Mak, S., Clements, R., Schorlemmer, D. (2017): Empirical Evaluation of Hierarchical Ground-Motion Models: Score Uncertainty and Model Weighting. - *Bulletin of the Seismological Society of America*, *107*, 2, pp. 949—965.

DOI: http://doi.org/10.1785/0120160232

THE SEISMOLOGICAL SOCIETY OF AMERICA
400 Evelyn Ave., Suite 201
Albany, CA 94706-1375
(510) 525-5474; FAX (510) 525-7204
www.seismosoc.org

Bulletin of the Seismological Society of America, Vol. 107, No. 2, pp. 949–965, April 2017, doi: 10.1785/0120160232

Ⓔ

Empirical Evaluation of Hierarchical Ground-Motion Models: Score Uncertainty and Model Weighting

by Sum Mak, Robert Alan Clements, and Danijel Schorlemmer

Abstract Using a score to generalize the model performance into one numeric value has been one of the most popular approaches to empirically evaluate ground-motion models (GMMs). This approach has an advantage of simplifying model comparison. We study the effects of data correlation and score variability on the evaluation of GMMs. Most modern GMMs are hierarchical, in which ground motions from the same earthquake are modeled as correlated. We demonstrate, with examples, that incorrect results could occur if such hierarchical GMMs are evaluated by a score that does not duly address the data correlation. We propose to use the multivariate logarithmic score, a natural extension of the widely used univariate logarithmic score (referred to as LLH in the seismological literature), to correctly score hierarchical GMMs. The score variability affects the interpretation of model ranking. We demonstrate that the cluster bootstrap is a better bootstrap strategy, compared with other strategies proposed in the literature, to study the score variability. The bootstrap allows computing two useful quantities: the distinctness index that indicates if two models are truly different given the score variability and the frequency weight, a data-driven weighting scheme that represents the frequentist's interpretation of the weight of a logic-tree branch. The frequency weight has a direct link to the current practice of using multiple GMMs in a probabilistic seismic hazard assessment.

Electronic Supplement: Python script to compute the multivariate logarithmic score for a hierarchical lognormal ground-motion model.

Introduction

Empirical evaluation of ground-motion models (GMMs; synonyms include ground-motion prediction equations [GMPEs] and attenuation relationships) has been attracting increasing attention (Table 1; for evaluating intensity prediction equations, see Mak *et al.*, 2015). Such an evaluation assesses the relative performance among multiple models by comparing the model predictions with the corresponding observations. This is necessary for properly selecting the suitable GMM(s) to be included in a seismic-hazard assessment, for demonstrating the superiority of a newly proposed model, or, ideally, for understanding how a model can be improved. An empirical evaluation could also aid experts in assigning weights to multiple suitable models when using a logic tree (e.g., Reiter, 1991, pp. 220–222) in a probabilistic seismic-hazard assessment (PSHA).

Among proposed evaluation methods, that of using a score to generalize the model performance into one numeric value has an advantage of simplifying comparisons among models. Scores have been widely used in forecast evaluation (e.g., Armstrong, 2001, for economic forecast; Bröcker, 2012, section 7.3 for weather forecast).

The purpose of this article is to explore two unresolved issues on scoring GMMs, namely the treatments of data correlation and score variability. We point out potential problems when these two issues are not properly considered and suggest solutions. This article is arranged as follows:

1. In the Logarithmic Scoring Rule section, we review, in the broad context of the literature in scoring probabilistic forecasts, the score introduced by Scherbaum *et al.* (2009, often referred to as LLH in the literature of GMM evaluation), the most popular score for GMM evaluation so far (see Table 1). We discuss its foundation and properties.
2. In the Scoring Hierarchical Models section, we first briefly review the use of hierarchical (similar terms include multilevel, nested, and mixed-effects) GMMs, which model ground motions from the same earthquake as correlated (often through an event term in a mixed-effects model). Such a correlation structure has been

Table 1
Empirical Evaluation Studies of Ground-Motion Models (GMM) and the Methods They Used

Study	Region	Evaluation Method			
		RA	LH*	LLH†,‡	Others
Bindi *et al.* (2006)	Umbra-Marche, Italy	○	○		
Drouet *et al.* (2007)	Pyrences, Spain		○		
Hintersberger *et al.* (2007)	Central Europe		○		
Stafford *et al.* (2008)	Euro-Mediterranean		○		
Delavaud *et al.* (2009)	California			○	
Douglas and Mohais (2009)	French Antilles		○		
Scasserra *et al.* (2009)	Italy	○			ANOVA
Nishimura (2010)	Japan	○			
Shoja-Taheri *et al.* (2010)	Iran	○			
Kaklamanos and Baise (2011)	California		○		NS‡
Uchiyama and Midorikawa (2011)	Japan	○			
Arango *et al.* (2012)	South and Central America		○		
Beauval *et al.* (2012)	Southern and Eastern France			○	
Delavaud *et al.* (2012)	Global			○	DSI‡
Massa *et al.* (2012)	Italy	○			
Mousavi *et al.* (2012)	Iran		○	○	
Vilanova *et al.* (2012)	Southwestern Iberia		○		
Edwards and Douglas (2013)	Cooper Basin, Australia			○	EDR‡
Vacareanu *et al.* (2013)	Eastern Romania	○	○	○	DSI‡
Mousavi *et al.* (2014)	Iran	○	○	○	NS‡, DSI‡
Ogweno and Cramer (2014)	Central and Eastern United States	○		○	EDR‡
Zafarani and Mousavi (2014)	Ahar-Varzaghan, Iran		○	○	
Allen and Brillon (2015)	Haida Gwaii, Canada	○	○		
Drouet and Cotton (2015)	French Alps	○	○	○	
Haendel *et al.* (2015)	Northern Chile			○	
Roselli *et al.* (2016)	Italy				BIC‡
Van Houtte (2016)	New Zealand	○		○	

Some cited studies may have used methods not mentioned here. RA, residual analysis; ANOVA, analysis of variance; NS, Nash–Sutcliffe efficiency coefficient (see Scherbaum *et al.*, 2004); DSI, data support index (Delavaud *et al.*, 2012; see equation 11); EDR, Euclidean distance-based ranking (Kale and Akkar, 2013); BIC, Bayesian information criteria.

*The Scherbaum *et al.* (2004) goodness of fit.

†The Scherbaum *et al.* (2009) likelihood-based score (equation 2).

‡This is a score.

considered to be an indispensable component of a good GMM (e.g., Bommer *et al.*, 2010, p. 791, point 7). The scoring of hierarchical GMMs has not been fully explored in the literature. We here demonstrate, using examples, some potential problems of scoring GMMs without due respect to the correlation structure and suggest remedies. The multivariate logarithmic score is introduced here as a suitable score for hierarchical GMMs.

3. In the Score Variability and Bootstrap section, we discuss the variability (or uncertainty) of scores. Early studies sometimes took models with different scores as truly different and proceeded to interpret the meaning of such a difference. We acknowledge that a score is fundamentally a random variable. We explore the use of bootstrap resampling on hierarchical data and demonstrate how to take the variability of a score into consideration to assess whether two models are truly different. We provide here practical examples, using the Next Generation Attenuation (NGA) dataset, to show that the result of an empirical evaluation of GMMs taking full account of data correlation can be different from a conventional one without considering the data correlation.
4. In the Frequency Weight: A Data-Driven Weighting Scheme section, we first review existing data-driven weighting schemes of GMMs. We then present a weighting scheme that is a natural result of studying the score variability. Compared with existing schemes, the proposed weight has a mathematical meaning more relevant to the use of the logic tree in PSHA.

This article is followed by a discussion on a few cautions and limitations of scoring GMMs and ends with a summary.

The Logarithmic Scoring Rule

The history of scoring GMMs began only after Scherbaum *et al.* (2009) introduced the score LLH to the seismological community:

$$\mathrm{LLH}(\ell, q) = -\log_2 \ell(q), \tag{1}$$

in which q is an observed ground motion and $\ell(\cdot)$ is the likelihood function for the prediction model. When there are N pairs of observation and prediction, the score becomes the average LLH, which is equivalent to the logarithmic likelihood (hence the name LLH) of the model, given the independent observations, divided by the sample size:

$$\overline{\mathrm{LLH}}(\ell, q_1, q_2, \ldots, q_N) = -\frac{1}{N}\sum_i^N \log_2 \ell(q_i). \qquad (2)$$

The information-theoretic meaning of the score can be stated in multiple ways: it is the cross entropy (or relative entropy) between the prediction and the observation; it represents the amount of information deficit of the prediction with respect to the observation; it describes how more uncertain the prediction is compared with the observation.

The same score has been widely used for weather forecast evaluation. Roulston and Smith (2002) named it the ignorance score, vividly describing its meaning under information theory: a larger score means that the modeler is more ignorant of the data-generating mechanism. This score, regardless of its name, is the implementation of the logarithmic scoring rule (e.g., Lindley, 1991, p. 163) on evaluating a forecast of a continuous quantity. The logarithmic score is the logarithm of the predicted probability (or likelihood) for the observed event to occur, multiplied by −1 to make it negatively oriented (i.e., a penalty score, so that a better model has a smaller score). For predictions of a continuous variable, the density function replaces the probability. The earliest use of the logarithmic score dates back at least to Good (1952, section 8, for evaluating binary predictions).

The logarithmic score (and therefore the LLH) is a proper score (Winkler and Murphy, 1968, pp. 754–755), meaning that the expected value of the score will not signal a model to be better than the actual data-generating mechanism (Bröcker and Smith, 2007, section 4). An improper score encourages hedging: modelers may deliberately make an incorrect forecast to score better (i.e., to make the forecast appear better than it is). Although this logic seems straightforward, improper scores were occasionally used (see e.g., Mak *et al.*, 2014a).

The logarithmic score is local, meaning that its value depends only on the forecasted probability for the event that eventually materialized. This can be seen from equation (1) that the score depends on the likelihood only at q; likelihoods of all other outcomes, regardless of how close they are to the actual outcome, do not affect the score. Bernardo (1979, theorem 2) proved that all local proper scores are affine transformations of the logarithmic score (see also Benedetti, 2010, section 3, for a less technical proof). Locality has a special meaning in gambling because all bets other than that hit are lost; only the predicted probability for the event that is eventually observed matters. Gambling strategy is closely related to the logarithmic score (Roulston and Smith, 2002, section 4). The local score is arguably the only score that strictly complies with the scientific method: it assesses a model with only the observed and not the unobserved (Benedetti, 2010, section 2). Nevertheless, nonlocal proper scores (e.g., Hersbach, 2000) have also been widely used in forecast evaluation.

For an observed (natural logarithmic) ground motion q_i, a conventional lognormal GMM predicts the ground motion as $\mathcal{N}(p_i, \sigma_t^2)$, a normal distribution with mean p_i, and standard deviation σ_t. The LLH score for N observations, following equation (2), is represented as

$$\begin{aligned}
&\overline{\mathrm{LLH}}[\mathcal{N}(p_i, \sigma_t), q_i] \\
&= \frac{1}{N}\sum_{i=1}^N -\log_2\left\{\frac{1}{\sigma_t\sqrt{2\pi}}\exp\left[-\frac{(q_i-p_i)^2}{2\sigma_t^2}\right]\right\} \\
&= \frac{1}{N}\sum_{i=1}^N \frac{1}{\ln 2}\left[\ln(\sigma_t\sqrt{2\pi}) + \frac{(q_i-p_i)^2}{2\sigma_t^2}\right] \\
&= \frac{1}{\ln 2}\left[\frac{1}{N}\sum_{i=1}^N \ln(\sigma_t\sqrt{2\pi}) + \frac{1}{2\sigma_t^2 N}\sum_{i=1}^N (q_i-p_i)^2\right] \\
&= \log_2(\sigma_t\sqrt{2\pi}) + \frac{\mathrm{MSE}}{2\sigma_t^2(\ln 2)}, \qquad (3)
\end{aligned}$$

in which MSE is the mean square error between the predicted means and observed ground motions (both in logarithmic scale). A similar but more rigorous analysis on the ignorance score (also known as the LLH) was given by Roulston and Smith (2002, their equations 9–11). Equation (3) shows that for conventional GMMs, the LLH score is a function of MSE. The first term is the smallest possible score, achieved if the predicted means p_i are identical to the observed values q_i. Beauval *et al.* (2012) used a Monte Carlo simulation to estimate the range of LLH values for good models. The analysis given in equation (3) is an analytic alternative.

Kale and Akkar (2013, p. 1071) considered the LLH score "may favor GMPEs with larger standard deviations as they can predict outlier observations with higher probabilities." Their statement is analytically described by equation (3): a larger predicted standard deviation (σ_t) reduces the second term. Because the larger predicted standard deviation also increases the first term, the LLH score does not unconditionally favor GMMs with larger standard deviations. Because errors are squared, outliers have large effects on the score. This is, however, not a property of the score itself but of the lognormal model. For example, the LLH score for a log-Laplacian GMM will be a function of mean absolute error, in which outliers are weighted less than in the MSE. It is the modeler, instead of the model evaluator, who decides how outliers should be treated. Even without bias and outliers, MSE could still be large, due to the intrinsic randomness of ground motions. Simple calculus shows that the LLH score achieves the minimum when $\sigma_t^2 = \mathrm{MSE}$. Therefore, if the error is normally distributed as assumed by the model, the best score occurs when the predicted standard deviation is identical to the sample standard deviation (regardless of large or small). This again shows the score is proper.

Scoring Hierarchical Models

Ground-Motion Observations as Hierarchical Data

Ground motions from the same earthquake have been considered as correlated. Consequently, many modern GMMs predict the (natural logarithmic) ground motion y_{ij} from earthquake i at site j as

$$y_{ij} = p_{ij} + \eta_i + \epsilon_{ij} \tag{4}$$

(e.g., Abrahamson and Youngs, 1992, their equation 2), in which p_{ij} is the predicted mean (natural logarithmic) ground motion (which is a function of various predictor variables such as earthquake magnitude and source-site distance), η_i is the event term, and ϵ_{ij} is the leftover residual. GMM modelers often consider $\eta_i \sim \mathcal{N}(0, \sigma_\mathrm{b}^2)$ and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\mathrm{w}^2)$. σ_b is called the between-event (or interevent) sigma and σ_w the within-event (or intraevent) sigma. See Youngs *et al.* (1995, their fig. 4) or Strasser *et al.* (2009, their fig. 3) for a graphical expression of this data hierarchy. This is the simplest but widely used form of hierarchical structure for ground-motion modeling. More complicated hierarchical structures, such as that with an additional station term to account for the correlation of records from the same station (e.g., Kuehn and Scherbaum, 2015), may become more popular.

Another source of data correlation is the vicinity of recording sites. The degree of correlation is inversely proportional to the distance between the recording sites. Goda and Hong (2008, their fig. 1) showed that two peak ground accelerations (PGAs) recorded more than 10 km apart in California are largely uncorrelated (correlation coefficient <0.1). This source of correlation is often not explicitly modeled in a GMM. It is therefore acceptable to evaluate GMMs without specifically considering this correlation.

Ground-Motion Predictions as a Multivariate Vector

To correctly score a hierarchical GMM, ground motions predicted by the GMM for multiple earthquakes should be taken as a single multivariate random variable, instead of multiple independent random scalars. Because the event terms and leftover residuals of a GMM (see equation 4) are often modeled as uncorrelated normal random variables, the corresponding mixed-effects model neatly describes the ground-motion prediction as a multivariate normal random vector (e.g., Jiang, 2007, pp. 6–7). Following equation (4), the N_i (natural logarithmic) ground motions predicted for earthquake i are

$$\begin{aligned}
\underset{N_i\times 1}{\mathbf{y}_i} &= \underset{N_i\times 1}{\mathbf{p}_i} + \overset{N_i\times a}{\mathbf{Z}_i}\underset{a\times 1}{\boldsymbol{\alpha}_i} + \underset{N_i\times 1}{\boldsymbol{\epsilon}_i} \\
\boldsymbol{\epsilon}_i &\sim \mathcal{N}(\mathbf{0}, \underset{N_i\times N_i}{\mathbf{R}_i}) \\
\boldsymbol{\alpha}_i &\sim \mathcal{N}(\mathbf{0}, \underset{a\times a}{\mathbf{G}_i}),
\end{aligned} \tag{5}$$

in which $\mathbf{p}_i$ is the predicted mean (natural logarithmic) ground motions for the earthquake, $\boldsymbol{\alpha}_i$ is the random effects, $\mathbf{Z}_i$ is a known matrix describing the linear relation of random effects, and $\mathbf{R}_i$ and $\mathbf{G}_i$ are the covariance matrices of the leftover residuals and the random effects, respectively. The dimensions for the vectors and matrices are indicated below the corresponding symbol. For the simplest form of the mixed-effects model as described by equation (4), a equals one, $\mathbf{Z}_i$ is a vector of ones, and $\mathbf{R}_i$ and $\mathbf{G}_i$ are diagonal matrices with σ_w^2 and σ_b^2 on the diagonal, respectively. Consequently, the N ground motions predicted for M earthquakes are

$$\begin{aligned}
\underset{N\times 1}{\mathbf{y}} &= \mathcal{N}(\underset{N\times 1}{\mathbf{p}}, \underset{N\times N}{\mathbf{V}}) \\
N &= \sum_i^M N_i \\
\underset{N\times N}{\mathbf{V}} &= \underset{N\times N}{\mathbf{R}} + \underset{N\times M}{\mathbf{Z}}\,\underset{M\times M}{\mathbf{G}}\,\underset{}{\mathbf{Z}'} \\
\mathbf{R} &= \mathrm{diag}(\mathbf{R}, \ldots, \mathbf{R}_M); \quad \text{similarly for } \mathbf{G} \text{ and } \mathbf{Z}
\end{aligned} \tag{6}$$

(Jiang, 2007, his equation 1.4), in which diag(·) denotes a block diagonal matrix, and $\mathbf{Z}'$ is the transpose of $\mathbf{Z}$. Two examples for computing the covariance matrix $\mathbf{V}$ are given in Appendix A.

Equations (5) and (6) are capable of describing GMMs with situation-dependent between-event and within-event sigmas. A special form appeared in the literature (e.g., Joyner and Boore, 1993, their equations 9–11) that requires constant between-event and within-event sigmas and so cannot be applied to some GMMs when the between-event sigma is allowed to vary (e.g., the implementation by Arroyo *et al.*, 2014, p. 1866, on the model of McVerry *et al.*, 2006); the between-event sigma is often modeled as magnitude dependent (Youngs *et al.*, 1995). Recently, Stafford (2015) pointed out the more general form (his equation 9, equivalent to equations 5 and 6 here) that would solve some difficulties encountered in model fitting when the specific form of Abrahamson and Youngs (1992, equivalent to equation 4 here) is used.

Multivariate Logarithmic Score

The LLH score (Scherbaum *et al.*, 2009) is currently the most popular score for GMM evaluation (Table 1). The current practice of computing the LLH score is to take ℓ (see equation 2) as the density function of $\mathcal{N}(p_i, \sigma_\mathrm{t}^2)$, in which $\sigma_\mathrm{t} = \sqrt{\sigma_\mathrm{b}^2 + \sigma_\mathrm{w}^2}$ is the total sigma of the model. Scherbaum *et al.* (2009, p. 3239) was aware of the effect of the event term on the LLH score and suggested a remedy:

> *Because of the interevent variability component in the ground-motion model, it is not to be expected that the median of the ground-motion observations and the model median would match. In certain instances, it may therefore be justified to subtract the difference*

between these medians from the observations before the likelihoods are calculated.

By first removing the event terms, the observations become independent, and the use of equation (2) is valid. If the event terms can be accurately calculated, it is of course valid to separately evaluate the between-event sigma and the GMM with the event term removed. In fact, this approach has been used by quite a few workers (Bindi *et al.*, 2006; Scasserra *et al.*, 2009; Bradley, 2010; Shoja-Taheri *et al.*, 2010; Uchiyama and Midorikawa, 2011; Vacareanu *et al.*, 2013; Azarbakht *et al.*, 2014; Van Houtte, 2016). There are two complications for this approach. First, event terms are fundamentally unobservable and can be meaningfully computed only during the modeling process (see Appendix B). Second, by removing the event terms, the score will then lose its ability to describe the overall performance of the model. One major merit of scoring GMMs for model evaluation is the convenience of using single numeric value of the score for model comparison. It is desirable to have a score that represents the overall performance of a hierarchical GMM.

Based on the multivariate form of lognormal GMM (equation 6), the multivariate logarithmic score can be directly computed:

$$\ell\ell(\mathbf{p},\mathbf{V},\mathbf{q}) = [N\log(2\pi)+\log|\mathbf{V}|+(\mathbf{q}-\mathbf{p})'\mathbf{V}^{-1}(\mathbf{q}-\mathbf{p})]/2, \quad (7)$$

in which $\mathbf{p}$ and $\mathbf{q}$ are the vectors of the predicted mean and observed (natural logarithmic) ground motions, respectively, and $|\mathbf{V}|$ and $\mathbf{V}^{-1}$ are the determinant and the inverse of the covariance matrix $\mathbf{V}$, respectively. Equation (7) is the negative logarithmic likelihood of the multivariate lognormal model. Seismologists could consider this score as the implementation of the LLH score (equation 1) on hierarchical GMMs. Interestingly, this score applies not only to evaluating the forecast based on the multivariate normal distribution, but also to that of a large number of other multivariate distributions with finite second moments (Gneiting and Raftery, 2007, section 4.4). Another reason for equation (7), instead of equation (2), to be the correct logarithmic score for a GMM in the form of equation (4) is that equation (2) (without dividing by the sample size; natural instead of binary logarithm) is a special case of equation (7) for substituting $\mathbf{V}_\mathrm{d}$ into $\mathbf{V}$, when $\mathbf{V}_\mathrm{d}$ is identical to $\mathbf{V}$, except that all off-diagonal elements are set to zero; this is a misformulation of the covariance matrix.

We provide a python implementation of equation (7) to compute the multivariate logarithmic score in the Ⓔ electronic supplement to this article.

Effects of Data Hierarchy on GMM Evaluation

Abrahamson and Wooddell (2010) pointed out that, when evaluating a GMM, the assumption of independence for the test data when such correlation has been included in the GMM would lead to incorrect results, although their approach for evaluating GMMs was very different from the scoring approach discussed in this article. Existing studies for scoring GMMs (see Table 1) did not explicitly include the hierarchical structure of the model (equation 4) into the score implemented. To illustrate this problem, we provide four examples of possible incorrect results when such hierarchical structure is excluded in the evaluation of GMMs.

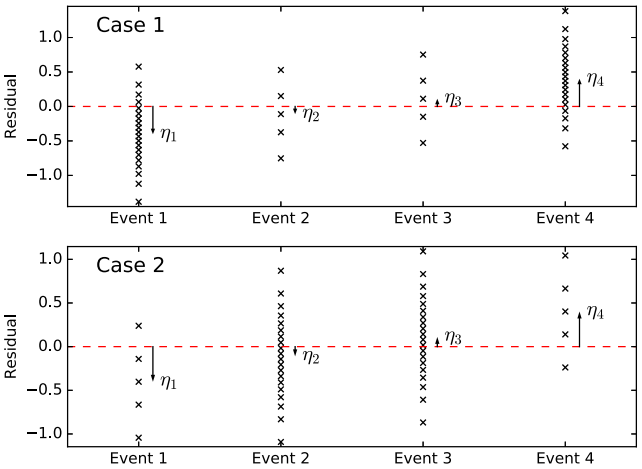

Figure 1. Residuals used in example 1 in the Effects of Data Hierarchy on GMM Evaluation section. The residuals follow $\mathcal{N}(\eta_i,\sigma_\mathrm{w})$, in which η_i is the event term, $i\in\{1,2,3,4\}$. The color version of this figure is available only in the electronic edition.

Example 1: Inflated Score Variability This example demonstrates that ignoring data hierarchy leads to a more unstable score. We use the data structure as expressed by equation (4) and a model that perfectly describes the probability distribution of data. Without loss of generality, we set $p_{ij}\equiv 0$ and focus on the residuals only. Suppose there are four earthquakes, with event terms η_i, $i\in\{1,2,3,4\}$, sampling uniformly the distribution $\mathcal{N}(0,\sigma_\mathrm{b})$. Within an earthquake, the N_i records sample uniformly the distribution $\mathcal{N}(\eta_i,\sigma_\mathrm{w})$. The residual j for event i is therefore

$$r_{ij} = Q_\mathrm{w}(y_j)+\eta_i \quad \text{with } y_j=\frac{2j-1}{2N_i}, j\in\{1,2,\ldots,N_i\}$$
$$\eta_i = Q_\mathrm{b}(x_i) \quad \text{with } x_i=\frac{2i-1}{8}, \quad (8)$$

in which $Q_z(\cdot)$ is the quantile function for $\mathcal{N}(0,\sigma_z)$. The purpose of x_i and y_i is to sample uniformly the distribution $\mathcal{N}(0,\sigma_x)$, based on the inverse probability integral transform (e.g., DeGroot and Schervish, 2012, corollary 3.8.1).

Consider two cases: (1) $N_i=\{20,5,5,20\}$ and (2) $N_i=\{5,20,20,5\}$. These two cases are identical, except that the two earthquakes with event terms farther (respectively, nearer) from zero produce more records for case 1 (respectively, case 2). A graphical expression of the residual distributions is given in Figure 1. The number of records an earthquake produces is completely unrelated to the predictive power of a GMM; it purely depends on how many accelerometers are installed near an earthquake, which is an envi-

Table 2
Logarithmic Scores Computed for Example 1 in the Effects of Data Hierarchy on GMM Evaluation Section

Data Hierarchy	Case 1	Case 2
Ignored*	45.3	39.3
Included†	38.8	38.5

Smaller score means better model.
*Univariate (equation 2 without division by sample size; natural instead of binary logarithm).
†Multivariate (equation 7).

Table 3
Logarithmic Scores Computed for Example 2 in the Effects of Data Hierarchy on GMM Evaluation Section

Data Hierarchy	Correct Model	Biased Model
Ignored*	72.6	68.3
Included†	61.2	61.5

Smaller score means better model.
*Univariate (equation 2 without division by sample size; natural instead of binary logarithm).
†Multivariate (equation 7).

Table 4
Logarithmic Scores Computed for Example 3 in the Effects of Data Hierarchy on GMM Evaluation Section

Data Hierarchy	Correct Model	Inflated σ_b*	Deflated σ_b*
Ignored†	45.3	45.3	45.3
Included‡	38.8	39.6	39.1

Smaller score means better model.
*Between-event sigma inflated (deflated) by 20%, keeping the total sigma unchanged.
†Univariate (equation 2 without division by sample size; natural instead of binary logarithm).
‡Multivariate (equation 7).

ronmental setting. In this example, we used realistic numbers of $\sigma_b = 0.35$ and $\sigma_w = 0.5$ (i.e., $\sigma_t = 0.61$).

Because the two cases are highly similar, it is reasonable to expect that the two sets of data should result in similar scores. The logarithmic scores computed for the two datasets are shown in Table 2. When the data hierarchy is ignored using a univariate logarithmic score, the score varies a lot over the two environmental settings, undesirably inflating the instability of the score.

Example 2: Favoring a Biased Model This example demonstrates that a score ignoring data hierarchy incorrectly favors a biased model. Similar to the setting in example 1, the residuals for four earthquakes are generated according to equation (8). In this example, $N_i = \{10, 10, 10, 50\}$, meaning that one event with an unfortunately large event term produces the most records. The logarithmic scores computed for two models are shown in Table 3. Both models use the correct uncertainty (i.e., $\sigma_b = 0.35$ and $\sigma_w = 0.5$). The first model is unbiased (i.e., $p_{ij} \equiv 0$), whereas the second one is biased toward the event that produced the most records (i.e., $p_{ij} \equiv k > 0$; here $k = \sigma_b/2$). When the data hierarchy is ignored, the score favors the biased model, a wrong result.

The degree of data unbalance may be somewhat overstretched in example 2 to highlight its effect. Unbalanced data, however, are quite ubiquitous. For example, although the NGA flatfile (2008 version) includes ground-motion records from more than 160 earthquakes, 15% of the records come from one earthquake (the 1999 Chi-Chi, Taiwan, earthquake). It is of course possible to apply expert judgment to manually select a balanced dataset. It is, however, still desirable to have a score directly applicable to hierarchical and unbalanced data.

Example 3: Insensitive to Sigma Partition This example demonstrates that a score ignoring data hierarchy could not identify models that incorrectly partition the total sigma into between-event and within-event sigmas. Using the setting of case 1 of example 1, two additional wrong models are created: one model has a between-event sigma inflated (respectively, deflated) by 20%, whereas the within-event is deflated (respectively, inflated) accordingly to keep the total sigma unchanged. The multivariate logarithmic score, taking data hierarchy into account, correctly attributes larger scores to the wrong models (Table 4). On the other hand, the univariate logarithmic score does not distinguish the three models.

The scores computed for the above three examples (Table 4) also show one common feature: under the same dataset, the multivariate logarithmic score (equation 7), which fully represents a hierarchical model, is always smaller (i.e., better) than the corresponding univariate logarithmic score (equation 2), which discards the hierarchical structure of the model. This feature, on the one hand, is straightforward because a hierarchical data-generating mechanism must be more correctly described by a hierarchical model. On the other hand, this feature points out an important fact that, when using a score that does not include the hierarchical structure of the model, the model evaluator has not used all information provided by the modeler to evaluate the model; this is unfair to the modeler.

Finally, we discuss the effect of data hierarchy on a popular graphical method on GMM evaluation proposed by Scherbaum *et al.* (2004, Goodness-of-Fit Measures section, A New Goodness-of-Fit Measure subsection; often referred to as the LH method in the literature of GMM evaluation; see Table 1 for its popularity). The LH method is identical to the probability integral transform (Rosenblatt, 1952; DeGroot and Schervish, 2012, theorem 3.8.3), a common diagnostic tool (e.g., Dawid, 1984, section 5.3). The foundation for the LH method is that a random variable of a certain distribution will become uniformly distributed after being mapped by its own distribution function. Although the probability integral transform works for multivariate models for evaluating hierarchical GMM, the hierarchical observation

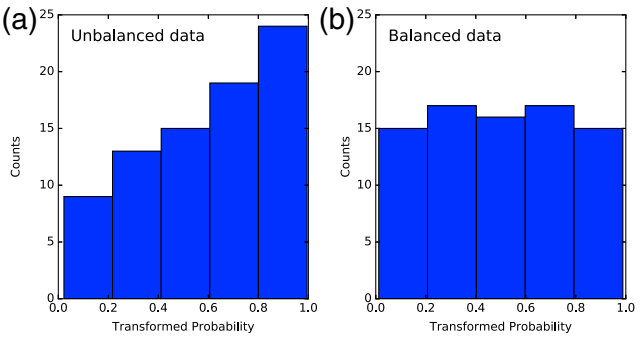

Figure 2. Probability integral transform (i.e., the LH method) histograms for example 4 in the Effects of Data Hierarchy on GMM Evaluation section. The model is correctly specified, and so the histogram should show a uniform distribution if the data are independent. The color version of this figure is available only in the electronic edition.

(e.g., equation 4) is a single vector; there will not be a set of independent and identically distributed random vectors to satisfy the requirement of the transform.

The use of the LH method in published studies (Table 1) did not explicitly consider the hierarchical structure of the model and the observation. We demonstrate here a potential outcome.

Example 4: The LH Method and Unbalanced Data For a set of unbalanced data like the one used in example 2, the probability integral transform histogram (i.e., the LH method), based on the distribution function of a univariate normal distribution, will not be uniform even if the predicted mean and standard deviation are correctly specified (Fig. 2a; in this case, $p_{ij} \equiv 0$ and $\sigma_t = 0.61$). A well-recorded earthquake with an unfortunately large event term will ruin the result. The method works better for a balanced dataset (i.e., each earthquake produced the same number of observations; for Fig. 2b, $N_i = \{20, 20, 20, 20\}$).

Score Variability and Bootstrap

Ground-motion generation is modeled as a random process. A score computed from a set of ground-motion data is, therefore, a random variable. Even if the underlying data-generating mechanism is unchanged, a score fluctuates if different datasets are used. To compare models by their scores, it is natural to ask whether two models of different scores (say, 159 and 161) are really different. It is neither the absolute difference ($161 - 159 = 2$) nor the difference ratio ($2/159 = 1.26\%$) that tells if the models are different. It is the combination of the score difference and the intrinsic randomness of the score that decides if the two models are truly different. Jolliffe (2007, p. 637) pointed out that "the value of a verification measure on its own is of little use; it also needs some quantification of the uncertainty associated with the observed value." If a model truly scores better than another, such a relative performance should persist over datasets of the same random property, although the score fluctuates.

It is impractical to keep waiting for new data to assess the variability of model performance because well-recorded earthquakes of moderate sizes often occur only once per several years even in an active tectonic region. Resampling (Efron and Tibshirani, 1993), which exploits the data variability within a fixed dataset, has been widely used to study the variability of a random variable. The simplest form of resampling, the naive resampling, samples with replacement the original dataset to generate new datasets of the same size (or without replacement, if the size of the resampled dataset is smaller than the original one). Some GMM evaluation studies (Mousavi *et al.*, 2012; Edwards and Douglas, 2013; Azarbakht *et al.*, 2014; Mousavi *et al.*, 2014; Van Houtte, 2016) used the dispersion of the scores computed from resampled datasets to represent of the variability of the score.

We first discuss two complications of resampling related to scoring GMMs. The first is about the resampling on hierarchical data (in the Bootstrap on Hierarchical Data section). The second is about the interpretation of scores computed on different datasets (in the Scores Based on Distinct Datasets Are Not Comparable section). We then demonstrate how the score variability can be assessed using the distinctness index (DI; in the Distinctness Index section).

Bootstrap on Hierarchical Data

A naive resampling does not consider the data hierarchy. There have been studies on bootstrap strategies on hierarchical data. For the hierarchical structure as described by equation (4), Davison and Hinkley (1997, pp. 100–102) described three viable bootstrap strategies. For in-sample bootstrap, in which the same set of data is used for both model fitting and bootstrap, Field and Welsh (2007) analyzed the performance of various multilevel bootstrap strategies. GMM evaluation is often an out-of-sample analysis (i.e., the model is fitted on one dataset and evaluated using another dataset), and we are not aware of any out-of-sample analyses of multilevel bootstrap strategies. We here describe a strategy called cluster bootstrap (terminology from Field and Welsh, 2007; Davison and Hinkley, 1997, called it strategy 1), which is easy to implement for GMM evaluation. In Appendix C and Appendix D, we give a further discussion on other proposed strategies in the literature and explain why they are not as desirable as the cluster bootstrap.

The cluster bootstrap samples with replacement the first level of a multilevel dataset (i.e., selects an earthquake) and then takes all data from the sampled event, thus preserving the data hierarchy. The number of earthquakes sampled is identical to the number of earthquakes in the original dataset. When the data within a cluster are also permuted, Field and Welsh (2007) called it randomized cluster bootstrap. Data permutation does not affect scoring so we focus on the cluster bootstrap. The implicit assumption of this bootstrap is that an earthquake of the same characteristics (leading to the

Table 5
Hypothetical Scores for Example 5 in the Scores Based on Distinct Datasets Are Not Comparable Section

Dataset	Score for Model	
	A	B
1	1.0	2.0
2	1.0	2.0
3	1.0	2.0
4	1.0	2.0
5	1.0	2.0
6	3.0	3.1
7	3.0	3.1
8	3.0	3.1
9	3.0	3.1
10	3.0	3.1
Mean	2.0	2.55
Standard deviation	1.05	0.58

same event term) could occur again, each time recorded by the identical environment. This strategy is the most conservative among proposed strategies (see Appendix C) because it allows the fewest variations. For meaningful use, the number of clusters (i.e., earthquakes) cannot be too small. For out-of-sample evaluation of forecasts using hierarchical data, this is the only kind of bootstrap we could find in the literature (e.g., Candille *et al.*, 2007, section 2c). It is conceptually similar to the moving blocks bootstrap (Efron and Tibshirani, 1993, section 8.6), widely used for time-series analysis.

Scores Based on Distinct Datasets Are Not Comparable

Because of simplicity, it appears attractive to assess the score variability by inspecting the dispersion of the scores computed from resampled datasets. Such a measure of score variability, however, implicitly assumes that the involved scores are independent. In fact, the two sets of resampled scores for two models are closely correlated because each pair of resampled scores is computed from the same resampled dataset. A fallacy may appear if the variability of a score is assessed by simply measuring how the numeric value of the score changes among resampled datasets, illustrated by the following example:

Example 5: False Impression of Score Variability Suppose the scores of two models are computed based on 10 sets of data (Table 5). Model A performs better than model B (assume that a smaller score is better) under all cases, but a simplistic comparison between the two models using the sample mean and standard deviation of the scores provides a false impression of score variability, which may mislead one to believe that the performance of the models are too uncertain to tell if model A is clearly better than model B.

Some workers tried to divide the score by the sample size, hoping to make scores based on datasets of different sizes comparable. Such a normalization does not address the concern raised here. In addition, this normalization assumes that each record carries the same amount of information, which is not true for ground-motion data. Two ground-motion records from two earthquakes, in fact, do not contain the same amount of information as two from one earthquake: the former case contains information about the variability between earthquakes, whereas the latter case contains information about the variability within an earthquake.

Distinctness Index

Although the dispersion of scores among resampled datasets does not correctly provide the information on score variability, relative rankings of models among the resampled datasets provide useful information. To assess if models i and j are distinct, one could count how often, among the N_{bs} bootstrap samples, model i scores better than model j. The result is a DI:

$$d_{ij} = \frac{1}{N_{\text{bs}}} \sum_{k}^{N_{\text{bs}}} \tilde{\mathbf{1}}(s_i^{(k)}, s_j^{(k)})$$

$$\tilde{\mathbf{1}}(x, y) = \begin{cases} 1 & \text{when } x < y \\ -1 & \text{when } x > y \\ 0 & \text{when } x = y \end{cases}, \tag{9}$$

in which $s_i^{(k)}$ is the score (assume a smaller score for a better model) for model i for the bootstrap sample k and $\tilde{\mathbf{1}}(\cdot)$ is a modified indicator function. The DI has the following properties:

1. $d_{ij} > 0$ if model i more often scores better than model j. The extreme case is $d_{ij} = 1$ if model i always scores better than model j.
2. $d_{ij} < 0$ if model i more often scores worse than model j. The extreme case is $d_{ij} = -1$ if model i always scores worse than model j.
3. $d_{ij} = 0$ if model i scores better than model j in half occasions, and worse for the remaining half. This is an extreme case in which two models are completely indistinguishable, due to the variability of the score. It is also possible to obtain $d_{ij} = 0$ if two models score exactly the same.
4. $d_{ij} = -d_{ji}$.

The DI can be computed from any scores and resampling methods of choice, not only from the logarithmic score and the cluster bootstrap used in this article. For multiple models, a set of DIs can be computed, one for each model pair, forming a distinctness table. Example 6 shows the distinctness table of four NGA GMMs.

Example 6: Distinctness Table Figure 3 shows the distinctness table for four NGA GMMs (Abrahamson and Silva, 2008; Boore and Atkinson, 2008; Campbell and Bozorgnia, 2008; Chiou and Youngs, 2008; these GMMs are hereafter referred to as models AS, BA, CB, and CY, respectively), based on the PGA observations of the NGA-West2 flatfile (see Data and Resources). Only records from

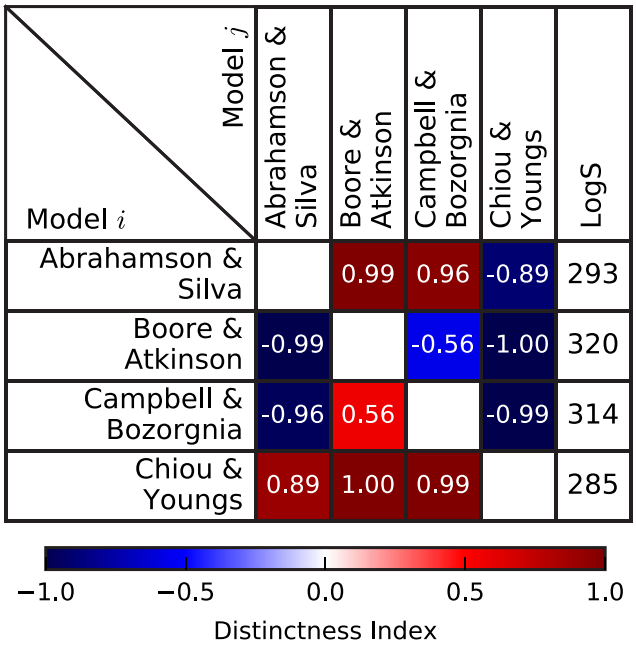

Figure 3. Distinctness table for Next Generation Attenuation (NGA) models (see example 6 in the Distinctness Index section). The distinctness index (DI) of each pairwise comparison is given in the intersecting box of a model pair. The multivariate logarithmic score (LogS) follows equation (7), computed using the whole dataset (i.e., no bootstrap). The color version of this figure is available only in the electronic edition.

prospective mainshocks (i.e., EQID >175 and class 1) of moment magnitudes ≥ 5 and rupture distances ≤ 40 km were used. Records with incomplete metadata so that the GMMs cannot be implemented were also discarded. The filtered dataset consisted of 365 records from 13 earthquakes with moment magnitudes up to 7.2. The DIs were computed based on 1000 samples of cluster bootstrap and the multivariate logarithmic score. A simple way to identify the best model is to locate the row of all positive (or the column of all negative) DIs. In this case, model CY is the best model. Most DIs are far from zero, meaning that most models are clearly different from other models as far as the available data can tell. Models BA and CB are the least dissimilar because their DI is the closest to zero.

We acknowledge that the definition of peak motions used by the NGA GMMs is GMrotI50 (Boore *et al.*, 2006), whereas that used in the NGA-West 2 flatfile is rotD50 (Boore, 2010). We assumed the difference in definitions did not affect the results. We acknowledge that the four NGA models evaluated here have been superseded by their newer versions in 2014. The 2008 versions are used here because prospective observations are available for them but not for the newer version, and only evaluations based on prospective observations can tell the predictive power of models.

The following two examples show that the result of a GMM evaluation using the DI (equation 9) and the multivariate logarithmic score (equation 7), which incorporate data correlation and score variability, can be different from that using scores and bootstrap strategies that do not consider data correlation.

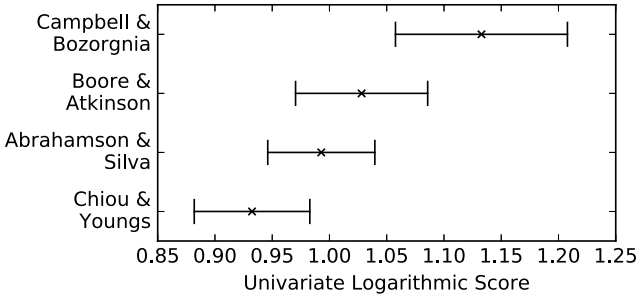

Figure 4. Univariate logarithmic scores for example 7 in the Distinctness Index section. Cross denotes the mean of 300 bootstrap samples. Interval denotes the mean $\pm$ one standard error. The univariate logarithmic score follows equation (2), using natural instead of binary logarithm.

Example 7: Effects of Data Correlation on Score Variability The logarithmic score and bootstrap have often been implemented without considering the data correlation. In that case, each ground-motion record is considered as independent; only the total sigma of the GMM is involved; and the bootstrap is done by sampling with replacement the whole dataset (i.e., a naive bootstrap). In addition, the dispersion of resampled scores were used to represent the score variability (Mousavi *et al.*, 2012; Edwards and Douglas, 2013; Azarbakht *et al.*, 2014; Mousavi *et al.*, 2014; Van Houtte, 2016). Such a use of logarithmic score and bootstrap (Fig. 4) may lead one to believe that model pairs of CY and AS, AS and BA, and BA and CB are comparable in performance, because the intervals of mean $\pm$ one standard error for each pair overlapped with each other. Models AS and BA appeared to be the most similar.

A different conclusion, however, will be drawn on the same dataset when the data correlation is given due respect using the multivariate logarithmic score and cluster bootstrap, and the DI is used to represent the score variability (Fig. 3): models BA and CB are the most similar, whereas all other model pairs are fairly different because their DIs are close to 1 or -1.

Example 8: Ranking Models In example 6 and Figure 3, the four models can be ranked as CY, AS, CB, and BA, for which CY is the best model. Model CY has positive DIs with respect to all other models, meaning that CY is usually better than other models. Similarly, model AS has positive DIs with respect to CB and BA but a negative DI with respective to CY, meaning that AS is usually better than CB and BA but not CY. Finally, model BA is usually worse than all other models. This ranking based on the DIs takes into account the score variability. A ranking based on the logarithmic score computed by the whole dataset (i.e., the last column in Fig. 3) does not consider score variability, although the resulting ranks are the same in this case.

Table 6
Hypothetical Scores for Example 9 in the Distinctness Index Section

Sample	Score for Model		
	A	B	C
1	10	20	30
2	10	20	30
3	10	20	30
4	10	20	30
5	20	30	10
6	20	30	10
7	20	30	10
8	30	10	20
9	30	10	20
10	30	10	20

The ranking result based on the univariate logarithmic score and the naive bootstrap (Fig. 4) is different. The ranks for models CB and BA are swapped compared with the ranking based on the DI.

Example 9: Unrankable Models This example demonstrates that it is not always possible to unambiguously rank multiple models. Table 6 shows a hypothetical example of scores for three models and 10 bootstrap samples. The corresponding distinctness table (Fig. 5) shows that although model A is often better than model B and model B is often better than model C, model A is not often better than model C. Therefore, no model can be ranked as the best. Such an unrankable situation is expected to occur more often as the number of models increases.

Frequency Weight: A Data-Driven Weighting Scheme

A score provides a data-driven basis for ranking GMMs, as well as assigning weights for GMMs in a logic tree of a PSHA. The first data-driven weighting scheme was suggested by Scherbaum *et al.* (2009, their equation 14):

$$w_{\overline{\mathrm{LLH}}i} = \frac{2^{-\overline{\mathrm{LLH}}_i}}{\sum_j 2^{-\overline{\mathrm{LLH}}_j}}. \tag{10}$$

See equation (2) for $\overline{\mathrm{LLH}}_i$. Delavaud *et al.* (2012, their equation 4) further extended this to a score called data support index (DSI):

$$\mathrm{DSI}_i = 100\frac{w_{\overline{\mathrm{LLH}}i} - 1/N_\mathrm{m}}{1/N_\mathrm{m}}, \tag{11}$$

in which N_m is the total number of evaluated models. They attributed it a meaning as the empirical support for assigning a weight to model i higher than a noninformative weight (i.e., $1/N_\mathrm{m}$).

In fact, a form slightly different from equation (10) has a well-founded meaning and widespread use in Bayesian inference:

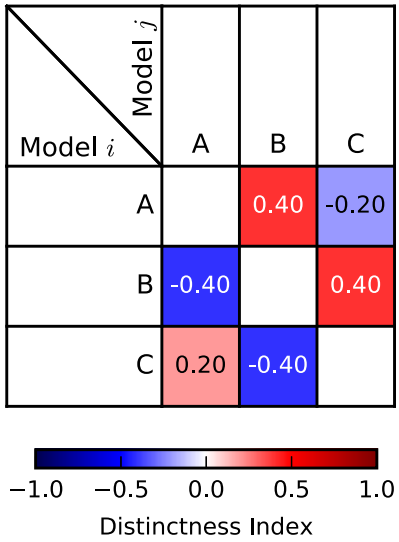

Figure 5. Distinctness table for example 9 in the Distinctness Index section. The DI of each pairwise comparison is given in the intersecting box of a model pair. The color version of this figure is available only in the electronic edition.

$$w_{\mathrm{b}i} = \frac{\ell_i}{\sum_j \ell_j}, \tag{12}$$

in which ℓ_i is the likelihood of model i given the observations. This is simply the posterior weight from a one-step Bayesian updating assuming noninformative (i.e., uniform) prior weights; weights here are taken as probabilities. Equation (12) is different from equation (10) in that the likelihood ℓ in equation (12) is replaced in equation (10) by the Nth root of the likelihood (N being the sample size); such an additional step disconnects equation (10) from the Bayesian meaning. A similar analysis was provided by Roselli *et al.* (2016, pp. 721–722). It may be a concern that because GMMs are not mutually exclusive, the simple and well-founded Bayesian approach of equation (12) for model selection is not applicable (e.g., Scherbaum *et al.*, 2009, p. 3235).

We here propose a weighting scheme that is a natural extension of the bootstrap process for assessing the variability of a score (see the Bootstrap on Hierarchical Data section). Musson (2012) considered the weight of a logic-tree branch as the probability for the branch to be better than other branches; Scherbaum and Kuehn (2011, p. 1238) phrased it slightly differently, that the weight is the degree-of-belief that the branch should be used. The weighting scheme we propose agrees well with this widely adopted interpretation of the weight of a logic-tree branch and can be understood as the implementation of the weight from a frequentist's point of view. We therefore call it the frequency weight:

$$w_i = \frac{1}{N_\mathrm{bs}}\sum_k^{N_\mathrm{bs}} \mathbf{1}\left(s_i^{(k)} = \min_j\{s_j^{(k)}\}\right), \tag{13}$$

in which $\mathbf{1}(\cdot)$ is the usual indicator function that takes the value of 1 when the bracketed statement is true, and zero

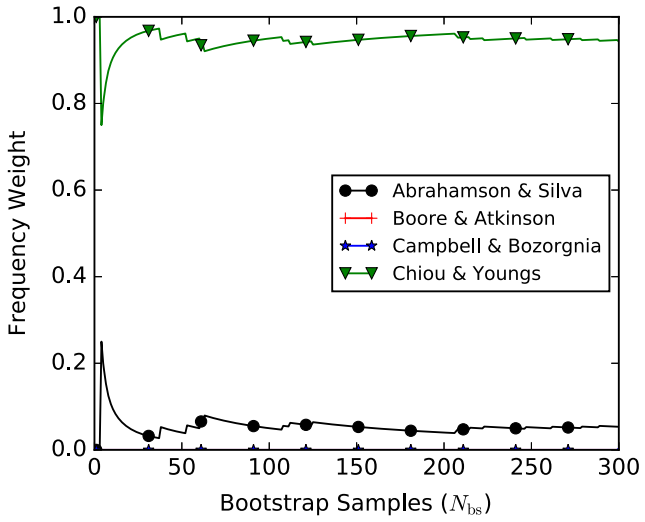

Figure 6. Frequency weights for NGA models (see example 10 in the Frequency Weight: A Data-Driven Weighting Scheme section) against N_{bs}. The color version of this figure is available only in the electronic edition.

otherwise. See equation (9) for the definitions of other notations. Note the similarity between equations (9) and (13).

The frequency weight is the relative frequency for a model to score the best, which essentially means the probability of the model to best describe the data; a model that best describes the data is naturally the model that should be used. A model selection method used by Burnham and Anderson (2002, section 2.13) is the in-sample version of the frequency weight. The frequency weight can be computed from any scores and resampling methods of choice, not only from the logarithmic score and the cluster bootstrap used in this article. The frequency weights of the models used in example 6 are given in the following example.

Example 10: Frequency Weight Using the same data and models as in example 6, Figure 6 shows the frequency weights for various choices of N_{bs}. The weights were stable for $N_{bs} \geq 50$, and so more bootstrap sampling is not necessary. Model CY scored the best (see Fig. 3) and was assigned the highest weight. Such a weight close to one means that the available data provide little empirical support to use models other than CY.

The weights of the models under different weighting schemes (equations 10, 12, and 13) are shown in Table 7. As pointed out by Arroyo *et al.* (2014, p. 1861), the weighting scheme based on the LLH (equation 10) tends to assign similar weights to the candidates even if one is known to be better than the others. The weighting scheme based on the Bayesian inference (equation 12; multivariate likelihoods were used), as well as the frequency weight proposed in the current study (equation 13; together with the multivariate logarithmic score and the cluster bootstrap), however, appeared to assign a high weight to the best model. The Bayesian weights for all models other than CY were zero, whereas the frequency weight of the model AS was nonzero, although small. We attribute this difference to the score variability taken into account by the frequency weight but not the Bayesian weight: the frequency weight can incorporate the uncommon cases that model AS scores the best. We acknowledge that the available data can seldom cover all concerns of a PSHA modeler. Therefore, by expert judgment, it is unusual to use only one GMM, regardless of the empirical evidence available.

Cautions for Scoring GMMs

We discussed two issues about scoring GMMs: (1) the treatment of data correlation and (2) the score variability. Under certain conditions, these issues may not severely affect the evaluation result. If the test data consist of a large number of earthquakes, each providing an equally small number of records (ideally only one record), then the data correlation may not affect the result much. If one GMM outperforms the others a lot, then the score variability may not affect the result much. In addition, if the GMM under evaluation fits well the test data, it is acceptable to decompose the misfits into event terms and leftover residuals, as discussed in Appendix B, and evaluate them separately.

The widely used logarithmic score can easily incorporate the correlation structure of the model, making it a desirable score. Because most modern GMMs model ground motions as correlated, such a correlation structure should be honored during the evaluation in order to be fair to the modeler. Nevertheless, the techniques we proposed to assess the score variability, namely the cluster bootstrap, the DI, and the frequency weight, work for any scores of choice.

Table 7
Weights Computed Using Different Weighting Schemes in Example 10 in the Frequency Weight: A Data-Driven Weighting Scheme Section

Model	LLH Weight*	Bayesian Weight†	Frequency Weight‡
Abrahamson and Silva (2008)	0.26	0.00	0.05
Boore and Atkinson (2008)	0.25	0.00	0.00
Campbell and Bozorgnia (2008)	0.22	0.00	0.00
Chiou and Youngs (2008)	0.27	1.00	0.94

*Equation (10).
†Equation (12); likelihoods are multivariate (values given in the last column of Fig. 3).
‡Equation (13); $N_{bs} = 300$.

The logarithmic score does not include a judgment of utility. For example, it is possible for a model that consistently overestimates the ground motion to score the same as another model that consistently underestimates. The use of the two seemingly equivalent models, however, may lead to different consequences of a PSHA project. Although scoring is an objective tool for model selection, choosing an appropriate score inevitably involves subjectivity in deciding what appropriate means.

The choice of the test data demands the evaluator to think in the context of the designated use of the GMMs under evaluation. For example, a model may clearly score better than other models, except for ground motions recorded at very soft sites. If the number of such records is small in the test data, they may not affect much the overall score. One may worry, however, if such an evaluation result is meaningful for a site-specific PSHA project on a very soft site.

The amount of test data is also critical to the usefulness of the model evaluation. Similar to the concept of statistical power in conventional hypothesis testing (e.g., Mak *et al.*, 2014b), a small dataset is less likely to reveal the difference between two models. A small dataset may also inadequately represent the between-event variability, making a bootstrap resampling less meaningful. Whether the size of a dataset is sufficient is a judgment of the evaluator.

When the logarithmic score is the only measure of model performance, and the NGA-West2 dataset (after filtered by prospectiveness, earthquake class, magnitude, and distance) is the only source of information for model performance, the Chiou and Young (2008) model is better than other NGA models under almost all situations extractable from the fixed dataset through the cluster bootstrap. This result, however, does not provide information on how much better the best model is in an engineering sense. An analog is that, suppose multiple trains, running in parallel, are identical (in terms of punctuality, reliability, comfort, etc.) except that the ticket for one of them is always 10 cents cheaper than others, then there is no logical reason to take any trains other than the cheapest one, although the value of 10 cents may not be too high for some people.

One may argue that given the epistemic uncertainty, a PSHA should not use only a single GMM. We emphasize that the empirical evaluation of GMMs relies on a predesignated metric of model performance and a fixed dataset that is assumed to provide all information for model performance, whereas the epistemic uncertainty is fundamentally about unmeasurable (or, at least, unmeasured) "unknown unknowns." The merit of scoring GMMs is that it separates the data-driven portion of model selection from the expert judgment. This is beneficial to transparently communicating uncertainties between modelers and users of a PSHA model.

Summary

- Modern GMMs are often hierarchical. To evaluate such models using a score, it is necessary for the score to duly incorporate the model hierarchy. This is both to fully utilize the information provided by the model under evaluation and to avoid potential fallacies due to the ignorance of data correlation (examples 1–4). We propose using the multivariate logarithmic score to represent the performance of hierarchical GMMs.
- Because the ground-motion observation is a random realization of the underlying data-generating mechanism, a score computed based on the observation is also a random variable. To compare the performance of two models by their scores, it is necessary to consider the variability of the score, instead of simply taking different values of the score to be truly different. Incorrect use of resampling to assess the score variability could lead to fallacy (example 5). We propose to use the distinctness index, together with the cluster bootstrap and the multivariate logarithmic score, to represent the difference between two models (example 6). Such a method of model comparison correctly extracts the information of score variability from a resampled dataset and could give a result different than that found by conventional methods used in the literature (examples 7–8). The distinctness index applies to any scores and resampling methods of choice.
- We propose to use the frequency weight (example 10) as a data-driven weighting scheme of GMMs. The frequency weight has a clear meaning as the frequentist's interpretation of the weight of a logic-tree branch and is therefore more directly linked to the practice of PSHA, compared with existing weighting schemes. The frequency weight applies to any scores and resampling methods of choice.
- Multiple models are not always unambiguously rankable (example 9).

Data and Resources

The Next Generation Attenuation-West (NGA-West2) flatfile we used in example 6 was downloaded from peer.berkeley.edu/ngawest2/databases (last accessed May 2016). The predicted ground motions of the four NGA ground-motion models were computed using the Openquake Hazard Library (v.0.20; github.com/gem/oq-hazardlib, last accessed May 2016).

Acknowledgments

We thank our colleague Fabrice Cotton as well as two anonymous reviewers, for providing useful comments. This study was supported by the Global Earthquake Model Foundation.

References

Abrahamson, N., and W. Silva (2008). Summary of the Abrahamson & Silva NGA ground-motion relations, *Earthq. Spectra* **24,** no. 1, 67–97, doi: 10.1193/1.2924360.

Abrahamson, N., and K. Wooddell (2010). Evaluation of evidence for inhibition of very strong ground motions in the Abrahamson and Silva Next Generation Attenuation ground-motion model, *Bull. Seismol. Soc. Am.* **100,** no. 5A, 2174–2184, doi: 10.1785/0120080278.

Abrahamson, N. A., and R. R. Youngs (1992). A stable algorithm for regression analyses using the random effects model, *Bull. Seismol. Soc. Am.* **82,** no. 1, 505–510.

Al Atik, L., and N. Abrahamson (2010). Nonlinear site response effects on the standard deviations of predicted ground motions, *Bull. Seismol. Soc. Am.* **100,** no. 3, 1288–1292, doi: 10.1785/0120090154.

Allen, T. I., and C. Brillon (2015). Assessment of ground-motion models for use in the British Columbia north coast region, Canada, *Bull. Seismol. Soc. Am.* **105,** no. 28, 1193–1205, doi: 10.1785/0120140266.

Arango, M. C., F. O. Strasser, J. J. Bommer, J. M. Cepeda, R. Boroschek, D. A. Hernandez, and H. Tavera (2012). An evaluation of the applicability of current ground-motion models to the South and Central American subduction zones, *Bull. Seismol. Soc. Am.* **102,** no. 1, 143–168, doi: 10.1785/0120110078.

Armstrong, J. S. (2001). Evaluating forecasting methods, in *Principles of Forecasting: A Handbook for Researchers and Practitioners*, J. S. Armstrong (Editor), Chap. 14, Kluwer Academic Publishers, Norwell, Massachusetts, 443–472.

Arroyo, D., M. Ordaz, and R. Rueda (2014). On the selection of ground-motion prediction equations for probabilistic seismic-hazard analysis, *Bull. Seismol. Soc. Am.* **104,** no. 4, 1860–1875, doi: 10.1785/0120130264.

Azarbakht, A., S. Rahpeyma, and M. Mousavi (2014). A new methodology for assessment of the stability of ground-motion prediction equations, *Bull. Seismol. Soc. Am.* **104,** no. 3, 1447–1457, doi: 10.1785/0120130212.

Beauval, C., H. Tasan, A. Laurendeau, E. Delavaud, F. Cotton, P. Guéguen, and N. Kuehn (2012). On the testing of ground-motion prediction equations against small-magnitude data, *Bull. Seismol. Soc. Am.* **102,** no. 5, 1994–2007, doi: 10.1785/0120110271.

Benedetti, R. (2010). Scoring rules for forecast verification, *Mon. Weather Rev.* **138,** 203–211, doi: 10.1175/2009MWR2945.1.

Bernardo, J. M. (1979). Expected information as expected utility, *Ann. Stat.* **7,** no. 3, 686–690.

Bindi, D., L. Luzi, F. Pacor, G. Franceschina, and R. R. Castro (2006). Ground-motion predictions from empirical attenuation relationships versus recorded data: The case of the 1997–1998 Umbria-Marche, central Italy, strong-motion data set, *Bull. Seismol. Soc. Am.* **96,** no. 3, 984–1002, doi: 10.1785/0120050102.

Bommer, J. J., J. Douglas, F. Scherbaum, F. Cotton, H. Bungum, and D. Fäh (2010). On the selection of ground-motion prediction equations for seismic hazard analysis, *Seismol. Res. Lett.* **81,** no. 5, 783–793, doi: 10.1785/gssrl.81.5.783.

Boore, D. M. (2010). Orientation-independent, nongeometric-mean measures of seismic intensity from two horizontal components of motion, *Bull. Seismol. Soc. Am.* **100,** no. 4, 1830–1835, doi: 10.1785/0120090400.

Boore, D. M., and G. M. Atkinson (2008). Ground-motion prediction equations for the average horizontal component of PGA, PGV, and 5%-damped PSA at spectral periods between 0.01 s and 10.0 s, *Earthq. Spectra* **24,** no. 1, 99–138, doi: 10.1193/1.2830434.

Boore, D. M., J. Watson-Lamprey, and N. A. Abrahamson (2006). Orientation-independent measures of ground motion, *Bull. Seismol. Soc. Am.* **96,** no. 4A, 1502–1511, doi: 10.1785/0120050209.

Bradley, B. A. (2010). NZ-specific pseudo-spectral acceleration ground motion prediction equations based on foreign models, *Research Report 2010-03*, Department of Civil Engineering, University of Canterbury, Christchurch, New Zealand.

Bröcker, J. (2012). Probability forecasts, in *Forecast Verification—A Practitioner's Guide in Atmospheric Science*, I. T. Jolliffe and D. B. Stephenson (Editors), Second Ed., Wiley-Blackwell, Chichester, United Kingdom, ISBN: 978-0-470-66071-3.

Bröcker, J., and L. A. Smith (2007). Scoring probabilistic forecasts: The importance of being proper, *Weather Forecast.* **22,** no. 2, 382–388, doi: 10.1175/WAF966.1.

Burnham, K. P., and D. R. Anderson (2002). *Model Selection and Multimodel Inference*, Second Ed., Springer, New York, New York, ISBN: 0-387-95364-7.

Campbell, K. W., and Y. Bozorgnia (2008). NGA ground motion model for the geometric mean horizontal component of PGA, PGV, PGD and 5% damped linear elastic response spectra for periods ranging from 0.01 to 10 s, *Earthq. Spectra* **24,** no. 1, 139–171, doi: 10.1193/1.2857546.

Candille, G., C. Côté, P. L. Houtekamer, and G. Pellerin (2007). Verification of an ensemble prediction system against observations, *Mon. Weather Rev.* **135,** 2688–2699, doi: 10.1175/MWR3414.1.

Chiou, B. S.-J., and R. R. Youngs (2008). An NGA model for the average horizontal component of peak ground motion and response spectra, *Earthq. Spectra* **24,** no. 1, 173–215, doi: 10.1193/1.2894832.

Davison, A. C., and D. V. Hinkley (1997). Bootstrap methods and their application, in *Cambridge Series in Statistical and Probabilistic Mathematics (No. 1)*, R. Gill, B. D. Ripley, S. Ross, M. Stein, and D. Williams (Series Editors), Cambridge University Press, Cambridge, United Kingdom, ISBN: 0521574714.

Dawid, P. (1984). Present position and potential development: Some personal views: Statistical theory: The prequential approach, *J. Roy. Stat. Soc. A* **147,** no. 2, 278–292, doi: 10.2307/2981683.

DeGroot, M. H., and M. J. Schervish (2012). *Probability and Statistics*, Fourth Ed., Pearson, Boston, Massachusetts, ISBN: 978-0-321-50046-5.

Delavaud, E., F. Scherbaum, N. Kuehn, and T. Allen (2012). Testing the global applicability of ground-motion prediction equations for active shallow crustal regions, *Bull. Seismol. Soc. Am.* **102,** no. 2, 707–721, doi: 10.1785/0120110113.

Delavaud, E., F. Scherbaum, N. Kuehn, and C. Riggelsen (2009). Information-theoretic selection of ground-motion prediction equations for seismic hazard analysis: An applicability study using Californian data, *Bull. Seismol. Soc. Am.* **99,** no. 6, 3248–3263, doi: 10.1785/0120090055.

Douglas, J., and R. Mohais (2009). Comparing predicted and observed ground motions from subduction earthquakes in the Lesser Antilles, *J. Seismol.* **13,** 577–587, doi: 10.1007/s10950-008-9150-y.

Drouet, S., and F. Cotton (2015). Regional stochastic GMPEs in low-seismicity areas: Scaling and aleatory variability analysis—Application to the French Alps, *Bull. Seismol. Soc. Am.* **105,** no. 4, 1883–1902, doi: 10.1785/0120140240.

Drouet, S., F. Scherbaum, F. Cotton, and A. Souriau (2007). Selection and ranking of ground motion models for seismic hazard analysis in the Pyrenees, *J. Seismol.* **11,** 87–100, doi: 10.1007/s10950-006-9039-6.

Edwards, B., and J. Douglas (2013). Selecting ground-motion models developed for induced seismicity in geothermal areas, *Geophys. J. Int.* **195,** no. 2, 1314–1322, doi: 10.1093/gji/ggt310.

Efron, B., and R. J. Tibshirani (1993). An introduction to the bootstrap, *Monographs on Statistics & Applied Probability (Book 57)*, Chapman and Hall/CRC, New York, New York, ISBN: 0412042312.

Field, C. A., and A. H. Welsh (2007). Bootstrapping clustered data, *J. Roy. Stat. Soc. B* **69,** no. 3, 369–390, doi: 10.1111/j.1467-9868.2007.00593.x.

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

Goda, K., and H. P. Hong (2008). Spatial correlation of peak ground motions and response spectra, *Bull. Seismol. Soc. Am.* **98,** no. 1, 354–365, doi: 10.1785/0120070078.

Good, I. J. (1952). Rational decisions, *J. Roy. Stat. Soc. B* **14,** no. 1, 107–114.

Haendel, A., S. Specht, N. M. Kuehn, and F. Scherbaum (2015). Mixtures of ground-motion prediction equations as backbone models for a logic tree: An application to the subduction zone in Northern Chile, *Bull. Earthq. Eng.* **13,** 483–501, doi: 10.1007/s10518-014-9636-7.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems, *Weather Forecast.* **15,** 559–570.

Hintersberger, E., F. Scherbaum, and S. Hainzl (2007). Update of likelihood-based ground-motion model selection for seismic hazard analysis in western central Europe, *Bull. Earthq. Eng.* **5,** 1–16, doi: 10.1007/s10518-006-9018-x.

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*, Springer, New York, New York, ISBN: 0-387-47941-4.

Jolliffe, I. T. (2007). Uncertainty and inference for verification measures, *Weather Forecast.* **22,** no. 3, 637–650, doi: 10.1175/WAF989.1.

Joyner, W. B., and D. M. Boore (1993). Methods for regression analysis of strong-motion data, *Bull. Seismol. Soc. Am.* **83,** no. 2, 469–487.

Kaklamanos, J., and L. G. Baise (2011). Model validations and comparisons of the Next Generation Attenuation of ground motions (NGA-west) project, *Bull. Seismol. Soc. Am.* **101,** no. 1, 160–175, doi: 10.1785/0120100038.

Kale, Ö., and S. Akkar (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,** no. 2A, 1069–1084, doi: 10.1785/0120120134.

Kuehn, N. M., and D. Scherbaum (2015). Ground-motion prediction model building: A multilevel approach, *Bull. Earthq. Eng.* **13,** no. 9, 2481–3491, doi: 10.1007/s10518-015-9732-3.

Lindley, D. V. (1991). *Making Decisions*, Second Ed., Wiley, Chichester, United Kingdom, ISBN: 0471908088.

Mak, S., R. A. Clements, and D. Schorlemmer (2014a). 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,** no. 6, 3139–3140, doi: 10.1785/0120140106.

Mak, S., R. A. Clements, and D. Schorlemmer (2014b). The statistical power of testing probabilistic seismic-hazard assessments, *Seismol. Res. Lett.* **85,** no. 4, 781–783, doi: 10.1785/0220140012.

Mak, S., R. A. Clements, and D. Schorlemmer (2015). Validating intensity prediction equations for Italy by observations, *Bull. Seismol. Soc. Am.* **105,** no. 6, 2942–2954, doi: 10.1785/0120150070.

Massa, M., L. Luzi, F. Pacor, D. Bindi, and G. Ameri (2012). Comparison between empirical predictive equations calibrated at global and national scale and Italian strong-motion data, *B. Geofis. Teor. Appl.* **53,** no. 1, 3753, doi: 10.4430/bgta0018.

McVerry, G., J. Zhao, N. Abrahamson, and P. Somerville (2006). New Zealand acceleration response spectrum attenuation relation for crustal and subduction zone earthquakes, *Bull. New Zeal. Soc. Earthq. Eng.* **39,** no. 1, 1–58.

Mousavi, M., A. Ansari, H. Zafarani, and A. Azarbakht (2012). Selection of ground motion prediction models for seismic hazard analysis in the Zagros region, Iran, *J. Earthq. Eng.* **16,** no. 8, 1184–1207, doi: 10.1080/13632469.2012.685568.

Mousavi, M., H. Zafarani, S. Rahpeyma, and A. Azarbakht (2014). Test of goodness of the NGA ground-motion equations to predict the strong motions of the 2012 Ahar-Varzaghan dual earthquakes in northwestern Iran, *Bull. Seismol. Soc. Am.* **104,** no. 5, 2512–2528, doi: 10.1785/0120130302.

Musson, R. (2012). On the nature of logic trees in probabilistic seismic hazard assessment, *Earthq. Spectra* **28,** no. 3, 1291–1296, doi: 10.1193/1.4000062.

Nishimura, T. (2010). Conformity of the attenuation relationships in Japan with those by the NGA-project, *Summaries of Technical Papers of Annual Meeting of Architectural Institute of Japan*, Toyama, Japan, 9–11 September 2010 (in Japanese).

Ogweno, L. P., and C. H. Cramer (2014). Comparing the CENA GMPEs using NGA-East ground-motion database, *Seismol. Res. Lett.* **85,** no. 6, 1377–1393, doi: 10.1785/0220140045.

Reiter, L. (1991). *Earthquake Hazard Analysis: Issues and Insights*, Columbia University Press, New York, New York, ISBN: 9780231065344.

Roselli, P., W. Marzocchi, and L. Faenza (2016). Toward a new probabilistic framework to score and merge ground-motion prediction equations: The case of the Italian region, *Bull. Seismol. Soc. Am.* **106,** no. 2, 720–733, doi: 10.1785/0120150057.

Rosenblatt, M. (1952). Remarks on a multivariate transformation, *Ann. Math. Stat.* **23,** no. 3, 470–472, doi: 10.1214/aoms/1177729394.

Roulston, M. S., and L. A. Smith (2002). Evaluating probabilistic forecasts using information theory, *Mon. Weather Rev.* **130,** 1653–1660.

Scasserra, G., J. P. Stewart, P. Bazzurro, G. Lanzo, and F. Mollaioli (2009). A comparison of NGA ground-motion prediction equations to Italian data, *Bull. Seismol. Soc. Am.* **99,** no. 5, 2961–2978, doi: 10.1785/0120080133.

Scherbaum, F., and N. M. Kuehn (2011). Logic tree branch weights and probabilities: Summing up to one is not enough, *Earthq. Spectra* **27,** no. 4, 1237–1251, doi: 10.1193/1.3652744.

Scherbaum, F., F. Cotton, and P. Smit (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,** no. 6, 2164–2185, doi: 10.1785/0120030147.

Scherbaum, F., E. Delavaud, and C. Riggelsen (2009). Model selection in seismic hazard analysis: An information-theoretic perspective, *Bull. Seismol. Soc. Am.* **99,** no. 6, 3234–3247, doi: 10.1785/0120080347.

Shoja-Taheri, J., S. Naserieh, and G. Hadi (2010). A test of the applicability of NGA models to the strong ground-motion data in the Iranian plateau, *J. Earthq. Eng.* **14,** 278–292, doi: 10.1080/13632460903086051.

Skrondal, A., and S. Rabe-Hesketh (2009). Prediction in multilevel generalized linear models, *J. Roy. Stat. Soc. A* **172,** no. 3, 659–687, doi: 10.1111/j.1467-985X.2009.00587.x.

Stafford, P. J. (2015). Extension of the random-effects regression algorithm to account for the effects of nonlinear site response, *Bull. Seismol. Soc. Am.* **105,** no. 6, 3196–3202, doi: 10.1785/0120140368.

Stafford, P. J., F. O. Strasser, and J. J. Bommer (2008). An evaluation of the applicability of the NGA models to ground-motion prediction in the Euro-Mediterranean region, *Bull. Earthq. Eng.* **6,** 149–177, doi: 10.1007/s10518-007-9053-2.

Strasser, F. O., N. A. Abrahamson, and J. J. Bommer (2009). Sigma: Issues, insights, and challenges, *Seismol. Res. Lett.* **80,** no. 1, 40–56, doi: 10.1785/gssrl.80.1.40.

Uchiyama, Y., and S. Midorikawa (2011). A study of the applicability of NGA models to strike-slip earthquakes in Japan, *Summaries of Technical Papers of Annual Meeting of Architectural Institute of Japan*, Tokyo, Japan, 23–25 August 2011 (in Japanese).

Vacareanu, R., F. Pavel, and A. Aldea (2013). On the selection of GMPEs for Vrancea subcrustal seismic source, *Bull. Earthq. Eng.* **11,** no. 6, 1867–1884, doi: 10.1007/s10518-013-9515-7.

Van Houtte, C. (2016). Performance of response spectral ground-motion models against New Zealand data, *Bull. New Zeal. Soc. Earthq. Eng.* **50,** no. 1.

Vilanova, S. P., J. F. B. D. Fonseca, and C. S. Oliveira (2012). Ground-motion models for seismic-hazard assessment in western Iberia: Constraints from instrumental data and intensity observations, *Bull. Seismol. Soc. Am.* **102,** no. 1, 169–184, doi: 10.1785/0120110097.

Winkler, R. L., and A. H. Murphy (1968). "Good" probability assessors, *J. Appl. Meteorol.* **7,** 751–758.

Youngs, R. R., N. Abrahamson, F. I. Makdisi, and K. Sadigh (1995). Magnitude-dependent variance of peak ground acceleration, *Bull. Seismol. Soc. Am.* **85,** no. 4, 1161–1176.

Zafarani, H., and M. Mousavi (2014). Applicability of different ground-motion prediction models for northern Iran, *Nat. Hazards* **73,** no. 3, 1199–1228, doi: 10.1007/s11069-014-1151-2.

Appendix A

Computing the Covariance Matrix of the Multivariate Normal Distribution

Computing the multivariate logarithmic score (equation 7) requires first computing the covariance matrix **V** (equation 6). Two examples are given here. For simplicity, the synthetic data for both examples contain only two events, and they produced two and three records, respectively.

Example A1 This example addresses the situation in which the between-event sigma ($\sigma_b^{(i)}$ for event i) is constant

within an event. The within-event sigma ($\sigma_w^{(ij)}$ for event i and record j) could be different for each record, due to the effect of the site amplification. Boore and Atkinson (2008) and Campbell and Bozorgnia (2008) are examples of ground-motion models (GMMs) with this structure.

Suppose $\{\sigma_b^{(1)},\sigma_b^{(2)}\}=\{0.3,0.32\}$ and $\{\sigma_w^{(11)},\sigma_w^{(12)},\sigma_w^{(21)},\sigma_w^{(22)},\sigma_w^{(23)}\}=\{0.42,0.43,0.44,0.45,0.47\}$. By equation (6):

$$\mathbf{R}=\begin{pmatrix}0.42^2&0&0&0&0\\0&0.43^2&0&0&0\\0&0&0.44^2&0&0\\0&0&0&0.45^2&0\\0&0&0&0&0.47^2\end{pmatrix}$$

$$\mathbf{Z}'=\begin{pmatrix}1&1&0&0&0\\0&0&1&1&1\end{pmatrix}$$

$$\mathbf{G}=\begin{pmatrix}0.3^2&0\\0&0.32^2\end{pmatrix}$$

$$\mathbf{V}=\mathbf{R}+\mathbf{Z}\mathbf{G}\mathbf{Z}'=\begin{pmatrix}0.2664&0.09&0&0&0\\0.09&0.2749&0&0&0\\0&0&0.296&0.1024&0.1024\\0&0&0.1024&0.3049&0.1024\\0&0&0.1024&0.1024&0.3233\end{pmatrix}. \quad \text{(A1)}$$

Example A2 This example addresses the situation when both the within-event sigma and the between-event sigma for each record are different. Abrahamson and Silva (2008) and Chiou and Youngs (2008) are examples of GMMs with this structure. The major reason for records of the same event to have different between-event sigmas is that the uncertainty of the nonlinear site amplification propagates to the between-event sigma (Al Atik and Abrahamson, 2010).

The treatment in example A1 requires explicitly specifying the matrix $\mathbf{Z}$, which has the advantage of clearly spelling out the hierarchical structure of the model (equation 5). In practice, however, GMM predictions often provide only the sigma values (e.g., when the prediction is computed by published computer codes). Therefore, we provide here an example of calculation that requires only $\sigma_w^{(ij)}$ and $\sigma_b^{(ij)}$ but not explicitly $\mathbf{Z}$. The treatment in this example applied also to the case of constant sigmas within an event.

Suppose $\{\sigma_b^{(11)},\sigma_b^{(12)},\sigma_b^{(21)},\sigma_b^{(22)},\sigma_b^{(23)}\}=\{0.3,0.31,0.32,0.32,0.33\}$ and $\{\sigma_w^{(11)},\sigma_w^{(12)},\sigma_w^{(21)},\sigma_w^{(22)},\sigma_w^{(23)}\}=\{0.45,0.46,0.47,0.47,0.48\}$:

$$\mathbf{R}=\begin{pmatrix}0.45^2&0&0&0&0\\0&0.46^2&0&0&0\\0&0&0.47^2&0&0\\0&0&0&0.47^2&0\\0&0&0&0&0.48^2\end{pmatrix}$$

$$\mathbf{Z}_g'=\begin{pmatrix}0.3&0.31&0&0&0\\0&0&0.32&0.32&0.33\end{pmatrix}$$

$$\mathbf{V}=\mathbf{R}+\mathbf{Z}_g\mathbf{Z}_g'=\begin{pmatrix}0.2925&0.093&0&0&0\\0.093&0.3077&0&0&0\\0&0&0.3233&0.1024&0.1056\\0&0&0.1024&0.3233&0.1056\\0&0&0.1056&0.1056&0.3393\end{pmatrix}. \quad \text{(A2)}$$

These two examples demonstrate that $\mathbf{V}$ is often sparse, because records from different events are often modeled as uncorrelated. If the data size is large, the computation will be more efficiently handled by numerical methods designed for sparse matrices. The python implementation of equation (7), provided in the Ⓔ electronic supplement to this article, uses sparse matrix operations.

Appendix B

Estimating Event Terms

The event term of an earthquake (i.e., η_i in equation 4) is needed for various purposes, such as using residual analysis to evaluate GMMs, a popular method (see Table 1). The event term (and the corresponding leftover residuals) is not directly observed and has to be estimated based on a prescribed model. In the literature of GMM evaluation, we found two ways to estimate an event term.

Expected Random Effects The event term η_i for earthquake i (with N_i records) was taken as

$$\eta_i=\frac{\sigma_b^2}{\sigma_b^2+\sigma_w^2/N_i}\left[\frac{1}{N_i}\sum_j^{N_i}(y_{ij}-p_{ij})\right] \quad \text{(B1)}$$

(Abrahamson and Youngs, 1992, their equation 10; implemented by, e.g., Bindi *et al.*, 2006; Bradley, 2010; Azarbakht *et al.*, 2014; Van Houtte, 2016). See equation (4) for the definitions of the notations. This equation comes from the expected value of random effects given the observations and the model parameters (Jiang, 2007, pp. 74–75). The general case is:

$$\boldsymbol{\eta}=\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{p}), \quad \text{(B2)}$$

where $\boldsymbol{\eta}$ is a vector of event terms. See equations (5) and (6) for the definitions of other notations. This estimation is the best linear unbiased predictor for random effects when the model is fitted to the data (Skrondal and Rabe-Hesketh, 2009, section 4.2). For constant σ_b and σ_w, it becomes equation B1. The fundamental assumption for this estimation of event terms is that the model is calibrated (or fitted) to the data so that residuals purely consist of random errors but not systematic bias. For out-of-sample analysis (i.e., evaluation using a dataset other than the one for the model development), the GMM under evaluation almost always does not fit, and so this estimation loses its rigorous meaning.

Mean Residual The event term for an earthquake was taken as the mean residual (e.g., Scasserra *et al.*, 2009; Shoja-Taheri *et al.*, 2010, their equations 1–4; Uchiyama and Midorikawa, 2011, their equations 1–2; Vacareanu *et al.*, 2013), probably by assuming that the zero-mean leftover residuals will cancel out each other through the summation. This estimation is actually identical to equation (B1) for large N_i. Similar to the expected random effects, the implicit assumption for this estimation is that the model fits the data.

For an out-of-sample analysis, the model may not fit the data. The residual is therefore a combination of the event term, the leftover residual, and the model bias. An example of the model bias is a wrong attenuation rate. This happens, for example, when a GMM designed for a warm-crust region (e.g., California) is applied to a cold-crust region (e.g., eastern United States). The components of the residual cannot be accurately separated unless the bias is precisely known. Attempting to estimate the event term will allocate part of the model bias to the event term and the remaining to the leftover residual; such separation depends on the features of the bias, which is in general unknown. Some of the residual analyses conducted for GMM evaluations (see Table 1) involved estimating event terms. The estimation was not strictly valid unless the evaluated model has been fitted to the test data; their results should therefore be interpreted with care.

Appendix C

Further Strategies for Bootstrap on Hierarchical Data

Two additional strategies for bootstrap on hierarchical data have been proposed in the literature. The terminology used here follows Field and Welsh (2007).

Two-Stage Bootstrap Davison and Hinkley (1997) called this "strategy 2." This is identical to the cluster bootstrap (see the Bootstrap on Hierarchical Data section) except that after sampling with replacement the first level, the second level is also sampled with replacement. For in-sample analysis, both Davison and Hinkley (1997, p. 101) and Field and Welsh (2007, Section 3.4) found this method inferior to the cluster bootstrap because it demands a larger amount of data to achieve consistency. For out-of-sample bootstrap, in the ideal case that the new data are probabilistically identical to the data that the model has been fitted to, the result of their analyses carries over, and so the two-stage bootstrap is not as good as the cluster bootstrap.

Random-Effect Bootstrap Davison and Hinkley (1997) described but did not name this strategy. For a hierarchical structure as described by equation (4), this strategy separately resamples random effects and leftover residuals, using the following three-step procedure:

1. form an empirical distribution for the event terms η_i and an empirical distribution for the leftover residuals ϵ_{ij};
2. resample the event terms and leftover residuals using the two empirical distribution functions;
3. combine the sampled event terms and residuals according to equation (4) to form a set of resampled data.

Step 1 requires estimating the event terms. Because of the complications of this estimation (see Appendix B), we consider this the random-effect bootstrap less desirable than the cluster bootstrap.

The simulation study in Appendix D also shows that the cluster bootstrap is a better choice than the two-stage bootstrap and the naive bootstrap.

Appendix D

A Simulation Study on Bootstrap Strategies on Hierarchical Data

Bootstrap utilizes the intrinsic variability of a fixed dataset to assess the true variability of the sample space. A good bootstrap strategy should reproduce, as close as possible, the true variability of the sample space. We here present a simulation study of the performance of the cluster bootstrap, the two-stage bootstrap, and the naive bootstrap. The former two are described in the Bootstrap on Hierarchical Data section and Appendix C, respectively. The naive bootstrap simply samples with replacement a given dataset to produce a resampled dataset of the same size, without special treatment of the hierarchical structure. The multivariate logarithmic score can be computed by bookkeeping which resampled records are from the same earthquake.

If the data-generating mechanism is known, a parametric simulation (Davison and Hinkley, 1997, chapter 2.2) can replace the nonparametric bootstrap to generate sample datasets for the computation of the frequency weight. If the event terms and the leftover residuals of a given dataset are known, and we know that they are both normally distributed, then we can use the given event terms and leftover residuals to fit two normal models, and then perform a parametric simulation using the two fitted models to generate further event terms and leftover residuals. A sample dataset is then formed by combining the generated event terms and leftover residuals according to equation (4).

As explained in Appendix B, the event terms and the leftover residuals are not observed in real-world data. We

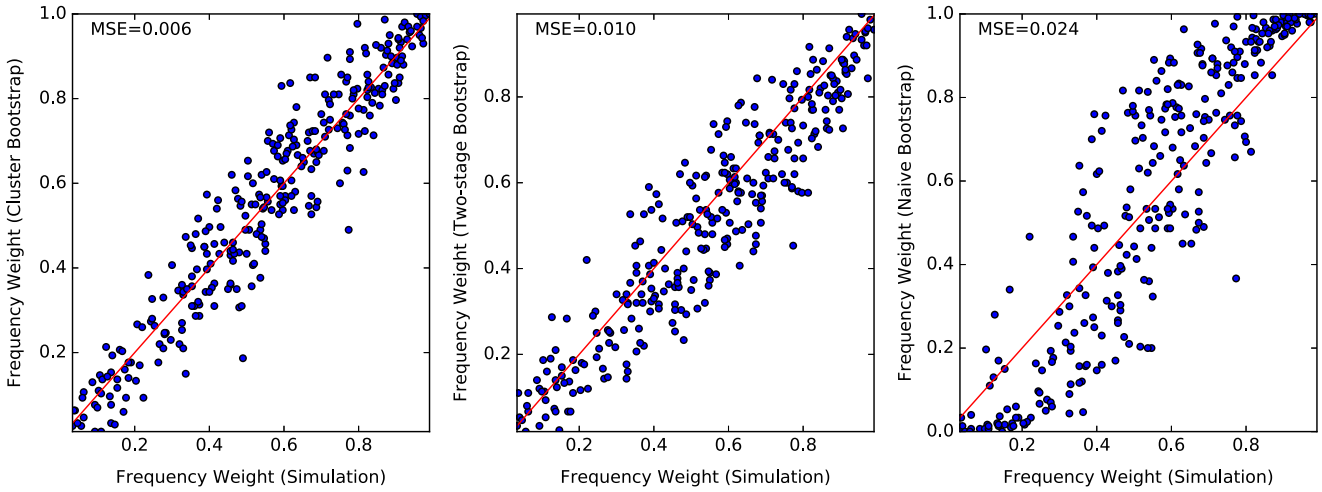

Figure D1. Frequency weights for model A (see Table D1 and Appendix D), computed from three bootstrap strategies and the parametric simulation. The mean square error (MSE) was calculated from the 1:1 line. The color version of this figure is available only in the electronic edition.

Table D1
Models Used in Appendix D

Model	p_{ij}	σ_b	σ_w
A	0.15	0.3	0.5
B	0.15	0.35	0.6
C	−0.15	0.25	0.45
D	−0.15	0.4	0.65

See equation (4) for the hierarchical structure.

therefore in practice can only use a bootstrap, instead of a parametric simulation, to generate sample datasets for the computation of the frequency weight. A good bootstrap strategy should result in a frequency weight similar to that produced by the parametric simulation. We investigated which of the three bootstrap strategies can better fulfill this task.

Four hierarchical models, in the form of equation (4), are given in Table D1. We calculated their frequency weights based on a mildly unbalanced synthetic dataset (with both the event terms and leftover residuals known) that consisted of 15 earthquakes, each having $i + 4$ recordings, in which $i = 1, 2, \ldots, 15$ is the earthquake index. Each bootstrap or parametric simulation used 300 bootstrap samples. The process was repeated for 300 synthetic datasets, generated using $p_{ij} = 0$, $\sigma_b = 0.3$, and $\sigma_w = 0.5$. The result (Fig. D1) shows that the cluster bootstrap was the most similar to the parametric simulation. The naive bootstrap, which did not honor data correlation, was the least similar to the parametric simulation. This study supports our assertion that the cluster bootstrap is a better bootstrap strategy for the purpose of our study.

Helmholtz-Zentrum Potsdam Deutsches GeoForschungsZentrum (GFZ)
Section 2.6
Helmholtzstraße 6
14467 Potsdam, Germany
smak@gfz-potsdam.de

Manuscript received 18 July 2016;
Published Online 21 February 2017