shown that the USGS hazard maps significantly overpredict the CHIMP dataset. This work aims to investigate whether a difference in the minimum magnitude between CHIMP and the hazard maps is the cause of the discrepancy. The paper is interesting and worthy of being published.

I note that both myself and the other reviewer have indicated that while the paper calculates the statistical likelihood that these results are not due to (random) chance, it does not consider the random uncertainty in their chosen GMICE and the model uncertainty of the NSHMP. The authors' response to the reviews (but not included in the paper) asserts that these uncertainties are not believed to be significant and that uncertainties in the NSHMP are not available (I would note that some of the (implicit) epistemic uncertainty in the NSHMP in the region can be approximated from the literature; e.g., https://doi.org/10.1785/0120170338).

Ideally, the paper would include a full uncertainty analysis which demonstrates their view (and thus provides a clearer picture of the likelihood of a bias). While these additional results may represent "future work" as it is a non-trivial task, it is reasonable to expect that the paper should include a more concrete discussion (1-2 paragraphs) on these potential sources of error, their potential significance and the limitations of the approach (e.g., using only mean hazard, only considering the Worden d-m independent median GMICE, PGA-only). This is important so that a reader can frame the results in the greater context of the issues surrounding the large uncertainties in ground motion and intensity estimation.

Minor and editorial comments:

L98-95: It would be helpful to the reader to elaborate slightly on how the +- 1 unit was chosen (e.g., judgement, average of standard deviation, etc.).

L134: median ground motions?

L139: "hazard distribution" may not be the best word here as to me that implies more of a change in epistemic uncertainty.

L141: Reference so to why you believe these are most likely? It's also not appropriate to include Mmin of the maps as a source of why the maps are biased high. It's the Mmin of the CHIMP dataset that is in question. This can be solved by rewording the first sentence to focus on these three factors being possible causes of the relative difference of the two predictions (rather than on potential bias in the maps).

Figure 2: reduce median ground motion. I would also suggest retitled to: "reconciling hazard map with CHIMP data.

L151: site effects (as predicted by NGA-West2).

L152: high-frequency ground motions

L175: should this not be in the discussion? Seems premature to list the effect of a larger mmin here.

Figure 4: some of the bars seem to be floating and the visual overlap (and rendering of the shading) of the bars seems incorrect. This is a known issue in matplotlib (which is what this looks like to me). You can fix this by explicitly plotting each bar in the correct order. https://stackoverflow.com/questions/18602660/matplotlib-bar3d-clipping-problems

L236 (and elsewhere): I would recommend not using the word "hazard" in this way. Either refer to a reduction in ground motion at a specified probability level or a change in the exceedance probability.

L238 – L246: While I appreciate the intent of this paragraph it is a confusing especially the parts on variance. I would recommend simply stating that the method assumes that the sources and their relative contributions to the target exceedance rate does not appreciably change between the original and assumed target exceedance rate.

L235: There is somewhat of a spatial pattern in the map (i.e., values are closer along the western margin and in the SW). Thoughts as to why this is the case? Should be discussed.

Figure 7 caption: “mapped values even more relative to the reference mapped values”. I’m not sure what “more relative” means here.

Reviewer Comments

For author and editor

Firstly, please accept my apologies for being slow with this review.

Overall, I think that the authors have done a good job of addressing the comments from the previous reviews, and the paper is well-presente and reading very clearly. I have a couple of additional comments below that should not take long to address.

Thank you for the detailed response to my comments about the statistical analysis. I agree that the Poisson approach to testing is what many previous authors have done, and there is consistency with the PSHA paradigm, so this is a good approach. I also agree that comparing the area (or number of sites) where the hazard is exceeded is suitable within this framework. Where I am a little concerned is around adding formal confidence limits to this analysis, particularly in the way that each intensity exceedance is treated as an individual independent event to do this. I do still feel that this is stretching things a little. After considering your responses and the other analysis presented, I do not think that this affects the conclusions of the paper, so possibly I am being a little pedantic. Given this, I will leave it up to the authors to decide whether they want to reconsider whether formal confidence intervals should be presented.

A couple of other comments:

L20: I am concerned about the blanket statement that hazard maps overpredict hazard as I am still not convinced it is universal. I also worry that it could be used as a justification for ignoring current seismic hazard maps, and I think this could be dangerous. I would strongly suggest adding some nuance to this statement. The rest of the abstract reads really well.

L74-75: I don't understand the argument being presented here. These studies didn't choose to look at locations with the highest intensities or only choose the largest intensity values. True, by considering completeness, the analysis is for rarer, large shaking events, but if an incomplete record of shaking is used then the comparison with the hazard map isn't fair. For the studies considering individual sites, these sites are typically chosen because they are of general interest (e.g. significant towns and cities, or locations where there is instrumental data), not because they have high shaking values, as seems to be implied the response to the review.

Furthermore, in the review response you state '*Our study considers the maximum observed intensity at every location and does not take into consideration the completeness value of the intensities. Using the largest observations (as done in the mentioned studies) will make the map better align with the data, because it only considers the strongest, rarest, ground motion. This eliminates the locations where observed shaking was low but was expected to be high – that are the places where the discrepancy develops.*' Surely consideration of the completeness of the observed intensity values is absolutely critical – if the intensity dataset is incomplete then it should be no surprise if the hazard map is higher!

All that said, we've recently published another paper that confirms with the generalisation of hazard maps over-predicting hazard compared with intensity datasets (please don't feel obligated to cite this):

Allen, T.I., Ghasemi, H. and Griffin, J.D., 2023. Exploring Australian hazard map exceedance using an Atlas of historical ShakeMaps. *Earthquake Spectra*.

Jonathan Griffin