DOI: http://doi.org/10.1785/0120150323

SSA

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. 106, No. 4, pp. 1817–1831, August 2016, doi: 10.1785/0120150323

A Comparison between the Forecast by the United States National Seismic Hazard Maps with Recent Ground-Motion Records

by Sum Mak and Danijel Schorlemmer

Abstract Confidence in scientific models accumulates by continuously validating the model's predictions by observations. We compared the seismic-hazard forecasts of the four published versions of the U.S. Geological Survey National Seismic Hazard Maps with observed ground motions. A large dataset is necessary for a statistically meaningful comparison, and so our comparison was based on an aggregated approach such that the observations and the forecast in a region (California, and the central and eastern United States [CEUS]) were combined and compared as a whole. We used instrumental records in California and macroseismic intensity in the CEUS since 2000 as the observation, which was largely prospective to the hazard maps. We verified that the observed seismic hazard based on macroseismic intensity was consistent with that based on instrumental records, making model evaluation in the CEUS, for which instrumental records were almost nonexistent, viable. The observed hazard was found to be generally consistent with the forecasted one for peak ground acceleration (PGA) in California and for both PGA and spectral acceleration at 1 s (SA1) in the CEUS. Forecasted hazard for SA1 for California appeared to be conservative. Recent versions of the hazard map were in better agreement with observations in California. Small earthquakes, as expected, were found to have insignificant impact on SA1. Induced earthquakes in the CEUS have increased the observed seismic hazard but did not invalidate the hazard model as a whole. We examined the resolving power of the test by computing its statistical power.

Introduction

The necessity to validate forecasts of a probabilistic seismic-hazard assessment (PSHA) model using observed ground motion has long been recognized (McGuire, 1979). This validation exercise (validation, verification, and testing are used in this article as synonyms, although workers in different fields may assign different meanings to these terms) has been receiving increasing attention in recent years (Ordaz and Reyes, 1999; Stirling and Petersen, 2006; Albarello and D'Amico, 2008; Fujiwara *et al.*, 2009; Miyazawa and Mori, 2009; Stirling and Gerstenberger, 2010; Mezcua *et al.*, 2013; Tasan *et al.*, 2014; Brooks *et al.*, 2016), although it has not yet been a routine and standard process such as how meteorologists treat the weather forecast (e.g., the quality of weather forecasts is regularly published by the European Centre for Medium-Range Weather Forecasts [ECMWF], see Data and Resources). The lack of a standard validation may have induced arguments over whether a PSHA model is serving its purpose, especially after a significant earthquake occurs not in the most expected place and/or is not of the most expected size (Stein *et al.*, 2011, 2012, 2013; Hanks *et al.*, 2012; Stirling, 2012; Frankel, 2013a,b). Extreme views such as "The problem in earthquake forecast is that models . . . have not been tested against relevant data, . . . so there is little reason to believe the probability estimates" (Stark and Freedman, 2003, p. 205) exist, often from workers with background primarily in mathematical statistics.

For the United States, comparisons between the forecasted hazard of the U.S. Geological Survey (USGS) National Seismic Hazard Maps (NSHM, see Data and Resources) were made using historical macroseismic intensity (Stirling and Petersen, 2006) and instrumental records (Stirling and Gerstenberger, 2011; see Data and Resources). Certain issues critical to the statistical comparison, including the treatment of data completeness, correlation between observations, dependence between observations and models, and the formality and resolving power of statistical methods, have not been fully explored in previous studies. In the current study, we compared the forecasted time-independent seismic hazard by the NSHM with the observed hazard derived from ground-motion records since 2000. We targeted two regions of the United States: California, where instrumental records were relatively abundant, and the central and eastern United States (CEUS, defined as east of 100°E), where instrumental records were nearly nonexistent and we had to resort to macroseismic intensity records. The two regions represent the two extremes of seismic hazard in the United States. We paid special attention to the data completeness, the quality of macroseismic intensity records, and their conformity with instrumental records, as well as the sufficiency of data quantity, represented by the statistical power of tests; the latter issue is especially important but has not been sufficiently addressed in early studies.

The evaluation of a PSHA model includes two approaches. The component-based evaluation of PSHA (e.g., Schorlemmer *et al.*, 2007, for seismicity forecast; Scherbaum *et al.*, 2009, for ground-motion forecast) addresses the performance of an individual component of a model. By deductive reasoning, if all components of a model are in good quality, the output of the model should be in good quality. On the other hand, the holistic approach, such as that used in the current study, provides direct inductive evidence as to the correctness of the model. It serves the same purpose as the clinical trial for medical research. Both approaches are necessary for a complete evaluation of a PSHA model.

Method and Data

Aggregated Hazard Curve

An ideal test of a PSHA model, just like an ideal PSHA model, should be site-specific so that the forecasted hazard of a site is tested against the observed hazard of that site. This approach was used in some early studies (Ordaz and Reyes, 1999; Stirling and Petersen, 2006; Stirling and Gerstenberger, 2010; Mezcua *et al.*, 2013), although it is now known that their statistical powers are unlikely to be high, so that the test is unlikely to find inconsistency between the observed and forecasted hazards, unless the model is grossly wrong (Mak *et al.*, 2014). To include more data to achieve a higher power, one can sacrifice the spatial resolution and compare the performance of the PSHA model for a region as a whole, using all observations in that region. Tasan *et al.* (2014, pp. 1555–1556) summarized this aggregated approach. Some of the main points are repeated below. The computational details are given in Appendix A.

The essence of testing a PSHA model is to compile the observed hazard by counting the recorded ground motions and to compare it with the forecasted hazard. The forecasted hazard, a random variable, is displayed in this article as its expected value (i.e., expected hazard) associated with an interval bounded by its 5% and 95% quantiles (hereafter referred to as the 5%–95% forecast interval). If the observed hazard falls outside this interval, the forecast is judged as inconsistent with the observation. This is equivalent to the conventional hypothesis test for model rejection with $\alpha = 5\%$ for a one-tail test, or 10% for a two-tail test; the null hypothesis is that the observed and forecasted aggregated hazards are the same.

An aggregated hazard curve represents the summation of observed or forecasted hazards over multiple sites. The conventional site-specific hazard curve uses the annual rate of exceedance as the ordinate and the ground-motion level (e.g., peak ground acceleration, known as PGA) as the abscissa. One method for computing the aggregated hazard curve is to use the total number of ground-motion exceedances over multiple sites as the ordinate. The corresponding expected hazard curve is simply the sum of all site-specific hazard curves, each multiplied by its own period of observation. Because the number of ground-motion exceedances at a single site is modeled as a Poisson random variable, the sum of them is also a Poisson random variable. The distribution function of the aggregated forecast can be computed accordingly.

Another method for computing the aggregate hazard curve is to use the number of sites with exceedances (i.e., sites that have experienced at least one exceedance during the observation period) as the ordinate. This is in line with the conventional practice of seismic hazard maps to express the seismic hazard in terms of probability of exceedance. The corresponding forecasted aggregated hazard is the sum of multiple Bernoulli trials, each having a different probability of success, which is the probability of observing at least one exceedance during the observation period, based on a Poisson model (Albarello and D'Amico, 2008, section 2.1). Such a sum of heterogeneous Bernoulli trials results in a Poisson-binomial distribution (Wang, 1993). The distribution function of the aggregated forecast can be computed accordingly. Albarello and D'Amico (2008) computed this distribution by normal approximation and Tasan *et al.* (2014) computed it by Monte Carlo simulations. Both are not strictly necessary.

The above-mentioned computation of the distribution function of the forecasted hazard requires the sites to be independent. When they are not, the actual variance will be larger than the computed one; note that the mean is not affected by site dependence. It is difficult to accurately estimate the dependence of the seismic hazards among sites. Sites sufficiently apart can be assumed to be independent, but it is unknown precisely how far is sufficient. Applying a conservative minimum intersite distance to ensure independence will lead to discarding a lot of data, which leads to low statistical power. We elaborate more on this issue in the Decisions Required to Conduct Testing section.

A PSHA model provides the annual rates of exceedance, the reciprocal of return period, at various ground-motion levels. For a given PSHA model, ground-motion level and return period are interchangeable. Therefore, the abscissa of an aggregated hazard curve can be expressed in either the ground-motion level or the return period; the latter requires specifying a PSHA model as the means of conversion. The same return period will represent different ground-motion levels at different sites. Together with the two choices of ordinate, there can be four forms of aggregated hazard curve (Fig. 1; see also Appendix A). When the abscissa is expressed in ground-motion levels (Fig. 1a,b), the observed hazard is model independent, whereas the expected (or forecasted) hazard is model dependent. On the other hand, when the abscissa is expressed in return periods (Fig. 1c,d), the observed hazard is model dependent, whereas the expected hazard is model independent. See Appendix A for details.

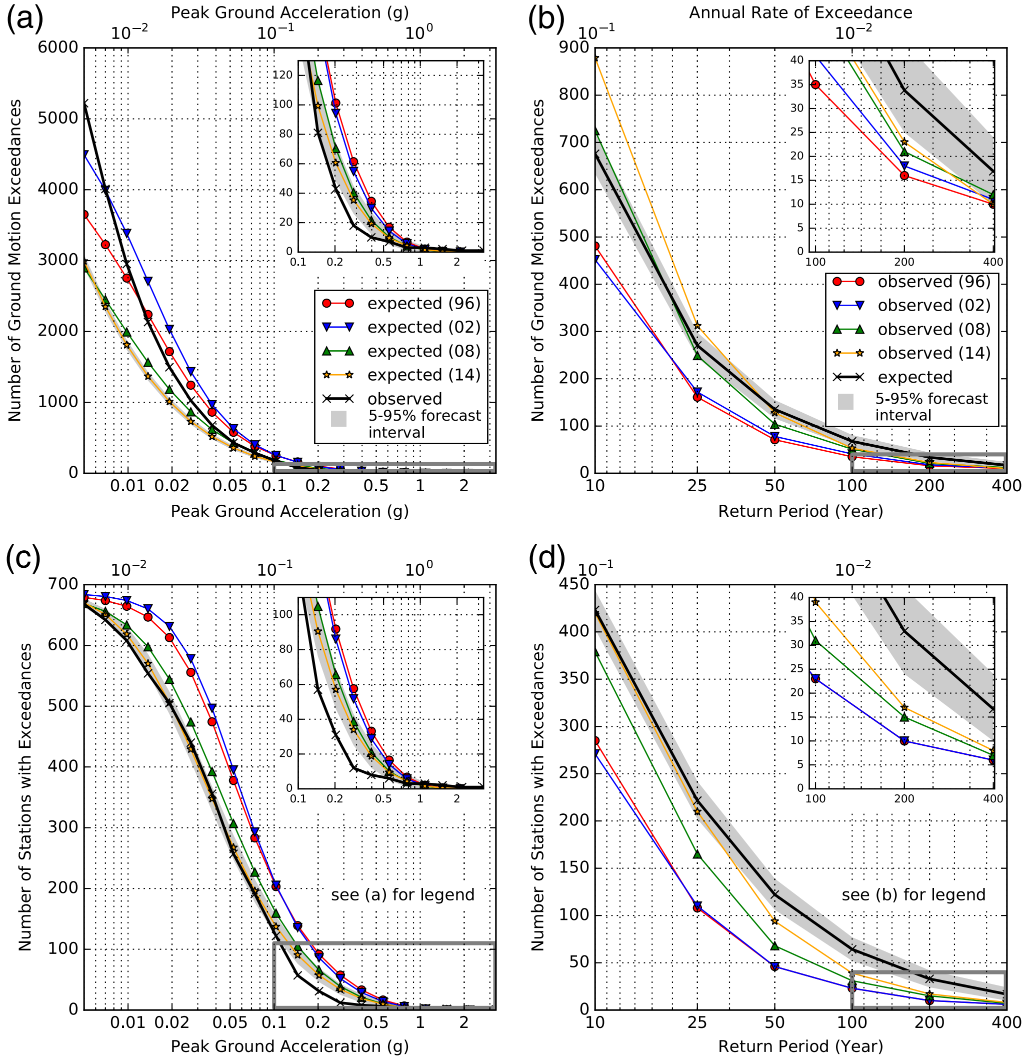

Figure 1. Four forms of aggregated hazard curve based on all available California instrumental peak ground acceleration (PGA) records. See Appendix A for the computational details. (a) Form 1: number of ground-motion exceedances versus PGA. The shaded region surrounding the expected hazard curve based on the National Seismic Hazard Maps (NSHM) 2014 is the 5%–95% forecast interval that the probability of the observed hazards to fall within is 90% if the forecast is correct, assuming the sites are independent (for which are not; see the Suppressing Data Correlation section). (b) Form 3: number of ground-motion exceedances versus return period. (c) Form 2: number of stations that have experienced at least one ground-motion exceedance versus PGA. (d) Form 4: number of stations that have experienced at least one ground-motion exceedance versus return period. Boxed regions at the tails are enlarged to be insets. See Figure 2a for station locations. The color version of this figure is available only in the electronic edition.

In this study, we express aggregated hazard curves as the number of sites with exceedances versus return period. The reason for not using absolute ground-motion levels as the abscissa is that the same ground-motion level could carry very different meanings between California and the CEUS, whereas the use of return period is not regional dependent. The reason for not using the number of ground-motion exceedances as the ordinate is that aftershocks and small earthquakes, which are excluded in the NSHM, will almost certainly generate a number of observed ground-motion exceedances at a low ground-motion level larger than the forecasted number (see Fig. 1a); on the other hand, aftershocks do not affect the counting of the number of sites with exceedances. The modeler's decision to exclude small earthquakes is based on the assumption that they do not contribute significantly to seismic hazards, especially for spectral periods of engineering interest. We put this assumption to test.

The four versions of the NSHM (1996, 2002, 2008, and 2014) provide the seismic hazards in a number of ground-motion units. The PGA and the spectral acceleration at 1 s (SA1) are common among them. Observations in these two units are also available from instrumental records of ShakeMap stations and intensity records of the "Did You Feel It?" (DYFI) system via conversion (explained below). Therefore, the comparison in this study was based on PGA and SA1. The mean hazard curve was used as the basis of comparison because it is the one most often published and used in practice (McGuire *et al.*, 2005; Musson, 2005). We

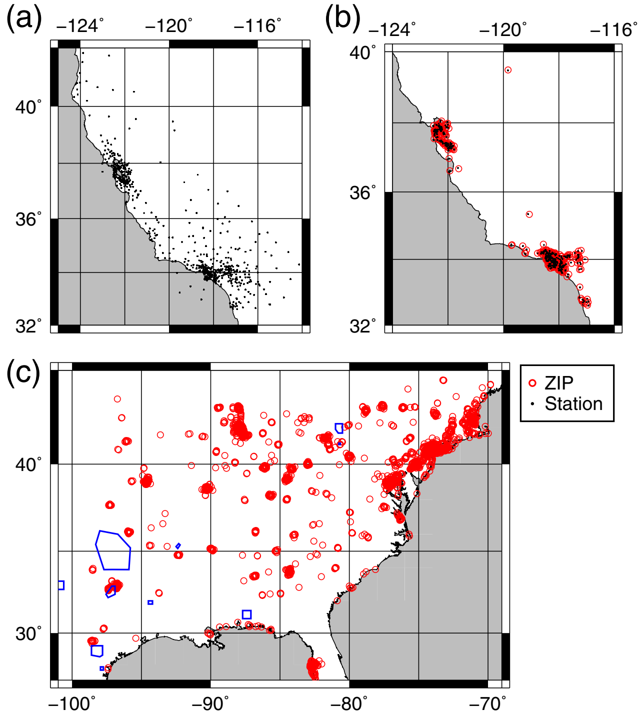

Figure 2. Locations for ShakeMap stations (dots) and ZIP regions (hollow circles) used in the current study. (a) All California stations. The corresponding aggregated hazard curves are given in Figure 1. (b) Pairs of California stations and ZIP regions that are within 5 km from each other. The corresponding aggregated hazard curves are given in Figure 3. (c) All central and eastern United States (CEUS) ZIP regions. Polygons denote identified zones of induced seismicity (see Data and Resources). The color version of this figure is available only in the electronic edition.

discuss the evaluation of ensemble forecasts in the Beyond the Mean Forecast section.

Instrumental Records for California

The record from an accelerometer is the best scientific observation of strong ground motions. We used the records provided by ShakeMap stations (see Data and Resources) as the data source for California. ShakeMap is designed for earthquake-hazard information dissemination. It has archived a comprehensive dataset of strong motion compiled from records from various seismic networks, but the data are not optimized for scientific analysis of strong motion. In the following, we describe two measures we implemented to ensure the data validity and completeness of ShakeMap records.

First, the station name for the ShakeMap database is not unique. Stations of the same name are sometimes distinct, usually because the same station name is used by different agencies. In addition, the same station occasionally has different names at different times. We manually checked the names and locations of each ShakeMap station to ensure that records of the same station name came for the same station, as well as grouped together the records for the same station with different names.

Table 1
Empirical Period of Completeness (years) for ShakeMap Stations in California

Period (year)	Number of Stations
5–8	182
8–10	161
10–12	173
12–13.1	174

Second, ShakeMap does not document the operational history of the seismic networks it uses. Consequently, there is no rigorous way to guarantee the data completeness. We estimated empirically the data completeness by checking if a station has reported all ground motions that are believed to be sufficiently strong by the ground-motion prediction equation (GMPE) of Boore *et al.* (2014). We considered the observation from a ShakeMap station of the time period between the first and the last records of that station to be potentially complete. We then predicted the mean-minus-one-standard-deviation logarithmic PGA of that station for each earthquake in the ShakeMap catalog; we assumed that the ShakeMap catalog was complete for California earthquakes of engineering significance. If the ShakeMap database had included all records of that station when the corresponding predicted PGA was larger than $0.02g$, we considered the records for that station to be complete for the above-mentioned time period. This empirical completeness means that if the actual PGA was likely ($>68\%$ by the lognormal distribution adopted by GMPEs) to be at least $0.02g$, the ground motion would have been included in the ShakeMap database. A similar approach of empirical completeness estimation was used by Tasan *et al.* (2014, section 3.1). The use of GMPE required parameters such as the station V_{S30}. We obtained the parameters from the Next Generation Attenuation (NGA)-West2 flatfile (see Data and Resources) by matching the locations of the ShakeMap stations with those of the NGA-West2 stations. Because of round-off errors, the locations for the same station in the two databases might not be identical; we tolerated a location mismatch of 0.5 km. ShakeMap stations not included in the NGA-West2 flatfile were not used. Only ShakeMap stations of an estimated completeness period of at least five years were used for analysis to avoid temporary stations, for which the data quality might be less satisfactory.

The above data selection resulted in 690 empirically complete stations (Fig. 2a), producing 99,885 PGA records from 1775 earthquakes during the period 22 February 2003 to 28 March 2016. Each station had a different estimated period of completeness (Table 1). The number of records for SA1 was slightly less because some stations produced only PGA but not SA1.

Because the NSHM predicts seismic hazard on rock ($V_{S30} = 760$ km/s), it is necessary to remove the site effect of the observed ground motions for fair comparison. We deamplified the observed ground motions using the nonlinear site amplification component of Boore *et al.* (2014, their equations 3.8–3.11). Boore *et al.* (2014) is the only NGA-West2 GMPE that uses the same rock definition as that of the NSHM, making it a convenient choice for the purpose of this study.

Intensity Records for the CEUS

For the CEUS, accelerometers were very local in coverage, and/or installed only very recently or temporarily, such that the duration of data availability was short; instrumental records suitable for the purpose of the current study therefore did not exist. We used macroseismic intensity data collected by the DYFI system (Wald *et al.*, 2011; see Data and Resources) as ground-motion observations for the CEUS. DYFI intensity records are similar to conventional macroseismic intensity reports, except that the intensity value is computed automatically based on the online questionnaires received from voluntary Internet users. A postal ZIP region is assigned an intensity value, representing the ground-shaking level of the region. Fundamental differences between macroseismic intensity and instrumental records demand additional treatments for using intensity data, compared with using instrumental records.

Intensity-Based Hazard Observation. To compare observed hazard in intensity with the forecasted hazard in acceleration, the use of an intensity-to-ground-motion conversion equation (IGMCE or GMICE) is necessary. We used the GMICE of Atkinson and Kaka (2007), which was specifically designed for eastern North America. We implemented a probabilistic conversion in which an intensity value was converted into a normal distribution of ground motion, with the mean as the converted value and the standard deviation as the reported standard deviation. The GMICE of Atkinson and Kaka (2007) was designed as a one-way GMICE, converting from ground motion to intensity; we used it for reverse conversion from intensity to ground motion. Atkinson and Kaka (2007) did not report the standard deviation of such a reverse conversion. Without a better way to quantify the conversion uncertainty, we assumed it to be the same as that reported in Worden *et al.* (2012), a more recently developed two-way GMICE for California.

There is no generally recognized method for site-effect adjustment for macroseismic intensity. We deamplified the converted ground motions as we did for instrumental records, assuming all sites to be National Earthquake Hazards Reduction Program (NEHRP) class C/D (i.e., $V_{S30} = 366$ m/s).

For N observed intensity values, the GMICE provides N converted ground motions μ_i (with site effects removed) and a standard deviation σ. We took the total number of ground motions exceeding ground-motion level g as N_g,

$$N_g = \sum_{i=1}^{N} \Pr(X \geq g|\mu_i, \sigma), \qquad (1)$$

in which Pr denotes probability and X is normally distributed with mean μ_i and standard deviation σ. This probabilistic conversion of intensity leads to fraction numbers, while the counting of instrumental records always results in integers.

Similarly, for M sites, each having N_i observed intensity values, converted by a GMICE into μ_{ij}, we took the total number of sites that have experienced at least one exceedance of ground-motion level g during the observation period as $\tilde{N}_g$:

$$\tilde{N}_g = \sum_{i=1}^{M} \left[1 - \prod_{j=1}^{N_i} \Pr(X < g|\mu_{ij}, \sigma) \right]. \qquad (2)$$

This again leads to fraction counts.

Quality Control. We implemented the following measures, in sequential order, to ensure the quality of the intensity data points (IDPs) used to represent the observed hazard in the CEUS.

Data Completeness. It is known that the quality of an IDP is proportional to the number of responses from which it is compiled (Worden *et al.*, 2012, their fig. 3). Mak and Schorlemmer (2016) created a statistical model to describe the probability for each ZIP region receiving a certain number of responses conditioned by various earthquake and socioeconomic properties. In the current study, we took only the ZIP regions that have at least 70% chance of producing at least 10 responses for a reported intensity value of 4, coming from a hypothetical earthquake of magnitude 5 located at 30 km from the site. Because the selected ZIP regions will very likely produce a large number of responses to the DYFI system under a nontrivial ground shaking, we assumed the records of a selected ZIP region to be complete, from the first time an intensity report had been generated for that ZIP region until the end of our data collection period (end of March 2016).

Induced Earthquakes. A large number of earthquakes in the CEUS were induced by fluid injection associated with petroleum extraction activities (Ellsworth, 2013), which were not considered in the NSHM. We excluded the ZIP regions and records from earthquakes that were located within identified zones of induced seismicity (Fig. 2c).

Areal Ground Shaking. Macroseismic intensity fundamentally represents an areal ground shaking, whereas the NSHM predicts point seismic hazard. We assumed IDPs to be point records at their reported geographic coordinates (usually the centroid of the ZIP region). For a large ZIP region, such a point approximation may be less accurate; we therefore excluded ZIP regions larger than 30 km^2.

Continual Recording. Consistent with the treatment of instrumental records, we took only the ZIP regions that have produced intensity reports for at least five years.

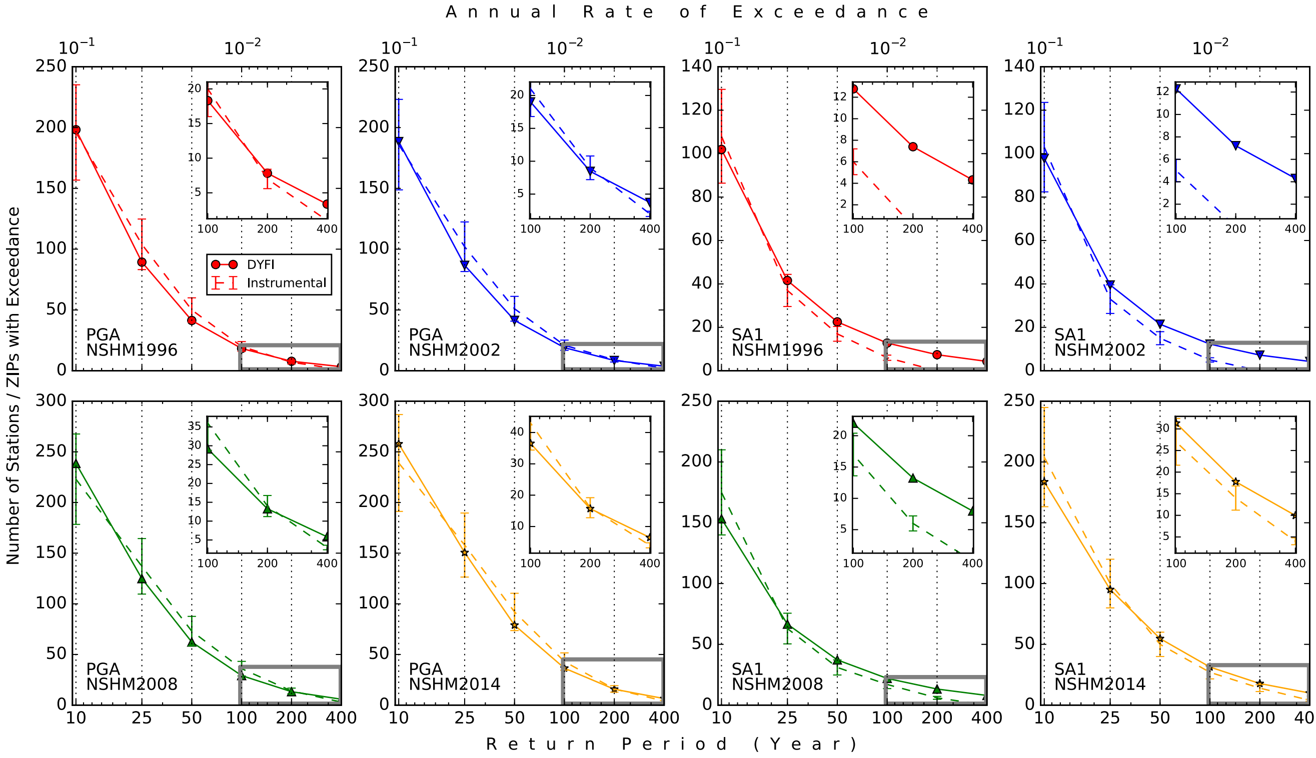

Figure 3. Observed hazard curves for California compiled from instrumental records (dashed line) and "Did You Feel It?" (DYFI) records (solid line). Vertical bars on dashed lines represent a ±20% range. Boxed regions at the tails are enlarged to be insets. See Figure 2b for the locations of the 307 pairs of ShakeMap stations and ZIP regions. The color version of this figure is available only in the electronic edition.

Suspicious Records. We excluded IDPs compiled from fewer than five responses or from earthquakes located at more than 200 km from the site. Large intensity values based on few or distant respondents are suspicious, and are likely erroneous.

The above data selection resulted in 2191 ZIP regions (Fig. 2c), producing 1858 IDPs from 98 earthquakes since 27 June 2000; the end of the data collection period was the end of March 2016. Each ZIP region had a different period of data availability (Table 2). Some ZIP regions contained no usable IDPs due to the above data winnowing and were treated as regions with no observed seismic hazard.

Validating Intensity-Based Hazard. Although DYFI intensity records were the only data source for the CEUS, both intensity and instrumental records existed for California. To ensure that macroseismic intensity could accurately represent the observed hazard, we compared the observed aggregated hazard curves compiled from DYFI records with those from instrumental records for California. DYFI IDPs and accelerometers never exactly sample the same points; we used all pairs of ZIP regions and accelerometers (307 in total) that are within 5 km from each other in California (Fig. 2b). The data selection procedure for California intensity records, as well as the calculation of hazard, was the same as that described above for the CEUS, except that here we used a GMICE specifically designed for California (Worden *et al.*, 2012). The databases of ShakeMap and DYFI did not cover exactly the same set of earthquakes and likely had different degrees of completeness for weak ground motions. In spite of these mismatches, we considered the agreement of the two observed hazard curves sufficient to warrant the use of DYFI records (Fig. 3).

Table 2
Period of Data Availability (years) for ZIP Regions in the Central and Eastern United States (CEUS)

Period (year)	Number of ZIPs
5–7.9	466
7.9–12.3	573
12.3–13.9	524
13.9–15.8	628

Results and Statistical Analysis

Suppressing Data Correlation

The observed and forecasted hazard curves computed from all selected instrumental records for California are shown in Figure 1d. Conventional hypothesis tests can be done by using the 5%–95% forecast interval (the shaded region in Fig. 1d; see the Aggregated Hazard Curve section), assuming the sites are independent. Because the sites are spatially clustered (Fig. 2a,c), they are not independent, and so the actual forecast interval will be at least as wide as the displayed one. It is necessary to account for the data correlation to conduct a meaningful statistical comparison. Explicitly

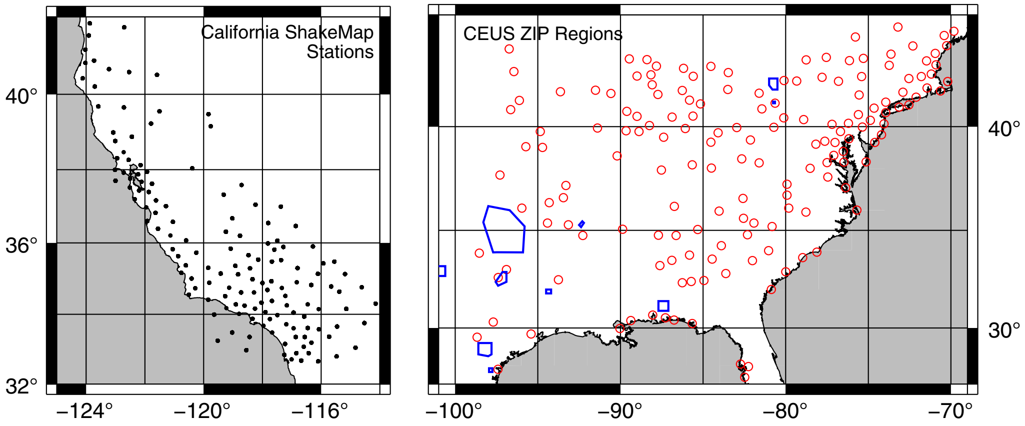

Figure 4. Example of declustered California ShakeMap stations (dots) and CEUS ZIP regions (hollow circles). Polygons denote identified zones of induced seismicity (see Data and Resources). The color version of this figure is available only in the electronic edition.

including the data correlation into the computation of forecast interval is difficult. We took the alternative way by suppressing the data correlation through discarding sites too close to each other, because seismic hazards at sites sufficiently apart are independent.

There are different ways to decluster the sites; each will result in a different dataset and may lead to a different interpretation of results. Without a generally recognized way of declustering, we applied the following Monte Carlo approach to randomly select sites with a predesignated minimum intersite distance to ensure site independence:

1. randomly pick one site from the 10% westmost sites,
2. discard all sites within the minimum intersite distance from the picked site,
3. randomly pick one site from the 10% nearest sites to the picked site,
4. repeat steps 2 and 3 until the pool of available sites is exhausted, and
5. repeat steps 1–4 400 times to generate 400 Monte Carlo datasets.

For ShakeMap stations in California, we took the minimum intersite distance to be 25 km. For ZIP regions in the CEUS, because the ground-motion attenuation is weaker, we considered a larger minimum intersite distance necessary, so we took 50 km. The numbers of sites included in each set of randomly generated data were slightly different, ranging from 142 to 155 ShakeMap stations (for California) and from 159 to 171 ZIP regions (for the CEUS); these small differences in the amount of sampled data were tolerated. The spatial distribution of the declustered sites for one Monte Carlo sample is shown in Figure 4 as an example.

Results

The observed and forecasted aggregated hazards for the 400 Monte Carlo samples are shown in Figures 5 and 6 for six return periods (10, 25, 50, 100, 200, and 400 years). Assuming the data in a sampled dataset to be mutually independent, hypothesis testing can be performed by checking if the observed hazard falls within the forecast interval. Different Monte Carlo samples could produce different results, but the general trend among all samples indicates how far the observed hazard is from the forecasted hazard by the NSHM.

For PGA in California (Fig. 5a), a lot of samples fell within the forecast intervals for all return periods, indicating that the observed hazard agreed well with the forecasted hazard. For SA1 in California (Fig. 5b), the observed hazards were mostly smaller than the forecasted hazard, implying a conservative forecast, except for the return period of 400 years. For California and small return periods, the observed hazard computed based on the two more recent versions (2008 and 2014) of the NSHM was closer to the forecasted hazard.

For PGA and SA1 in the CEUS (Fig. 6), the observed hazard was mostly smaller than the forecasted hazard for small return periods (10–50 years for PGA and 10–100 years for SA1). For larger return periods, the observed hazard was similar to the forecasted hazard. The different versions of the NSHM produced similar forecasts.

Discussion

Prospective Test

A true test of a forecast must use prospective data (i.e., data collected after the forecast has been made). The rarity of engineering ground motions often renders such a rigorous test impossible. Therefore, most published studies on PSHA model validation did not emphasize the use of prospective data. The observations in the current study were truly prospective for the two early versions of the NSHM (1996 and 2002). For the two more recent versions (2008 and 2014), we consider our data from 2000 to March 2016 still fairly independent of the models. First, merely a few more years of instrumental records seldom warrant a substantial adjustment of a long-term seismicity model. Second, GMPEs used by the NSHM are global models in which the recent California records have slight influence for the distance and magnitude

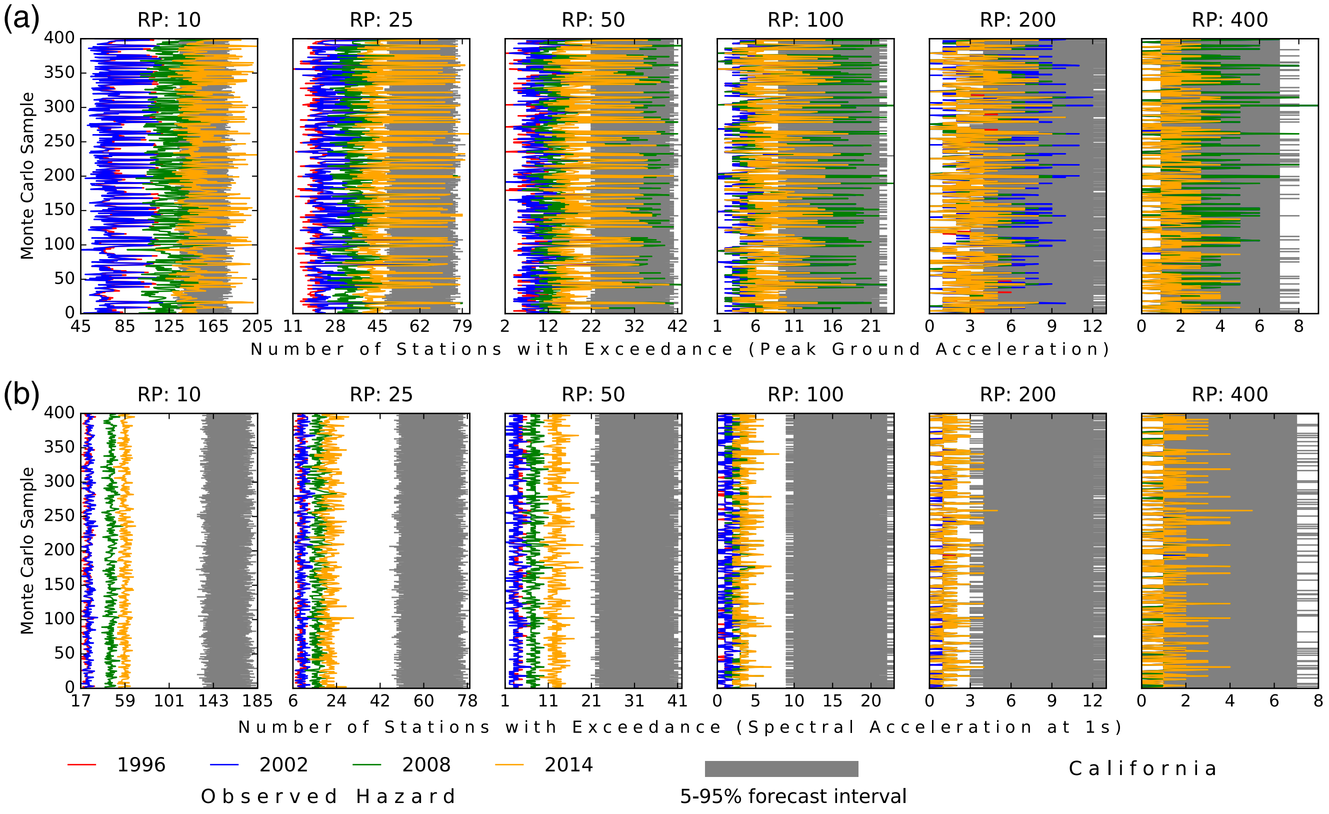

Figure 5. Observed and forecasted aggregated hazards for 400 Monte Carlo samplings (ordinate) and 6 return periods (RP, in years) in California for (a) PGA and (b) spectral acceleration at 1 s. The abscissa is the same as the ordinate of Figure 1d. The color version of this figure is available only in the electronic edition.

range of the most engineering significance under common situations (i.e., near-field, moderate-to-strong magnitudes).

Decisions Required to Conduct Testing

PSHA models are fundamentally not created to be tested empirically. The model and the observation often do not naturally represent the seismic hazard in exactly the same way. It is often necessary for testers to make decisions to reformat the forecasted and observed hazards into a comparable form. We explicitly state these decisions, and argue that they did not affect the validity of the result. It is, however, impossible to completely avoid the effects of these decisions.

Small Earthquakes. Most PSHAs discard small earthquakes and aftershocks in the modeling process for two reasons: to render the earthquake occurrence a manageable memoryless process and to save computational effort by not spending time on microseismicity that is believed to pose no threat to buildings. Model testers, however, have no reason to assume that nature should follow the modeler's assumptions. We did not specifically discard small earthquakes and aftershocks. The effect of earthquakes with magnitude smaller than 4.5 is shown in Figure 7. These earthquakes had some impact on the observed hazard of PGA for small return periods, and almost no effect on that of SA1. The modeler's decision to exclude small earthquakes was justified.

Site Effects. Most PSHAs forecast seismic hazard on rock, while the observed hazard always includes site effects. Although the conversion from observed ground motions to rock ground motions for instrumental records using an empirical amplification factor based on V_{S30} is a standard practice, the same use on ground motions converted from intensity records is tentative; we do not see a better measure currently available. The use of a different site amplification model may lead to a different result.

Site Independence. Site independence is a fundamental assumption for conducting statistical comparison between the observed and forecasted hazards (Figs. 5 and 6). We ensured site independence by declustering the sites. The choice of minimum intersite distances of 25 km (for California) and 50 km (for the CEUS) was fundamentally arbitrary (for reference, Albarello and D'Amico, 2008, used 50 km, and Tasan *et al.*, 2014, used 10 km). In general, using a different minimum intersite distance will discard a different amount of data, and so the width of the forecast interval varies; hypothesis tests could therefore have different result on different choices of threshold distance, especially if the forecasted and the actual seismic hazard are already different enough to be a marginal case. A sharp hypothesis rejection based on a fundamentally arbitrary α value (5% here for a one-tail test, or 10% for a two-tail test) is therefore inappropriate. The

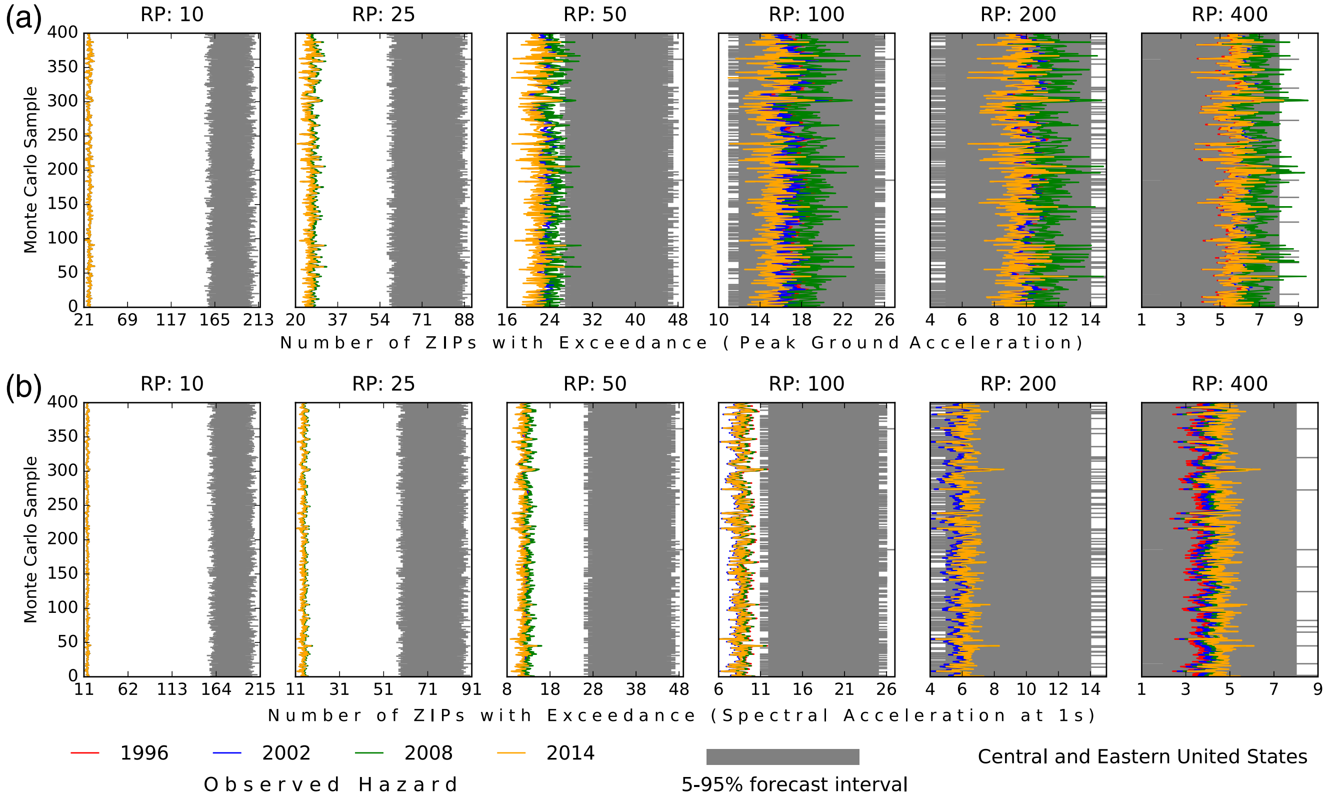

Figure 6. Observed and forecasted aggregated hazards for 400 Monte Carlo samplings (ordinate) and 6 return periods (RP, in years) in the CEUS for (a) PGA and (b) spectral acceleration at 1 s. The abscissa is the same as the ordinate of Figure 1d. The color version of this figure is available only in the electronic edition.

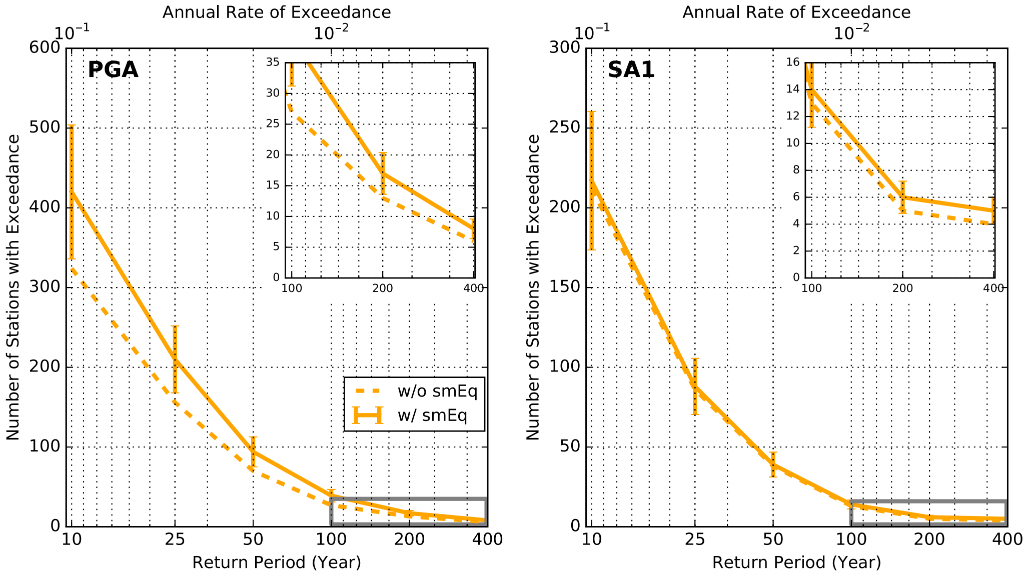

Figure 7. Observed hazard curves based on the NSHM 2014 for California, using all available ShakeMap stations, including (solid) and excluding (dotted) small earthquakes (magnitude smaller than 4.5). The solid curve for PGA is identical to the corresponding curve given in Figure 1d. Vertical bars on solid lines represent a ±20% range. Boxed regions at the tails are enlarged to be insets. The color version of this figure is available only in the electronic edition.

forecast interval shown above better serves as a reference for the difference between the observed and forecasted hazards. The empirical evaluation of the NSHM should be based on a qualitative comparison by looking at all sampling results, as well as by considering the statistical power discussed below.

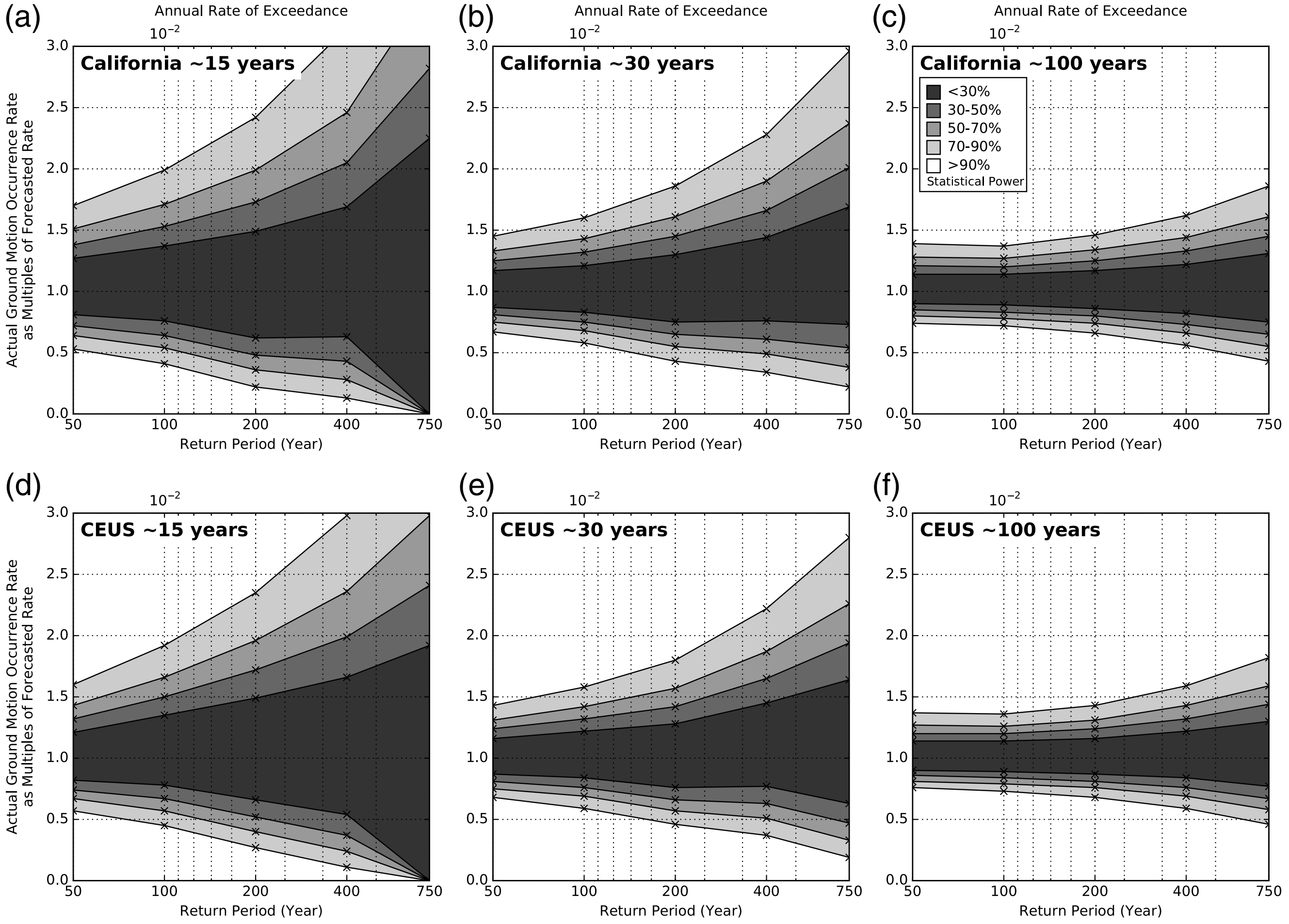

Figure 8. Statistical powers for the hypothesis tests for (a–c) California and (d–f) the CEUS. The observation used in the current study spanned approximately 15 years. The statistical power increases if longer periods of observation are available.

Statistical Power

For PGA in California, SA1 in California for the return period of 400 years, PGA in the CEUS for the return periods of 100–400 years, and SA1 in the CEUS for the return periods of 200–400 years, the hypothesis tests based on Figures 5 and 6 will lead to a conclusion that the observed and forecasted hazards are not significantly different. Two quantities not significantly different in a statistical sense do not mean they are not different. The statistical result could be due to chance, and the chance will become higher if the two quantities are not different enough with respect to the available amount of data. A quantitative inspection on this issue requires the calculation of the statistical power of the test (Mak *et al.*, 2014). It is common in the literature of laboratory science that the statistical power is not explicitly reported because it is often assumed that a well-designed experiment should result in high power, such that any meaningful difference between the observed and predicted quantities will be likely revealed by the experiment. Such assumption is often harmful in observational science, in which the experiment is often not designed by the experimenter but a process of nature, for which the availability of data is limited by the frequency of occurrence of natural phenomena.

The statistical power is the probability for a test to reveal that two quantities with a predefined degree of difference are significantly different in a statistical sense (i.e., not committing a type II error). Figure 8a and 8d shows the statistical power of the hypothesis tests for California (i.e., Fig. 5) and the CEUS (i.e., Fig. 6), respectively, computed as described in Mak *et al.* (2014) but for a Poisson-binomial distribution (see also Appendix B). The power shown is for one Monte Carlo sample of dataset, but those for other samples are similar because the forecasted hazards among the samples are similar (see the highly similar forecast intervals among the samples in Figs. 5 and 6).

The statistical power provides particularly useful information when a null hypothesis is not rejected, such as the case for California PGA of 100-year return period (Fig. 5a). The statistical power of this case (Fig. 8a) shows that if the actual hazard (in terms of occurrence rate) is <0.4 or >2 (respectively, <0.6 or >1.5) times the forecasted hazard, the test will have a $>90\%$ (respectively, $>50\%$) chance to reveal that the two hazards are different (for $\alpha = 5\%$ and one-tail test). In other words, the likelihood for a result of nonrejection is <0.1 (respectively, <0.5) if the actual hazard is <0.4 or >2 (respectively, <0.6 or >1.5) times the forecasted one. This is the resolving power of the test given the available data; the statistical power increases with the amount of available data. On the other hand, when the amount of data is limited and statistical power low, if a test still finds the observed hazard significantly different from the forecasted one, such a difference is likely large enough to warrant an in-depth investigation.

For empirical evaluation of PSHA models, the statistical power depends only on the period of data availability, not the actual data (Mak *et al.*, 2014). It is therefore possible to foresee how much future data would improve the statistical power. Figure 8b and 8e shows the statistical powers for the California and CEUS tests, respectively, assuming that 15 more years of observation (i.e., up to approximately 30 years of observation in total) is available. A test based on such an enhanced dataset (for California or for the CEUS) will likely ($\geq 90\%$) detect the inconsistency between the observed and forecasted hazards, if the actual hazard at the return period of 400 years is at least about 2.2 times the forecasted one. This is better than the currently available resolving power of about 3 times. If we take 30 years of observations to be the upper limit of the availability of prospective data, we can conclude that it is quite unlikely ($\leq 30\%$) that an aggregated test of the scale of the California or the CEUS case can reveal the difference between the actual and forecasted hazards of the return period of 400 years, if the actual hazard is within 0.75–1.5 times the forecasted hazard. This provides a rule of thumb for the limit of empirical validation. Albarello and D'Amico (2015, pp. 275) considered a low-power test acceptable for testing PSHA models because they considered that the aim of a test was to demonstrate if a PSHA model is somehow compatible with the reality. The current study demonstrated quantitatively how high (or how low) the statistical power could be under a realistic environment. Including historical records could lead to a higher power (see Fig. 8c,f for 100 years of available observation), but the independence of the data with respect to the model, as well as the completeness of the data, will then become a concern.

The limit of the empirical validation of PSHA models could be taken as the limit of how reliable, based on direct empirical evidence, a model is. It does not address the theoretical correctness of the physics on which a model is based. A user of a model should understand how much his decision to adopt a model is based on direct empirical evidence (an inductive reasoning), the verifiable part of the model, and how much is based on the model's physical correctness (a deductive reasoning), the theoretical part of the model. Both components are crucial for scientific decisions and should be explicitly addressed.

Result Interpretation

The observed hazards in the CEUS at small return periods were apparently much lower than the forecasted one (Fig. 6). This is an artifact. Figure 6 shows that the observed hazards are similar for return periods ≤ 50 years. This is because, for the CEUS, earthquakes are so infrequent that even a weak ground motion has a return period larger than about 50 years. The NSHM does not model ground motions smaller than $0.5\%g$ (for PGA). Therefore, there is no way to count the exceedance of ground motions for short return periods. Forecasted hazards of return periods smaller than about 50 years are not modeled in the CEUS and so cannot be tested.

Observing the limitations in the treatment (described in the Decisions Required to Conduct Testing section) and amount (in terms of statistical power, explained in the Statistical Power section) of observations, the current study shows that observed seismic hazards generally agreed with those forecasted by the NSHM for PGA for both California and the CEUS, and SA1 for the CEUS (Figs. 5a and 6). The models appeared to be conservative for SA1 in California at return periods ≤ 200 years (Fig. 5b). Delavaud *et al.* (2012, their table 7) found that GMPEs performing well for PGA did not necessary perform well for SA1, and vice versa. This could be a reason for the different results between PGA and SA1 in California.

Jaiswal *et al.* (2015) described the change of hazard estimates among different versions of the NSHM. The current study provided additional evidence that the forecast by the two more recent versions of the NSHM (2008 and 2014) was closer to the observed hazard for California (Fig. 5). For the CEUS, the corresponding trend of change was less clear (Fig. 6). Compared with California, the seismic hazard for the CEUS is always less understood; a larger degree of expert judgment is believed to be necessary to estimate the seismic hazard for the CEUS. The current study showed that the available empirical evidence did not provide an overall disagreement with those judgments.

Beyond the Mean Forecast

The mean hazard curve was used as the forecast in the current study. The goal for the NSHM is to reflect the center, the body, and the range of the estimate of seismic hazard (Kammerer and Ake, 2012, section 3.1). In practice, this is implemented by experts assigning weights to a spectrum of input parameters and components, resulting in an ensemble forecast that includes a number of hazard curves associated with weights. If the weight is taken as a probability, so that the annual rate of exceedance on a hazard curve is conditioned by a probability identical to the weight assigned to that curve, then statistical comparisons with observations like those presented in the current study can be readily conducted for an ensemble forecast. In other words, the center, the body, and the range can be tested as a whole. Marzocchi and Jordan (2014, pp. 11,975–11,976) described a one-station example of testing an ensemble forecast.

In the context of the current study, when the abscissa of the aggregated hazard curve is expressed in ground-motion levels (e.g., Fig. 1a,b), testing an ensemble forecast requires

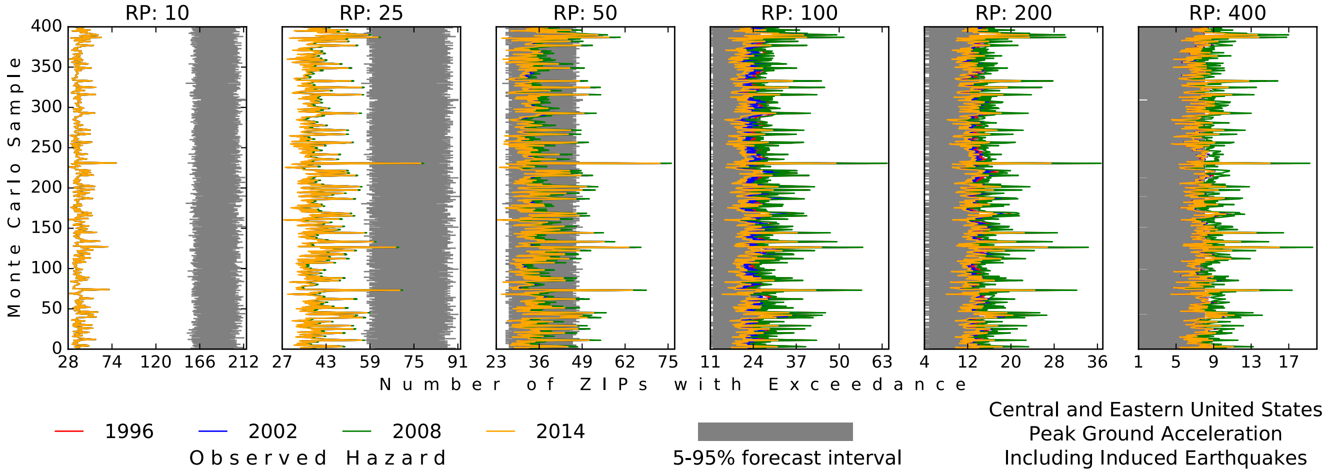

Figure 9. Observed and forecasted aggregated hazards for 400 Monte Carlo samplings (ordinate) and 6 return periods (RP, in years) for PGA for the CEUS, including induced earthquakes. The abscissa is the same as the ordinate of Figure 1d. This figure is identical to Figure 6a, except that induced earthquakes are not removed. The color version of this figure is available only in the electronic edition.

computing the forecast as the weighted sum of distribution functions, based on the law of total probability. As pointed out by Marzocchi and Jordan (2014), the result will be generally different from that based only on the mean hazard curve. When the abscissa of the aggregated hazard curve is expressed in return periods (e.g., Fig. 1c,d), one needs to specify a PSHA model to convert a return period into ground-motion level for exceedance counting. An ensemble forecast will lead to a probabilistic conversion in which the converted ground-motion level is conditioned by a probability identical to the weight assigned to the ensemble member. The law of total probability is again involved in conducting a hypothesis test. Mathematically, the aggregated approach used in the current study can be used to test an ensemble forecast.

There are, however, other complications in testing an ensemble forecast. For example, there may not have been a general consent on whether the weight assigned to a hazard curve in an ensemble forecast is identical to a probability. In addition, it is easy to create an infallible model to subdue testing by including an indefinitely large epistemic uncertainty, so that even the most inconceivable observation falls within the forecast interval. These issues need to be resolved before proceeding to the use of empirically evaluating ensemble models.

Induced Earthquakes

Finally, we give a brief note on the highly concerned effect of induced seismicity to the seismic hazard of the CEUS. Figure 9 shows the observed and forecasted PGA aggregated hazards for the 400 Monte Carlo samples for the CEUS, with the induced earthquakes not excluded. It is therefore the same as Figure 6a, except that earthquakes and ZIP regions within the identified zones (polygons in Fig. 2c) were not excluded. The conclusion that the observed hazard agreed with the forecasted one is not changed by including the induced earthquakes, although the observed hazard has obviously been increased (compared with that in Fig. 6a). The sampled ZIP regions (with a minimum intersite distance of 50 km) spanned the whole CEUS. Therefore, induced seismicity did not render the hazard forecast in the CEUS invalid as a whole. The effect of induced seismicity for a single site locating within a zone of induced seismicity could be very different, but the limited amount of data for such a refined region makes empirical evaluation of hazard forecast difficult.

Summary

We presented the first empirical prospective test for the U.S. NSHM with the following major findings.

1. Aggregating all observations since 2000 and avoiding the effects of aftershocks and induced earthquakes, the observed hazard was found to be compatible (subjected to the limitation by the statistical power) with the forecasted hazard for PGA for California and PGA and SA1 for the CEUS.
2. The NSHM for SA1 for California appeared to be conservative.
3. Recent versions of the NSHM appeared to be more consistent with the observed hazard for California, whereas the corresponding trend for the CEUS was less obvious.
4. In the CEUS, induced seismicity has increased the observed seismic hazard but did not invalidate the hazard forecast as a whole.
5. It is quite unlikely ($\leq 30\%$) that a prospective aggregated test of the scale of the California or the CEUS case can reveal the difference between the actual and forecasted hazards of the return period of 400 years, if the actual hazard is within 0.75–1.5 times the forecasted hazard. The limit of an empirical evaluation of a PSHA model is reflected by the statistical power of the test.
6. Macroseismic intensity data from DYFI, after proper probabilistic conversion, could adequately represent the observed hazard. This makes testing hazard models for

a region with poor instrumental coverage like the CEUS feasible.

7. The forecast of the total number of sites that have experienced at least one ground-motion exceedance is a Poisson-binomial random variable. Hypothesis tests can be conducted based on this distribution.

Data and Resources

Regular updates of weather forecast performance can be found on the webpage of the European Centre for Medium-Range Weather Forecasts (ECMWF (www.ecmwf.int/en/forecasts/quality-our-forecasts). The U.S. Geological Survey (USGS) National Seismic Hazard Maps were downloaded from earthquake.usgs.gov/hazards/products/conterminous. The USGS report of Stirling and Gerstenberger (2011) was downloaded from earthquake.usgs.gov/research/external/reports/G11AP20024.pdf. Instrumental records from ShakeMap stations were downloaded from earthquake.usgs.gov/earthquakes/shakemap. Macroseismic intensity records from "Did You Feel It?" (DYFI) were downloaded from earthquake.usgs.gov/earthquakes/dyfi. The Next Generation Attenuation (NGA)-West2 flatfile was downloaded from peer.berkeley.edu/ngawest2/databases. The zones of induced seismicity displayed in Figure 2c were identical to those used in the NSHM 2014, obtained through direct communication with Charles Mueller (USGS). Figures 2 and 4 were prepared using Generic Mapping Tools (Wessel *et al.*, 2013). Computations related to the Poisson-binomial distribution were conducted using the package poibin (cran.r-project.org/web/packages/poibin) of R (www.R-project.org). All webpages were last accessed on March 2016.

Acknowledgments

S. M. would like to thank Ross Stein for his encouragement to begin this study, Bruce Wordern for his discussion on GMICE during the Seismological Society of America Annual Meeting in 2015, and Vince Quitoriano and David Wald for their crucial help in extracting "Did You Feel It?"(DYFI) data. This article has been substantially improved based on the comments from Fabrice Cotton, reviewer Céline Beauval, and Associate Editor Matt Gerstenberger. This study was supported by the Global Earthquake Model Foundation.

References

Albarello, D., and V. D'Amico (2008). Testing probabilistic seismic hazard estimates by comparison with observations: An example in Italy, *Geophys. J. Int.* **175,** no. 3, 1088–1094, doi: 10.1111/j.1365-246X.2008.03928.x.

Albarello, D., and V. D'Amico (2015). Scoring and testing procedures devoted to probabilistic seismic hazard assessment, *Surv. Geophys.* **36,** no. 2, 269–293, doi: 10.1007/s10712-015-9316-4.

Atkinson, G. M., and S. I. Kaka (2007). Relationships between felt intensity and instrumental ground motion in the Central United States and California, *Bull. Seismol. Soc. Am.* **97,** no. 2, 497–510, doi: 10.1785/0120060154.

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.

Brooks, E. M., S. Stein, and B. D. Spencer (2016). Comparing the performance of Japan's earthquake hazard maps to uniform and randomized maps, *Seismol. Res. Lett.* **87,** no. 1, 90–102, doi: 10.1785/0220150100.

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.

Ellsworth, W. L. (2013). Injection-induced earthquakes, *Science* **341,** no. 6142, doi: 10.1126/science.1225942.

Frankel, A. (2013a). Comment on "Why earthquake hazard maps often fail and what to do about it" by S. Stein, R. Geller, and M. Liu, *Tectonophysics* **592,** 200–206, doi: 10.1016/j.tecto.2012.11.032.

Frankel, A. (2013b). Corrigendum to comment on "Why earthquake hazard maps often fail and what to do about it" by S. Stein, R. Geller, and M. Liu [TECTO 592(2013) 200-206], *Tectonophysics* **608,** 1453–1454, doi: 10.1016/j.tecto.2013.08.010.

Fujiwara, H., N. Morikawa, Y. Ishikawa, T. Okumura, J. Miyakoshi, N. Nojima, and Y. Fukushima (2009). Statistical comparison of national probabilistic seismic hazard maps and frequency of recorded JMA seismic intensities from the K-NET strong-motion observation network in Japan during 1997–2006, *Seismol. Res. Lett.* **80,** no. 3, 458–464, doi: 10.1785/gssrl.80.3.458.

Hanks, T. C., G. C. Beroza, and S. Toda (2012). Have recent earthquakes exposed flaws in or misunderstandings of probabilistic seismic hazard analysis? *Seismol. Res. Lett.* **83,** no. 5, 759–764, doi: 10.1785/0220120043.

Jaiswal, K. S., M. D. Petersen, K. Rukstales, and W. S. Leith (2015). Earthquake shaking hazard estimates and exposure changes in the conterminous United States, *Earthq. Spectra* **31,** no. S1, S201–S220, doi: 10.1193/111814EQS195M.

Kammerer, A. M., and J. P. Ake (2012). Practical implementation guidelines for SSHAC level 3 and 4 hazard studies, *Tech. Rept. NUREG-2117, Rev. 1*, Office of Nuclear Regulatory Research.

Mak, S., and D. Schorlemmer (2016). What makes people respond to "Did You Feel It?"? *Seismol. Res. Lett.* **87,** no. 1, 119–131, doi: 10.1785/0220150056.

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

Marzocchi, W., and T. H. Jordan (2014). Testing for ontological errors in probabilistic forecasting models of natural systems, *Proc. Natl. Acad. Sci. Unit. States Am.* **111,** no. 33, 11,973–11,978, doi: 10.1073/pnas.1410183111.

McGuire, R. K. (1979). Adequacy of simple probability models for calculating felt-shaking hazard, using the Chinese earthquake catalog, *Bull. Seismol. Soc. Am.* **69,** no. 3, 877–892.

McGuire, R. K., C. A. Cornell, and G. R. Toro (2005). The case for using mean seismic hazard, *Earthq. Spectra* **21,** no. 3, 879–886, doi: 10.1193/1.1985447.

Mezcua, J., J. Rueda, and R. M. G. Blanco (2013). Observed and calculated intensities as a test of a probabilistic seismic-hazard analysis of Spain, *Seismol. Res. Lett.* **84,** no. 5, 772–780, doi: 10.1785/0220130020.

Miyazawa, M., and J. Mori (2009). Test of seismic hazard map from 500 years of recorded intensity data in Japan, *Bull. Seismol. Soc. Am.* **99,** no. 6, 3140–3149, doi: 10.1785/0120080262.

Musson, R. M. W. (2005). Against fractiles, *Earthq. Spectra* **21,** no. 3, 887–891, doi: 10.1193/1.1985445.

Ordaz, M., and C. Reyes (1999). Earthquake hazard in Mexico City: Observations versus computations, *Bull. Seismol. Soc. Am.* **89,** no. 5, 1379–1383.

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.

Schorlemmer, D., M. Gerstenberger, S. Wiemer, D. D. Jackson, and D. A. Rhoades (2007). Earthquake likelihood model testing, *Seismol. Res. Lett.* **78,** no. 1, 17–29, doi: 10.1785/gssrl.78.1.17.

Stark, P. B., and D. A. Freedman (2003). What is the chance of an earthquake? in *Earthquake Science and Seismic Risk Reduction, NATO Science Series*, F. Mulargia and R. K. Geller (Editors), Chapter 5.3. Kluwer Academic Publishers, Dordrecht, The Netherlands, ISBN: 1402017782.

Stein, S., R. J. Geller, and M. Liu (2011). Bad assumptions or bad luck: Why earthquake hazard maps need objective testing, *Seismol. Res. Lett.* **82,** no. 5, 623–626, doi: 10.1785/gssrl.82.5.623.

Stein, S., R. J. Geller, and M. Liu (2012). Why earthquake hazard maps often fail and what to do about it, *Tectonophysics* **562/563,** 1–25, doi: 10.1016/j.tecto.2012.06.047.

Stein, S., R. J. Geller, and M. Liu (2013). Reply to comment by Arthur Frankel on "Why earthquake hazard maps often fail and what to do about it", *Tectonophysics* **592,** 207–209, doi: 10.1016/j.tecto.2013.01.024.

Stirling, M., and M. Gerstenberger (2010). Ground motion-based testing of seismic hazard models in New Zealand, *Bull. Seismol. Soc. Am.* **100,** no. 4, 1407–1414, doi: 10.1785/0120090336.

Stirling, M., and M. Petersen (2006). Comparison of the historical record of earthquake hazard with seismic-hazard models for New Zealand and the continental United States, *Bull. Seismol. Soc. Am.* **96,** no. 6, 1978–1994, doi: 10.1785/0120050176.

Stirling, M. W. (2012). Earthquake hazard maps and objective testing: The hazard mapper's point of view, *Seismol. Res. Lett.* **83,** no. 2, 231–232, doi: 10.1785/gssrl.83.2.231.

Tasan, H., C. Beauval, A. Helmstetter, A. Sandikkaya, and P. Guéguen (2014). Testing probabilistic seismic hazard estimates against accelerometric data in two countries: France and Turkey, *Geophys. J. Int.* **198,** no. 3, 1554–1571, doi: 10.1093/gji/ggu191.

Wald, D. J., V. Quitoriano, B. Worden, M. Hopper, and J. W. Dewey (2011). USGS "Did You Feel It?" internet-based macroseismic intensity maps, *Ann. Geophys.* **54,** no. 6, 688–707, doi: 10.4401/ag-5354.

Wang, Y. H. (1993). On the number of successes in independent trials, *Stat. Sinica* **3,** no. 2, 295–312.

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,** no. 45, 409–410, doi: 10.1002/2013EO450001.

Worden, C. B., M. C. Gerstenberger, D. A. Rhoades, and D. J. Wald (2012). Probabilistic relationships between ground-motion parameters and modified Mercalli intensity in California, *Bull. Seismol. Soc. Am.* **102,** no. 1, 204–221, doi: 10.1785/0120110156.

Appendix A

Four Forms of Aggregated Hazard Curve

Notations used throughout the two appendixes are defined here. Suppose that there are M sites, each has N_i records observed throughout t_i years, and the record j of site i has a value of o_{ij}. For a given ground-motion level g, a given probabilistic seismic-hazard assessment (PSHA) model gives λ_i as the corresponding annual rate of exceedance for site i. A_x is a Poisson random variable with mean x. B_p is a Bernoulli random variable with parameter p (i.e., p is the probability for success). $\mathbb{1}(\cdot)$ is the usual indicator function that takes the value one when the bracketed statement is true, or zero otherwise.

Form 1

Ordinate: Total number of exceedances over multiple sites.

Abscissa: Ground-motion level.

Example: Figure 1a.

For a given ground-motion level g, the observed hazard is

$$h_{\text{obs}}^{(1)} = \sum_i^M \sum_j^{N_i} \mathbb{1}(o_{ij} \geq g). \tag{A1}$$

The counting method when the observation is macroseismic intensity is different (see equation 1).

The corresponding model-dependent forecasted hazard is

$$H_f^{(1)} = \sum_i^M A_{\lambda_i t_i} \tag{A2}$$

with the mean as the expected hazard:

$$h_{\text{ex}}^{(1)} = \sum_i^M \lambda_i t_i. \tag{A3}$$

$H_f^{(1)}$ is a Poisson random variable because it is the sum of Poisson random variables.

Form 2

Ordinate: Number of sites with at least one exceedance.

Abscissa: Ground-motion level.

Example: Figure 1b.

For a given ground-motion level g, the observed hazard is

$$h_{\text{obs}}^{(2)} = \sum_i^M \mathbb{1}(\exists j : o_{ij} \geq g). \tag{A4}$$

The counting method is different when the records are given in macroseismic intensity (see equation 2).

The corresponding model-dependent forecasted hazard is

$$H_f^{(2)} = \sum_i^M B_{p(i)} \quad \text{with } p(i) = \Pr(A_{\lambda_i t_i} > 0) \tag{A5}$$

with the mean as the expected hazard:

$$h_{\text{ex}}^{(2)} = \sum_i^M \Pr(A_{\lambda_i t_i} > 0). \tag{A6}$$

Pr(.) literally means the probability of observing at least one exceedance at site i. $H_f^{(2)}$ is a Poisson-binomial random variable with parameters $p(i)$ because it is a sum of heterogeneous Bernoulli random variables. There is no closed-form expression for the distribution function of a Poisson-binomial distribution (see Wang, 1993).

Form 3

Ordinate: Total number of exceedances over multiple sites.

Abscissa: Return period.

Example: Figure 1c.

For a given annual rate of exceedance r and a given PSHA model, the ground-motion level at return period $1/r$ for site i is g_i. The observed hazard is

$$h_{\mathrm{obs}}^{(3)} = \sum_{i}^{M} \sum_{j}^{N_i} \mathbb{1}(o_{ij} \geq g_i), \tag{A7}$$

which is model dependent because g_i is given by a PSHA model. The counting method is different when the records are given in macroseismic intensity (see equation 1).

The corresponding model-independent forecasted hazard is

$$H_f^{(3)} = \sum_{i}^{M} A_{rt_i} \tag{A8}$$

with the mean as the expected hazard:

$$h_{\mathrm{ex}}^{(3)} = \sum_{i}^{M} rt_i. \tag{A9}$$

$H_f^{(3)}$ is a Poisson random variable because it is the sum of Poisson random variables.

Form 4

Ordinate: Number of sites with at least one exceedance.
Abscissa: Return period.
Examples: Figures 1d, 3, 5, 6, 7, and 9.

For a given annual rate of exceedance r and a given PSHA model, the ground-motion level at return period $1/r$ for site i is g_i. The observed hazard is

$$h_{\mathrm{obs}}^{(4)} = \sum_{i}^{M} \mathbb{1}(\exists j : o_{ij} \geq g_i), \tag{A10}$$

which is model dependent because g_i is given by a PSHA model. The counting method when the observation in macroseismic intensity is different (see equation 2).

The corresponding forecasted hazard is independent of both the model and observed ground motions:

$$H_f^{(4)} = \sum_{i}^{M} B_{p(i)} \quad \text{with } p(i) = \Pr(A_{rt_i} > 0) \tag{A11}$$

with the mean as the expected hazard:

$$h_{\mathrm{ex}}^{(4)} = \sum_{i}^{M} \Pr(A_{rt_i} > 0). \tag{A12}$$

$H_f^{(4)}$ is a Poisson-binomial random variable with parameters $p(i)$ because it is a sum of heterogeneous Bernoulli random variables. It becomes binomial if the durations of observation for all sites are the same (i.e., t_i is constant for all i).

Appendix B

Calculation of Statistical Power

The statistical powers for tests shown in Figure 8 were computed as follows. (See Appendix A for the definitions of notations.) For a given annual rate of exceedance r, the forecasted aggregated hazard (form 4), $H_f^{(4)}$, is a Poisson-binomial random variable with parameters $\Pr(A_{rt_i} > 0)$ (equation A11). Assuming that the actual annual rate of exceedance for the same ground-motion levels is Kr (K is the ordinate of Fig. 8), for an upper one-tail test, the statistical power is the probability:

$$\Pr[H_f^{(4*)} > q(1-\alpha)], \tag{B1}$$

in which $H_f^{(4*)}$ is the Poisson-binomial random variable with parameters $\Pr(A_{Krt_i} > 0)$:

$$H_f^{(4*)} = \sum_{i}^{M} B_{p^*(i)} \quad \text{with } p^*(i) = \Pr(A_{Krt_i} > 0) \tag{B2}$$

and q is the quantile function (or inverse cumulative distribution function) of $H_f^{(4)}$. For the corresponding lower one-tail test, replace $1-\alpha$ by α and replace the $>$ by $<$ in equation (B1) to obtain the power. It can be seen from equations (B1) and (B2) that for a fixed r, the power increases with t_i, K, M, and α.

Another equivalent interpretation of the statistical power involves the concept of confidence interval, illustrated here using an example. If the annual rate of exceedance for the ground-motion level of the return period of 100 years (i.e., $r = 0.01$), defined by the NSHM 2014 for the CEUS, is estimated from the available observations, the estimated rate will have a 90% confidence interval of $(K_1 r, K_2 r)$, in which K_1 and K_2 are about 0.4 and 2 (read from Fig. 8d), respectively. The corresponding 30% confidence interval would have K_1, K_2 of about 0.75, 1.4, meaning that the estimated return period for those ground-motion levels has a confidence of 30% to lie between $1/(0.01 \times 1.4) = 71$ and $1/(0.01 \times 0.75) = 133$ years. Certainly, it is not necessary for the actual annual rate of exceedance of all sites to be the same multiples of r, so the description here is a special case. A more rigorous treatment for the equivalence between hypothesis test and confidence interval is given by DeGroot and Schervish (2012, pp. 540–543).

Helmholtz-Zentrum Potsdam Deutsches GeoForschungsZentrum (GFZ)
Section 2.6
Helmholtzstraße 6
14467 Potsdam, Germany

Manuscript received 20 November 2015;
Published Online 28 June 2016