Ranking and Selection of Earthquake Ground-Motion Models Using the Stochastic Area Metric.

Jaleena Sunny *[1] [2], Marco De Angelis[2], and Benjamin Edwards[1]

[1]School of Environmental Sciences, University of Liverpool, UK

[2]Institute for Risk and Uncertainty, University of Liverpool, UK

November 11, 2021

Declaration of Competing Interests - The authors acknowledge there are no conflicts of interest recorded.

*Corresponding Author

Jane Herdman Building

School of Environmental Sciences

University of Liverpool

Liverpool

L69 3GP

United Kingdom

jaleena.sunny@liverpool.ac.uk

Abstract

We introduce the cumulative-distribution-based *area metric* (AM)–a.k.a. stochastic area metric–as a scoring metric for earthquake ground-motion models. The AM quantitatively informs the user of the degree to which observed or test data fit with a given model, providing a rankable absolute measure of misfit. The AM considers underlying data distributions and model uncertainties without any assumption of form. We apply this metric, along with existing testing methods, to four ground-motion models in order to test their performance using earthquake ground motion data from the Preston New Road (UK) induced seismicity sequences in 2018 and 2019. An advantage of the proposed approach is its applicability to sparse datasets. We therefore focus on the ranking of models for discrete ranges of magnitude and distance, some of which have few data points. The variable performance of models in different ranges of the data reveals the importance of considering alternative models. We extend the ranking of ground-motion models (GMMs) through analysis of inter-model variations of the candidate models over different ranges of magnitude and distance using the AM. We find the inter-model AM can be a useful tool for selection of models for the logic tree framework in seismic hazard analysis. Overall, the AM is shown to be efficient and robust in the process of selection and ranking of GMMs for various applications, particularly for sparse and small-sized datasets.

Introduction

Ground-motion models (GMMs), which are used for predicting intensity measures (IM), such as peak ground acceleration (PGA), velocity or displacement, have been extensively studied and developed in recent years (Douglas et al., 2013). This is due to the fact that GMMs have significant influence on the modelling of seismic hazard and risk: obtaining more accurate predictions from GMMs results in more accurate seismic hazard and risk estimates, particularly at long return periods. GMMs describe the ground motion field in terms of a particular IM for given characteristics of earthquake source (e.g., magnitude, fault mech-

anism), wave propagation (e.g., epicentral distance), and site effects (e.g. site class or V_{S30}). Due to the complexity of the earthquake process, wave propagation and site effects, tens or even hundreds of candidate models are available for various tectonic regions; examples can be found in Douglas (2020). Assessing the predictive capability of a GMM and the selection of the most appropriate GMMs for a given application from this growing suite of predictive models, therefore poses many challenges. In addition, the ranking of models for particular applications, which involves consideration of data from limited ranges of magnitude and distance (especially in low seismicity regions) are error-prone because of the smaller sample size. Appropriate selection and subsequent ranking of GMMs is considered to be a critical step in the development of hazard and risk models because of the dependency of predicted spectra on the chosen GMMs (Stewart et al., 2015). It is, therefore, very important that the metrics used in the ranking and validation of models should be able to perform well, are mathematically justified and transparent.

Various statistical and probabilistic methods have been introduced to make the selection of GMMs more robust. Some tests, such as the chi-square and Kolmogorov-Smirnov (K-S) tests, analyse the shape and distribution of model misfit residuals, while others, such as the Pearson correlation coefficient and Chi Square Misfit (CHISQ-MF), use direct observations for the selection of models. Recently developed methods include the likelihood (LH) value test (Scherbaum et al., 2004) which utilises the the assumption of log-normal distribution for each GMM and calculates the probabilities of residuals; the use of information theory in the log-likelihood (LLH) test, which uses probability of the data under a model to determine how likely the model is for the given data (Scherbaum et al., 2009; Delavaud et al., 2009); a multivariate logarithmic score (Mak et al., 2017), in contrast to univariate measures (such as LLH), which considers the correlation and score variability in the ranking process; and a Euclidean distance method (Kale and Akkar, 2013). There are several studies—e.g., (Delavaud et al., 2012b; Stafford et al., 2008; Delavaud et al., 2012a)—which have implemented the above methods for the ranking of models. These have shown the methods to be

useful quantitative metrics for ranking GMMs. However, there are significant limitations to be considered. For example, the K-S test is more sensitive to the centre of the distribution than at the tails, whilst the chi-square test has a tendency to obtaining biased results when the sample size is low. Furthermore, LLH may sometimes provide a good fit for models with wider distributions (higher sigma) and the LH test requires subjective decisions for ranking. The case of Euclidean-distance-based ranking favors a smaller modelled uncertainty when two predictions give the same mean (Mak et al., 2014).

A direct and efficient method to compare observed IMs to modelled intensities and for ranking of models, even in low seismicity regions or for sparse datasets, is discussed in this article. We introduce a cumulative distribution based *area metric* (AM), which is developed by considering predicted data as a probability distribution. The AM has been adopted for model validation (Ferson et al., 2009), and for model calibration (Gray et al., 2022). The AM can be seen as the extension of the Minkowski distance to probabilistic metric spaces, and often goes by the name of Wasserstein distance for continuous random variables and for the case of order 1 (De Angelis and Gray, 2021). The proposed metric neither involves the calculation of residuals, nor testing the residual distribution shape. A comparison against alternative metrics currently used will be given in the discussion section, whilst highlighting the benefits of adopting such metric. Generally, a qualitative comparison can be derived from analysing the shape of the residuals while a robust quantitative comparison is essential for the ranking of models. We quantify the degree to which observed reference data fits the predicted data from different models. The use of the best GMM from alternative models for applications that involve a specific range of data can sometimes be misleading. The model which gives the best fit to the entire dataset may not be the one with the best performance over the range of data relevant to the application. Therefore, we have ranked models in different ranges of magnitude and distance in order to select the best model for the given scenario of earthquakes. We also used the AM to analyse the inter-model variations of the candidate models using the complete dataset and also for different ranges of magnitude and

distance.

One of the most commonly used tools in capturing the epistemic uncertainty associated with the GMMs in hazard analysis, is the logic tree (Kulkarni et al., 1984; Bommer and Scherbaum, 2008). However, with the increasing number of GMMs, the selection and consistency of weight assignments in logic trees for seismic hazard analysis has become a cumbersome process. Comparing models against one another in order to find similar and dissimilar behaviour can help in appropriately representing the epistemic uncertainty in final hazard estimates. This has been performed in recent studies using the Sammon's mapping technique (Scherbaum et al., 2010), which uses the average Euclidean distances between the model representation vectors to find the proximity between ground-motion models. In this work, we extend the use of the AM as a tool not only for ranking, but also for inter-model comparison. The AM is shown to be helpful for computing the weights on logic trees and thus appropriately representing the total epistemic uncertainty of GMMs in seismic hazard analysis.

The dataset from the hydrocarbon site at Preston New Road (PNR), in Blackpool (UK), is considered in this study (Clarke et al., 2019). The GMMs used for the ranking are Edwards et al. (2021), Atkinson (2015), Douglas et al. (2013) and Rietbrock et al. (2013) owing to their suitability for predicting IMs at the PNR site. The following sections first introduce the fundamental concepts of our proposed ranking method and follow with more details on what are the advantages and disadvantages of using such metric. Application of the proposed method to data from the PNR site is explained towards the end of the paper. Existing methods, specifically the CHISQ-MF and LLH tests, have been also used alongside the AM when ranking the GMMs. Overall, the AM is shown to be convenient and robust in the process of selection and ranking of GMMs for various applications.

The Proposed Ranking Metric

We propose the AM as a measure of mismatch between the marginal distribution of the data and the marginal distribution of the model. This is possible under the working

assumption that the data can be treated as samples drawn at random (*iid*) from an unknown stationary probability distribution.This assumption permits the construction of an *empirical cumulative distribution function* (ECDF), and the assignment of equal probability mass to each datum. It is important to note that this ECDF makes no assumption on the distribution type (e.g. Gaussian, Lognormal, etc.), but it requires the data distribution to be stationary.

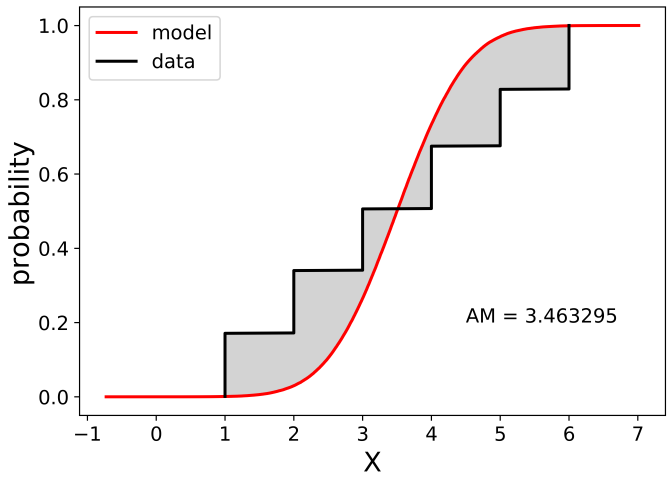

Figure 1: *Graphical representation which shows the CDF (red) and the ECDF (black). The shaded area gives the AM which quantifies the fit of the model distribution (red) and an empirical data set (black)*

The AM need not involve any kind of residual analysis, instead it directly makes use of the observed and modelled IMs. We consider the model distribution as a cumulative distribution function (CDF) and the data distribution as an ECDF (Figure 1). The CDF is the probability measure assigned to the event that the random variable X takes a value less than or equal to x:

$$M(x) = P(X \leq x). \tag{1}$$

The ECDF, just like the CDF, is the probability measure assigned to the event that the discrete random variable $X_i,\ i = 1, ..., n$ takes a value less than or equal to x, however this time the random variable is discrete instead of continuous. The ECDF can be represented as a non-decreasing step function that jumps n times in steps of height $1/n$, at each data

point, where n is the size of the dataset. The ECDF for the data X_i, $i = 1, ..., n$ is given by:

$$D_n(x) = \frac{\sum_{i=1}^{n} I(X_i, x)}{n}, \tag{2}$$

where, $I(X_i, x) = 0$, for $X_i < x$ and $I(X_i, x) = 1$, for $X_i \geq x$, is the indicator random variable. This distribution preserves all the statistical information of the dataset, therefore can be used as a proxy for the dataset.

The AM is the area between M and D_n, or equivalently the integral of the absolute difference between the CDF values:

$$d(M, D_n) = \int_{-\infty}^{\infty} |M(x) - D_n(x)| dx. \tag{3}$$

Without loss of generality, we can compute the integral of (3) horizontally by means of the *generalised inverses*, M^{-1}, and D_n^{-1}:

$$d(M, D_n) = \int_{0}^{1} |M^{-1}(u) - D_n^{-1}(v)| du, \tag{4}$$

where $u = M(x)$, $v = D_n(x)$, and so $x = M^{-1}(u)$, and $x = D_n^{-1}(v)$, thus avoiding slow numerical quadrature integration (De Angelis and Sunny, 2021; De Angelis and Gray, 2021). The smaller the distance $d(M, D_n)$ the better is the match between the model and the data.

Advantages and Disadvantages of the Proposed Metric

There are several advantages for AM and one of them is that it can be computed for very small data sets, or even a single data point, in which case the $D_n(x)$ function of equation (3) would be the unit step function at that value. This is very important when analyzing the predictive capability of the models for high consequence large magnitude earthquakes, where we have very limited data available. Secondly, this metric gives full consideration to the differences in the whole distribution. This implies that a distributional comparison should not just be sensitive to the differences in mean or variances but it should be able

to consider the whole statistical distribution. The next advantage is that it expresses the mismatch in physical units rather than in arbitrary statistical units. In the case of GMMs, we have the observed and the predicted data in logarithms of units of ground motion, such as cm/s^2 for peak ground acceleration (PGA): the AM is therefore also represented by the same units. Another feature is that the AM is unbounded. If the prediction is far from the observation, the AM shows the full extent of this difference, rather than being limited to a particular range. Finally we note that the AM obeys all four axioms of a mathematical metric, i.e, it is (i) symmetric, (ii) non-negative, (iii) follows triangular inequality, and (iv) zero between two identical entries (Ferson et al., 2009; Gray et al., 2022; De Angelis and Gray, 2021).

The explicit and appropriate consideration of uncertainties is a required property for the metric used in ranking GMMs. The uncertainty in the modelled data can come from inherent randomness (aleatory) and/or from lack of knowledge (epistemic). In case of GMMs, the ground motion is described in terms of a median and a logarithmic standard deviation, sigma (σ) [e.g., Strasser et al. (2009)] as shown in equation (4),

$$log(Y_{obs}) = log(Y_{pred}) + N(0, \sigma) \tag{5}$$

where Y_{obs} is the observed data and Y_{pred} is the median prediction. The sigma comes from the assumption of the normal distribution of ground motion residuals (difference between the observed and the predicted ground motions) and it defines the scatter associated with the ground motion prediction. Here, $M(x)$ will also include the aleatory uncertainty (inherent randomness), which is modelled as a normal distribution of mean zero and the standard deviation σ.

In this proposed methodology, the modelled IM for a single data point (that consists of a specific magnitude and distance) is assumed to take a set of values that is computed from a pre-established sigma value (σ, standard deviation) of the considered GMM. Hence

a smooth distribution of modelled IMs is obtained for each scenario (M, R, etc.) due to the consideration of aleatory uncertainty . One of the major assumptions in Probabilistic Seismic Hazard Analysis is that the ground motion residuals are log-normally distributed. Some tests, such as the LLH test, rely on this assumption. However, the assumption of log-normally distributed residuals has become a de-facto standard and, as a result, usually is not tested routinely for new datasets, but is accepted as a given (Pavlenko, 2015; Raschke, 2013). We were able to exclude this assumption, as the AM does not involve any kind of residual calculations or assumption of data distribution. We have considered the IM distribution as a CDF without any kind of underlying assumption on distribution. In the case of the AM, the steepness of the ECDF and CDF quantifies the aleatory uncertainty. As the model aleatory uncertainty increases (the sigma), we obtain a wider CDF and thus account for the aleatory uncertainty in ranking of GMMs. Models that do not reflect the aleatory variability in the tested dataset (whether too small or too large), are penalised.

The AM's accuracy is sensitive to sample size, i.e., having more data will allow us to assess the model more confidently. But we should understand that only the evidence for the apparent fit between the model and the data will increase with the increase in sample size and not the accuracy of the model (Ferson et al., 2009). This limitation leads to the fact that while we can calculate the area metric for different ranges of magnitude and distance to asses the best performing model in the given range, we cannot obtain a conclusion about the best performing range of a single model. A further limitation is that the AM tends to be less sensitive to the tails of the distribution. Finally, the AM also depends on the scale in which the distributions are represented, although this is somewhat alleviated in the case of GMMs through the use of logarithmic values.

Application of Area Metric to the PNR Dataset

The proposed metric is used to analyse the performance of four different models on the hydrocarbon site in the north of England at Preston New Road (PNR), Blackpool (Clarke et al., 2019). We also used the existing methods such as LLH and CHISQ-MF test along

with the AM to analyse and compare the ranking results from each test and the AM.

Data and GMMs Used

Hydraulic fracturing at the PNR site for shale gas extraction was undertaken by Cuadrilla Resources Ltd. in 2018 and 2019, during which 57 and 137 events ($-0.2 < M_W < 2.7$) were recorded at the surface, respectively. M_L is converted to M_W according to Edwards et al. (2021). Characteristics of the magnitude-distance data distribution can be seen in Figure 2. Events were recorded and located by using several surface sensors operated by the British Geological Survey (BGS), Cuadrilla Resources and University of Liverpool (Edwards et al., 2021). We have analysed the performance of four GMMs at the PNR site and ranked them using the AM. In addition, we rank models' performance within a hypocentral distance of 1-10 km and 10-25 km and moment magnitude (M_W) 0 - 1 and 1 - 4 separately, in-order to understand the differences in prediction for different distance and magnitude ranges. The data distribution at the Preston New Road site is highly relevant due to the dense distribution of data with $R < 10$ km and $M_W < 1$ and sparse distribution for $R > 10$ km and $M_W > 2$. This brings us to the question of whether the model which gives the best fit for the entire dataset is the one with the best performance for a specific range of data (i.e., at larger magnitude and up to greater distances)?

The candidate models selected for the ranking are Atkinson (2015), Douglas et al. (2013), Rietbrock et al. (2013) and Edwards et al. (2021). A brief summary of the models and justification for their selection is provided in the following. Atkinson (2015) (A2015) is an empirical GMM derived for small to moderate earthquakes for induced seismicity applications. This model has been derived using a subset of the NGA-West2 database (Ancheta et al., 2014), consisting of ground motions with magnitude range of 3.0 to 6.0 within a hypocentral distance of 40km. Douglas et al. (2013) (D2013) derived GMMs focusing on induced events in geothermal areas. Models have been developed by assembling data from induced and natural seismicity datasets (from Basel, Geysers, Hengill, Roswinkel, Soultz, and Voerendaal). Most of the recordings are between magnitudes 1 and 4 and within a

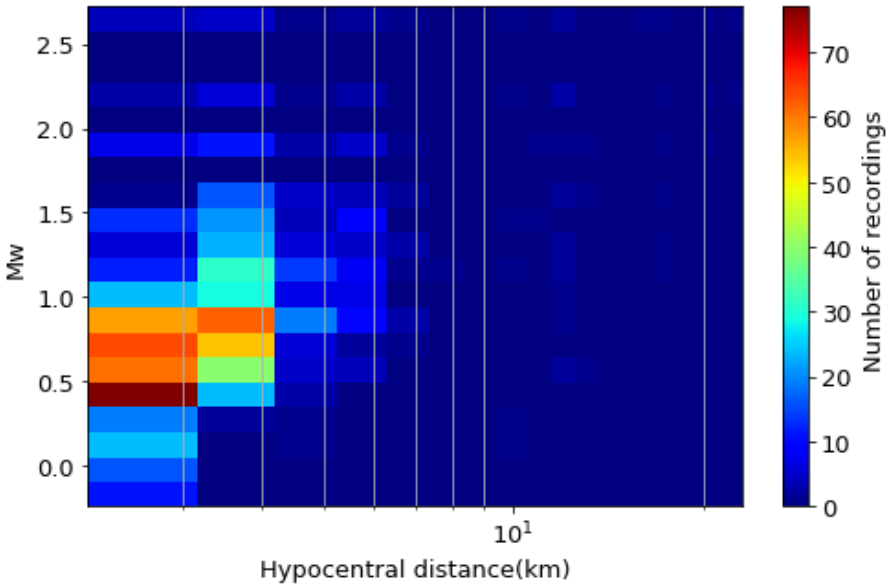

Figure 2: *Hit counts computed for the data distribution, dividing the distance range (1–23km) into 20 equally spaced bins over a logarithmic scale and considering 0.125 magnitude unit intervals. The dark blue regions denotes the sparse data (especially after 10 km).*

hypocentral distance of 20 km. We used their Model 1 (uncorrected for site effects) with the site effects accounted for by considering Boore et al. (2014) site amplification by using the site specific V_{S30} values and a reference velocity of 540 m/s. Rietbrock et al. (2013) (R2013) is a model developed for the UK, which used numerical simulations based on the stochastic point-source model with magnitude-dependent stress drop to derive GMM within a magnitude range of 3-7 and epicentral distance below 300 km. Edwards et al. (2021) (E2021) is the calibrated GMM derived for the Preston New Road dataset using A015 as the starting model. They have developed the model for M_w above 1 and within an epicentral distance of 24 km. Site modifications are adopted from Boore et al. (2014) for site specific V_{S30} values. The summarised details are given in Table 1.

Table 1: *Table shows the features of the models used (A2015, Atkinson (2015),D2013, Douglas et al. (2013), E2021, Edwards et al. (2021) and R2013, Rietbrock et al. (2013)) on PNR dataset.*

Model feature	A2015	D2013	E2021	R2013
Magnitude scale	Mw	Mw	Mw	Mw
Minimum magnitude	3	1	1	3
Maximum magnitude	6	4	2.7	7
Hypocentral distance	<40 km	<20 km	<24 km	<300 km
Site Corrections	Boore et al., 2014 Site specific	Boore et al., 2014 Site specific	Boore et al., 2014 Site specific	Not used
Reference Vs30	760 m/s	540 m/s	760 m/s	2310 m/s
Region	subset of NGA-West2	Basel, Geysers, Hengill, Roswinkel, Soultz and Voerendaal	Preston New Road	United Kingdom

Results and Discussion

The four GMMs mentioned in the previous section have been ranked using the PNR PGA dataset. Table 2 shows the results of AM along with other commonly used ranking methods, LLH (Scherbaum et al., 2009) and the CHISQ-MF test. Figure 3 provides AM plots for the complete dataset and for different ranges of hypocentral distance. We can infer from Table 2, that the best performing model is E2021 given the smaller value of AM. LLH and CHISQ-MF tests also provides a minimum value for this model, which supports the proposed metric.

A minimum value of AM for E2021 is expected as it is the model developed for the target region and an AM value of 0.153 implies that the data and the model is different by 0.153 log10 PGA units. The reason for the small difference is that E2021 is calibrated for

Table 2: *Table shows results of the tests (log-likelihood (LLH) and chisquare-misfit (CHISQ-MF)) including the Area Metric (AM) values for the ranking of four different models (E2021, Edwards et al. (2021), R2013, Rietbrock et al. (2013), D2013, Douglas et al. (2013) and A2015, Atkinson (2015)) on PNR dataset.*

Metric	E2021	R2013	D2013	A2015
AM	0.153	1.069	0.258	0.573
LLH	0.611	9.742	0.847	2.395
CHISQ- MF	1.257	13.854	0.731	3.472

magnitudes above 1 and here we have used the complete dataset for ranking, which contains many recordings below magnitude 1. After removing recordings from $M_W < 1$ events it is found that the AM value is reduced to 0.06. The effect of extrapolation of E2021 to magnitudes below 1 can therefore be inferred from the AM results. It can be also seen from Table 2 that D2013 shows better performance than A2015 for the complete dataset while evaluating the AM results. The low performance of R2013 on the PNR dataset is because of the fact that it is developed by using magnitudes above 3 for the whole UK and the model (not specific for induced seismicity regions). Furthermore, the model predicts using the Joyner-Boore distance (depth to surface extent of rupture) without consideration of depth. This means that shallow source depths seen at PNR (and other induced seismicity sequences) cannot be accounted for, but are rather assumed to be similar to tectonic earthquakes occurring at depths of 10 km or more. Plot (i) of Figure 3 shows the ECDF and CDF of the data and the predicted PGA values of different models. The location of the CDF to the right or left of the ECDF gives us an idea about the residual (observed - predicted) distribution. If the CDF is located right of the ECDF, as in the case of D2013, the residuals will mostly be distributed below the zero horizontal line and if the CDF is to the left of the ECDF, as in all other models, the residuals will be clustered above the zero horizontal line. It is also interesting to note that the width between the ECDF and CDF of different models gives us an idea about the amount of shift of residuals (the mean) from being unbiased, whenever the curves are not intersecting. An increase in distance between ECDF and CDF shows the increase in the mean of residuals from zero. But whenever the curves are intersecting, it will be difficult

to analyse the shifts because the intersection of curves shows the symmetric distribution of residuals above and below the zero horizontal line.

Table 3 shows the performance of models for different ranges of hypocentral distance and magnitude. We chose the distance ranges 1-10m km and 10-25 km for model ranking because of the data distribution of PNR dataset. The consideration of M_W below and above 1 is because of the fact that all candidate models used here are derived from events with magnitudes above 1. Ranking them separately will, therefore, help us to understand the effect of model extrapolation. We have also presented magnitude ranges 1 - 2 and 2 - 3 M_W separately. When models are ranked using data below 10 km, the results seem to be similar to the results provided in Table 2. The performance of the models below 10 km and for the complete PNR dataset are therefore comparable. This is because of the high density of observations at short distance with 99 percent of the recordings below hypocentral distance of 10 km. It is also interesting to note that these results are similar to the results by Cremen et al. (2020) in which they have ranked models using data recorded at distances below 10 km. A2015 and D2013 are two of the models included in their study and the results are similar with this study, i.e., D2013 has a better performance overall than A2015. The AM plots shown in Figure 3 for this distance range are also comparable with the complete dataset plot. When models are analysed for 10-25 km with the AM metric, the results are different from the other two cases (complete dataset and 0-10 km). E2021 itself is again the best performing model in all the ranges of hypocentral distance while the performance of A2015 and D2013 differs. From Table 3, we can infer that the A2015 performs better than D2013 for 10-25 km. For applications above 10 km, it is better to use A2015 rather than using D2013 based on the performance with the complete dataset. Interestingly, the behavior of D2013 and R2013 for magnitudes below 1 and distance greater than 10 km, is different from all other models, D2013 and R2013 fits better for these ranges compared to the other models. Even though E2021 is the region specific calibrated model, it is ranked slightly below A2015 for M_W range of 2-3 within the distance between 10-25 km. This may due to the fact that

Table 3: *Table showing the Area Metric (AM), log-likelihood (LLH) and chisquare-misfit (CHISQ-MF) values for different ranges of hypocentral distance and magnitude (1-10 km and 10-24 km, 0-1 M_w, 1-2 M_w, 2-3 M_w and 0-4M_w) for four different models (E2021, Edwards et al. (2021), R2013, Rietbrock et al. (2013), D2013, Douglas et al. (2013), A2015, Atkinson (2015)) on the PNR dataset.*

Range		Metric	E2021	R2013	D2013	A2015
R0 - 10 km	Mw0 - 3	AM	0.155	1.127	0.257	0.588
		LLH	0.551	10.006	0.846	2.409
		CHISQ-MF	1.175	14.221	0.730	3.490
	Mw0 - 1	AM	0.199	1.271	0.325	0.655
		LLH	0.574	11.758	0.854	2.749
		CHISQ-MF	1.206	16.649	0.741	3.961
	Mw1 - 2	AM	0.058	0.798	0.109	0.435
		LLH	0.475	5.982	0.768	1.612
		CHISQ-MF	1.068	8.643	0.621	2.386
	Mw2 - 3	AM	0.147	0.850	0.409	0.464
		LLH	0.732	6.706	1.293	1.914
		CHISQ-MF	1.425	9.646	1.349	2.804
R10 - 25 km	Mw0 - 3	AM	0.192	0.441	0.305	0.251
		LLH	2.014	3.498	1.078	2.080
		CHISQ-MF	3.202	5.199	1.051	3.035
	Mw0 - 1	AM	1.010	0.241	0.193	1.191
		LLH	11.410	5.192	1.711	11.317
		CHISQ-MF	16.228	7.547	1.929	15.840
	Mw1 - 2	AM	0.081	0.632	0.503	0.092
		LLH	0.461	4.143	1.326	0.536
		CHISQ-MF	1.049	6.093	1.395	0.893
	Mw2 - 3	AM	0.135	0.424	0.210	0.105
		LLH	0.126	1.989	0.521	0.001
		CHISQ-MF	0.234	3.106	0.279	0.148

E2021 is designed to converge to the A2015 model for larger magnitudes and hypocentral distances, limiting its degree of freedom at the higher magnitude and distance. Furthermore, E2021 is not the best performing model for magnitudes below 1 at larger distances (note that the model is calibrated only for magnitudes above 1). From the plots in Figure 3, it can also be seen that a general trend of under-prediction of GMMs is mainly because of the effect of events with magnitudes below 1.

When the models are ranked using LLH and CHISQ-MF for different ranges of magnitude and distance, the results are different from the ranking provided by the AM, especially for smaller sample size. For example, E2021 (AM = 0.192) is showing better performance than D2013 (AM = 0.305) when ranked using AM within a distance and magnitude range of 10-25 km and 0-4 M_w, while D2013 (LLH = 1.078, CHISQ-MF = 1.051) is ranked higher than E2021 (LLH = 2.014, CHISQ-MF = 3.202) when analysed using LLH and CHISQ-MF for the same data range. This can be because of the higher standard deviation of D2013 (sigma = 0.498) compared to E2021 (sigma. = 0.325), LLH may sometimes provide a good fit for models with higher sigma (Kale and Akkar, 2013), but in case of the AM, the metric properly considers the aleatory uncertainty while ranking. The aleatory uncertainty of both the observed data and the model is quantified using the AM while both the LLH and CHISQ-MF tend to favour models with higher sigma. This can be a reason why sometimes the older models with simpler functional forms and larger standard deviations tend to provide better fit compared to the new partially non-ergodic models. This is also evident in the results of E2021 and A2015 within a distance and magnitude range of 10-25 km and 0-1 M_w, LLH and CHISQ-MF favour A2015 above E2021 because of the larger sigma of A2015 (sigma = 0.37), while AM provides higher rank for E2021. In order to test the impact of model sigma, the performance of D2013 and E2021 is analysed in the distance and magnitude range of 10-25 km and 0-4 M_w using observed (i.e. data-based) sigma, i.e., 0.427 for D2013 and 0.517 for E2021. In this case the AM of D2013 changes only slightly to 0.387 and E2021 to 0.262 and both the LLH and CHISQ-MF supported the result, i.e, E2021 is ranked higher than D2013

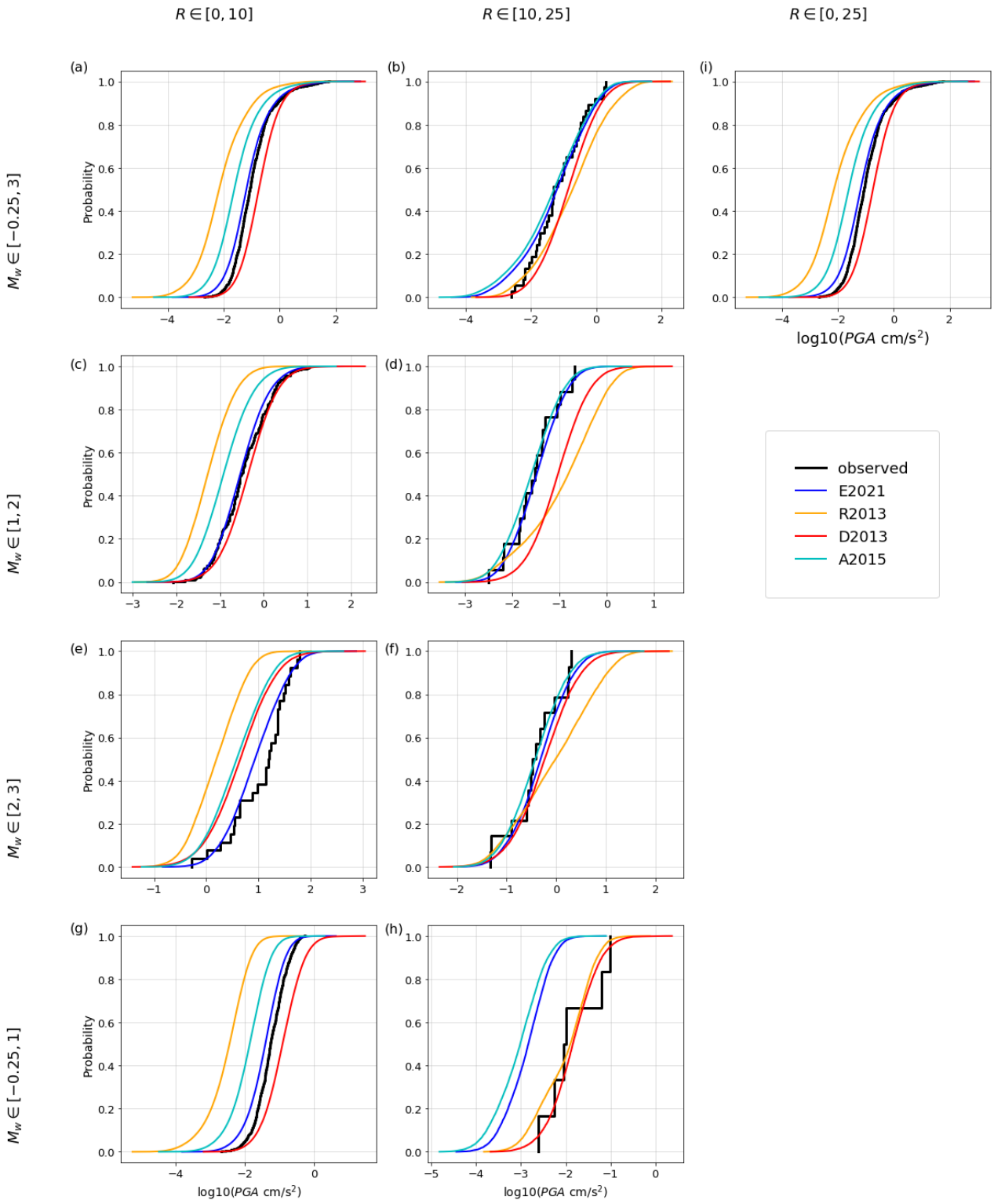

Figure 3: *Plots showing the Area Metric of four different models (E2021, Edwards et al. (2021), R2013, Rietbrock et al. (2013) D2013, Douglas et al. (2013) and A2015, Atkinson (2015)) on PNR dataset. The black curve denotes the empirical data distribution.*

using all the metrics. This clearly shows that model sigma has an impact on the metrics used in ranking. We have also tried a high resolution analysis by dividing the data into smaller bins below 10 km and provided the results as a supplementary document for the paper.

Inter-model variations are analysed by calculating the mutual AM between the models. Here, rather than a CDF and ECDF we have have CDFs, computed over the parameter space of the complete dataset. As for the previous analyses, we also calculate results over different ranges of distance and magnitude. An AM matrix is created after calculating the AM values between the models. This matrix will visually provide an idea about the similarity of models used. For example, in Figure 4, the lighter shades indicate similar models (high inter-model proximity) and darker shades, representing a larger AM value between the models, indicate dissimilar models. Plot (i) of Figure 4 shows the that the similar models are E2021 - A2015 and E2021 - D2013. This trend is similar for all recording below 10 km. Beyond 10 km, the proximity between the A2015 and E2021 increases compared to D2013 and E2021, this can be because of the reason explained earlier, i.e, E2021 is designed to converge to the A2015 model for larger magnitudes and hypocentral distances. A high disagreement between models E2021 and R2013 is seen for all ranges of magnitude and distance and most of the models show higher variations to one another in the range R greater than 10 km and M_w less than 1 compared to other distance and magnitude ranges.

While we have shown that the AM performs well and demonstrates advantages over alternative ranking methods, the AM and inter-model AM matrix should not be considered self-sufficient for the selection of ground-motion models for seismic hazard assessment. The smallest AM value certainly aids in the process of selection of models, however both expert opinion and the results from various testing metrics, including AM, are required for the appropriate selection of GMMs. In particular, when we have insufficient data or unavailability of data, it is not possible to decide the models only using AM (or indeed any data-driven ranking metric). Nevertheless, the use of AM along with expert guidance will facilitate a more robust and defensible ranking according to the available data. For example, for the

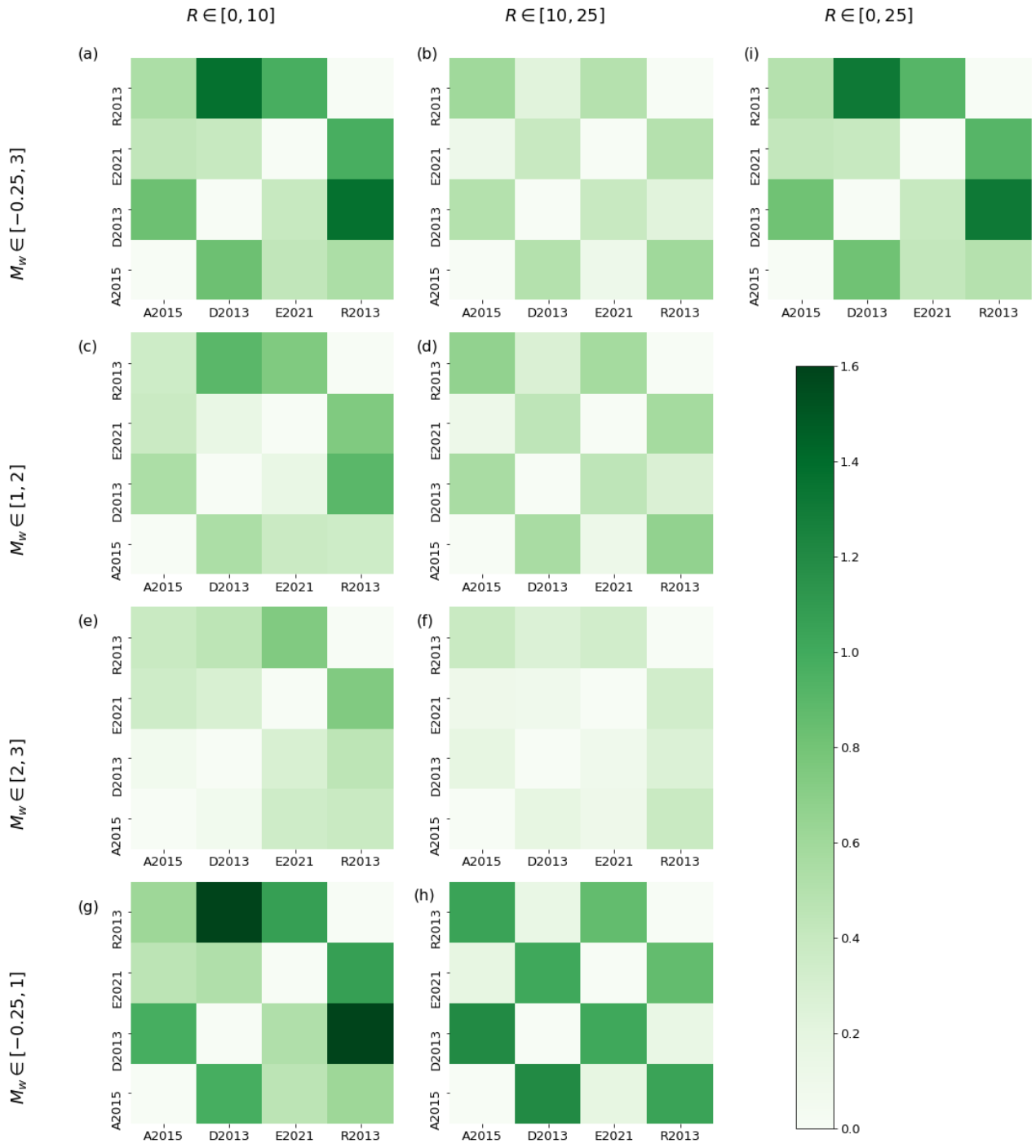

Figure 4: *AM matrix showing the inter-model variations of four different models in different ranges of magnitude and distance. Dark green indicate large distances between models, light green indicate smaller distances.*

complete dataset investigated here, a practitioner may have chosen E2021 as the best and exclusive model, as it is the GMM developed for the given dataset. From our results, AM

seems to agree with the opinion, but for smaller magnitudes ($M_w < 1$) and larger distances ($R > 10$km), E2021 is not the best performing model according to the available data. In this range, better results are observed from D2013 and R2013. Expert guidance is therefore essential, as while E2021 is shown to not be the best model at $M_w < 1$, it was not calibrated in this range, and furthermore, these events do not impact the seismic hazard. We may therefore consider neglecting the AM information for this parameter range. Finally, we note that while building a logic tree using AM matrix, weights should be given by considering not only the AM, but also the characteristics and parametric basis (form, validity, etc.) of the models, and expert opinion.

Conclusion

This paper introduced a new metric for model testing and ranking against a given regional dataset. The AM can be used for selection of models in very low seismicity regions without any assumptions on the distribution of data and also properly accounts for the aleatory uncertainty. There are several other advantages for the proposed metric over the existing methods of ranking. The AM is represented in the same physical units rather than ambiguous statistical units. This is useful to understand the misfit between the model and data more distinctly. The proposed procedure also accounts for aleatory variability in ground-motion estimations by considering the standard deviations of the GMMs, without any underlying assumption on the data distribution. The sensitivity of the AM to the sample size can be a concern when a single model is ranked for different ranges of magnitude and distance. This means that we can only rank different models for a common range of magnitude or distance, while the best performing range for a single model cannot be assessed. However, ranking of a single model for different ranges of dataset is usually irrelevant for applications in probabilistic seismic hazard analysis.

We used the AM to evaluate the performance of four different GMMs (E2021, R2013, D2013, A2015) on the PNR dataset as a case study and obtained the highest ranked model, which is E2021. E2021 is the model calibrated for the PNR region. We also ranked models

in different ranges of both distance and magnitude in-order to understand the variations in prediction. D2013 shows better performance than A2015 when analyzed with the complete dataset (including $M_W < 1$) but when ranking is focused on data range 10-25 km, the misfit of D2013 is larger than the misfit of A2015. The performance of models significantly improved when magnitudes below 1 are removed. The AM provides reliable ranking results compared to the LLH and CHISQ-MF. The dependency of these latter metrics on the model sigma is shown to lead to variability in ranking, i.e, metrics tend to provide better performance for the model with larger sigma. On the other hand, AM takes into account aleatory uncertainty, penalising models that not only present high misfit, but also that do not reflect the distribution evidenced in the data. Finally, the inter-model proximity analyses using AM seems to be an efficient and direct approach to find the models with similar prediction distribution for several applications in probabilistic seismic hazard analysis, especially for the proper assignment of logic tree weights.

Data and Resources

No new data were created as part of this study. The data used in this study is available on request from the British Geological Survey (BGS and operator data) and University of Liverpool (UoL data). All other data used in this study are from the sources listed in the references. The code to calculate the AM is available at `https://doi.org/10.5281/zenodo.4419644` (De Angelis and Sunny, 2021). To maximise reproducibility, all the calculations performed in this paper have been made available at `https://github.com/Jaleena/Ranking-and-Selection-of-Earthquake-Ground-Motion-Models-Using-the-Stochastic-Area-Metric`. Additional figures are provided in the supplementary document for the paper.

Acknowledgements

The authors thank the Editor in Chief Allison Bent, the Managing Editor, Annastasia Pratt, the reviewer Sreeram Reddy Kotha, as well as the anonymous reviewer for their very

helpful feedback and comments. This research is funded by the European Commission's ITN Marie-Sklodowska-Curie New Challenges for Urban Engineering Seismology under the URBASIS-EU project, under Grant Agreement 813137 and the European Union's Horizon 2020 research, and was undertaken with the assistance of resources provided at the University of Liverpool. We thank the British Geological Survey, who provided data for the surface seismometers operated by both the BGS and by Cuadrilla Resources.

References

Ancheta, T. D., Darragh, R. B., Stewart, J. P., Seyhan, E., Silva, W. J., Chiou, B. S.-J., Wooddell, K. E., Graves, R. W., Kottke, A. R., Boore, D. M., et al. (2014). Nga-west2 database. *Earthquake Spectra*, 30(3):989–1005.

Atkinson, G. M. (2015). Ground-motion prediction equation for small-to-moderate events at short hypocentral distances, with application to induced-seismicity hazards. *Bull Seismol Soc Am*, 105(2A):981–992.

Bommer, J. J. and Scherbaum, F. (2008). The use and misuse of logic trees in probabilistic seismic hazard analysis. *Earthquake Spectra*, 24(4):997–1009.

Boore, D. M., Stewart, J. P., Seyhan, E., and Atkinson, G. M. (2014). Nga-west2 equations for predicting pga, pgv, and 5% damped psa for shallow crustal earthquakes. *Earthquake Spectra*, 30(3):1057–1085.

Clarke, H., Verdon, J. P., Kettlety, T., Baird, A. F., and Kendall, J.-M. (2019). Real-time imaging, forecasting, and management of human-induced seismicity at preston new road, lancashire, england. *Seismol Res Lett*, 90(5):1902–1915.

Cremen, G., Werner, M. J., and Baptie, B. (2020). A new procedure for evaluating ground-motion models, with application to hydraulic-fracture-induced seismicity in the united kingdom. *Bull Seismol Soc Am*, 110(5):2380–2397.

De Angelis, M., Sunny J. (2021). The stochastic area metric. *Github repository*, 10.5281/zenodo.4419645.

De Angelis, M., Gray A. (2021). Why the 1-Wasserstein distance is the area between the two marginal CDFs. *arXiv math.ST*, 2111.03570.

Delavaud, E., Cotton, F., Akkar, S., Scherbaum, F., Danciu, L., Beauval, C., Drouet, S., Douglas, J., Basili, R., Sandikkaya, M. A., et al. (2012a). Toward a ground-motion

logic tree for probabilistic seismic hazard assessment in europe. *Journal of Seismology*, 16(3):451–473.

Delavaud, E., Scherbaum, F., Kuehn, N., and Allen, T. (2012b). Testing the global applicability of ground-motion prediction equations for active shallow crustal regions. *Bull Seismol Soc Am*, 102(2):707–721.

Delavaud, E., Scherbaum, F., Kuehn, N., and Riggelsen, C. (2009). Information-theoretic selection of ground-motion prediction equations for seismic hazard analysis: An applicability study using californian data. *Bull Seismol Soc Am*, 99(6):3248–3263.

Douglas, J. (2020). Ground motion prediction equations 1964-2020.

Douglas, J., Edwards, B., Convertito, V., Sharma, N., Tramelli, A., Kraaijpoel, D., Cabrera, B. M., Maercklin, N., and Troise, C. (2013). Predicting ground motion from induced earthquakes in geothermal areas. *Bull Seismol Soc Am*, 103(3):1875–1897.

Edwards, B., Crowley, H., Pinho, R., and Bommer, J. J. (2021). Seismic hazard and risk due to induced earthquakes at a shale gas site. *Bull Seismol Soc Am*, 111(2):875–897.

Ferson, S., Oberkampf, W. L., and Ginzburg, L. (2009). Validation of imprecise probability models. *Int J Reliab Saf*, 3(1-3):3–22.

Gray, A., Wimbush, A., de Angelis, M., Hristov, P.O., Calleja, D., Miralles-Dolz, E., Rocchetta, R. From inference to design: A comprehensive framework for uncertainty quantification in engineering with limited information. *Mech Syst Signal Process*, 195:108210.

Kale, Ö. and Akkar, S. (2013). A new procedure for selecting and ranking ground-motion prediction equations (gmpes): The euclidean distance-based ranking (edr) method. *Bull Seismol Soc Am*, 103(2A):1069–1084.

Kulkarni, R., Youngs, R., and Coppersmith, K. (1984). Assessment of confidence intervals

for results of seismic hazard analysis. In *Proceedings of the eighth world conference on earthquake engineering*, volume 1, pages 263–270.

Mak, S., Clements, R. A., and Schorlemmer, D. (2014). Comment on "a new procedure for selecting and ranking ground-motion prediction equations (gmpes): The euclidean distance-based ranking (edr) method" by özkan kale and sinan akkar. *Bull Seismol Soc Am*, 104(6):3139–3140.

Mak, S., Clements, R. A., and Schorlemmer, D. (2017). Empirical evaluation of hierarchical ground-motion models: Score uncertainty and model weighting. *Bull Seismol Soc Am*, 107(2):949–965.

Pavlenko, V. (2015). Effect of alternative distributions of ground motion variability on results of probabilistic seismic hazard analysis. *Nat Hazards*, 78(3):1917–1930.

Raschke, M. (2013). Statistical modeling of ground motion relations for seismic hazard analysis. *Journal of seismology*, 17(4):1157–1182.

Rietbrock, A., Strasser, F., and Edwards, B. (2013). A stochastic earthquake ground-motion prediction model for the united kingdom. *Bull Seismol Soc Am*, 103(1):57–77.

Scherbaum, F., Cotton, F., and Smit, P. (2004). On the use of response spectral-reference data for the selection and ranking of ground-motion models for seismic-hazard analysis in regions of moderate seismicity: The case of rock motion. *Bull Seismol Soc Am*, 94(6):2164–2185.

Scherbaum, F., Delavaud, E., and Riggelsen, C. (2009). Model selection in seismic hazard analysis: An information-theoretic perspective. *Bull Seismol Soc Am*, 99(6):3234–3247.

Scherbaum, F., Kuehn, N. M., Ohrnberger, M., and Koehler, A. (2010). Exploring the proximity of ground-motion models using high-dimensional visualization techniques. *Earthquake Spectra*, 26(4):1117–1138.

Stafford, P. J., Strasser, F. O., and Bommer, J. J. (2008). An evaluation of the applicability of the nga models to ground-motion prediction in the euro-mediterranean region. *Bull Earthquake Eng*, 6(2):149–177.

Stewart, J. P., Douglas, J., Javanbarg, M., Bozorgnia, Y., Abrahamson, N. A., Boore, D. M., Campbell, K. W., Delavaud, E., Erdik, M., and Stafford, P. J. (2015). Selection of ground motion prediction equations for the global earthquake model. *Earthquake Spectra*, 31(1):19–45.

Strasser, F. O., Abrahamson, N. A., and Bommer, J. J. (2009). Sigma: Issues, insights, and challenges. *Seismol Res Lett*, 80(1):40–56.

Mailing address of authors

Jaleena Sunny, Jane Herdman Building, School of Environmental Sciences, University of Liverpool, Liverpool L69 3GP, United Kingdom, - jaleena.sunny@liverpool.ac.uk

Marco De Angelis, Institute for Risk and Uncertainty, University of Liverpool, Liverpool L69 3GP, United Kingdom - mda@liverpool.ac.uk

Benjamin Edwards, Jane Herdman Building, School of Environmental Sciences, University of 775 Liverpool, Liverpool L69 3GP, United Kingdom - ben.edwards@liverpool.ac.uk

List of Figures