Mak, S., Cotton, F., Gerstenberger, M., Schorlemmer, D. (2018): An Evaluation of the Applicability of NGA-West2 Ground-Motion Models for Japan and New Zealand. - *Bulletin of the Seismological Society of America*, *108*, 2, pp. 836—856.

DOI: http://doi.org/10.1785/0120170146

Bulletin of the Seismological Society of America, Vol. XX, No. XX, pp. –, – 2018, doi: 10.1785/0120170146

Ⓔ

An Evaluation of the Applicability of NGA-West2 Ground-Motion Models for Japan and New Zealand

by Sum Mak, Fabrice Cotton, Matthew Gerstenberger, and Danijel Schorlemmer

Abstract We compared the accuracies of the probabilistic predictions of strong ground motions made by ground-motion models (GMMs) using the observed ground motions from 13 Japanese and 14 New Zealand shallow crustal earthquakes with moderate-to-large magnitude (5.5–6.6 for Japan and 5.07–7.85 for New Zealand). The data are independent of the GMMs so only the predictive power, instead of the explanatory power, of the models is evaluated. We examined the performance gains of state-of-the-art GMMs developed under the Next Generation Attenuation-West2 (NGA-West2) project over widely adopted regional GMMs for Japan and New Zealand. The large global dataset used by NGA-West2 GMMs allows sophisticated modeling, whereas the regional datasets used by regional GMMs may more directly represent region-specific ground-motion features. We measured the model performance by a newly developed method based on the multivariate logarithmic score, an extension of the widely used univariate logarithmic score (LLH) method. Our method measures the relative performance of models, taking into account the effects of data correlation, unbalanced data, and result variability. For the Japan case, we evaluated the model predictions for peak ground velocity (PGV) and found that NGA-West2 GMMs unambiguously performed better than regional GMMs and the superseded NGA GMMs. Proposed regional optimizations implemented in NGA-West2 GMMs improved the predictions for some models but had adverse effects for others. For the New Zealand case, we evaluated the model predictions for peak ground acceleration (PGA) and spectral accelerations at 0.3, 1, and 3 s and found that a recently developed regional GMM performed well, but NGA-West2 GMMs with performance comparable to or better than the regional model can also be identified. There appears to be no general answer as to whether a regional or global model should be preferred or whether a newer model is always better than the superseded model. This highlights the importance of evaluating the predictive power of GMMs using independent data.

Electronic Supplement: Tables of all data and metadata necessary to reproduce the case study for Japan.

Introduction

Empirical evaluation of ground-motion models (GMMs; also known as ground-motion relations, attenuation relationships, and ground-motion prediction equations) has attracted much attention in the recent decade (see Mak, Clements, and Schorlemmer, 2017, their table 1). It provides empirical support for expert judgment for selecting suitable GMMs to be used in a seismic hazard model. One dilemma often encountered by modelers is the decision of whether to use a GMM derived from a large dataset of global strong motions or a GMM derived from a regional dataset of smaller size. Some studies conclude that ground motions of engineering interest do not show regional differences (see a review by Douglas, 2011). Specifically, ground motions from small earthquakes often show regional differences, but small earthquakes slightly affect the seismic hazard. The regional difference of crustal attenuation (i.e., the Q factor) is a physical reality, but its effect in the near field, which usually dominates the hazard calculation, is often negligible. Nevertheless, it is not uncommon for national seismic hazard maps to be predominately based on regional GMMs (e.g., Japan, Fujiwara *et al.*, 2009; Italy, Stucchi *et al.*, 2011; New Zealand, Stirling *et al.*, 2012; and Taiwan, Wang *et al.*, 2016).

The advent of the Next Generation Attenuation-West2 (NGA-West2) GMMs provides a good opportunity to revisit this issue. The NGA-West2 database (Ancheta *et al.*, 2014), from which the GMMs are derived, contains 600 earthquakes, of which only 5 are from Japan and 3 from New Zealand. It is, therefore, possible to assemble sets of Japanese and New Zealand strong-motion data that are completely independent (i.e., data not used in developing the GMMs) of the NGA-West2 GMMs. For both Japan and New Zealand, there are widely used national GMMs derived mainly using regional data. This study compares the performances between NGA-West2 and regional models using independent data, which assess the predictive instead of the explanatory power of a model (Mak, Cotton, and Schorlemmer, 2017). The result could inform decisions on how these GMMs could be used in the two regions and can be qualitatively generalized to understand the advantages (or disadvantages) of using global GMMs over regional ones.

Besides the dichotomy of regional and global GMM, global GMMs with regional modifications (e.g., Scasserra *et al.*, 2009) have been developed, in the hopes of capturing the advantages of both ends. NGA-West2 GMMs provide both a generic version and a regional version specifically optimized for Japan. This study also evaluates the performance gain of such regional optimizations with respect to their corresponding generic versions, using independent data.

Table 1
Ground-Motion Models (GMMs) Evaluated

Model	ID	Peak*	Origin	Predecessor
Molas and Yamazaki (1995)	1995MY	Larger2	Japan	
Si and Midorikawa (1999)†,‡	1999SM	Larger2	Japan	
Midorikawa and Ohtake (2002)‡	2002MO	Larger2	Japan	1999SM
Kanno *et al.* (2006)§	2006KNM	vec	Japan	
Abrahamson and Silva (2008)	2008AS	GMRotI50	NGA	
Boore and Atkinson (2008)	2008BA	GMRotI50	NGA	
Campbell and Bozorgnia (2008)	2008CB	GMRotI50	NGA	
Chiou and Youngs (2008)	2008CY	GMRotI50	NGA	
Abrahamson *et al.* (2014)	2014ASK	RotD50	NGA-West2	2008AS
Boore *et al.* (2014)	2014BSS	RotD50	NGA-West2	2008BA
Campbell and Bozorgnia (2014)	2014CB	RotD50	NGA-West2	2008CB
Chiou and Youngs (2014)	2014CY	RotD50	NGA-West2	2008CY
McVerry *et al.* (2006)	2006M	GeoMean	New Zealand	Abrahamson and Silva (1997)
Bradley (2013)	2013B	GMRotI50	New Zealand	Chiou *et al.* (2010)

*Definition of peak motion. Larger2, larger of the two peak horizontal values; vec, peak of the vector sum of the two horizontal time histories; GMRotI50, orientation-independent geometric mean (Boore *et al.*, 2006); NGA, Next Generation Attenuation; RotD50, orientation-independent non-geometric-mean measure (Boore, 2010); GeoMean, geometric mean of the two peak horizontal values.
†Sigma values not completely reported; assumed to be the same as those of 2002MO.
‡Ground motions at basement ($V_S = 600$ m/s) was converted to ground motion at surface using Fujimoto and Modorikawa (2006).
§Sigma values not completely reported; see Data and Resources.

Ground-Motion Models

Japan

The 2014 version of the National Seismic Hazard Maps for Japan (hereafter, NSHMJ14; see Data and Resources) forecasts the seismic hazard in Japan Meteorological Agency intensity, which is based on peak ground velocity (PGV; see equation 7.2-7 of NSHMJ14). PGV is, therefore, a widely used intensity measure in seismic hazard analysis in Japan. In this study, we evaluated the performance of the PGV predictions of 12 GMMs, including 4 NGA GMMs, 4 NGA-West2 GMMs, and 4 GMMs developed using only or mainly Japanese data (hereafter, Japanese models). These GMMs are hereafter referred to as their IDs given in Table 1. The four NGA-West2 GMMs also provide regional versions: Abrahamson *et al.* (2014; ID: 2014ASK) provide the coefficients for shallow site effect, anelastic attenuation, and within-event sigma optimized for Japan; Boore *et al.* (2014; ID: 2014BSS) and Campbell and Bozorgnia (2014; ID: 2014CB) provide the anelastic attenuation coefficient optimized for low-Q regions, such as Japan, and a term for the Japan basin; and Chiou and Youngs (2014; ID: 2014CY) provide the coefficients for site effects (shallow and deep), anelastic attenuation, and within-event sigma optimized for Japan. We evaluated both the generic and Japan-optimized versions of these four GMMs; the IDs of the latter are appended by jp (e.g., 2014ASK becomes 2014ASKjp).

The four Japanese GMMs are:

Molas and Yamazaki (1995; ID: 1995MY). As far as we are aware, this is the first Japanese GMM that reported both the within-event and between-event sigmas. Among the selected Japanese GMMs, this is the only one that accounts for the site effect using site class.

Si and Midorikawa (1999; ID: 1999SM). This is a widely used and important GMM for Japan because it has been adopted by the NSHMJ14 as the default GMM (their equation 7.2-1); it receives much less attention outside of Japan, probably because it was published in Japanese. It reported only the total sigma, not separately the within-event and between-event sigmas, although it was based on a two-stage regression, so the hierarchy of data (see the Method section for further explanations) has been addressed. Because only the total sigma enters the NSHMJ14 (its equations 7.3-1 and 7.3-2; this is a common practice in probabilistic seismic hazard analysis), 1999SM is sufficient for the use of hazard modeling. Its performance in an empirical evaluation using hierarchical observations, however, will be limited because its correlation structure (for further

explanations, see the Method section) cannot be fully assessed. Because of its importance, we still included it in our comparison, assuming its within-event and between-event sigmas were the same as those of its successor, Midorikawa and Ohtake (2002; ID: 2002MO). Two supplements to this model were made to account for the special attenuation of deep earthquakes in northeast and southwest Japan (Morikawa *et al.*, 2003, 2006, equations 7.2-3 and 7.2-4 of NSHMJ14). Because our testing dataset included only shallow earthquakes, these two supplementary components were not involved in the current study.

2002MO. This is a successor of 1999SM, developed by the same research group, using the same method and a slightly expanded dataset. It is, however, not as widely used in Japan as its predecessor.

Kanno *et al.* (2006; ID: 2006KNM): This is one of the best-known Japanese GMMs outside of Japan. It is one of the very few modern GMMs for active shallow crust that uses the hypocentral distance, instead of some kind of rupture-plane distance, as the source-station distance measure. As far as we are aware, 2006KNM is the first Japanese GMM that includes a site correction term based on V_{S30}.

We inspected Horike and Nishimura (2004) but did not include it in this study because it did not report sigma values. Both 1999SM and 2002MO predict PGV at the basement ($V_S = 600$ m/s). Because the observed ground motion always includes a shallow site effect, to meaningfully compare GMMs with observations, an additional site correction is needed. This complicates the comparison because the site correction model itself will affect the result. We implemented Fujimoto and Modorikawa (2006) as the site correction (based on V_{S30}) in this study because it has been adopted by the NSHMJ14 (their equation 7.2-2). We added the reported variance of the site correction model to the predicted within-event variance of 2002MO (and therefore also 1999SM).

New Zealand

For New Zealand, we compared the performance of the four NGA GMMs, the four NGA-West2 GMMs, and two GMMs specifically designed for New Zealand. These GMMs are hereafter referred to as their IDs given in Table 1. The NGA-West2 GMMs do not include regional optimization for New Zealand, and so we used the generic versions. We evaluated the performance of the GMMs for the predictions of peak ground acceleration (PGA) and spectral accelerations (SAs) at 0.5, 1, and 3 s. The upper limit of 3 s was selected because it is the upper limit of one of the selected models (McVerry *et al.*, 2006). The SA at 0.5 s has a special role in the design of earthquake-resistant structures in New Zealand because it has a direct relation to the hazard factor that represents the regional hazard (see equation 3.1(1) of NZS 1170.5:2004, 2004, and section C3.1.4 of NZS 1170.5 Supp 1:2004, 2004).

The two New Zealand GMMs are:

McVerry *et al.* (2006; ID: 2006M). The National Seismic Hazard Model for New Zealand (Stirling *et al.*, 2012; hereafter, NSHMNZ) is entirely based on this GMM. This GMM predicts the ground motions for both crustal and subduction earthquakes; we evaluated only the portion about crustal earthquakes. This GMM was developed based on Abrahamson and Silva (1997). McVerry *et al.* (2006) used the same coefficients for magnitude scaling, reverse-fault adjustment, and hanging-wall adjustment as Abrahamson and Silva (1997) but adjusted some other coefficients, including those for geometric spreading and site amplification (based on site classes), using local data. They also simplified the functional form of the magnitude scaling of Abrahamson and Silva (1997) and added additional terms for anelastic attenuation, volcanic paths, and normal-fault adjustment. We omitted the attenuation term for volcanic paths for this model (also for Bradley, 2013; ID: 2013B) in our study because our data contain very few records with volcanic paths.

2013B. This GMM is based on Chiou *et al.* (2010), an extension of Chiou and Youngs (2008; hereafter, 2008CY) to small earthquakes ($3 \le M_w \le 5.5$). Bradley (2013) modified the coefficients of Chiou *et al.* (2010) in five aspects, namely the scaling for small earthquakes, the factor for normal faulting, the amplification for hard-rock sites, the coefficient for anelastic attenuation, and an additional attenuation for volcanic paths. Unlike McVerry *et al.* (2006), Bradley (2013) did not implement the modifications through refitting the coefficients using local data but by manually adjusting the coefficients based on a residual analysis.

Strong-Motion Data

Japan

We obtained PGVs recorded by K-NET and KiK-net surface stations (see Data and Resources). The criteria for selecting earthquakes were:

1. We intended to use only independent data because only the predictive power, instead of the explanatory power, of a GMM is meaningful for its application in seismic hazard assessment. For the selected Japanese GMMs, this means earthquakes no earlier than 2004. For NGA and NGA-West2 models, this means any Japanese earthquakes except for the five events included in the NGA-West2 database (Ancheta *et al.*, 2014). Both of these two criteria were used in selecting data.
2. Small ($M_w < 5.5$) earthquakes are less likely to produce ground motions of engineering interest and were discarded. This study focused on active-shallow-crustal earthquakes. Noncrustal and deep (focal depth > 25 km) earthquakes were discarded. The natural minimum magnitude to use should be 5 because this is what is used in NSHMJ14. The criterion 5 (see below), however, practically set the minimum magnitude to be 5.5. Three of the four evaluated Japanese GMMs (1999SM, 2002MO, and 2006KNM) also did not consider earthquakes smaller than 5.5. Therefore, we set the minimum magnitude to be 5.5.

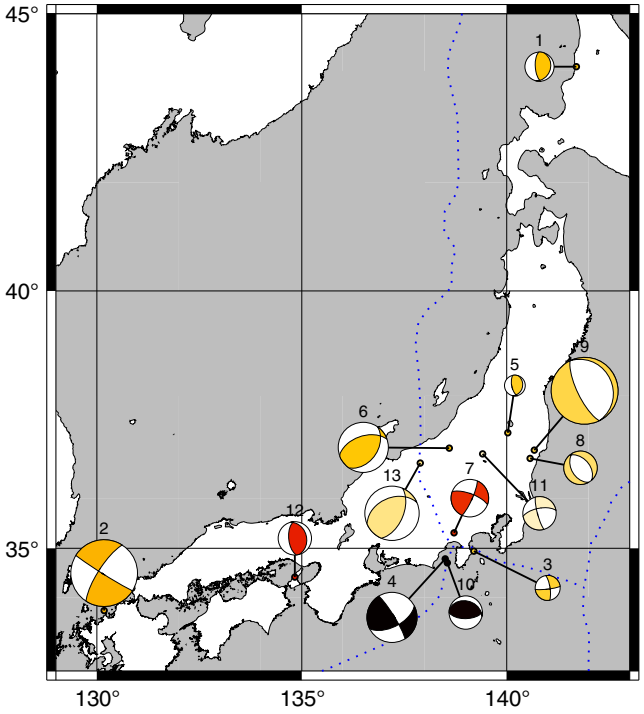

Figure 1. Selected Japanese earthquakes. An epicenter is denoted by a dot connected to a focal mechanism plot (see Ⓔ Table S1, for the focal parameters). The ID of an earthquake (Table 2) is given above the corresponding focal mechanism plot. Size of the focal mechanism plot scales with the magnitude (see Table 2). Darker color refers to deeper events (see Table 2). The plate boundaries shown (dotted lines) are from Bird (2003; see also Data and Resources). The color version of this figure is available only in the electronic edition.

3. Earthquakes that produced less than eight observations within 40 km to the epicenter were discarded. We wanted to compile a dataset with a large amount of near-field observations because near-field ground motions are usually the most important for seismic hazard analysis.
4. Clustered events were manually identified. Only the largest event within a cluster (assumed to be the mainshock) was taken. Some of the selected GMMs are derived from mainshocks only; also, seismic hazard assessments often consider only mainshocks. Some studies showed that ground motions generated from aftershocks were different from those from mainshocks (e.g., Chiou and Youngs, 2008, p. 179). Excluding aftershocks avoids this complication.
5. Earthquakes with no reported finite-fault rupture models were discarded. Most selected GMMs required the shortest distance from the rupture plane (i.e., the rupture distance R_{rup}) as an input; a finite-fault rupture model is necessary to compute this input. The difference between the rupture distance and the hypocentral distance could occasionally be more than 100% in the near field for earthquakes larger than upper 5, so a point-source assumption is not appropriate.

The 13 selected earthquakes are shown in Table 2 and Figure 1. We used only records with rupture distance less than 120 km. We divided the records into bins by the rupture distance and evaluated the performance of the GMMs in each bin separately; each bin had a width of 40 km. It is not uncommon that the performance of a GMM changes with distance (e.g., Kaklamanos and Baise, 2011); we inspected this potential change by binning the data. We did not bin the data by magnitude for two reasons. First, the magnitude range (5.5–6.6) is of significant engineering interest and sufficiently narrow. Second, the number of earthquakes falling into each magnitude bin would be so small that the GMM performance in each bin would not be representative.

The distributions of the numbers of records with respect to magnitude, distance, and V_{S30} are given in Figure 2. The

Table 2
Selected Japanese Earthquakes (See Also Fig. 1)

ID	Date and Time (yyyy/mm/dd hh:mm:ss.ss)	Location	Longitude (°)	Latitude (°)	Depth (km)	M_w	Rupture Plane
1	2004/12/14 14:56:10.54	Rumoi, Hokkaido	141.70	44.08	8	5.7	Maeda and Sasatani (2009)*
2	2005/03/20 10:53:40.32	Offshore Western Fukuoka-ken	130.18	33.74	9	6.6	Asano and Iwata (2006)*
3	2006/04/21 02:50:39.51	Offshore Eastern Izu Peninsular	139.20	34.94	7	5.6	Asano and Iwata (2006)*
4	2009/08/11 05:07:05.74	Suruga Bay	138.50	34.79	23	6.2	Aoi *et al.* (2010)*
5	2010/09/29 16:59:55.98	Fukushima-ken Nakadori	140.03	37.28	7	5.5	Kobayashi *et al.* (2011, their table 1)
6	2011/03/12 03:59:15.62	Northern Nagano-ken	138.60	36.99	8	6.2	JMA*
7	2011/03/15 22:31:46.34	Eastern Shizuoka-ken	138.71	35.31	14	5.9	Fujita *et al.* (2013, their table 1)
8	2011/03/19 18:56:48.06	Northern Ibaraki-ken	140.57	36.78	5	5.8	
9	2011/04/11 17:16:12.02	Fukushima-ken Hamadori	140.67	36.95	6	6.6	Fukushima *et al.* (2013)*
10	2011/08/01 23:58:11.04	Suruga Bay	138.55	34.71	23	5.8	JMA*
11	2013/02/25 16:23:53.58	Northern Tochigi-ken	139.41	36.87	2	5.8	Hikima (2014)
12	2013/04/13 05:33:17.75	Awaji Island	134.83	34.42	14	5.8	JMA*
13	2014/11/22 22:08:17.90	Kamishiro fault, Nagano-ken	137.89	36.69	4	6.3	GSI (2015, their fig. 11)*

Time from Japan Meteorological Agency (JMA) catalog (see Data and Resources), in Japan Standard Time. Moment magnitude (M_w) from F-net (see Data and Resources).

*See Data and Resources.

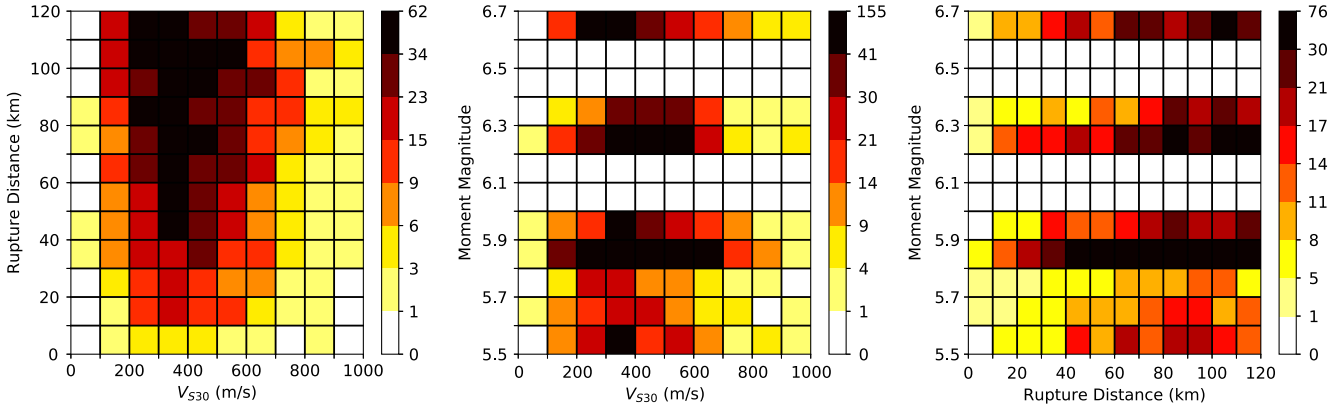

Figure 2. Distributions of the numbers of Japanese records with respect to magnitude, distance, and V_{S30}. The discrete scale was chosen so that approximately the same number of bins fall into each nonzero range. Records with $V_{S30} > 1100$ m/s are few and not shown. The color version of this figure is available only in the electronic edition.

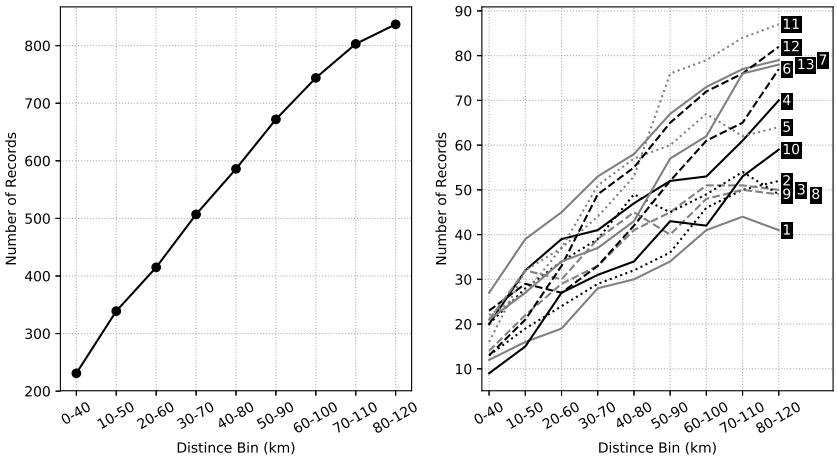

Figure 3. (Left) The number of records in each distance bin. (Right) The number of records in each distance bin for each earthquake. The earthquake IDs (see Table 2) are given at the end of each line.

number of records for each earthquake in each distance bin is given in Figure 3. The dataset, together with a description of how the metadata were obtained, is provided in Ⓔ Tables S1–S19, available in the electronic supplement to this article.

New Zealand

We used the ground-motion observations and metadata from the New Zealand Strong-Motion Database (Kaiser *et al.*, 2017; Van Houtte *et al.*, 2017; see Data and Resources). We used only crustal earthquakes with $M_w > 5$ (making the magnitude range 5.07–7.85). The lower limit of 5 is the same as the minimum magnitude considered in NSHMNZ. It is also the same as the lower limit of magnitude used in Van Houtte (2017), so that we can compare our result directly with theirs (see Comparison with Van Houtte (2017)). Similar to our treatment of the Japanese data, we manually identified clustered events in space and time and used only the largest event in a cluster. We made the dataset prospective to all NGA/NGA-West2 GMMs, as well as 2013B, by excluding events in our dataset that were also used in developing those models. The resulting dataset contains 14 earthquakes with 472 records at distances smaller than 120 km (for PGA; slightly less for other spectral periods; Table 3 and Fig. 4). The distributions of the numbers of records with respect to magnitude, distance, and V_{S30} are given in Figure 5.

Just as what we did for the Japanese data, we divided the New Zealand data into distance bins; the performance of the GMMs was evaluated separately for the three groups. The number of records for each earthquake in each distance bin is given in Figure 6.

Method

In this study, we measured the relative performance of GMMs by following a three-step approach:

1. measure the performance of a GMM by the multivariate logarithmic score (hereafter, mvLogS);
2. measure the variability of the scores of a set of GMMs by the cluster bootstrap (for each pair of models, summarize the bootstrap results as the distinctness index, hereafter DI, that represents the relative performance of the two models); graphically displays the DIs of all model pairs as a distinctness table;
3. based on the DIs, rank the GMMs.

The first two steps follow the scoring approach proposed by Mak, Clements, and Schorlemmer (2017). The model rank, although providing less complete information than the distinctness table, was used in this article as the primary

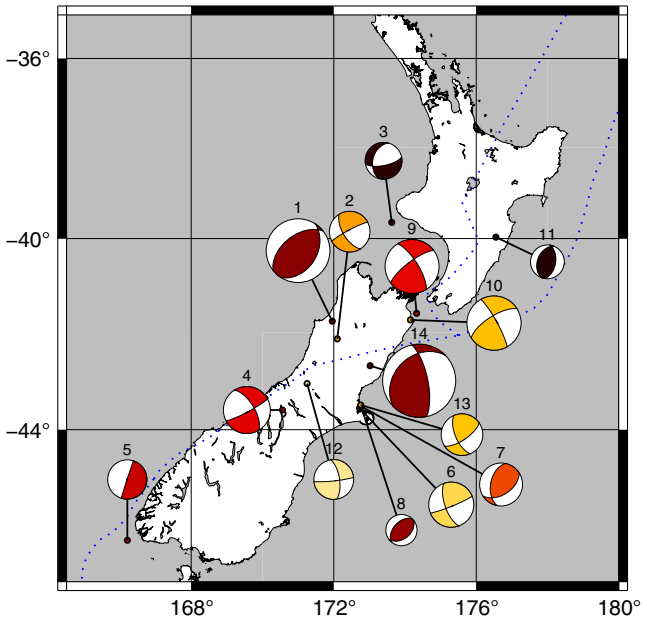

Figure 4. Selected New Zealand earthquakes. An epicenter is denoted by a dot connected to a focal mechanism plot. The ID of an earthquake (Table 3) is given above the corresponding focal mechanism plot. Size of the focal mechanism plot scales with the magnitude (see Table 3). Darker color refers to deeper events (see Table 3). The plate boundaries shown (dotted lines) are from Bird (2003; see also Data and Resources). The color version of this figure is available only in the electronic edition.

expression of the evaluation results because it is easier to inspect the change of model performance with distance from the model rank. We describe these three steps in the following three sections.

Multivariate Logarithmic Score

The mvLogS is an extension of the widely used univariate logarithmic score (known as the LLH in the seismological literature; see Scherbaum *et al.*, 2009) by taking the full effect of the correlation structure of a GMM into account for measuring the model's performance. GMMs often provide the correlation structure of their predictions by providing a set of sigma components, instead of a single total sigma; this is mathematically equivalent to prescribing some records to be correlated. For example, if the total sigma is partitioned into a between-event and a within-event sigma, records produced by the same earthquake are treated as correlated.

The advantage of using the mvLogS is that it fully utilizes all information of sigma components provided by the modeler to evaluate the performance of the model without invoking residual partitioning; residual partitioning could be problematic, see the Appendix. Compared with the original LLH, which does not use the information of data correction provided by the prediction, the mvLogS uses more the information contained in the prediction, which often leads to better identifying the difference among models. This advantage will likely become even more important in the future because the correlation structure of GMMs is becoming more complicated. GMMs in the past (e.g., Campbell, 1997) provided only a single-constant sigma value (i.e., the total sigma; equivalent to no correlation structure). More recent GMMs (e.g., Boore and Atkinson, 2008) partition the total sigma into the between-event and the within-event sigmas, equivalent to introducing a two-layer hierarchical structure to allow records produced by the same earthquake to be correlated; the between-event and within-event sigmas are constants for all records. State-of-the-art GMMs (e.g., Abrahamson *et al.*, 2014) allow the sigmas to vary with parameters such as magnitude, distance, and nonlinear site effects. Consequently, the sigmas for each record are generally different, representing a more complicated correlation structure than that of constant sigmas. As a result, the simple physical meaning of the constant sigma, which measures the spread of event terms and within-event residuals, has become abstract. GMMs of even more sophisticated correlation structures are being proposed. For example, Kotha *et al.* (2016) provided an additional treatment to station-to-station variability, essentially introducing an additional station term to allow ground motions recorded by the same station to be correlated. An even more complicated variability structure for GMMs was described by Al Atik *et al.* (2010, their table 1). All of the above-mentioned correlation structures were modeled as N-layer ($N \geq 2$) hierarchical structures using the mixed-effect model. The mvLogS utilizes all the information provided by the modeler through the data correlation in evaluating the model performance; this method can fairly measure the relative performance of GMMs developed in the past two decades and, also likely, those that will appear in the near future.

Table 3
Selected New Zealand Earthquakes (See Also Fig. 4)

ID	Date and Time (yyyy/mm/dd hh:mm:ss)	Longitude (°)	Latitude (°)	Depth (km)	M_w
1	1968/05/23 17:24:15	171.96	−41.76	15	7.23
2	1971/08/13 14:42:41	172.1	−42.13	9	5.7
3	1974/11/05 10:38:38	173.63	−39.65	17	5.44
4	1984/06/24 13:29:39	170.56	−43.6	13	6.12
5	2009/07/16 00:24:05	166.2051	−46.199	13.6	5.61
6	2011/06/13 02:20:49	172.724	−43.561	6.47	5.99
7	2011/12/23 02:18:03	172.763	−43.5208	11.21	5.85
8	2012/05/25 02:44:50	172.7723	−43.534	14.73	5.07
9	2013/07/21 05:09:30	174.3287	−41.5957	12.85	6.58
10	2013/08/16 02:31:05	174.1522	−41.734	8.16	6.6
11	2014/03/31 01:01:18	176.5519	−39.9703	17.3	5.25
12	2015/01/05 17:48:41	171.252	−43.0579	5.12	5.64
13	2016/02/14 00:13:43	172.7546	−43.4973	8	5.76
14	2016/11/13 11:02:56	173.02	−42.69	15	7.85

Time in UTC.

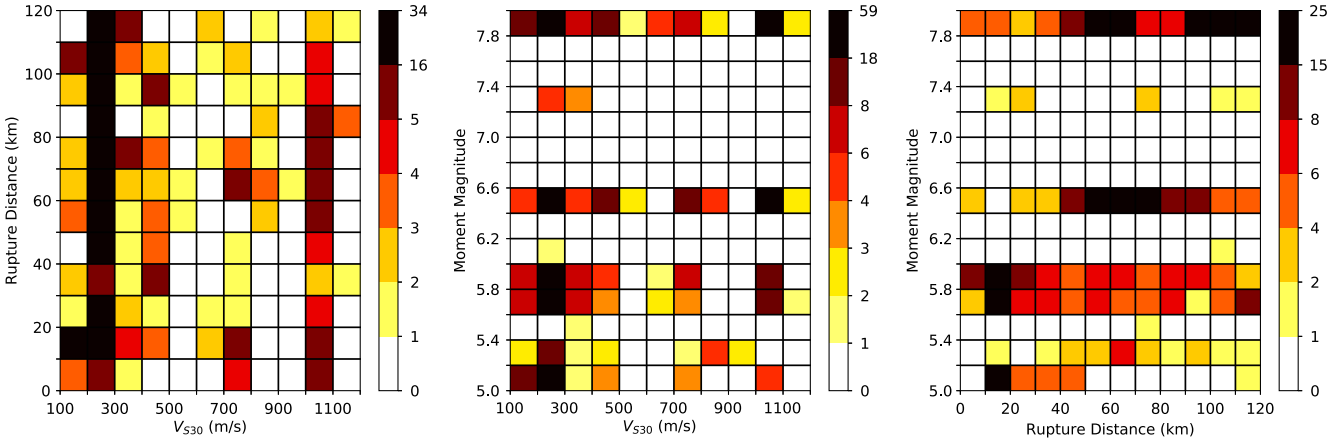

Figure 5. Distributions of the numbers of New Zealand peak ground acceleration (PGA) records with respect to magnitude, distance, and V_{S30}. The numbers of records for other spectral periods are slightly less (see Table 3). The discrete scale was chosen so that approximately the same number of bins fall into each nonzero range. Records with $V_{S30} > 1200$ m/s are few and not shown. The color version of this figure is available only in the electronic edition.

One may wonder why one should care about the details of the ground-motion variability because only the total sigma is used in most probabilistic seismic hazard studies. It is true that only the total sigma matters if one is concerned with the long-term hazard contributed by infinitely many earthquakes and ground-motion excitations, which is often the case for calculating a hazard curve. In reality, however, a GMM can only be empirically evaluated with finite data. The details of the ground-motion variability will, therefore, affect the evaluation result, as shown by various examples in Mak, Clements, and Schorlemmer (2017).

A desirable consequence of fully incorporating the correlation structure of a GMM using the mvLogS is that the evaluation result will be less affected by unbalanced data (see examples 1–3 of Mak, Clements, and Schorlemmer, 2017). Strong-motion observations are generally unbalanced. For example, the number of records of event 7 in the distance bin of 0–40 km of the Japanese data is about three times that of event 10 (Fig. 3). The New Zealand data used in this study are even more unbalanced; a few events (e.g., event 14) have produced far more records than others (e.g., events 1–5; see Fig. 6c–f). Evaluation methods that do not take into account the data correlation will likely bias toward well-recorded earthquakes (i.e., a GMM will appear to perform well even if it only predicts well the few well-recorded earthquakes).

Variability of Evaluation Results

Ground-motion observations are random variables and so is the resulting mvLogS computed from them. We quantified the variability of the results by the cluster bootstrap, a resampling technique at the event level. The bootstrap technique often used in other studies of GMM performance (e.g., Edwards and Douglas, 2013, their table 4; hereafter, simple bootstrap) resamples at the record level. This technique is not suitable for correlated ground-motion data because it does not sufficiently represent the variability at the event level. The cluster bootstrap utilizes the variability within the available data to assess the variability of the resulting score while preserving the correlation structure of the data.

The mvLogS is based on likelihood. Likelihoods computed from distinct samples (in this case, bootstrap samples) are not comparable, so it is not meaningful to compute a distribution of scores for a GMM and evaluate the difference between GMMs by their distributions of scores, as often seen in previous studies. The DI summarizes the results of the cluster bootstrap into a single quantity, ranging from −1 to 1, to represent how often one model is better than the other. A positive (or negative, respectively) DI means one GMM is more (or less, respectively) often better than the other, given the variability of the available data. The two extremes of 1 and −1 mean that one GMM is always better (or worse, respectively) than the other. We show pairwise DIs as a distinctness table to facilitate the comparison of multiple GMMs.

Model Ranking

A model can be ranked using its DIs with respect to all other models. A model with all positive DIs is one that is usually better than all other models and could be considered as the best model. The second-best model should have only one negative DI (i.e., that with respect to the best model) and so on. The rank of model i, R_i, over N models can be computed by counting how many negative DIs there are:

$$R_i = 1 + \sum_{j;j\neq i}^{N} \tilde{\mathbb{1}}(D_{ij})$$
$$\tilde{\mathbb{1}}(x) = \begin{cases} 1 & \text{if } x < 0 \\ 0 & \text{if } x > 0 \\ 0.5 & \text{if } x = 0, \end{cases} \quad (1)$$

in which D_{ij} is the DI of model i with respect to model j. A model with all positive DIs will be ranked 1 (i.e., the best). Because of the nature of multiple comparisons, the ranking is not necessarily unique (see example 9 of Mak, Clements, and

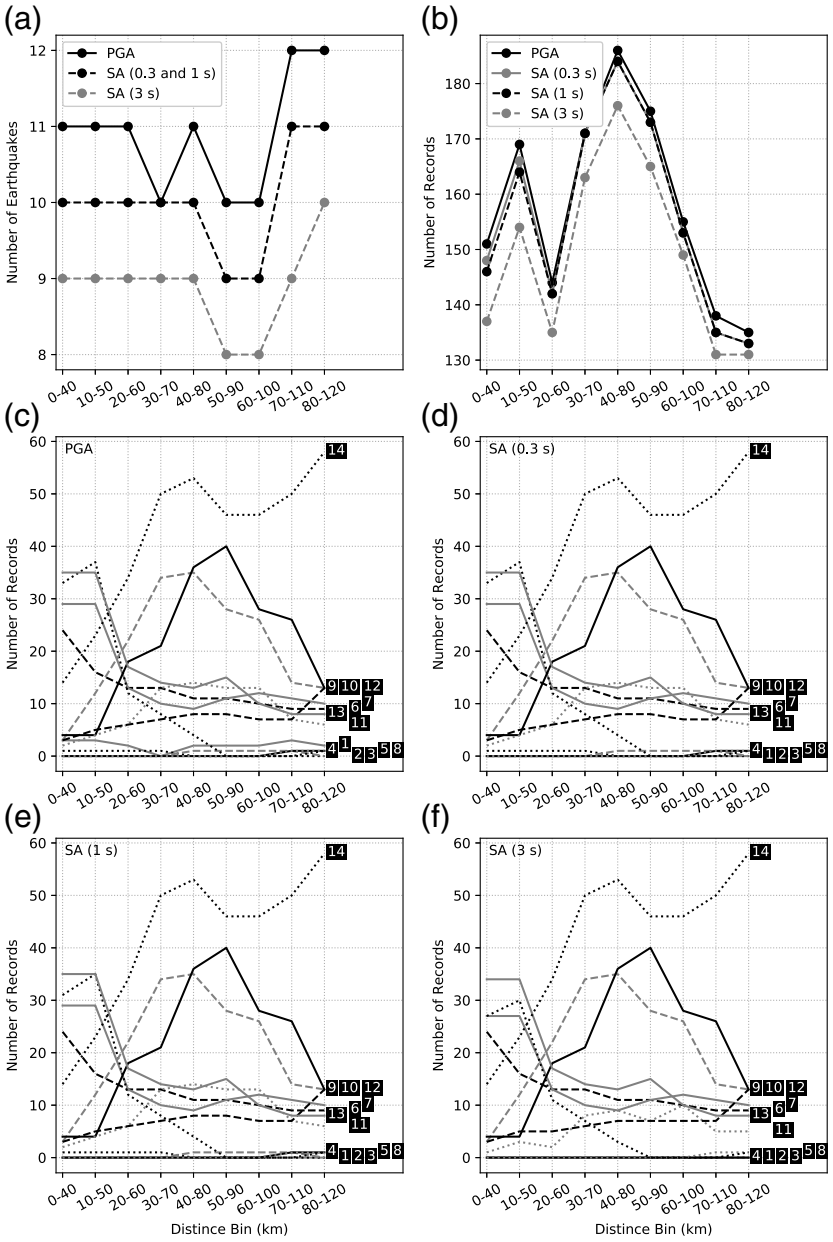

Figure 6. (a) The number of earthquakes in each distance bin. (b) The number of records in each distance bin. (c–f) The number of records in each distance bin for each earthquake for each spectral period. The earthquake IDs (see Table 3) are given at the end of each line. SA, spectral acceleration.

Schorlemmer, 2017); in the current study, however, we did not encounter this problem in our analysis.

We used the model rank as a summary statistic of the model performance. The distinctness table, explained earlier, provides the full information on the relative performance of a pair of models, including the stability of the performance. It, however, graphically expresses the evaluation results of only a single data subset (a certain distance bin, in our study). The model rank, on the other hand, can be plotted versus data subsets; it is, therefore, easier to graphically express the change of model performance over data subsets (i.e., distance) using the model rank.

Result Stability

The DI measures the relative performance of a model pair using a cluster bootstrap. The stability of the bootstrap itself, however, requires a sufficiently large amount of earthquakes. We tested if the number of earthquakes in our data was sufficient to produce a stable result. For the data in a distance bin, we picked subsets of data containing N earthquakes, in which N ranged from five to the number of available events. Then, we calculated model ranks using the data subsets containing different numbers of earthquakes. The stability of the resulting model ranks over increasing numbers of earthquakes is an indicator of the stability of the evaluation results.

Results and Discussions

Japan

The distinctness table for the GMMs of the distance bin 0–40 km (i.e., the nearest field) is given in Figure 7; the overall mvLogS calculated using the whole data subset (i.e., no bootstrap needed) are also shown in the same figure. The model ranks versus distance bins are given in Figure 8. The ranks for the distance bin 0–40 km in Figure 8 are identical to those shown in Figure 7; Figure 8 is essentially a summary of the model ranks shown in various figures such as Figure 7 but for different distance bins. In the near field ($R_{rup} < 50$ km), 2014CY appeared to be the best model; in the far field ($R_{rup} > 50$ km), 2014ASKjp performed the best. Each NGA-West2 GMM showed a general improvement over its predecessor; NGA-West2 GMMs as a group also performed better than NGA GMMs, especially in the far field. In the following, we discuss three issues based on the model ranks.

Regional versus Global GMMs. In the far field ($R_{rup} > 50$ km), three Japanese models (1999SM, 2002MO, and 2006KNM) outperformed the NGA models, although the newest NGA-West2 models were still better. Such improvement in performance of regional models in the far field is often interpreted as the anelastic attenuation dominating the regional characteristics of ground motions (see, e.g., fig. 11 of Boore *et al.*, 2014). In an outstanding case, 2014CB was found to have similar performance, in the far field, with the three above-mentioned Japanese models, although it is more sophisticated and more than a decade newer.

Model i \ Model j	1995MY	1999SM	2002MO	2006KNM	2008AS	2008BA	2008CB	2008CY	2014ASK	2014BSS	2014CB	2014CY	2014ASKjp	2014BSSjp	2014CBjp	2014CYjp	mvLogS	Rank
1995MY		0.71	0.55	0.63	0.20	-0.75	-0.74	-0.99	-0.99	-1.00	-1.00	-1.00	-0.99	-1.00	-1.00	-0.98	250	12
1999SM	-0.71		-0.91	0.09	-0.32	-1.00	-0.97	-1.00	-0.99	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	260	15
2002MO	-0.55	0.91		0.35	-0.14	-0.95	-0.92	-1.00	-0.99	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	257	14
2006KNM	-0.63	-0.09	-0.35		-0.41	-0.95	-0.93	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	267	16
2008AS	-0.20	0.32	0.14	0.41		-0.74	-0.85	-0.94	-1.00	-0.97	-1.00	-1.00	-1.00	-0.97	-1.00	-1.00	261	13
2008BA	0.75	1.00	0.95	0.95	0.74		-0.36	-0.93	-0.78	-0.87	-1.00	-1.00	-0.77	-0.87	-0.99	-0.71	240	11
2008CB	0.74	0.97	0.92	0.93	0.85	0.36		-0.62	-0.59	-0.41	-0.91	-0.94	-0.49	-0.42	-0.93	-0.55	233	10
2008CY	0.99	1.00	1.00	1.00	0.94	0.93	0.62		0.11	0.41	-0.79	-0.99	0.15	0.39	-0.59	0.22	233	4
2014ASK	0.99	0.99	0.99	1.00	1.00	0.78	0.59	-0.11		0.25	-0.67	-0.85	0.05	0.22	-0.89	0.25	232	5
2014BSS	1.00	1.00	1.00	1.00	0.97	0.87	0.41	-0.41	-0.25		-0.95	-1.00	-0.39	-0.83	-0.91	-0.09	230	9
2014CB	1.00	1.00	1.00	1.00	1.00	1.00	0.91	0.79	0.67	0.95		-0.91	0.68	0.93	-0.17	0.73	222	3
2014CY	1.00	1.00	1.00	1.00	1.00	1.00	0.94	0.99	0.85	1.00	0.91		0.91	1.00	0.36	0.86	224	1
2014ASKjp	0.99	1.00	1.00	1.00	1.00	0.77	0.49	-0.15	-0.05	0.39	-0.68	-0.91		0.31	-0.82	0.21	230	6
2014BSSjp	1.00	1.00	1.00	1.00	0.97	0.87	0.42	-0.39	-0.22	0.83	-0.93	-1.00	-0.31		-0.89	-0.05	230	8
2014CBjp	1.00	1.00	1.00	1.00	1.00	0.99	0.93	0.59	0.89	0.91	0.17	-0.36	0.82	0.89		0.92	218	2
2014CYjp	0.98	1.00	1.00	1.00	1.00	0.71	0.55	-0.22	-0.25	0.09	-0.73	-0.86	-0.21	0.05	-0.92		236	7

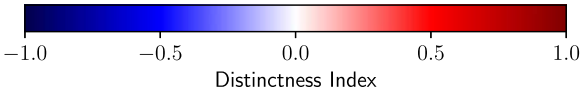

Figure 7. Distinctness table for the Japan case for the distance bin 0–40 km. See Table 1 for the IDs of the ground-motion models (GMMs). IDs for global GMMs optimized for Japan end with jp. The distinctness index of each pairwise comparison (based on 300 cluster bootstrap samples) is given in the intersecting box of a model pair. A positive value means model i (indicated in the leftmost column) is better than model j (indicated in the topmost row) when data correlation and result variability have been taken into account. Negative values are underlined. The multivariate logarithmic scores (mvLogS) given in the second-to-last column are computed using the whole dataset (i.e., no bootstrap). The rank of each model, calculated by equation (1), is given in the last column. The color version of this figure is available only in the electronic edition.

Previous studies on the applicability of NGA GMMs to Japan (Nishimura, 2010; Beauval *et al.*, 2012; Delavaud *et al.*, 2012) generally concluded that Japanese models outperformed foreign models when evaluated using Japanese data. Part of their results was confirmed in the current study: in the far field ($R_{rup} > 50$ km), three Japanese GMMs (1999SM, 2002MO, and 2006KNM) performed better than the four NGA GMMs. In the near field, however, Japanese GMMs did not perform better than the NGA GMMs. The previous studies did not separately evaluate near- and far-field data. Because far-field data often dominate the dataset (as can be seen in Fig. 3), conclusions from previous studies were probably restricted to the far field.

For the magnitude range of our data (M_w 5.5–6.6), the hazard is likely dominated by near-field ground motions. The performance for NSHMJ14, which is largely based on a modified form of 1999SM, may, therefore, improve if an NGA or NGA-West2 GMM has been used, even if those GMMs are not optimized for Japan. The performance gain in newer GMMs over dated Japanese GMMs could come from various factors, including the use of a larger dataset, the more sophisticated variability structure (sigmas), and the special treatments to near-field effects, such as a buried fault and hanging wall. More than half of the records in the 0–40 km bin had a horizontal distance within 10 km from the fault strike (the parameter R_x of the NGA-West2 flat file), and so near-field effects should be noticeable. A new Japanese GMM taking into account all the above factors may perform comparable to or better than the current state-of-the-art GMMs represented by the NGA-West2 GMMs. Prospective evaluation of such a new regional model, however, will not be possible in the near future.

Performance Gain for Regional Optimization. We found a mixed result regarding the performance gain of the regional optimization of NGA-West2 GMMs. For 2014ASK and 2014BSS, the corresponding regional optimization (i.e., 2014ASKjp and 2014BSSjp) clearly outperformed the unmodified version; the effort spent on regional optimization is justified. The performance gain for 2014CBjp over 2014CB, however, was not unambiguous and fluctuated over distance. For 2014CY, the regional optimization clearly showed an adverse effect. The regional optimization of 2014CY involves three factors: the anelastic attenuation, the site effects (shallow and deep), and the within-event sigma. We implemented these three factors separately and found that the adjusted anelastic attenuation and sigma actually improved the prediction in the far field ($R > 50$ km; Fig. 9). It appeared that it was the adjusted site effects that adversely affected the model performance. The nonimprovement for 2014CYjp may imply some overfitting (e.g., Bindi, 2017). Our results show that, although a modeler's decision on using a certain adjustment factor for ground-motion modeling may be physically sound, it is important to empirically verify the performance gain of a model complexity.

Stability of Results. Figure 10 shows the model ranks of the distance bin 0–40 km versus data subsets of increasing number of earthquakes. The ranks of GMMs with high ranks in this distance bin (2014CY, 2008CY, and 2014CB) were quite stable. The ranks for 2014ASK and 2014BSS exchanged a few times but kept staying at middle ranks. Models of low

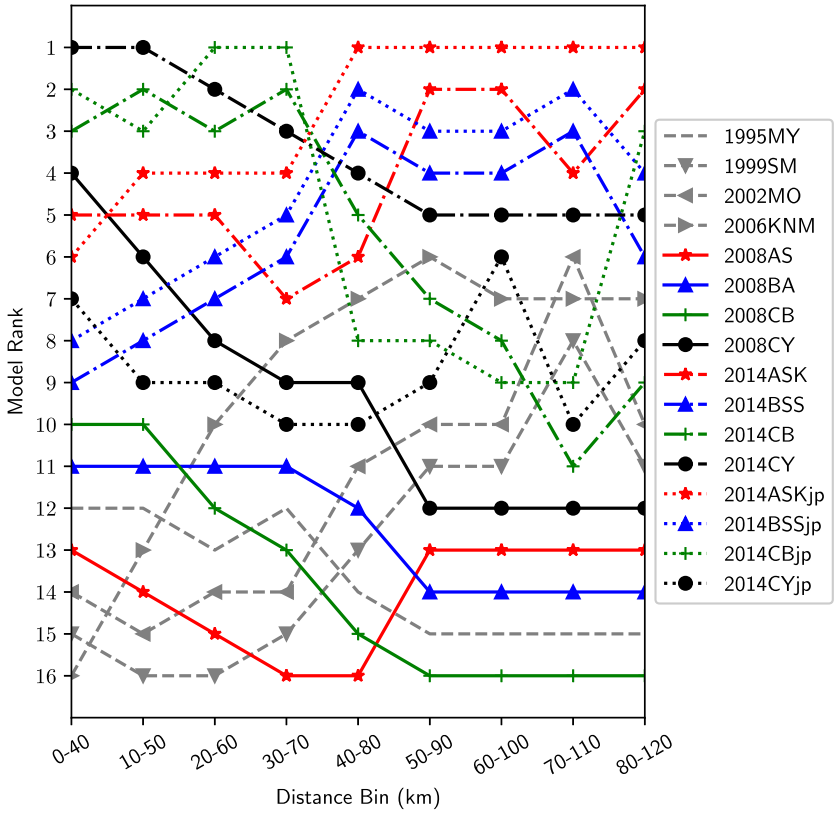

Figure 8. Model ranks for the Japan case versus distance bins. See Table 1 for the IDs of the GMMs. IDs for global GMMs optimized for Japan end with jp. The color version of this figure is available only in the electronic edition.

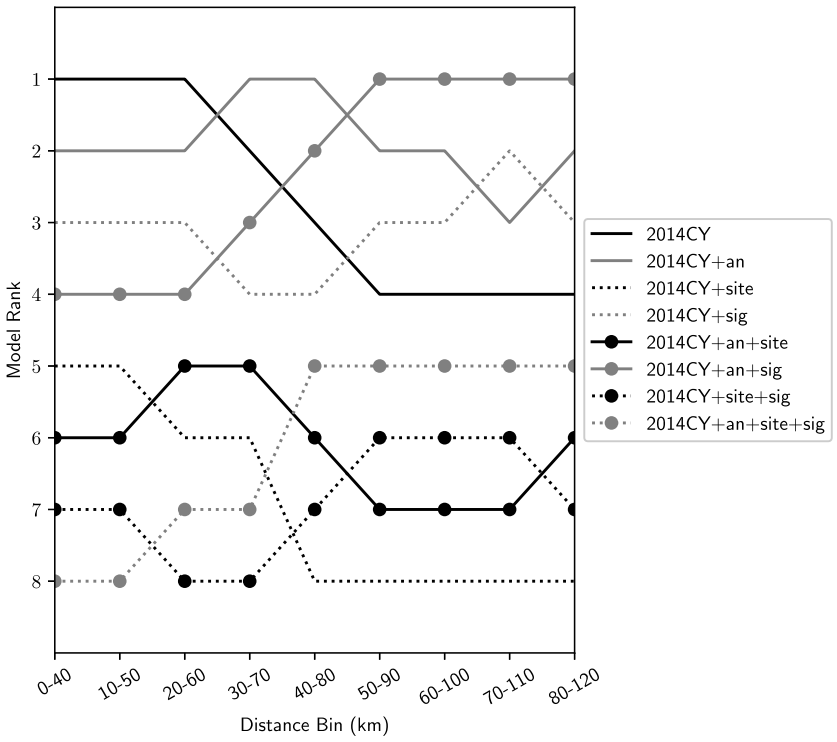

Figure 9. Model ranks for the Japan case versus distance bins. Only the various versions of 2014CY are shown. Each version is a combination of the three adjustments of anelastic attenuation (an), shallow soil amplification (site), and sigma (sig).

ranks also remained at low ranks. We consider that the results for the Japan case are stable enough for robust interpretation.

New Zealand

The distinctness table for the GMMs of the distance bin 0–40 km and the overall mvLogS calculated using the whole data subset are given in Figure 11. The model ranks versus distance bins are given in Figure 12. 2013B, a recent New Zealand model, performed well for PGA and SA at 0.3 s over a wide range of distance, and for SA at 1 and 3 s at the near field. The other New Zealand model, 2006M, which is the oldest among the evaluated GMMs, generally did not perform well, even though our data were not completely independent to it, and so the model should have had a slight benefit.

Similar to the finding in many other GMM evaluation studies (e.g., Delavaud *et al.*, 2012, their table 7), we also found the relative performance of GMMs at different spectral periods not entirely the same. For example, in the near field ($R_{rup} < 50$ km), 2014ASK and 2008AS performed less well for SA at 3 s than at shorter periods. We did not find any GMMs that had consistently good performance over all spectral periods. For PGA, 2013B, 2014CB, 2014CY, and 2008CY performed the best. For SA at 0.3 s, 2013B, 2014ASK, 2008AS, and 2008CY performed the best. For SA at 1 s, no GMMs consistently ranked high over distance. For SA at 3 s, 2014CY performed generally well over distance.

Comparison with the Japan Case. The model performances for New Zealand differed from those of Japan on two issues. First, the model ranks for New Zealand were more stable over distance; GMMs that ranked consistently well over distance could be found (e.g., 2013B and 2014CB for PGA, 2013B for SA at 0.3 s, and 2014CY for SA at 3 s), although GMMs of ranks that fluctuated much over distance were also seen (e.g., 2014ASK). Comparatively, the relative performances of GMMs over distance for the Japan case (Fig. 8) were more unstable. This could be a result of the general observation (e.g., as found in NGA-West2 GMMs) that the

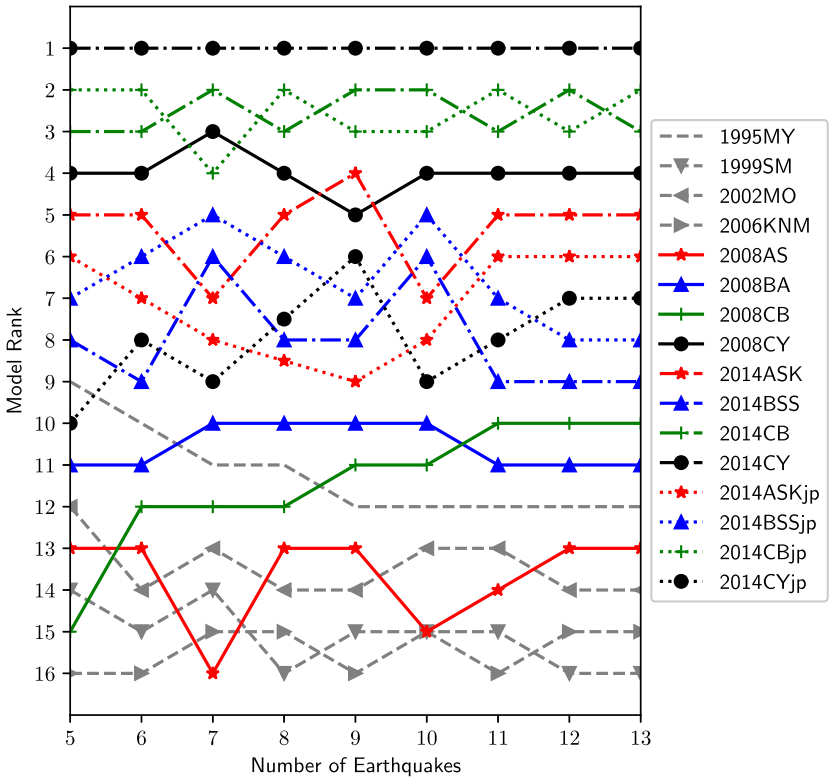

Figure 10. Model ranks of the Japan case for the distance bin 0–40 km versus data subsets of increasing number of earthquakes. See Table 1 for the IDs of the GMMs. IDs for global GMMs optimized for Japan end with jp. The color version of this figure is available only in the electronic edition.

anelastic attenuation for New Zealand is not particularly different from the global average, while that for Japan is more notable. The direct consequence of this observation is that regional models do not have particular advantage in performance over global models in the far field; this matches our results: the relative performances of 2006M and 2013B did not increase with distance. For SA at 1 s, the performance of 2013B, in fact, decreased with distance.

Second, we did not see an unambiguous performance gain of the NGA-West2 models over the NGA models they superseded in the New Zealand case, as we found in the Japan case. We saw a few cases that an NGA-West2 GMM consistently outperformed its predecessor, for example, 2014CB and 2014BSS for PGA, 2014CB for SA at 1 s, and 2014CY for SA at 3 s. There were also occasions that NGA-West2 GMMs did not perform much differently than its predecessor (e.g., 2014ASK for PGA and SAs at 0.3 and 1 s) and that NGA-West2 GMMs were outperformed by its predecessor (e.g., 2014CY for SA at 1 s and 2014BSS for SA at 3 s). Some workers recommend against using superseded GMMs (Cotton *et al.*, 2006, p. 139, point 4; Bommer *et al.*, 2010, p. 791, point 4). This recommendation is not always followed (e.g., Edwards *et al.*, 2016, p. 1842) because some workers are more confident in older models that have been tested and verified. Our results show that it is important to quantify the improvement of a new model over its predecessor by an empirical evaluation such as the current study, instead of following the philosophical argument that the new model is expected to be better than an old one.

To compare more directly the results between the New Zealand and Japan cases, we separately analyzed the PGV predictions for the New Zealand data, although PGV is not a metric commonly used in New Zealand for seismic hazard analysis. The two New Zealand models (2006M and 2013B) do not provide PGV predictions, and therefore were excluded from this analysis.

The results (Fig. 13) show that GMMs ranked differently in the two regions. 2014ASK, its predecessor 2008AS, and 2014BSS performed the best for PGV predictions for New Zealand. The good performance of 2014CY and 2004CB in the near field for Japan was not seen for New Zealand. Similar to the results for other spectral periods for New Zealand, the model performances over distance were stable for PGV for New Zealand; this stability was not found in the Japan case. Unlike the results for other spectral periods for New Zealand, the performance gain of NGA-West2 GMMs over their corresponding NGA GMMs was quite clear for PGV; this feature was also found in the Japan case.

Stability of Results. Figure 14 shows the model ranks of the distance bin 0–40 km versus data subsets of increasing number of earthquakes. Similar to what we found in the Japan case (Fig. 10), GMMs in groups of high, middle, and low ranks remained in the same group. This stability demonstrates the robustness of our results. These groups, however, appeared to contain more GMMs than those in the Japan case, indicating a smaller degree of stability. For example, for PGA, the performance of 2014ASK, 2008AS, 2014CB, 2014CY, and 2008CY were quite mixed; the same was found for 2014ASK, 2008AS, and 2013B for SA at 0.3 s, for 2014ASK, 2008AS, 2013B, and 2008CY for SA at 1 s, and for 2013B, 2014CY, and 2008CY for SA at 3 s. One reason for this is that the performance gain for NGA-West2 GMMs over their corresponding predecessors was found to be less clear. The ranks of an NGA/NGA-West2 pair, therefore, often exchanged when more data were used. We consider that this degree of stability is sufficient to support our interpretations of the results of the New Zealand case; it confirms again that NGA-West2 GMMs are not always better than their corresponding predecessors. A measure of result stability in a model evaluation study would demonstrate the epistemic uncertainty of

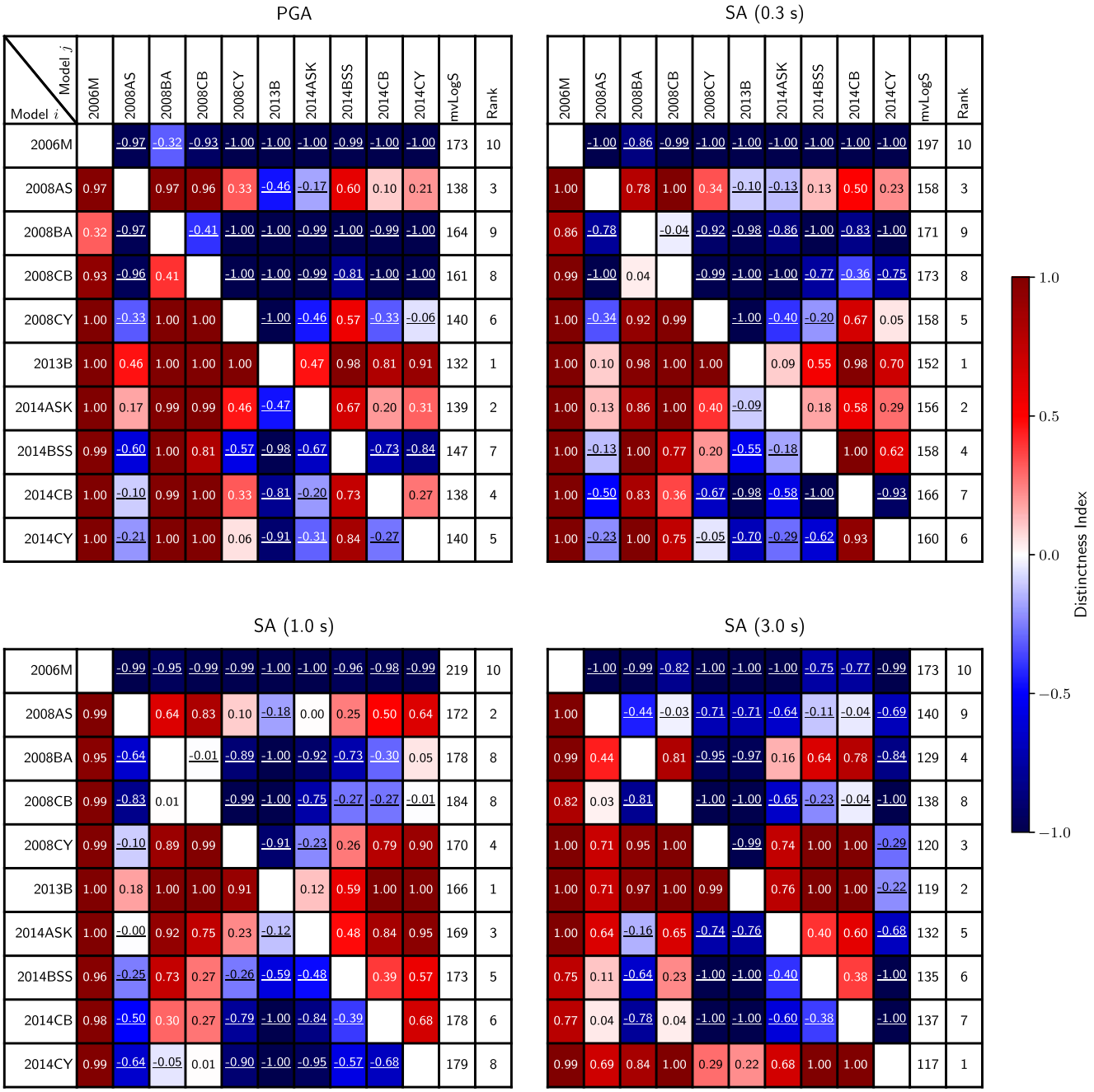

Figure 11. Distinctness table for the New Zealand case for the distance bin 0–40 km. See Table 1 for the IDs of the GMMs. The distinctness index of each pairwise comparison (based on 300 cluster bootstrap samples) is given in the intersecting box of a model pair. A positive value means model *i* (indicated in the leftmost column) is better than model *j* (indicated in the topmost row) when data correlation and result variability have been taken into account. Negative values are underlined. The mvLogS given in the second-to-last column are computed using the whole dataset (i.e., no bootstrap). The rank of each model, calculated by equation (1), is given in the last column. The color version of this figure is available only in the electronic edition.

GMMs; modelers should carefully consider this when analyzing the seismic hazard for New Zealand.

Comparison with Van Houtte (2017). Van Houtte (2017) compared the model performance between NGA-West2 models and the two New Zealand models (2006M and 2013B) based on the LLH score and residual analysis. His results are not directly comparable to ours for various reasons. First, he used ground motions with distances up to 200 km, whereas we evaluated the model performance by distance groups. Second, he used all events with $M_w > 5$ from the database of Van Houtte *et al.* (2017), whereas we used the same magnitude range but fewer earthquakes because we ensured only prospective (to all GMMs except

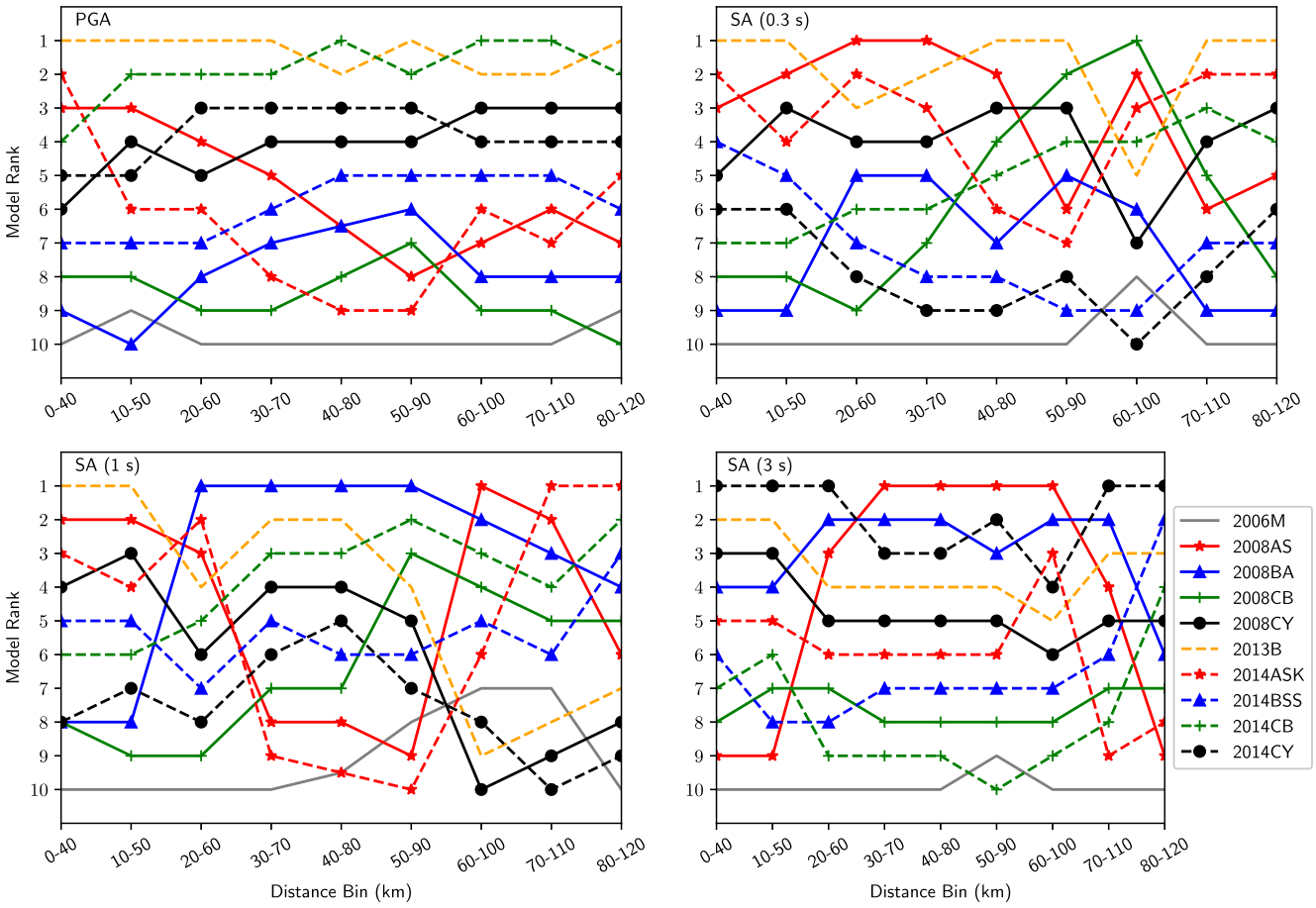

Figure 12. Model ranks of the New Zealand case versus distance bins. See Table 1 for the IDs of the GMMs. The color version of this figure is available only in the electronic edition.

for 2006M) mainshocks were used. Third, data correlation, which is a major reason for us to use the multivariate logarithmic score, was not explicitly taken into account in Van Houtte (2017). With all the differences between the two studies, both showed that 2006M generally did not perform well and 2013B generally performed well. This is, therefore, likely a robust result.

There are quite a few differences between the results of Van Houtte (2017) and our study. For example, Van Houtte (2017, his fig. 10a) showed that, for short spectral periods (≤ 0.3 s), 2014CY performed better than other NGA-West2 models; our results do not show 2014CY to be better than other NGA-West2 models, except for long periods (3 s) or PGA in the near field. Van Houtte (2017) showed that 2013B performed well for short spectral periods (≤ 0.4 s), whereas the good performance of 2013B was found to be more general in our results. The generally good performance of 2013B shown in our study is probably a more reasonable result because 2013B is a contemporary model made for the region. This more reasonable result may imply that our evaluation method has better utilized the information provided by the GMMs and the data (see also the Notes on the Evaluation Method section).

Insights from Hazard Model Disaggregation. The use of GMMs on seismic hazard modeling often involve extrapolation because the available data seldom cover the whole range of parameters of interest (e.g., magnitude and distance); the model behavior outside the range of available data is often constrained by only physics models. For empirical evaluation of GMMs, the GMMs are also seldom evaluated directly in the same parameter space that is used in a seismic hazard model. To assess how close the parameter space of the test data being used for model evaluation is to that used in a seismic hazard model, it is necessary to conduct the evaluation study together with the corresponding seismic hazard analysis. This was not often performed in published studies of GMM evaluation, presumably because published models for seismic hazard are often not easily reproducible. For example, we do not have access to the full model of NSHMJ14, so here we focus on NSHMNZ.

Figure 15 shows the disaggregation of NSHMNZ at the densely populated central areas of six New Zealand cities. For places such as Palmerston North and Wellington, the seismic hazard is dominated by large earthquake ($M \sim 7.5$). It may never be possible to empirically evaluate GMMs suitable for these places in a statistically rigorous

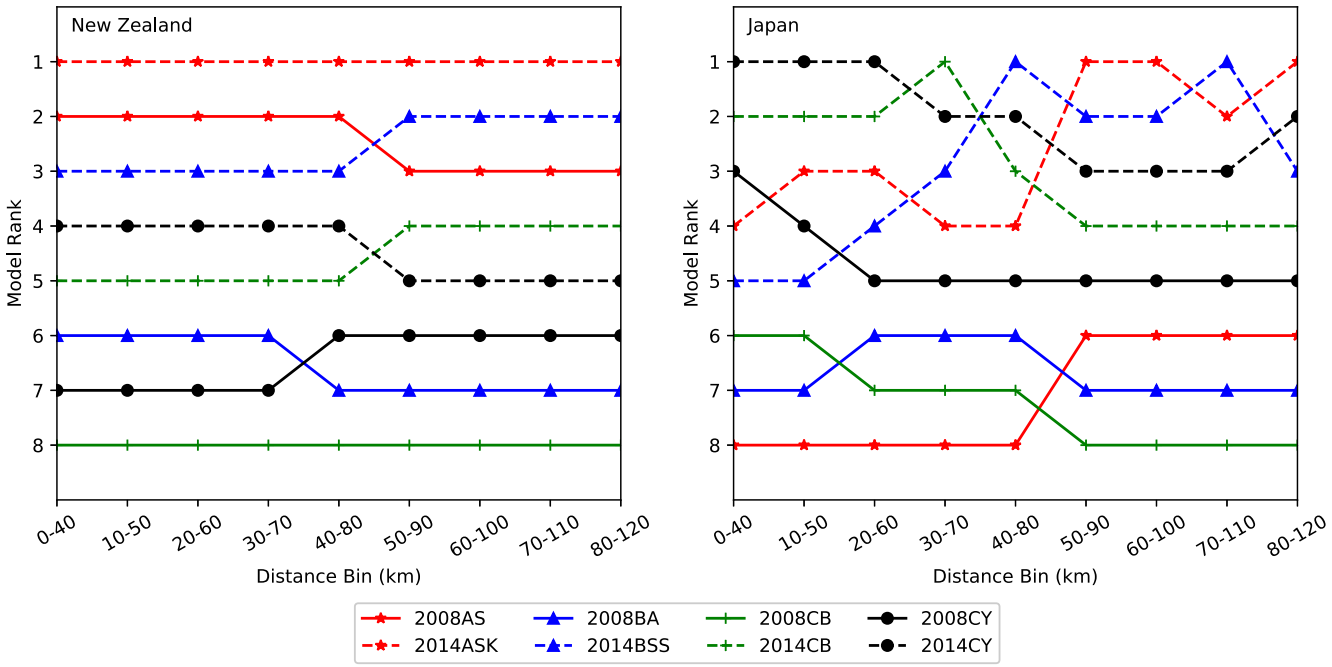

Figure 13. Model ranks of Next Generation Attenuation (NGA) and NGA-West2 GMMs for peak ground velocity (PGV) predictions for New Zealand (left) and Japan (right) versus distance bins. See Table 1 for the IDs of the GMMs. The color version of this figure is available only in the electronic edition.

sense because data from large earthquakes are always sparse, even in the future. GMMs suitable for these places are therefore the least constrained; the modeler may need to pay special attention to the epistemic uncertainty. For places such as Auckland and Tauranga, the seismic hazard is dominated by moderate earthquakes (M 5–6) at short distances ($R < 30$ km). Our available data (Fig. 5) cover this range, and so our evaluation results are more relevant to those places.

Effects of Peak Definition

The definition of peak motion used in the observed and predicted ground motions might not be the same in our analysis. The peak definitions of each GMM are given in Table 1. The observation for Japan (New Zealand, respectively) was based on the larger of the two horizontal components (RotD50, respectively). For the Japan case, the three GMMs with the same peak definition as the observation (i.e., 1995MY, 1999SM, and 2002MO) generally did not perform better. The inconsistency in peak definitions between the observed and predicted ground motions, therefore, do not affect our results.

Boore (2010, his fig. 3) reported that the difference between the RotD50 and the GMRotI50 was in average about 3% (or 0.03 log unit) for spectral periods up to about 3 s. Van Houtte *et al.* (2017, his fig. 1b) reported that the difference between the geometric mean and the RotD50 was also about this degree. Given that the effect of peak definition is likely small, we did not convert one peak definition to the other because we did not want to introduce another factor to affect the model performance. For New Zealand, the good-performing 2013B (based on GMRotI50) does not use the same peak definition as the observations (based on RotD50). The consistency of peak definition with the observation might have contributed to the usually better performance of NGA-West2 models (based on RotD50) over NGA models (based on GMRotI50).

Logically, the definition of peak motion used by a GMM should be the same as that of applications of GMMs, such as a hazard model or structural designs standard. This will make both applications and model testing easier. The current practice to use a GMM often involves conversions between peak definitions (e.g., Beyer and Bommer, 2006), which introduces uncertainty. We hope future ground-motion modelers will consider using a peak definition that is consistent with their target applications.

Notes on the Evaluation Method

Although the overall mvLogS provides a sense of how similar the two models are, we used the DI to measure the relative performance of a model pair because this metric takes variability of the score into account. These two metrics sometimes give apparently inconsistent results. For example, 2014CY has positive DIs with respective to all other GMMs, and so is ranked 1 (Fig. 7). Its overall mvLogS, however, is not the lowest; both 2014CJjp and 2014CB have slightly lower scores than that. We consider that 2014CY performs better than 2014CBjp and 2014CB despite its slightly higher score because its positive DIs with respect to the two indicate

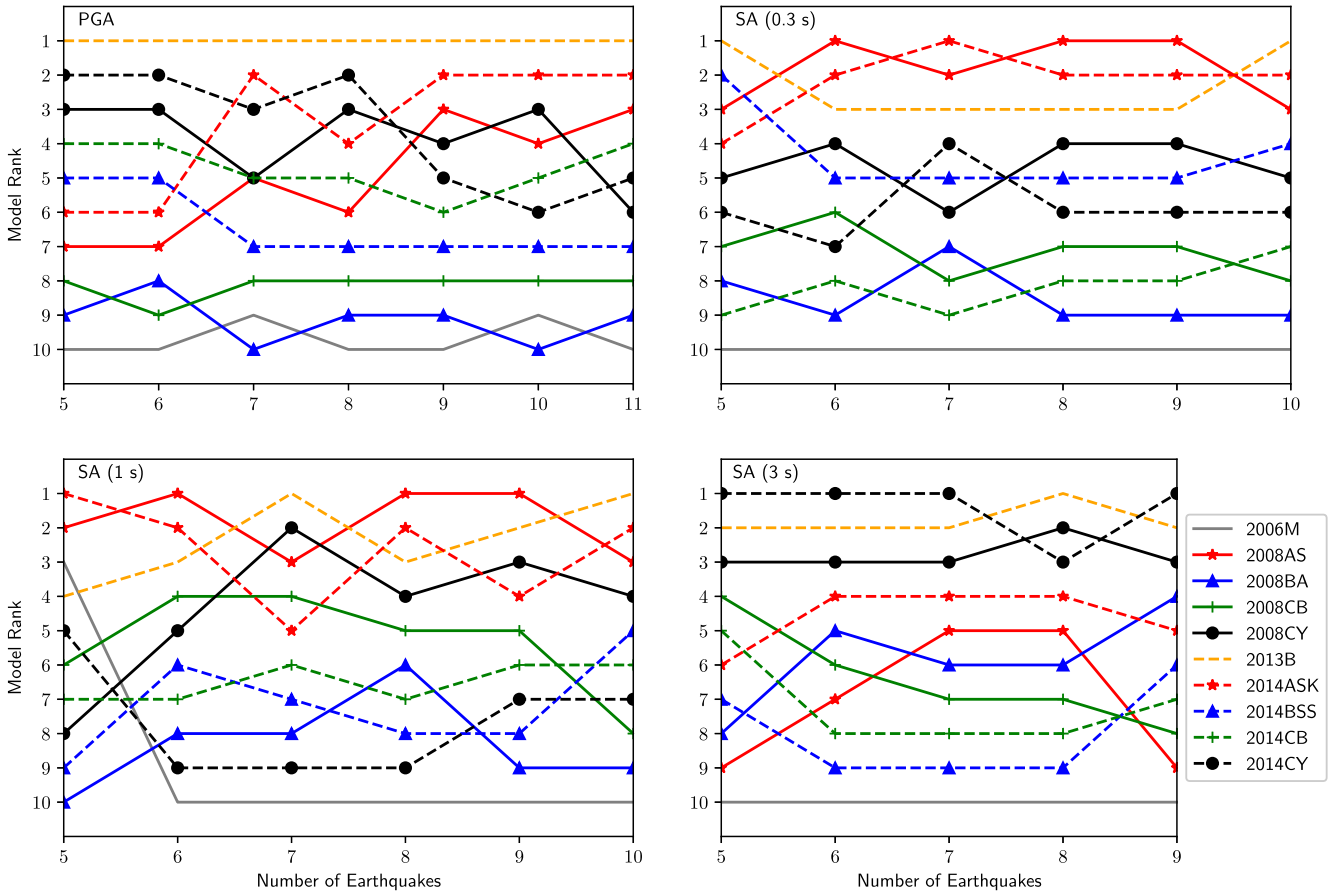

Figure 14. Model ranks of the New Zealand case of the distance bin 0–40 km versus data subsets of increasing number of earthquakes. See Table 1 for the IDs of the GMMs. The color version of this figure is available only in the electronic edition.

that, when the data change, 2014CY more often score better than them; the whole dataset is actually an uncommon case that 2014CBjp and 2014CB score better. Whether two models are similar in performance and whether one is better than the other are the two pieces of information indicated by the mvLogs and the DI, respectively. For example, the similar overall mvLogS indicate that 2014CY is similar to 2014CB. The DI of the former respective to the latter (close to 1), however, indicates that 2014CY is very likely better than 2014CB. We recommend that the relative performance of GMMs should be evaluated, taking into account the variability of the evaluation result.

Mak, Clements, and Schorlemmer (2017) pointed out some potential pitfalls of the widely used evaluation method based on the univariate logarithmic score (i.e., the LLH) and the simple bootstrap (hereafter, conventional approach). We investigated the difference between the results based on this conventional approach and our method (described in the Method section and Mak, Clements, and Schorlemmer, 2017) for the Japan case in the near field. We computed the LLH scores using 300 resampled datasets based on the simple bootstrap. The average model performance and its uncertainty are represented by the mean and standard error of the sample scores. Figure 16 shows that a large number of GMMs (e.g., 2014CB, 2014BSS, and 2008CY; 2014ASK and 2008BA) appear to have similar performance under the conventional approach; the ranges of their LLH overlap significantly with each other. Our new approach, however, shows that the mentioned GMMs are clearly different in performance from each other because the corresponding DIs are near 1 or −1 (2014CB vs. 2014BSS: DI = 0.95; 2014CB vs. 2008CY: DI = 0.79; 2014ASK vs. 2008BA: DI = 0.78; see Fig. 7). Conversely, the difference between 2002MO and 2006KNM appear to be clear by the conventional approach because their ranges of LLH separate (Fig. 16), while our new approach shows the opposite because their DI (= 0.35) is close to zero. Because the method we used takes into account the data correlation, preserves the data correlation during the bootstrap process, and compares likelihoods only when they are comparable, we consider the results based on our approach more logical.

For the New Zealand case, the results obtained by the conventional approach is summarized in Figure 17, extracted from Van Houtte (2017, his table 1; for PGA). In his results, 2013B and 2014CY appear to be somewhat comparable, so are 2014CY and 2014ASK, and 2014CB and 2014BSS; their ranges of scores overlap with each other. Our results based on the DI (Fig. 11, top left for PGA) would lead to different

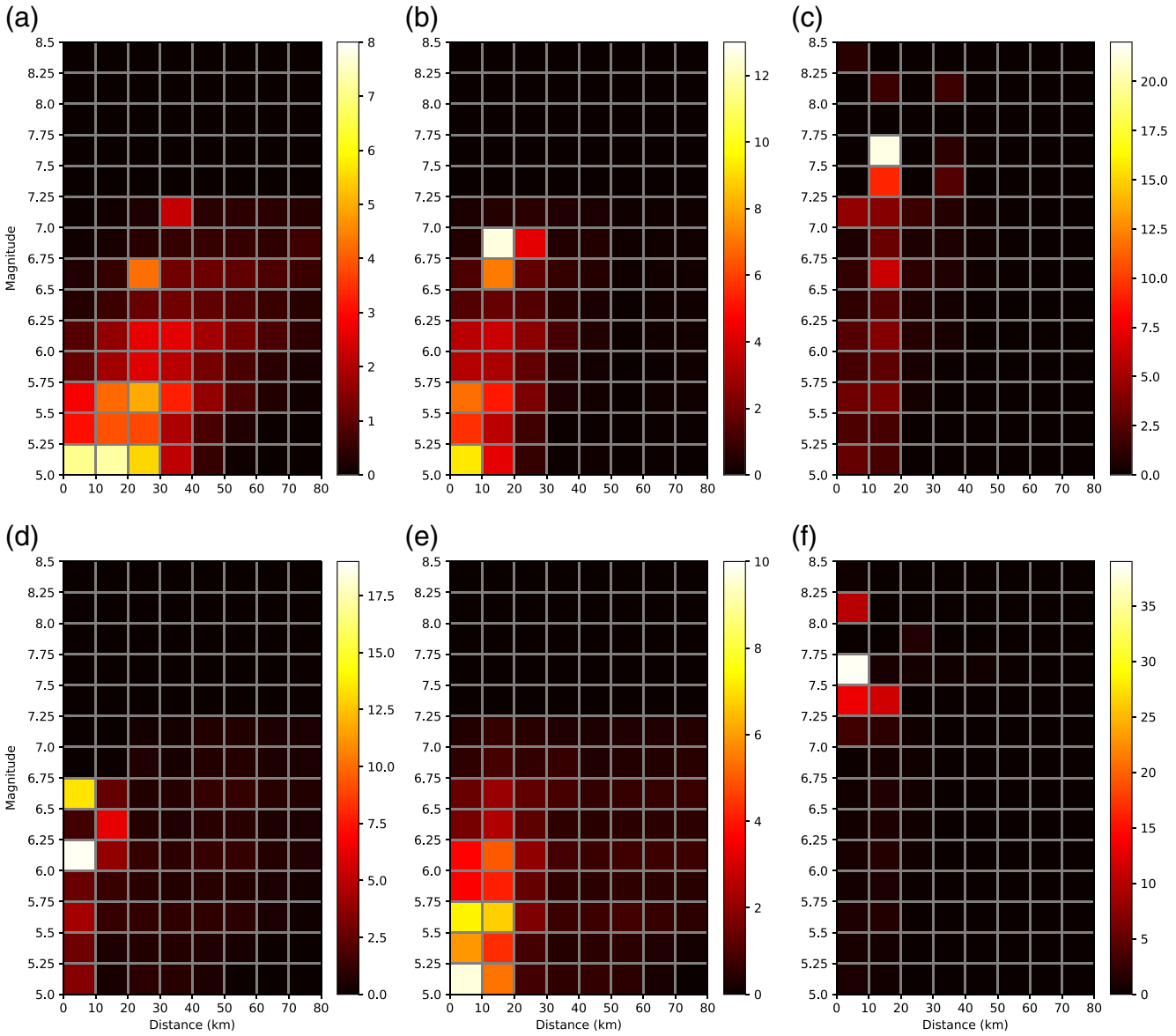

Figure 15. Disaggregation of the National Seismic Hazard Model for New Zealand of a 475-year return period based on 2013B, PGA, and $V_{S30} = 300$ m/s for (a) Auckland, (b) New Plymouth, (c) Palmerston North, (d) Rotorua, (e) Tauranga, and (f) Wellington. The color scale shows the contribution to hazard for earthquakes of each magnitude–distance bin. The color version of this figure is available only in the electronic edition.

conclusions: 2013B and 2014CY are clearly different (DI = 0.91, close to 1); 2014CB and 2014BSS are quite different (DI = 0.73); while 2014CY and 2014ASK are more similar (DI = −0.31).

To fully understand the meaning of the DI, one could refer to its original frequentist meaning. Technically, this is to replace the modified indicator function of equation 9 in Mak, Clements, and Schorlemmer (2017) by the usual indicator function, equivalent to computing $x = (\mathrm{DI} + 1)/2$, in which x is the frequency that one model performs better than the other. In this sense, a DI of 0.73 for 2014CB with respect to 2014BSS means that the former performed better than the latter in 87% of the bootstrap samples.

Summary

We empirically evaluated the performance of four Japanese GMMs, four NGA models, and four NGA-West2 models using the observed PGV of 13 Japanese shallow crustal earthquakes with magnitudes 5.5–6.6 that has not been used in the model fitting. We paid due respect to the correlation structure of the models and the variability of the results. We found that:

1. NGA-West2 models performed generally better in all the near medium and far fields than both NGA and Japanese models. In the near field ($R_{\mathrm{rup}} < 50$ km), Chiou and Youngs (2014) performed better than other NGA/NGA-

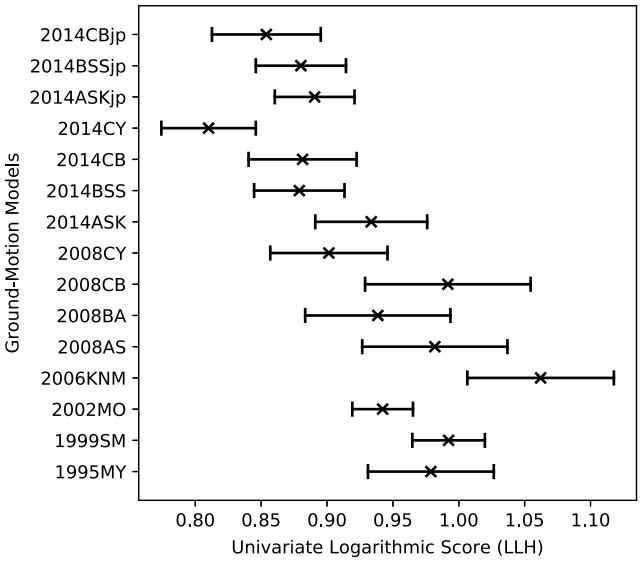

Figure 16. Univariate logarithmic scores (LLH) computed from the near-field data group for Japan. The mean LLH (based on 300 bootstrap samples) is denoted by a cross. The error bar denotes the standard error of the mean. Smaller values of LLH indicate better model performance. Refer to Table 1 for model IDs.

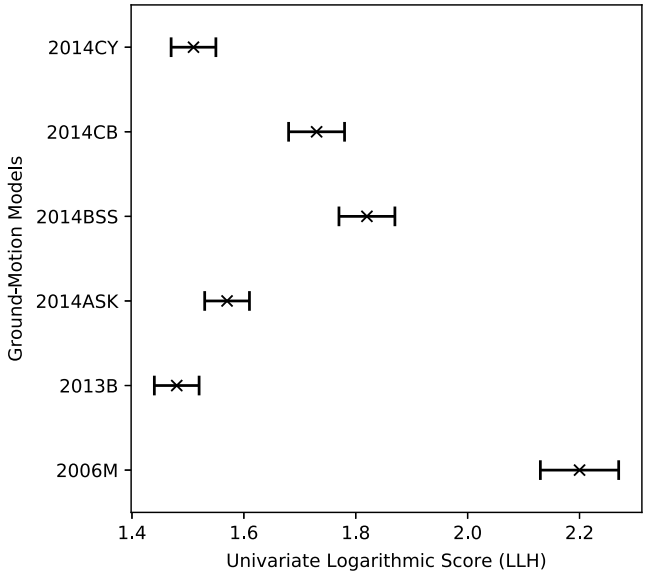

Figure 17. Univariate logarithmic scores (LLH) and their variability for PGA from Van Houtte (2017, his table 1). The standard error of the LLH is denoted by the error bar. Smaller values of LLH indicate better model performance. Refer to Table 1 for model IDs.

West2 and Japanese models. In the far field ($R_{rup} > 50$ km), Abrahamson *et al.* (2014), with coefficients optimized for Japan, performed better than other models.

2. Our results are stable.
3. Some Japanese models outperformed the NGA models in the far field, highlighting the significance of the regional effect of attenuation.
4. The regional optimizations for two NGA-West2 models (Abrahamson *et al.*, 2014; Boore *et al.*, 2014) improved the model over a wide distance range. The effect of the optimization for Campbell and Bozorgnia (2014) was somewhat mixed, while that for Chiou and Youngs (2014) had an adverse effect. This highlights the need to independently test a model before its implementation in seismic hazard analysis, even if the model is developed based on seemingly sound physics.

Our results imply that seismic hazard assessments for Japan can be improved if state-of-the-art NGA-West2 models are adopted, or new Japanese models that include the additional factors considered by NGA-West2 models, such as near-field effects and more sophisticated correlation structures, are being developed.

We empirically evaluated the performance of two New Zealand GMMs, four NGA models, and four NGA-West2 models using observed PGA and SAs (at 0.5, 1, and 3 s) of 14 New Zealand shallow crustal earthquakes with magnitudes 5.07–7.85. We paid due respect to the correlation structure of the models and the variability of the results. We found that:

1. Bradley (2013), a recently developed regional model for New Zealand, performed well over a wide range of distance for PGA and SA at 0.3 s, and for the near field for SAs at 1 and 3 s. In the near field, there are other models with performance comparable to Bradley (2013), such as Abrahamson *et al.* (2014) for SAs at 0.3 and 1 s and Chiou and Youngs (2014) for SA at 3 s.
2. Our results are stable but to a less degree compared with the results for Japan. We can identify groups of GMMs with good performance. The relative performance of the models within the group, however, may change when more data become available. This identifies an epistemic uncertainty for GMMs for New Zealand that hazard modelers should observe.
3. The performance gain of NGA-West2 models over their corresponding predecessors is mixed and depends on spectral period. There are occasions for which a superseded NGA model performed similar to or even better than the corresponding NGA-West2 model. This highlights the need to independently test a model before its implementation in seismic hazard analysis, even if the model is developed based on seemingly sound physics.

Our results imply that the epistemic uncertainty for the seismic hazard for New Zealand could be better captured if state-of-the-art global GMMs are adopted.

Our results imply that the question of whether a regional GMM should be preferred to a global model may not have a general answer. We found both the case of global models outperforming contemporary regional models (NGA vs. Japanese models) as well as the case of a regional model performing comparably with global models (NGA-West2 models vs. Bradley, 2013). It appears that the only clear answer is that GMMs are better first tested using independent data before being implemented in seismic hazard analysis.

Data and Resources

The NGA-West2 flatfile inspected in this study is available online (peer.berkeley.edu/ngawest2/databases/). The 2014 version of the National Seismic Hazard Maps for Japan (NSHMJ14), published (in Japanese) by the Earthquake Research Committee, The Headquarters for Earthquake Research Promotion, is available online (www.jishin.go.jp/evaluation/seismic_hazard_map/shm_report/shm_report_2014/). The Japan Meteorological Agency (JMA) catalog is available at www.jma.go.jp. Moment magnitudes for Japanese earthquakes were downloaded from F-net (www.fnet.bosai.go.jp/fnet). Strong-motion records for Japan were downloaded from K-NET/KiK-net (www.kyoshin.bosai.go.jp/kyoshin). The deep sediment model used to compute the basin depth (Ⓔ Table S2) was downloaded from the Japan Seismic Hazard Information System (JSHIS; www.j-shis.bosai.go.jp/map/JSHIS2/download.html?lang=en). The finite-fault model for Maeda and Sasatani (2009) was obtained directly from Takahiro Maeda (Natural Research Institute for Earth Science and Disaster Prevention [NIED]). The finite-fault model for Asano and Iwata (2006) was downloaded from SRCMOD (equakerc.info/SRCMOD). The finite-fault model for the 2006 Offshore Eastern Izu Peninsular earthquake was inferred from a figure produced by the Geospatial Information Authority of Japan (GSI; www.jishin.go.jp/main/chousa/06may_izu/p11.htm). The finite-fault model for Aoi *et al.* (2010) was obtained directly from Wataru Suzuki (NIED). The finite-fault parameters for the 2011 Northern Nagano-ken, the 2011 Northern Ibaraki-ken, the 2011 Suruga Bay, and the 2013 Awaji Island earthquakes were obtained partly from the JMA webpage (www.data.jma.go.jp/svd/eqev/data/sourceprocess) and partly from Koji Sakota (JMA) directly. The finite-fault model for Fukushima *et al.* (2013) was obtained directly from Yo Fukushima (Tohoku University). The finite-fault model for Hikima (2014) was obtained directly from Kazuhito Hikima Tokyo Electric Power Company (TEPCO). The finite-fault model for GSI (2015) was obtained directly from Tomokazu Kobayashi (GSI). The within-event and between-event sigmas for Kanno *et al.* (2006) were obtained them directly from Tatsuo Kanno (Kyushu University). The plate boundary data used in Figures 1 and 4 were download from peterbird.name/publications/2003_PB2002/2003_PB2002.htm. Predicted ground motions for some ground-motion models (GMMs) were computed using the OpenQuake Hazard Library (https://github.com/gem/oq-hazardlib). Figures 1 and 4 were prepared using Generic Mapping Tools (Wessel *et al.*, 2013). The above webpages were last accessed on January 2016. The strong-motion flat file for New Zealand was downloaded from GeoNet (info.geonet.org.nz/display/appdata/The+New+Zealand+Strong-Motion+Database, last accessed April 2017).

Acknowledgments

The authors thank all modelers who have provided unpublished information about their models (see Data and Resources). The authors thank Timothy Ancheta (Risk Management Solutions [RMS]) for explaining how some of the Next Generation Attenuation-West2 (NGA-West2) metadata were computed. The authors acknowledge the Natural Research Institute for Earth Science and Disaster Prevention (NIED) (F-net, K-NET, and KiK-net) and the Japan Meteorological Agency (JMA), Japan, for providing data used in this study. The authors acknowledge the New Zealand GeoNet project and its sponsors Earthquake Commission (EQC), GNS Science, and Land Information New Zealand (LINZ), for providing data used in this study. The authors thank Brendon Bradley (University of Canterbury) and an anonymous reviewer for their constructive comments that have substantially improved this article. The first author was supported by the ReThinking Probabilistic Seismic Hazard Analysis (PSHA) project financed by the New Zealand Natural Hazards Research Platform for a visit to GNS Science for developing this study. This study was supported by the Global Earthquake Model Foundation and the King Abdullah University of Science and Technology (KAUST) research Grant URF/1/2160-01-01.

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. A., and W. J. Silva (1997). Empirical response spectral attenuation relations for shallow crustal earthquakes, *Seismol. Res. Lett.* **68,** no. 1, 94–127, doi: 10.1785/gssrl.68.1.94.

Abrahamson, N. A., W. J. Silva, and R. Kamai (2014). Summary of the ASK14 ground motion relation for active crustal regions, *Earthq. Spectra* **30,** no. 3, 1025–1055, doi: 10.1193/070913EQS198M.

Al Atik, L., N. Abrahamson, J. J. Bommer, F. Scherbaum, F. Cotton, and N. Kuehn (2010). The variability of ground-motion prediction models and its components, *Seismol. Res. Lett.* **81,** no. 5, 794–801, doi: 10.1785/gssrl.81.5.794.

Ancheta, T. D., R. B. Darragh, J. P. Stewart, E. Seyhan, W. J. Silva, B. S.-J. Chiou, K. E. Wooddell, R. W. Graves, A. R. Kottke, D. M. Boore, *et al.* (2014). NGA-West2 database, *Earthq. Spectra* **30,** no. 3, 989–1005, doi: 10.1193/070913EQS197M.

Aoi, S., B. Enescu, W. Suzuki, Y. Asano, K. Obara, T. Kunugi, and K. Shiomi (2010). Stress transfer in the Tokai subduction zone from the 2009 Suruga Bay earthquake in Japan, *Nature Geosci.* **3,** 496–500, doi: 10.1038/NGEO885.

Asano, K., and T. Iwata (2006). Source process and near-source ground motions of the 2005 West Off Fukuoka Prefecture earthquake, *Earth Planets Space* **58,** 93–98, doi: 10.1186/BF03351920.

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.

Beyer, K., and J. J. Bommer (2006). Relationships between median values and between aleatory variabilities for different definitions of the horizontal component of motion, *Bull. Seismol. Soc. Am.* **96,** no. 4A, 1512–1522, doi: 10.1785/0120050210.

Bindi, D. (2017). The predictive power of ground-motion prediction equations, *Bull. Seismol. Soc. Am.* **107,** no. 2, 1005–1011, doi: 10.1785/0120160224.

Bird, P. (2003). An updated digital model of plate boundaries, *Geochem. Geophys. Geosys.* **4,** no. 3, 1027, doi: 10.1029/2001GC000252.

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. (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. P. Stewart, E. Seyhan, and G. M. Atkinson (2014). NGA-West2 equations for predicting PGA, PGV, and 5% damped PSA for shallow crustal earthquakes, *Earthq. Spectra* **30,** no. 30, 1057–1085, doi: 10.1193/070113EQS184M.

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. (2013). A New Zealand-specific pseudospectral acceleration ground-motion prediction equation for active shallow crustal earthquakes based on foreign models, *Bull. Seismol. Soc. Am.* **103,** no. 3, 1801–1822, doi: 10.1785/0120120021.

Campbell, K. W. (1997). Empirical near-source attenuation relationships for horizontal and vertical components of peak ground acceleration, peak ground velocity, and pseudo-absolute acceleration response spectra, *Seismol. Res. Lett.* **68,** no. 1, 154–179, doi: 10.1785/gssrl.68.1.154.

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.

Campbell, K. W., and Y. Bozorgnia (2014). NGA-West2 ground motion model for the average horizontal components of PGA, PGV, and 5% damped linear acceleration response spectra, *Earthq. Spectra* **30,** no. 30, 1087–1115, doi: 10.1193/062913EQS175M.

Chiou, B., R. Youngs, N. Abrahamson, and K. Addo (2010). Ground-motion attenuation model for small-to-moderate shallow crustal earthquakes in California and its implications on regionalization of ground-motion prediction models, *Earthq. Spectra* **26,** 907–926, doi: 10.1193/1.3479930.

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.

Chiou, B. S.-J., and R. R. Youngs (2014). Update of the Chiou and Youngs NGA model for the average horizontal component of peak ground motion and response spectra, *Earthq. Spectra* **30,** no. 30, 1117–1153, doi: 10.1193/072813EQS219M.

Cotton, F., F. Scherbaum, J. J. Bommer, and H. Bungum (2006). Criteria for selecting and adjusting ground-motion models for specific target regions: Application to Central Europe and rock sites, *J. Seismol.* **10,** no. 2, 137–156, doi: 10.1007/s10950-005-9006-7.

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.

Douglas, J. (2011). Investigating possible regional dependence in strong ground motions, in *Earthquake Data in Engineering Seismology*, S. Akkar, P. Gülkan, and T. van Eck (Editors), Springer, Dordrecht, The Netherlands, 29–38, ISBN: 978-94-007-0152-6.

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.

Edwards, B., C. Cauzzi, L. Danciu, and D. Fäh (2016). Region-specific assessment, adjustment, and weighting of ground-motion prediction models: Application to the 2015 Swiss seismic-hazard maps, *Bull. Seismol. Soc. Am.* **106,** no. 4, 1840–1857, doi: 10.1785/0120150367.

Fujimoto, K., and S. Modorikawa (2006). Relationship between average shear-wave velocity and site amplification inferred from strong motion records at nearby station pairs, *J. Japan Assoc. Earthq. Eng.* **6,** no. 1, 11–22, doi: 10.5610/jaee.6.11 (in Japanese with English abstract).

Fujita, E., T. Kozono, H. Ueda, Y. Kohno, S. Yoshioka, N. Toda, A. Kikuchi, and Y. Ida (2013). Stress field change around the Mount Fuji volcano magma system caused by the Tohoku megathrust earthquake, Japan, *Bull. Volcanol.* **75,** 679, doi: 10.1007/s00445-012-0679-9.

Fujiwara, H., S. Kawai, S. Aoi, N. Morikawa, S. Senna, N. Kudo, M. Ooi, K. X.-S. Hao, K. Wakamatsu, Y. Ishikawa, *et al.* (2009). *Technical Reports on National Seismic Hazard Maps for Japan*, Technical Note of the National Research Institute for Earth Science and Disaster Prevention No. 366, National Research Institute for Earth Science and Disaster Prevention, Ibaraki, Japan, ISSN: 0917-057X (in Japanese).

Fukushima, Y., Y. Takada, and M. Hashimoto (2013). Complex ruptures of the 11 April 2011 M_w 6.6 Iwaki earthquake triggered by the 11 March 2011 M_w 9.0 Tohoku earthquake, Japan, *Bull. Seismol. Soc. Am.* **103,** no. 2B, 1572–1583, doi: 10.1785/0120120140.

Geospatial Information Authority of Japan (GSI) (2015). Crustal movements in the Kanto district, in *Report of the Coordinating Committee for Earthquake Prediction*, Vol. 94, Geospatial Information Authority of Japan, Tsukuba, Japan, 112–125 (in Japanese).

Hikima, K. (2014). Source process of the February 25, 2013 Tochigi-ken Hokubu earthquake (M_j 6.3) - part 3—Analyses using empirical and theoretical Green's functions, *Abstracts of the Fall Meeting of the Seismological Society of Japan*, S15-P22, Niigata, Japan, 24–26 November 2014 (in Japanese).

Horike, M., and T. Nishimura (2004). Attenuation relationships of peak ground velocity inferred from the Kyoshin network data, *J. Struct. Constr. Eng., AIJ* **575,** 73–79 (in Japanese with English abstract).

Kaiser, A., C. Van Houtte, N. Perrin, L. Wotherspoon, and G. McVerry (2017). Site characterisation of GeoNet stations for the New Zealand strong motion database, *Bull. New Zeal. Soc. Earthq. Eng.* **50,** no. 1, 39–49.

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.

Kanno, T., A. Narita, N. Morikawa, H. Fujiwara, and Y. Fukushima (2006). A new attenuation relation for strong ground motion in Japan based on recorded data, *Bull. Seismol. Soc. Am.* **96,** no. 3, 879–897, doi: 10.1785/0120050138.

Kobayashi, T., M. Tobita, A. Suzuki, and Y. Noguchi (2011). InSAR analysis of the 2010 Fukushima-ken Nakadori earthquake (M5.7), *J. Geospatial Inf. Auth. Japan* **121,** 165–169 (in Japanese).

Kotha, S. R., D. Bindi, and F. Cotton (2016). Partially non-ergodic region specific GMPE for Europe and Middle-East, *Bull. Earthq. Eng.* **14,** 1245–1263, doi: 10.1007/s10518-016-9875-x.

Maeda, T., and T. Sasatani (2009). Strong ground motions from an M_j 6.1 inland crustal earthquake in Hokkaido, Japan: The 2004 Rumoi earthquake, *Earth Planets Space* **61,** 689–701.

Mak, S., R. A. Clements, and D. Schorlemmer (2017). Empirical evaluation of hierarchical ground motion models: Score uncertainty and model weighting, *Bull. Seismol. Soc. Am.* **107,** no. 2, 949–965, doi: 10.1785/0120160232.

Mak, S., F. Cotton, and D. Schorlemmer (2017). Measuring the performance of ground-motion models: The importance of being independent, *Seismol. Res. Lett.* **88,** no. 5, 1212–1217, doi: 10.1785/0220170097.

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.

Midorikawa, S., and Y. Ohtake (2002). Attenuation relationships of peak ground acceleration and velocity considering attenuation characteristics for shallow and deeper earthquakes, *Proc. of the 11th Japan Earthquake Engineering Symposium*, Number 117, Tokyo, Japan, 20–22 November 2002, 609–614 (in Japanese with English abstract).

Molas, G. L., and F. Yamazaki (1995). Attenuation of earthquake ground motion in Japan including deep focus events, *Bull. Seismol. Soc. Am.* **85,** no. 5, 1343–1358.

Morikawa, N., T. Kanno, A. Narita, H. Fujiwara, and Y. Fukushima (2003). Additional correction terms for attenuation relations corresponding to the anomalous seismic intensity in northeast Japan, *J. Japan Assoc. Earthq. Eng.* **3,** no. 4, 14–26 (in Japanese with English abstract).

Morikawa, N., T. Kanno, A. Narita, H. Fujiwara, and Y. Fukushima (2006). New additional correction terms for attenuation relations of peak amplitudes and response spectra corresponding to the anomalous seismic intensity in Northeast Japan, *J. Japan Assoc. Earthq. Eng.* **6,** no. 1, 23–41 (in Japanese with English abstract).

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).

NZS 1170.5 Supp 1:2004 (2004). *Structural Design Actions Part 5: Earthquake Actions—New Zealand—Commentary*, Standards New Zealand, Wellington, New Zealand, ISBN: 1-86975-019-5.

NZS 1170.5:2004 (2004). *Structural Design Actions Part 5: Earthquake Actions—New Zealand*, Standards New Zealand, ISBN: 1-86975-018-7.

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., 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.

Si, H., and S. Midorikawa (1999). New attenuation relationships for peak ground acceleration and velocity considering effects of fault type and site condition, *J. Struct. Constr. Eng., AIJ* **523,** 63–70 (in Japanese with English abstract).

Stirling, M., G. McVerry, M. Gerstenberger, N. Litchfield, R. V. Dissen, K. Berryman, P. Barnes, L. Wallace, P. Villamor, R. Langridge, *et al.* (2012). National seismic hazard model for New Zealand: 2010 update, *Bull. Seismol. Soc. Am.* **102,** no. 4, 1514–1542, doi: 10.1785/0120110170.

Stucchi, M., C. Meletti, V. Montaldo, H. Crowley, G. M. Calvi, and E. Boschi (2011). Seismic hazard assessment (2003-2009) for the Italian building code, *Bull. Seismol. Soc. Am.* **101,** no. 4, 1885–1911, doi: 10.1785/0120100130.

Van Houtte, C. (2017). Performance of response spectral ground-motion models against New Zealand data, *Bull. New Zeal. Soc. Earthq. Eng.* **50,** no. 1, 21–38.

Van Houtte, C., S. Bannister, C. Holden, S. Bourguignon, and G. McVerry (2017). The New Zealand strong motion database, *Bull. New Zeal. Soc. Earthq. Eng.* **50,** no. 1, 1–20.

Wang, Y.-J., C.-H. Chan, Y.-T. Lee, K.-F. Ma, J. B. H. Shyu, R.-J. Rau, and C.-T. Cheng (2016). Probabilistic seismic hazard assessment for Taiwan, *Terr. Atmos. Ocean. Sci.* **27,** no. 3, 325–340, doi: 10.3319/TAO.2016.05.03.01(TEM).

Wessel, P., W. H. F. Smith, R. Scharroo, J. F. Luis, and F. Wobbe (2013). Generic Mapping Tools: Improved version released, *Eos Trans. AGU* **94,** 409–410, doi: 10.1002/2013EO450001.

Appendix

Potential Problem in Partitioning Residuals of a Biased Model

Residual analyses in the literature often partition residuals of a ground-motion model (GMM) into the between-event and the leftover residuals (called the within-event residuals, when the correlation structure of the GMM includes only the correlation by event). Such partitioning requires subtracting the event terms from the residuals. It is worth noting that the event term is not directly observable. It is merely a construct used in a mixed-effect model. Event terms are estimated by assuming that the model is unbiased, a natural assumption during the model-creation process. In an empirical evaluation of GMMs for the purpose of assessing the relative performance of models using new data that the models are not fitted to, however, such an assumption could be inappropriate.

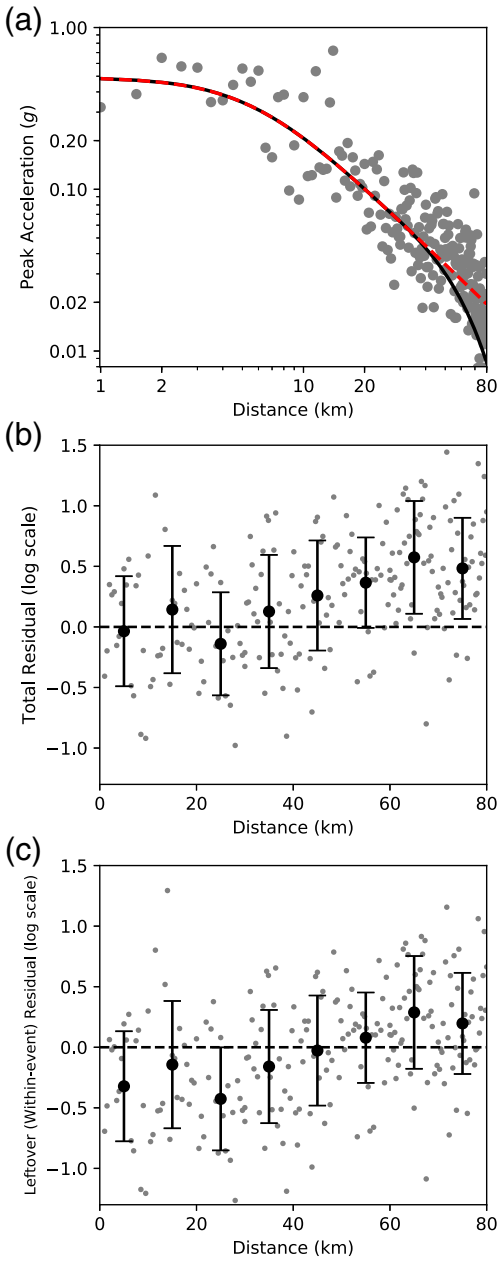

Figure A1. (a) Biased prediction (solid line) versus actual mean ground motion (dashed line). The synthetic observations (dots) were generated by adding random errors (zero-mean normally distributed with standard deviation of 0.48) to the mean of the true model. (b) Total residuals versus distance. The moving average is indicated by the solid line. (c) Leftover residuals versus distance. The leftover residual is the residual minus the estimated event term. The moving average is indicated by the solid line. The color version of this figure is available only in the electronic edition.

Figure A1 illustrates a possible consequence for a residual analysis based on partitioned residuals. Suppose a GMM can predict the actual mean ground model (in logarithmic scale) well, except for a bias in the far field (Fig. A1a).

The total residuals (Fig. A1b) demonstrate this bias correctly. To partition the residuals, one needs to estimate the event term. An approximation of the event term is the mean residual for the event (e.g., Shoja-Taheri *et al.*, 2010, their equations 1–4). The leftover residuals (Fig. A1c) will then show a bias in both the near and the far fields, an artifact due to the abovementioned bias. What we have shown here is a simple bias versus distance. A more complicated bias that involves multiple factors could lead to other apparent features. Interpretations (or even more dangerously, overinterpretations) of these apparent features are often not conducive to understanding the model performance.

Residual partitioning is even more problematic for GMMs of more complicated correlation structures (e.g., Al Atik *et al.*, 2010). Model bias could be distributed into more components, leading to even more uninterpretable patterns. In the current study, we did not use residual partitioning to evaluate the relative performance of GMMs.

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

GNS Science
1 Fairway Drive, Avalon 5010
Lower Hutt 5040, New Zealand
(M.G.)

Manuscript received 17 May 2017;
Published Online 30 January 2018