Response to Reviewer Comments

Nicole Paul Carmine Galasso Vitor Silva Jack Baker

September 4, 2024

Thank you to both of the reviewers and editor for constructive feedback regarding our manuscript. We agree that the original manuscript was confusing with respect to the relevance and interpretation of housing damage-based predictions of population displacement. In this revision, we've aimed to clarify the role of housing damage-based displaced estimates in rapidly determining potential long-term housing needs, as opposed to initial displacement estimates. Relevant portions of the abstract, introduction, and conclusion have been revised to clarify this. Additionally, we present additional comparisons for the 2016 Kumamoto earthquake across both space and time. Lastly, we have sought to elaborate on source and interpretation of the data sources we compare against (e.g., IDMC, mobile location data-based estimates from the literature), as some key context was previously missing.

The reviewer's specific comments have been listed in order below in italic text, followed by our responses in normal text. Relevant text snippets are included herein.

Reviewer 1

Q 1.1 *Regional damage predictions from risk models are pretty uncertain. I see the authors report comparisons of damaged buildings at the country level but not at more local levels, where displacement mechanisms occur, e.g., from one city to another. In my opinion, regional damage predictions from remote and close-range sensing techniques could help here a lot, but the authors do not explore such literature.*

Reply: Thank you for the comment. We acknowledge that country-level comparisons can obscure interesting findings about *where* people are moving to and from. To address this concern, we have added additional comparisons of damage and displacement estimates at the subnational level for the 2016 Kumamoto earthquake. The potential inclusion of remote and close-range sensing techniques for estimating damage (independent of risk models) is interesting. However, to our knowledge, there has been limited evidence that such damage proxy maps can meaningfully distinguish between varying levels of damage (e.g., Rao et al., 2023), which seems critical for estimating the population rendered homeless. If there is some specific literature that we are neglecting, please advise. That said, we do think there is good motivation to incorporate field surveys to better constrain engineering forecasts, and we have added a statement and reference to support that claim.

Please see the revised manuscript for the inserted table and below for the additional text provided.

> Since the JCO (2017) reports data at the subnational level, the damage and displacement estimates can be compared across prefectures as shown in Tab. 6. More variability across the damage and displacement estimates is evident at the prefecture-level, with the scenario model (OQ) notably overestimating impacts in Fukuoka relative to what was reported. This could be partially due to the conditioning of ground motion fields, where stations near to cities outlying the heavily-populated Fukuoka city recorded higher values of PGA than expected for the GMM employed, as can be observed in Fig. 3. More generally, the estimated damage and displacement is less concentrated in the scenario model as compared with what was reported. This may underscore the importance of incorporating on-the-ground observations to better-constrain engineering forecasts (Loos et al., 2020).

Q 1.2 *Similarly, the comparisons between displacements and damaged buildings are only given at the country level. Given that the author's damage and human data are at very fine resolutions, would it make more sense to do the analysis of those two variables at such scales, e.g., building, neighborhood, and zip code? That way, the authors can actually quantify what can be statistically explained with damaged buildings and support better their conclusions.*

Reply: Thank you for the comment. We believe our response to Q 1.1 also addresses this point, as include comparisons of damage and displacement estimates at the subnational level for 2016 Kumamoto.

Q 1.3 *The paper differentiates between traditional methods that estimate displacement based on building damage and alternative methods such as the analysis of mobile data. The rationale for exploring these alternative methods is their potential to estimate additional factors related to recovery, such as displacement duration and rate of return, which traditional methods do not address. However, the comparison in the study is confined to estimates of immediate displacement only. This focus was initially confusing, especially on first reading. If predicting factors beyond immediate displacement (e.g., displacement duration) is acknowledged as an important goal, the rationale for limiting the comparison to immediate displacement should be more clearly articulated, particularly in the abstract and introduction.*

Reply: Thank you, we agree that the original manuscript was confusing regarding the usefulness of building damage-based displacement estimates and misleadingly implied that such estimates might only be representative of immediate displacement. This is an important distinction for earthquakes, which usually estimate the population rendered homeless (due to housing destruction) versus other hazards such as hurricanes or storms, which usually estimate the population evacuated. In the case of earthquakes and for the population rendered homeless, these estimates are indicative of potential long-term housing needs or the population requiring housing assistance (Guadagno and Yonetani, 2023). We have aimed to clarify this throughout the abstract, introduction, and conclusions. Additionally, we have added an additional comparison of mobile location data-based estimates (over time) versus the scenario damage-based estimates for 2016 Kumamoto. Here, we can see that the OQ estimates seem to reasonably represent the long-term displacement rates from (Yabe et al., 2020). However, what the OQ model critically cannot inform us of is the timing of population return.

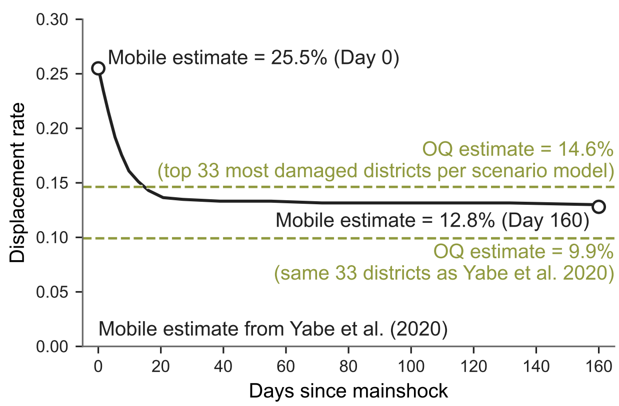


Figure. The estimated population displacement rate over time from mobile location data in Yabe et al. (2020) versus the estimated population displaced in this study (OQ) for the 2016 M_W7.0 Kumamoto earthquake in Japan.

The estimated displacement rate over time is also provided by Yabe et al. (2020) as shown in Fig. Q 1.3, which shows reasonable consistency between the scenario model and the estimated displacement rate from mobile location data at 160 days after the earthquake. This supports the assumption that housing destruction-based displacement estimates might reasonably estimate long-term housing needs. However, the scenario model provides no view on the time period of the displaced estimate.

***Q* 1.4** *When discussing the results from mobile data analysis, the discrepancies potentially caused by sampling issues are highlighted: in countries with lower mobile device ownership, the mobile device user sample may not represent the full population. This issue, however, is not addressed when mobile data use is first introduced (line 101). I recommend mentioning this limitation early in the discussion of mobile data (similar to how limitations are presented for other methods, such as school enrollment).*

Reply: Thank you, we agree that this potential limitation should be highlighted upfront. We have inserted the following sentences into our discussion of mobile location data in section 2.2.

> The validity of this proxy relies on the movements of those with mobile phones being similar to the movements of the overall disaster-affected population. Concerns regarding sample representativeness may be higher in countries where mobile phone ownership is limited or potentially correlated with factors such as housing quality, homeownership, or income level.

***Q* 1.5** *Additionally, in the section starting at line 108, the temporal limitations of mobile data in estimating displacement duration should be explored. In a recent discussion with Yabe, who authored some of the studies referenced here, it was noted that longitudinal mobile data are typically available for a limited duration (e.g., one year), after which IDs are reset, making it difficult to continue tracking displaced persons. This aspect should be addressed.*

Reply: Thank you, this is absolutely correct and confirmed by our personal correspondence with different providers of mobile location data. We have inserted two sentences to highlight this limitation: one for location intelligence firms (e.g., Spectus.ai shuffles device IDs every 12 months) and one for major tech firms (e.g., Meta Data for Good tracks data for up to 3 months max after a given disaster event)

> Additionally, data may not be available over the full disaster recovery timeline, as many location intelligence firms shuffle unique device identifiers after a set period of time (e.g., every few months, every year) to preserve data contributors' privacy.

> However, this data is often only tracked for more limited periods and may not cover the entire recovery timeline (e.g., Meta Data for Good records data for up to three months after a disaster event).

***Q* 1.6** *As mentioned earlier, while each alternative method is comprehensively described, the specific reasons why mobile data analysis is considered the most promising are not distinctly stated (line 154).*

Reply: Thank you for the comment. To address this, we have clarified why we find mobile location data to be a promising proxy for tracking population displacement and return after disasters in section 2.2 where we first introduce this data type.

> Mobile location data seems promising, as it can continuously capture how populations move across both time and space, potentially informing post-disaster planning needs in real-time and enabling data-driven retrospective studies.

Additionally, we are not currently aware of any studies that estimate population displacement using the other proxies for the considered earthquakes.

***Q* 1.7** *Regarding the benchmarks used for comparison, which include official statistics and the IDMC, more detail on how these estimates are generated would be beneficial. For instance, the IDMC reports vary, citing "Homeless" for Haiti and Nepal (tables 4 and 6) and "Sheltered" for Japan (table 5). The reasons for these differences should be explained. Additionally, if information is available, the methods by which official statistics (CDEMA for Haiti, JCO for Japan, and ICIMOD for Nepal) are produced should also be further specified.*

Reply: Thank you, this comment is synergistic with Q 2.1 from Reviewer 2. IDMC internally collects data on each of the different metrics (evacuated, sheltered, rendered homeless) as a part of their triangulation and quality assurance processes, but does not publicly report these values and their underlying sources in their Global Internal Displacement Database (GIDD). IDMC shared the internal data with us, and thus we were able to identify which metric was reported in the GIDD for each event.

To clarify this, we have added the following sentence when we introduce the IDMC's GIDD.

> While each of these metrics are recorded by IDMC to inform their internal triangulation and quality assurance processes, the specific metric reported and underlying data source are not made publicly available in the GIDD.

Where we reference the IDMC GIDD data for each event, we also have inserted further elaboration. For the 2021 Nippes earthquake in Haiti:

> For this event, the IDMC based the displaced estimate on the reported housing destruction count from CDEMA (2021) and multiplied that by an average household size of 4.08.

For the 2016 Kumamoto earthquake in Japan:

> In this case, the IDMC directly adopted the sheltered estimates from the JCO.

For the 2015 Gorkha earthquake in Nepal:

> For this event, the IDMC estimated displacement based on the number of households identified as eligible for receiving the housing reconstruction grant per Nepal's Housing Recovery and Reconstruction Platform (HRRP), multiplied by an average household size of 4.3.

Q 1.8 *In line 253, the term "Above normal level" is introduced but not clearly defined. It is presumed to refer to the outflow of people exceeding normal levels observed outside of disaster contexts, but clarification would ensure understanding.*

Reply: Thank you, we have added a sentence to clarify this. Your understanding is correct.

> This refers to the number of post-earthquake outflows in excess of the pre-earthquake outflows during the benchmark period (January 1, 2015 through April 7, 2015).

Q 1.9 *Between lines 273 and 278, the authors explain why mobile data might have sampling issues in the context of natural disaster displacement. However, I found the two sentences a little bit confusing. Initially, it was not clear to me why the second sentence opposes to the first one. I'd suggest rephrasing for better clarity, stating examples of when mobile data proved robust in terms of sampling, as opposed to the case of displacements due to natural disasters.*

Reply: Thank you, we have aimed to clarify this by adding the text **in bold** within the existing text.

> On one hand, **baseline** mobility estimates **(i.e., irrespective of disasters)** using CDRs have been found to be surprisingly robust despite these differences in mobile phone ownership (Wesolowski et al., 2013), especially as compared with smartphone-based estimates (Milusheva et al., 2021). On the other hand, these studies do not consider sudden-onset hazards and the potential role of damage in forcing movement away from habitual dwelling units (i.e., disaster displacement).

Q 1.10 *In lines 300 - 301, it would be beneficial to explicitly state why existing Ground Motion Models (GMMs) may be inapplicable to the case studies under consideration.*

Reply: Thank you for this comment, we have added the following text **in bold** to clarify this.

> These discrepancies could be exacerbated by the inapplicability of existing GMMs for this scenario: **this was a continent-continent subduction zone earthquake, whereas existing subduction GMMs are primarily derived from data in ocean-continent or ocean-ocean subduction zones (Rajaure et al., 2017).**

Reviewer 2

Q 2.1 *A key metric for benchmarking is the number of displaced residents from IDMC. While the methods for estimating the populations that are evacuated or sheltered are defined, there isn't a description for 'population rendered homeless'. The method(s) for estimating this value, based on housing damage, could have a significant influence on the tally. For example, is it based on engineering assessments, satellite imagery, or citizen science? Especially in the case of Haiti, the political situation there made information gathering difficult. Some detail on the method the IDMC used to reach their tallies, and perhaps a brief discussion on how that might influence the benchmarking, would provide some helpful information when considering the benchmarking results.*

Reply: Thank you, this comment is synergistic with Q 1.7 from Reviewer 1. In general, IDMC prioritizes data from government agencies, UN organizations, or local authorities. We have added a sentence to clarify this where we first introduce the metric.

> The IDMC typically estimates this value by multiplying the reported housing destruction counts from government agencies, UN organizations, or local authorities by the average household size.

Please also see our response to Q 1.7, as it elaborates on the details of the IDMC estimate for each event.

Q 2.2 *The assumption that 'extensive' or 'complete' damage states render a dwelling uninhabitable are reasonable. Is there a definition of 'moderate' damage that could be added? As you've noted, in many cases, 'moderate' is probably habitable (structurally at least), although in some situations it might not be (e.g. multi-storey buildings with stricter requirements, such as requiring water/power to be functional for fire safety). Adding the definition for 'moderate' could help clarify whether this consideration is likely to influence the benchmarking significantly, or not.*

Reply: Thank you for this comment. We have made three additions into section 3.3 to elaborate on the OQ damage state definitions and to highlight potential nuances between re-occupancy standards for different dwelling types. Please see the quotes below for these additions.

> The structural fragility models are defined for each building class within the exposure model for four discrete damage states: slight, moderate, extensive, and complete damage. These four damage states roughly correspond to the five damage states in the European macroseismic scale (EMS-98; Grünthal, 1998), except that the complete damage state in OQ encompasses both the fourth damage state and the fifth damage state in EMS-98 (i.e., heavy structural damage with very heavy nonstructural damage and very heavy structural damage).

> That is, dwellings in the extensive damage state (i.e., moderate structural damage and heavy non-structural damage) or complete damage state (i.e., heavy structural damage or beyond) are assumed to be "uninhabitable," thereby displacing their occupants.

> Further, different dwelling types could have stricter requirements for re-occupancy, such as requiring water and power availability for fire-safety in multi-storey apartment buildings.

Q 2.3 *The quality of the inputs for the building data, and whether the fragility functions are appropriate, could have a significant influence on the modelling results. I wonder how much this influences things - and whether* *building data/function uncertainty might be obscuring any potential insights into the question of whether modelling displaced population numbers driven by damage alone is sufficient, given other factors could be important (as they were in the Canterbury earthquake sequence). This is really a thought for any future work - reducing these sources of uncertainty, where possible, could help shed some light on this question.*

Reply: This is true, thank you for the comment. We have elaborated on this detail in our discussion of the limitations and challenges regarding validation of the risk results. Please see the inserted quote below.

> Additionally, scenario models require several assumptions across the rupture characterization, ground motion model selection, building and population exposure derivation, and fragility function assignment. Each of these model inputs influences the resulting risk estimates, and this epistemic uncertainty complicates comparisons.

Regarding the role of drivers beyond housing damage, we agree this is important. We emphasize this throughout the text, such as in the abstract:

> Furthermore, purely basing displacement estimates on housing damage offers no view on how the displaced population counts vary with time as compared to more comprehensive models that include other factors influencing population return or alternative approaches, such as using mobile location data.

Q 2.4 *Line 255: I think this should be referencing Fig. 3, not Fig. 4.*

Reply: Thank you, we have re-confirmed all references/labels for the figures and tables.

References

CDEMA. Haiti Earthquake: Final Situation Report #12. Technical report, St. Michael, Barbados, Sept. 2021. `https://www.cdema.org/images/2021/09/FINAL_CDEMA_Situation_Report_12_Haiti_Earthquake_14Sep2021.pdf`.

Grünthal, G. European macroseismic scale 1998 (EMS-98). 1998.

Guadagno, L. and Yonetani, M. Displacement risk: Unpacking a problematic concept for disaster risk reduction. *International Migration*, 61(5):13–28, 2023. doi:10.1111/imig.13004.

JCO. Disaster Report for 2016 Kumamoto earthquake. Technical report, Apr. 2017. `https://www.bousai.go.jp/updates/h280414jishin/pdf/h280414jishin_39.pdf`.

Loos, S., Lallemant, D., Baker, J., McCaughey, J., Yun, S.-H., Budhathoki, N., Khan, F., and Singh, R. G-DIF: A geospatial data integration framework to rapidly estimate post-earthquake damage. *Earthquake Spectra*, 36(4):1695–1718, Nov. 2020. doi:10.1177/8755293020926190.

Milusheva, S., Bjorkegren, D., and Viotti, L. Assessing Bias in Smartphone Mobility Estimates in Low Income Countries. In *ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS)*, pages 364–378, Virtual Event Australia, June 2021. ACM. doi:10.1145/3460112.3471968.

Rajaure, S., Asimaki, D., Thompson, E. M., Hough, S., Martin, S., Ampuero, J. P., Dhital, M. R., Inbal, A., Takai, N., Shigefuji, M., Bijukchhen, S., Ichiyanagi, M., Sasatani, T., and Paudel, L. Characterizing the Kathmandu Valley sediment response through strong motion recordings of the 2015 Gorkha earthquake sequence. *Tectonophysics*, 714-715:146–157, Sept. 2017. doi:10.1016/j.tecto.2016.09.030.

Rao, A., Jung, J., Silva, V., Molinario, G., and Yun, S.-H. Earthquake building damage detection based on synthetic-aperture-radar imagery and machine learning. *Natural Hazards and Earth System Sciences*, 23 (2):789–807, Feb. 2023. doi:10.5194/nhess-23-789-2023.

Wesolowski, A., Eagle, N., Noor, A. M., Snow, R. W., and Buckee, C. O. The impact of biases in mobile phone ownership on estimates of human mobility. *Journal of The Royal Society Interface*, 10(81):20120986, Apr. 2013. doi:10.1098/rsif.2012.0986.

Yabe, T., Tsubouchi, K., Fujiwara, N., Sekimoto, Y., and Ukkusuri, S. V. Understanding post-disaster population recovery patterns. *Journal of The Royal Society Interface*, 17(163):20190532, Feb. 2020. doi:10.1098/rsif.2019.0532.

Response to Reviewer Comments, Rev1

Nicole Paul Carmine Galasso Vitor Silva Jack Baker

October 11, 2024

We once again thank the reviewers for their time and effort in reviewing the manuscript. We have made some minor revisions based on this feedback to better clarify the model results and our conclusions. The reviewer's general comments are included below in italic text, followed by our responses in normal text. Relevant text snippets are included herein.

Reviewer 1

Thank you for taking the time to address my previous comments. The authors have made some progress in responding to most of the points raised. However, I still have reservations regarding the validity of the central hypothesis, specifically the assertion that scenario models can reliably estimate population displacement (as stated in the abstract). While I acknowledge the data limitations the authors face, and I find several interesting insights within this study, I remain unconvinced on this particular point. As I will not be able to review the paper again, I would like to offer some suggestions that could further highlight the contributions of this work and better situate it within the broader context.

Reply: Thank you for the time spent reviewing our responses and revisions, and for these suggestions.

We understand that a reader may have some reservations about the reliability of scenario models for estimating population displacement. Although we do believe the study results demonstrate that scenario models *can* reasonably estimate potential long-term housing needs, we also understand that there are many nuances. In particular, we raise concerns about the large uncertainty range and the lack of information on the time-component of displacement.

We highlight this nuanced take upfront in the abstract:

> The results highlight the promise of scenario models to realistically estimate population displacement and potential long-term housing needs after earthquakes, but also highlight a large range of uncertainty in the predicted values. Furthermore, purely basing displacement estimates on housing damage offers no view on how the displaced population counts vary with time as compared to more comprehensive models that include other factors influencing population return or alternative approaches, such as using mobile location data.

And we re-emphasize this in the conclusions:

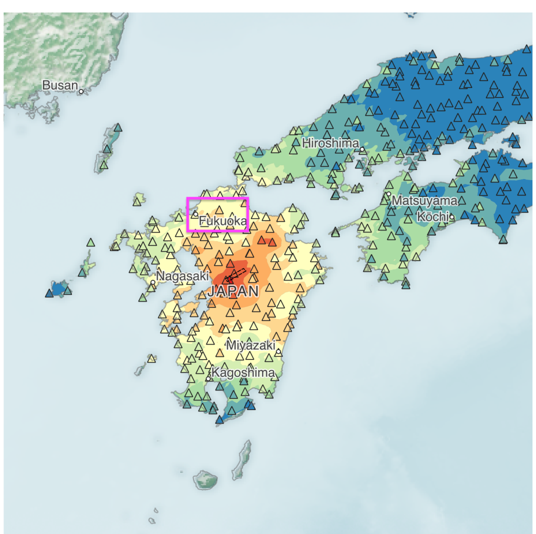
> The scenario model estimates are largely consistent with what was officially reported for these earthquake events, albeit with a large range of uncertainty[...] Additionally, scenario models require several assumptions across the rupture characterization, ground motion model selection, building and population exposure derivation, and fragility function assignment. Each of these model inputs influences the resulting risk estimates, and this epistemic uncertainty complicates comparisons.

That said, we do understand that perhaps there was not enough emphasis or discussion around the various sources of uncertainty, which you also raise again later. We have added a sentence in our scenario analysis methodology section to comment on the role of uncertainty (especially in the hazard component) on risk model predictions. See added text **in bold**.

The scenario analyses are performed using the OpenQuake Engine (OQ), an open-source seismic hazard and risk analysis software (Silva et al., 2014). To leverage observed data from recording stations, the scenario calculator within OQ has been extended to condition ground motion fields using data from seismic stations following the procedure proposed in Appendix B by Engler et al. (2022). **Uncertainties in the hazard component (i.e., source model and GMMs) have previously been identified to dominate the uncertainty in regional risk predictions (Kalakonas et al., 2020) and there is evidence that incorporating observational data from recording stations improves the accuracy and precision of scenario loss estimates (Silva and Horspool, 2019).**

Q1.1 and 1.2. A deeper discussion of the uncertainty of regional risk models could help. The authors discuss in a brief paragraph the uncertainties in their predictions. In the specific region of Fukuoka, they state that the reasons for disparities between the predictions and the reports were that ground motions might have been higher than in their models (actually, I was confused about the logic because then I would expect that predictions would be smaller instead of higher instead of the opposite). However, the authors did not discuss the implications of these sources of uncertainty for their main hypothesis.

Reply: Thank you for your feedback, we see that our description of the results for Fukuoka city could be better-clarified regarding disparities between ground motion predictions (from GMMs) versus observations (from seismic stations). The employed GMM typically estimates lower shaking intensities at that source-to-site distance; however, our ground motion fields (GMFs) are conditioned on available observations from seismic stations per Engler et al. (2022) Appendix B (see also: the OQ docs). Using the recordings at each seismic station, OQ corrects the inter-event term (bias) for all sites and reduces the intra-event term inversely proportional to the distance from each site and the neighbouring stations (using a cross-spatial correlation model). Once at a far enough distance, the total variability will match the one provided by the GMM. However, realizations of ground shaking at seismic station sites will have median values identical to what was recorded/observed, while neighboring sites will have similar median values to the recording and lower total variability than the GMM. In the case of Fukuoka city, there is a nearby seismic station that recorded a higher PGA than the selected GMM would suggest. Therefore, at that site and many of its neighboring sites, our scenario model predicts *higher* ground shaking values than the selected GMM would have otherwise. If we zoom into Fig. 3, we can see this more clearly: although the ground shaking generally attenuates with distance, there are some "islands" of isolated higher shaking near to station observations of that shaking. This results in higher damage counts near Fukuoka city than would have been predicted without ground motion conditioning. Conditioning GMFs on observations seems reasonable and reduces a major source of uncertainty, but there could also be cases where highly localized site conditions or equipment malfunction leads to issues.

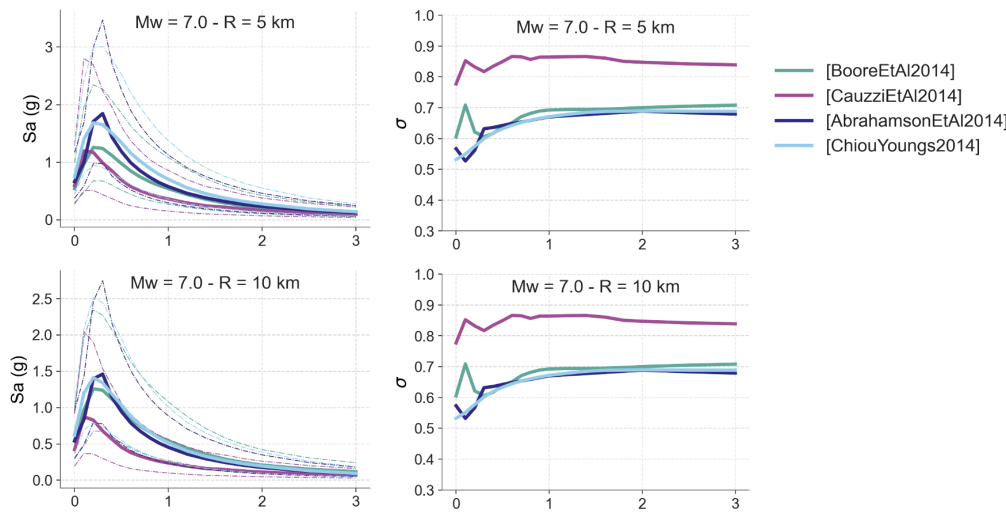


We've aimed to clarify this with the added text **in bold**.

> More variability across the estimated damage and displacement is evident at the prefecture-level, with the scenario model (OQ) notably overestimating impacts in Fukuoka relative to what was reported. This could be partially due to the conditioning of ground motion fields, where stations near to cities outlying the heavily-populated Fukuoka city recorded higher values of PGA than expected for the GMM employed. **Since the ground motion fields are conditioned on the recording station data, the model correspondingly adjusts the inter-event term (bias) in the GMM for all sites and reduces the intra-event term at other sites inversely proportional to distance (using a cross-spatial correlation model). This** can be observed in Fig. 3**: while the estimated ground shaking mostly attenuates with distance from the rupture, there are some "islands" of relatively higher (or lower) ground shaking near to station observations. In the case of Fukuoka city, the station recorded relatively higher shaking, leading to higher damage counts than would have predicted without conditioning. While conditioning the ground motion fields on available observational data is appealing, there could also be situations where highly localized site conditions or station malfunction yield unrealistic predictions at the station site and neighboring sites (due to cross-spatial correlation). A higher density of stations, especially near heavily populated areas, could help mitigate this issue.** More generally, the estimated damage and displacement is less concentrated in the scenario model as compared with what was reported. This may underscore the importance of incorporating on-the-ground observations to better-constrain engineering forecasts (Loos et al., 2020).

In fact, I think that we know more about ground motions than our buildings, so in my view, exposure uncertainty can easily dominate these risk predictions instead of ground motions. In one sentence, they argue that adding ground observations can help, which is central to the validity of their hypothesis. But how many observations do we need for the hypothesis to work? How do we "constrain" the risk models with the observations? Can we use the model without observations? The lack of details leaves me wondering if the risk models are actually needed in case there are enough ground truth observations.

Reply: First, there is the role of various source of uncertainty and their contribution to risk model predictions. There are sources of epistemic uncertainty: rupture characterization, ground motion model selection, site conditions, exposure characterization, and fragility model definition. These epistemic uncertainties are due to incomplete scientific knowledge, which in principle, can be reduced. Kalakonas et al. (2020) investigated the impact of these epistemic uncertainties on regional risk model predictions, highlighting that factors in the hazard component (source model and GMMs) dominated the uncertainty, and not the factors in the exposure component. Then, there is the aleatory variability (i.e., inherent randomness), which cannot be easily reduced. The GMMs are a significant source of this uncertainty, especially for these scenarios that feature a large magnitude with short source-to-site distances (see figure below). Potential GMMs for this event have σ in the range of 0.6 to 0.8 (right), which results in a significant variability of ground shaking intensity predictions as can be seen in the response spectra (left) that indicates the $\pm 1\sigma$ in dashed lines. This range of spectral acceleration values would have a dramatic impact on the damage predicted, regardless of whether a given building was incorrectly assigned as unreinforced masonry versus code-compliant reinforced concrete (a type of misclassification that should be less common than a misclassification between more similar building types).



Second, there is the question of how many (if any) observations are necessary for scenario models to meaningfully predict population displacement. On the ground motion side, we illustrate this by considering three case studies: Japan has a relatively high coverage of seismic stations (although not necessarily a high density near some populated areas), whereas Haiti and Nepal have practically no seismic stations. For Haiti, it seems that the scenario model predictions were reasonable despite the lack of observational data. For Nepal, this was not the case; however, a higher density of stations could have potentially partially compensated for any limitations/inappropriateness within the selected GMM.

Third, there is a question of how to "constrain" risk models with observational data. While we cite our approach for conditioning based on seismic stations (Engler et al., 2022), we do not take any approach to incorporate observed damage data. This is outside the scope of this paper, but could potentially be explored by future researchers.

Lastly, there is the question of whether we need risk models at all if there are enough observations. On one hand, we note that risk models are not solely useful for near-real-time predictions, but also for hypothetical "what if?" scenarios or calculating probabilistic losses (e.g., average annual losses, loss exceedance curves). On the other hand, for a near-real-time prediction, high-fidelity damage data (e.g., field surveys) takes time and rarely has full coverage over the affected area. As more comprehensive on-the-ground data becomes available, it could indeed replace the need for a risk estimate. However, there is potential value in having more timely estimates, even if they carry considerable uncertainty.

Also, it looks like the authors did not think remote sensing data would be useful, but maybe it could help you predict collapsed buildings. Based on that, you can create an approach to updating your risk models.

Reply: Thank you for the comment. We agree that predictions of collapsed buildings would be generally useful for near real-time risk predictions, especially of fatalities or injuries. However, there is evidence that population displacement is triggered at a lower threshold of damage than collapse, and remote sensing-derived damage predictions seem limited in this case (Rao et al., 2023). Nonetheless, we recognize that advancements could be made in the future and would be interested to see what those improvements or new approaches are. We have included an extra sentence in the conclusions to note this potential; please see added text **in bold**.

> Additionally, scenario models require several assumptions across the rupture characterization, ground motion model selection, building and population exposure derivation, and fragility function assignment. Each of these model inputs influences the resulting risk estimates, and this epistemic uncertainty complicates comparisons. **Various observational data could be used to better-constrain model predictions and reduce uncertainty: while this study only incorporated recorded**

ground shaking from seismic stations, other relevant sources of observational data such as field surveys or remote sensing-derived damage data, could also potentially be incorporated (Loos et al., 2023).

Reviewer 2

Thank you to the authors for the robust response and amendments addressing the comments from both reviewers. The amendments provide helpful clarity, and confidence that the authors are aware of the uncertainties within the benchmarking. This is a valuable study representing the type of benchmarking that provides very useful insights into model applicability.

Reply: Thank you for taking the time to review our responses and revisions. We appreciate the positive feedback.

References

Engler, D. T., Worden, C. B., Thompson, E. M., and Jaiswal, K. S. Partitioning Ground Motion Uncertainty When Conditioned on Station Data. *Bulletin of the Seismological Society of America*, 112(2):1060–1079, Jan. 2022. doi:10.1785/0120210177.

Kalakonas, P., Silva, V., Mouyiannou, A., and Rao, A. Exploring the impact of epistemic uncertainty on a regional probabilistic seismic risk assessment model. *Natural Hazards*, 104(1):997–1020, 2020.

Loos, S., Lallemant, D., Baker, J., McCaughey, J., Yun, S.-H., Budhathoki, N., Khan, F., and Singh, R. G-DIF: A geospatial data integration framework to rapidly estimate post-earthquake damage. *Earthquake Spectra*, 36(4):1695–1718, Nov. 2020. doi:10.1177/8755293020926190.

Loos, S., Lallemant, D., Khan, F., McCaughey, J. W., Banick, R., Budhathoki, N., and Baker, J. W. A data-driven approach to rapidly estimate recovery potential to go beyond building damage after disasters. *Communications Earth & Environment*, 4(1):1–12, Feb. 2023. doi:10.1038/s43247-023-00699-4.

Rao, A., Jung, J., Silva, V., Molinario, G., and Yun, S.-H. Earthquake building damage detection based on synthetic-aperture-radar imagery and machine learning. *Natural Hazards and Earth System Sciences*, 23 (2):789–807, Feb. 2023. doi:10.5194/nhess-23-789-2023.

Silva, V. and Horspool, N. Combining USGS ShakeMaps and the OpenQuake-engine for damage and loss assessment. *Earthquake Engineering & Structural Dynamics*, 48(6):634–652, 2019. doi:10.1002/eqe.3154.

Silva, V., Crowley, H., Pagani, M., Monelli, D., and Pinho, R. Development of the OpenQuake engine, the Global Earthquake Model's open-source software for seismic risk assessment. *Natural Hazards*, 72: 1409–1427, 2014.