Bulletin of the Seismological Society of America, Vol. 105, No. 4, pp. –, August 2015, doi: 10.1785/0120140164

Metrics for Assessing Earthquake-Hazard Map Performance

by Seth Stein, Bruce D. Spencer, and Edward M. Brooks

Abstract Recent large earthquakes that caused great damage in areas predicted to be relatively safe, illustrate the importance of criteria that assess how well earthquake hazard maps used to develop codes for earthquake-resistant construction are actually performing. At present, there is no agreed-upon way of assessing how well a map performed and thus determining whether one map performed better than another. The fractional site exceedance metric implicit in current maps, that during the chosen time interval the predicted ground motion will be exceeded only at a specific fraction of the sites, is useful but permits maps to be nominally successful even if they significantly underpredict or overpredict shaking, or permits them to be nominally unsuccessful but do well in terms of predicting shaking. We explore some possible metrics that better measure the effects of overprediction and underprediction and can be weighted to reflect the two differently and to reflect differences in populations and property at risk. Although no single metric alone fully characterizes map behavior, using several metrics can provide useful insight for comparing and improving hazard maps. For example, both probabilistic and deterministic hazard maps for Italy dramatically overpredict the recorded shaking in a 2200-yr-long historical intensity catalog, illustrating problems in the data (most likely), models, or both.

Introduction

As Hurricane Irene threatened the U.S. east coast in August 2011, meteorologist Kerry Emanuel explained to the public that “We do not know for sure whether Irene will make landfall in the Carolinas, on Long Island, or in New England, or stay far enough offshore to deliver little more than a windy, rainy day to East Coast residents. Nor do we have better than a passing ability to forecast how strong Irene will get. In spite of decades of research and greatly improved observations and computer models, our skill in forecasting hurricane strength is little better than it was decades ago.”

This example illustrates that the performance of forecasts has multiple aspects (in this case, a storm’s path and strength), each of which needs to be quantified. Metrics are numerical measures that describe a property of a system, so its performance can be quantified beyond terms like “good,” “fair,” or “bad.” For example, the performance of medical diagnostic tests is assessed using two metrics: specificity (the lack of false positives, or type I errors) and sensitivity (lack of false negatives, or type II errors). Similarly, a statistical estimate may be biased with high precision or unbiased with low precision; more generally, its performance is described by a probability distribution for its error.

Metrics describe how a system behaves but not why. A weather forecast or medical test may perform poorly because of problems with the model, the input data, or both. Similarly, although metrics measure relative performance, they do not themselves tell whether the differences are explicable solely by chance or instead are statistically significant. Assessing whether one model is significantly better than another requires assuming and applying a probability model to the data underlying the metric.

Metrics are crucial in assessing the past performance of forecasts. For example, weather forecasts are routinely evaluated to assess how well their predictions matched what actually occurred (Stephenson, 2000). This assessment involves adopting metrics. Murphy (1993) notes that “it is difficult to establish well-defined goals for any project designed to enhance forecasting performance without an unambiguous definition of what constitutes a good forecast.”

Figure 1 shows an example comparing the predicted probability of rain to that actually observed. National Weather Service forecasts have only a slight “wet bias” toward predicting rain more often than actually occurs. This bias is much greater for a local television station, where forecasts are much less accurate. A metric describing the misfit would quantify the difference but would not tell us why the television forecasts do worse. Silver (2012) suggests that television forecasters feel that viewers enjoy unexpectedly sunny weather but are annoyed by unexpected rain, and they therefore prefer the biased forecast. Other users, however, would likely prefer the less-biased forecast. Similarly, the metric does not quantify the possibility that the television station’s forecast is worse purely by chance, which requires

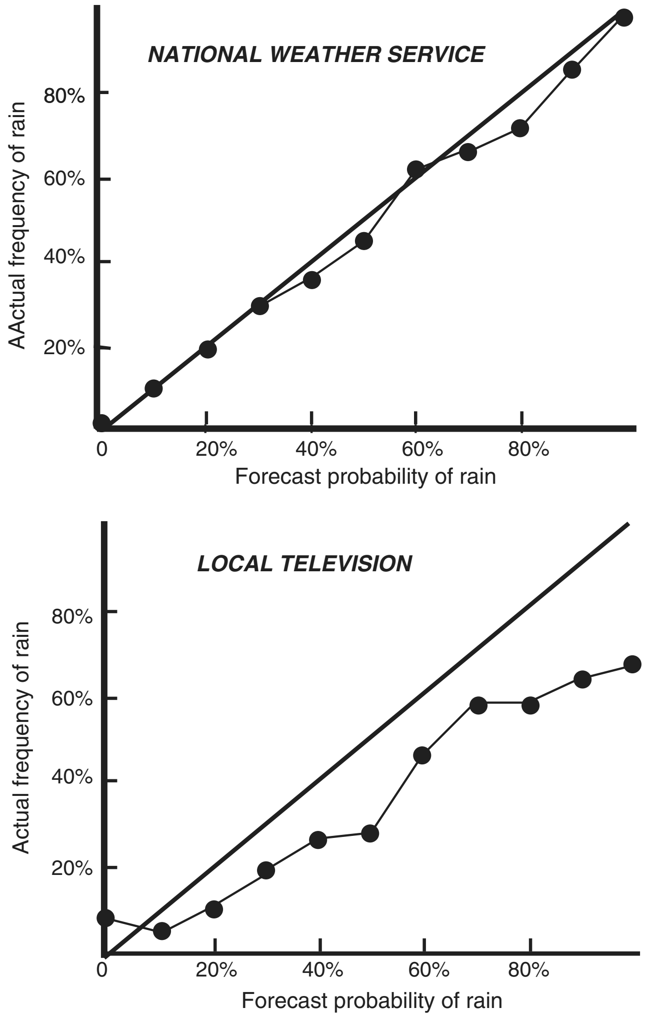

Figure 1. Comparison of the predicted probability of rain to that actually observed in National Weather Service and a local television station's forecasts. After Silver (2012).

assuming and applying a probability model to the data. Information about how a forecast performs is crucial in determining how best to use it. The better a weather forecast has worked to date, the more we factor it into our daily plans.

Similar issues arise for earthquake hazard maps that are used to develop codes for earthquake-resistant construction. These maps are derived by estimating a variety of parameters for selected models that are used to forecast future seismicity and the resulting shaking.

Recent destructive large earthquakes underscore the need for agreed metrics that measure how well earthquake hazard maps are actually performing. The 2011 M_w 9.1 Tohoku earthquake, and thus the resulting tsunami, was much larger than anticipated in the Japanese national earthquake hazard map (Geller, 2011). The 2008 M_w 7.9 Wenchuan, China, and 2010 M_w 7.1 Haiti earthquakes occurred on faults mapped as giving rise to low hazard (Stein *et al.*, 2012).

These events have catalyzed discussions among seismologists and earthquake engineers about commonly used earthquake-hazard mapping practices (Kerr, 2011; Stirling, 2012; Gulkan, 2013; Iervolinoa, 2013). The underlying question is the extent to which the occurrence of low-probability shaking indicates problems with the maps—either in the algorithm or the specific parameters used to generate them—or chance occurrences consistent with the maps. Several explanations have been offered.

One explanation (Hanks *et al.*, 2012; Frankel, 2013) is that these earthquakes are low-probability events allowed by probabilistic seismic-hazard maps, which use estimates of the probability of future earthquakes and the resulting shaking to predict the maximum shaking expected with a certain probability over a given time. The probabilistic algorithm anticipates that in a specified number of cases, shaking exceeding that mapped should occur (Cornell, 1968; Field, 2010). Hence, it is important to assess whether such high shaking events occur more or less often than anticipated.

However, the common practice of extensively remaking a map to show increased hazards after unexpected events or shaking (Fig. 2) is inconsistent with the interpretation that these were simply low-probability events consistent with the map. For example, although the chance that a given lottery ticket is a winner is low, the probability that some lottery ticket wins is high. Hence, the odds of winning are only reassigned after a winning ticket is picked when the operators think their prior model was wrong. The revised maps thus reflect both what occurred in these earthquakes and other information that was either unknown or not appreciated (e.g., Minoura *et al.*, 2001; Manaker *et al.*, 2008; Sagiya, 2011) when the earlier map was made (Stein *et al.*, 2012).

Choosing whether to remake the map in such cases is akin to deciding whether and how much to revise your estimate of the probability that a coin will land heads up after it landed heads four times in a row (Stein *et al.*, 2015). If, prior to the coin tosses, you had confidence that the coin was fair—equally likely to land heads or tails—and the person tossing it would not deliberately influence how it lands, you might regard the four heads as an unlikely event consistent with your probability model and so would not change it. However, if a magician was tossing the coin, your confidence in your prior model would be lower, and you would likely revise it. When and how to update forecasts as additional information becomes available, depending on one's confidence in the initial model, is extensively discussed in the statistical literature (e.g., Sivia, 2006; Rice, 2007) but beyond our scope here.

Another explanation that has been offered is that the probabilistic approach is flawed (Klügel *et al.*, 2006; Wang, 2011; Wang and Cobb, 2012), in that the expected value of shaking in a given time period is a mathematical quantity not corresponding to any specific earthquake that is inappropriate for designing earthquake-resistant structures, especially for rare large events that critical facilities like nuclear power plants should withstand. In this view, it is better to specify the largest earthquakes and resulting shaking that realistically could occur in a deterministic seismic-hazard assessment (Peresan and Panza, 2012). This approach avoids uncertainties from

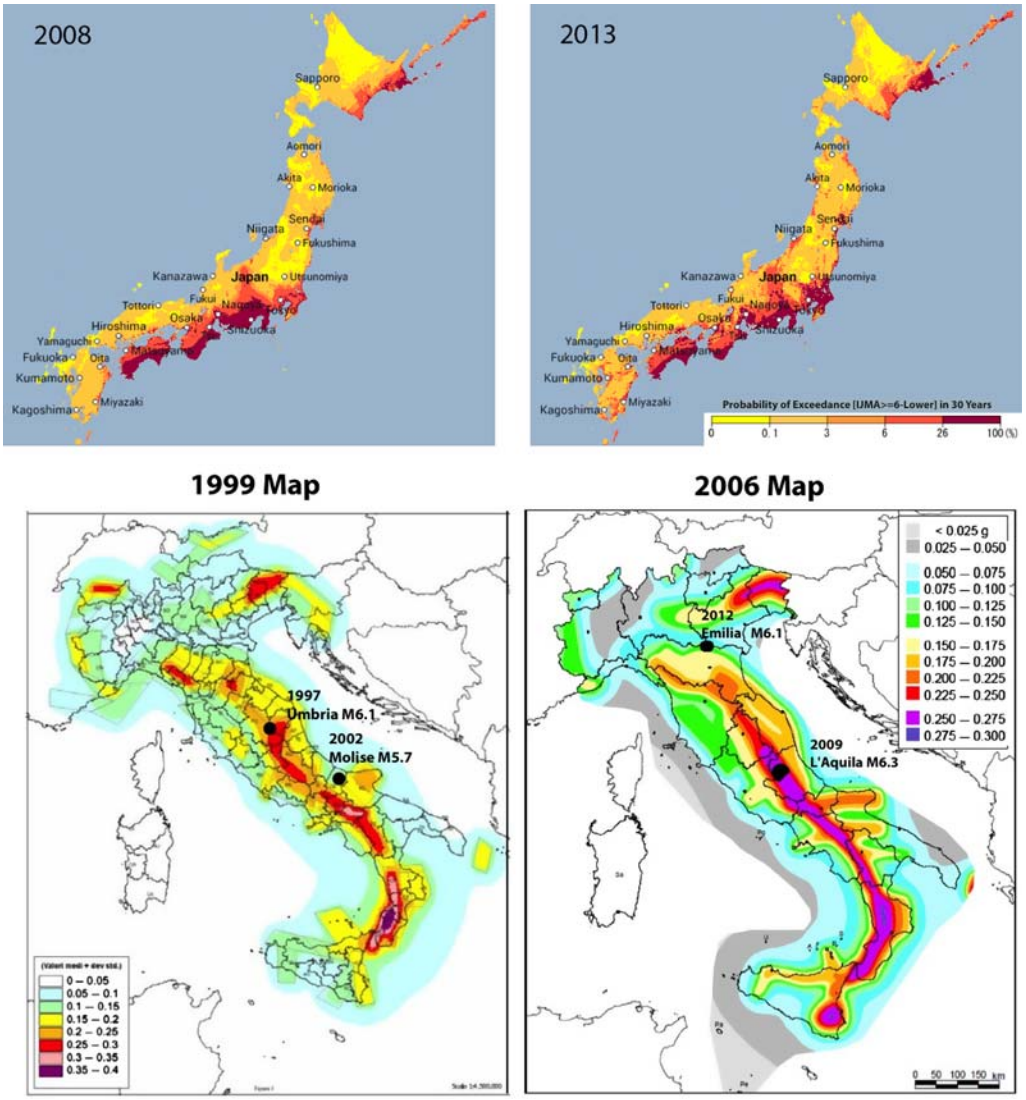

Figure 2. (Top) Comparison of Japanese national seismic hazard maps before and after the 2011 Tohoku earthquake. The predicted hazard has been increased both along the east coast, where the 2011 earthquake occurred, and on the west coast (J-SHIS, 2015; see Data and Resources). (Bottom) Comparison of successive Italian hazard maps (Stein *et al.*, 2013). The 1999 map was updated to reflect the 2002 Molise earthquake, and the 2006 map will likely be updated to include the 2012 Emilia earthquake. The color version of this figure is available only in the electronic edition.

assumptions about earthquake probabilities but otherwise faces the same uncertainties as a probabilistic approach. In some applications, probabilistic and deterministic approaches are combined.

In an intermediate view, both the probabilistic and deterministic algorithms are reasonable in principle, but in many cases key required parameters (such as the maximum earthquake magnitude) are poorly known, unknown, or unknowable (Stein *et al.*, 2012; Stein and Friedrich, 2014). This situation causes some maps to have large uncertainties, which could allow presumed low-probability events to occur more often than anticipated.

The importance of these issues is illustrated by Geller (2011), who noted that the Tohoku earthquake and the others that caused 10 or more fatalities in Japan since 1979 occurred in places assigned a relatively low probability. Hence, he argued that "all of Japan is at risk from earthquakes, and the present state of seismological science does not allow us to reliably differentiate the risk level in particular geographic areas," so a map showing uniform hazard would be preferable to the existing map.

Geller's proposal raises the question of how to quantify how well an earthquake-hazard map is performing. Because the maps influence policy decisions involving high costs to society, measuring how well they perform is important. At present, there are no generally accepted metrics to assess performance. Hence, there are no agreed ways of assessing how well a map performs, to what extent it should be viewed as a success or failure, or whether one map is better than another. Similarly, there is no agreed way of quantifying when and how new maps should be produced and the improvements that they should provide.

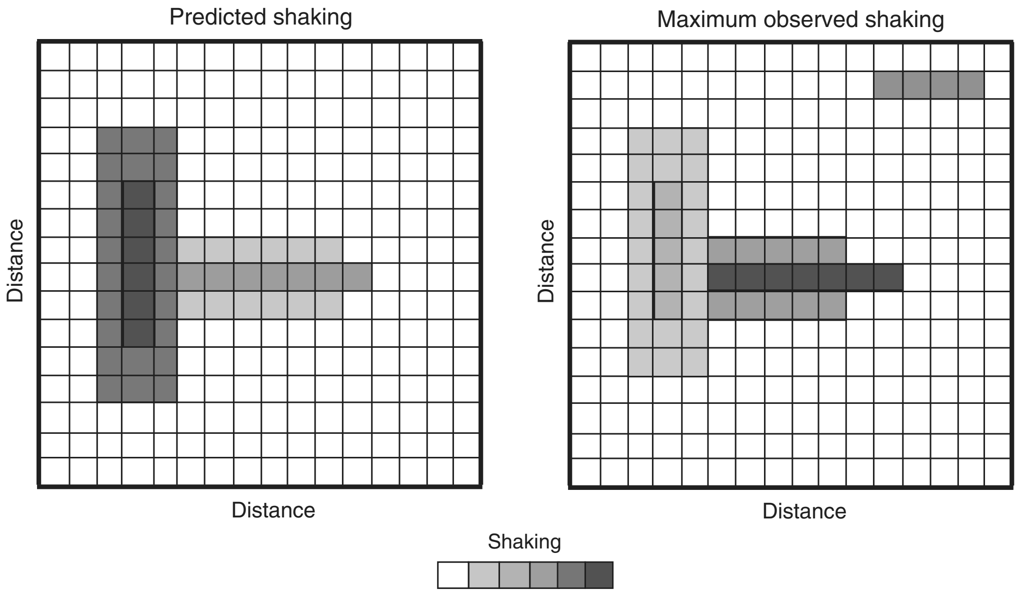

Figure 3. Schematic comparison of hazard map prediction to a map of the maximum observed shaking.

In this article, we explore some possible metrics. Although no single metric can fully characterize map behavior, examining map performance using several metrics can provide useful insight.

Hazard Maps

Conceptually, the issue is how to compare a map of predicted shaking to the maximum shaking observed at sites within it over a suitably long period of time after the map was made. There is increasing interest in this issue, and a variety of approaches have recently been used (Stirling and Peterson, 2006; Albarello and D'Amico, 2008; Mucciarelli *et al.*, 2008; Miyazawa and Mori, 2009; Stirling and Gerstenberger, 2010; Kossobokov and Nekrasova, 2012; Wyss *et al.*, 2012; Mak *et al.*, 2014; Nekrasova *et al.*, 2014) and are being developed under the auspices of the Global Earthquake Model project.

The natural first way to do this is to compare the observations and predictions in map view, as illustrated by the schematic maps in Figure 3, in which, for simplicity, we assume the observation time well exceeds the return time. Such maps could represent ground shaking as acceleration, velocity, or intensity.

In general, this map did reasonably well, in that it identified most of the areas that were subsequently shaken. However, it overpredicted the shaking associated with the north–south-striking fault and underpredicted that associated with the east–west-striking fault. It also did not predict the shaking caused by earthquakes on an unrecognized smaller fault to the northeast.

Quantifying these statements requires going beyond the visual comparison and depends on how the map was made and what it seeks to predict. Most seismic-hazard maps are made using probabilistic seismic-hazard assessment (Cornell, 1968; Field, 2010), which involves assuming the location and recurrence of earthquakes of various magnitudes and forecasting how much shaking will result. Summing the probabilities of ground motions exceeding a specific value yields an estimate of the combined hazard at a given point. The resulting hazard curve (Fig. 4a) shows the estimated probability that shaking will exceed a specific value during a certain time interval.

The predicted hazard in probabilistic maps depends on the probability or, equivalently, on the observation period t and return period T used. The Poisson (time-independent) probability p that earthquake shaking at a site will exceed some value in t years, assuming this occurs on average every T years, is assumed to be

$$p = 1 - \exp(-t/T),$$

which is approximately t/T for $t \ll T$. This probability is small for t/T small and grows with time (Fig. 5). For a given return period, higher probabilities occur for longer observation periods. For example, shaking with a 475 yr return period should have about a 10% chance of being exceeded in 50 yrs, 41% in 250 yrs, 65% in 500 yrs, and 88% in 1000 yrs. Thus, in 50 yrs, there should be only a 10% probability of exceeding the mapped shaking, whereas there is a 63% probability of doing so in an observation period equaling the return period.

The long times involved pose the major challenge for hazard map testing. The time horizon for weather forecasts matches the observation period, so forecasts can be tested at specific sites (Fig. 1). In contrast, as discussed shortly, earthquake-hazard maps are tested by analysis of shaking across many areas to compensate for the short time periods of data available (Ward, 1995).

Probabilistic hazard maps are developed by representing hazard curves for different sites, which is done in two ways. In constant probability hazard maps, the hazard curves for areas are sampled at a fixed probability p to predict the largest anticipated shaking in each area during a certain observation

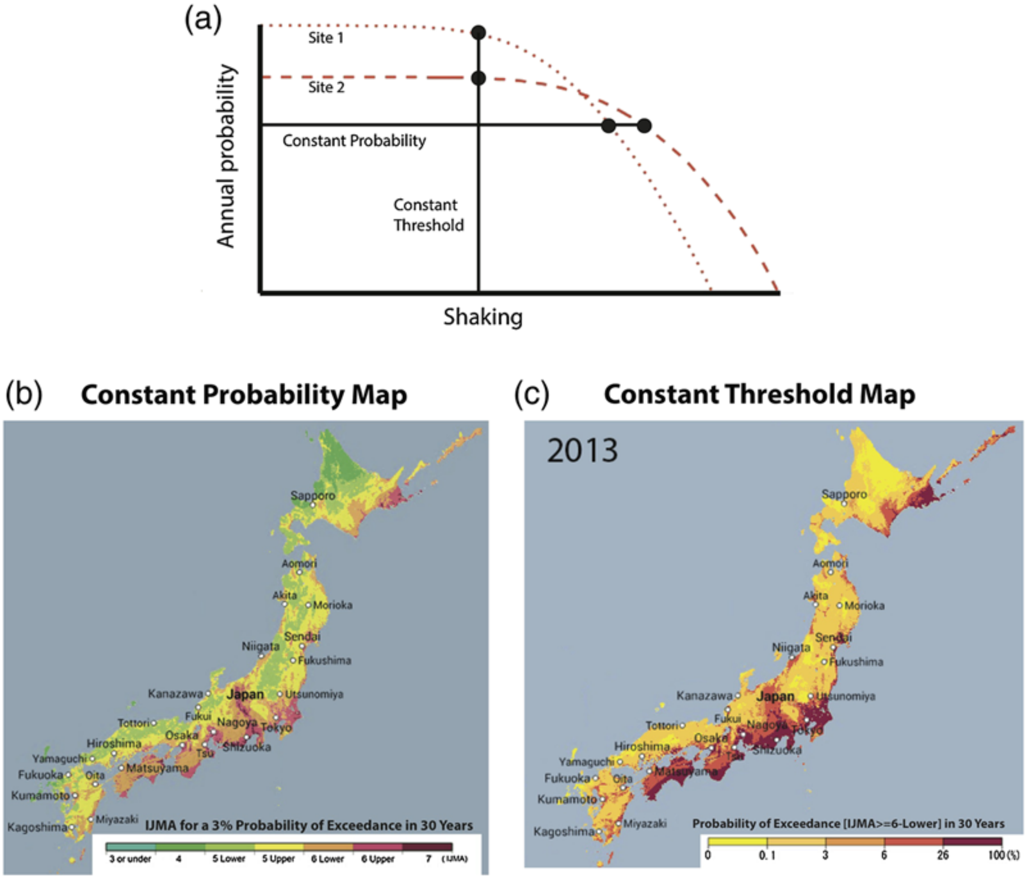

Figure 4. (a) Schematic hazard curves for two sites. Constant probability hazard maps like (b) are made by sampling the hazard curves at a fixed probability to predict that the largest shaking in each area will exceed a specific value with a certain probability during a certain time (observation period). Constant threshold maps like (c) are made by sampling the hazard curves at a fixed shaking level to predict the probability that this shaking level will be exceeded in a certain time. The color version of this figure is available only in the electronic edition.

period. Thus, the map shows the predicted shaking levels s_i for a given probability $p = \Pr(x_i \geq s_i)$ for all areas i. For example, Figure 4b shows the shaking intensity on the Japan Meteorological Agency scale that is anticipated to have only a 3% chance of exceedance in 30 yrs. This approach, termed "uniform hazard," is used in developing seismic design maps in the United States and Europe. An alternative is to present constant threshold hazard maps like that in Figure 4c. In these, the hazard curves are sampled at a fixed shaking level to estimate the probability that this shaking level will be exceeded. The resulting map shows the forecasted probabilities $p_i = \Pr(x_i \geq s)$ for all sites. This representation is commonly used in Japan to show the probability of shaking at or above a given intensity, in this case 6-lower on the Japan Meteorological Agency scale (corresponding approximately to modified Mercalli intensity VIII) in 30 yrs. Such maps show how the probability that a structure will be shaken at or above a certain threshold varies across locations.

Exceedance Metric

Because maps can be made in various ways and thus predict different aspects of the future shaking distribution, we can ask two questions:

1. How well does the map predict the aspects of the distribution of shaking that it was made to predict?
2. How well does the map predict other aspects of the distribution of shaking?

These are most easily explored for the commonly used constant probability maps. These maps predict that ground shaking at any given site will exceed a threshold only with probability p in a given time period. This prediction can be assessed by comparing the actual fraction f of sites with shaking exceeding the threshold to p. This approach, introduced by Ward (1995), considers a large number of sites to avoid the difficulty that large motions at any given site are rare. For example, suppose a map correctly specifies that for a given site there is a 2% chance of a given amount of shaking in a 50 yr period, corresponding to a 2475 yr return period. If the observation period is 250 yrs, Figure 5 shows that there is a 10% chance that the maximum shaking is as large or larger than predicted, and hence there is a 90% chance that it is less than predicted.

The longer the observation time compared with the return period assumed in making the map, the more information we have and the better we can evaluate the map (Beauval *et al.*, 2008, 2010). For example, if in a 50 yr period a large

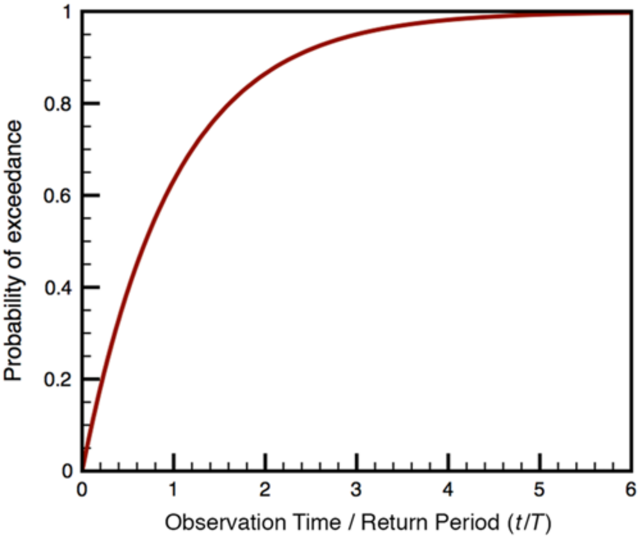

Figure 5. Assumed probability p that during a t year long observation period, shaking at a site will exceed a value that is expected on average once in a T yr return period, assuming $p = 1 - \exp(-t/T)$. The color version of this figure is available only in the electronic edition.

earthquake produced shaking exceeding that predicted at 10% of the sites, this situation could imply the map was not performing well. However, if in the subsequent 200 yrs no higher shaking occurred at the sites, the map would be performing as designed. The exceedance fraction can be thought of as a random variable, for which the expected value is better estimated with longer observation periods. As the length of the observation period as a fraction of the return period increases, the more likely it is that a difference between the predicted and observed exceedance fractions does not occur purely by chance, as discussed later.

This approach allows for the fact that both predictions and observations at nearby sites are correlated. The expected value of the empirical fraction of sites with shaking exceeding thresholds Ef always equals the average true probability of exceedance, regardless of any correlation between sites. This equality holds regardless of any correlation between sites, because the expected value of a sum always equals the sum of the expected values, provided the expected values are finite (as they are). However, as shown later, positive spatial correlation decreases the information available for evaluating maps.

The difference between the observed and predicted probabilities of exceedance, $f - p$, decomposes into a random component and a systematic component,

$$f - p = \underbrace{[f - Ef]}_{\text{random component}} + \underbrace{[Ef - p]}_{\text{systematic component}}.$$

The systematic component is the difference between the average true probability (which equals Ef) and the average predicted probability p of exceedance. If the map parameters do reasonably well in describing earthquakes in the area, Ef will be close to the average predicted probability of exceedance p, and the systematic error will be small. The remaining random component depends on the probability distribution of shaking, which includes both actual chance effects and unmodeled site effects that appear as random scatter. The magnitude of the random component is affected by correlation across sites, as shown in the example discussed later in the article.

Thus, the implicit criterion of success, which can be called a fractional site exceedance criterion, is that if the maximum-observed shaking is plotted as a function of the predicted shaking, only a fraction p (or percentage P) of sites or averaged regions should plot above a 45° line (Fig. 6), aside from chance effects and unmodeled site effects.

How well a map satisfies the fractional site exceedance criterion can be measured using a corresponding metric. A hazard map shows, for all N areas i within it, an estimate of the probability that the maximum observed ground shaking x_i in a time period of length t exceeds a shaking value s_i. This estimated probability can be written $p_i = \Pr(x_i \geq s_i)$. For a sufficiently large number of areas, the fraction f of areas in which $x_i > s_i$ should be approximately equal to the average probability for the areas, or $f \approx \bar{p}$ with $\bar{p} = N^{-1} \sum_{i=1}^{N} p_i$. For the commonly used constant probability maps, $\bar{p} = p$.

Hence, the simplest measure of how well such maps performed is to use a metric based on the fractional site exceedance criterion used in making the maps. This fractional site exceedance metric M0 can be written as

$$\mathrm{M0} = |f - p|,$$

in which f is the fraction of sites at which the predicted ground motion was exceeded during a time period for which p is the appropriate probability (Fig. 5). M0 ranges from 0 to 1, with an ideal map having M0 equal to zero. If M0 > 0, then the map has either positive fractional site exceedance, measured by

$$\mathrm{M0}^{+} = \begin{cases} |f - p| & \text{if } f > p \\ 0 & \text{otherwise,} \end{cases}$$

or negative fractional site exceedance, measured by

$$\mathrm{M0}^{-} = \begin{cases} |f - p| & \text{if } f < p \\ 0 & \text{otherwise.} \end{cases}$$

For any map, either $\mathrm{M0}^{+}$ or $\mathrm{M0}^{-}$ must equal zero, and $\mathrm{M0} = \mathrm{M0}^{+} + \mathrm{M0}^{-}$.

Limitations of Exceedance Metric

Although the exceedance metric is reasonable, it only reflects part of what we might want a probabilistic hazard map to do. This issue is illustrated by the results from four hypothetical probabilistic hazard maps (Fig. 6), all of which satisfy the criterion that the actual shaking exceeds that predicted for this observation period only at 10% of the sites. Thus, all these maps have zero fractional site exceedance,

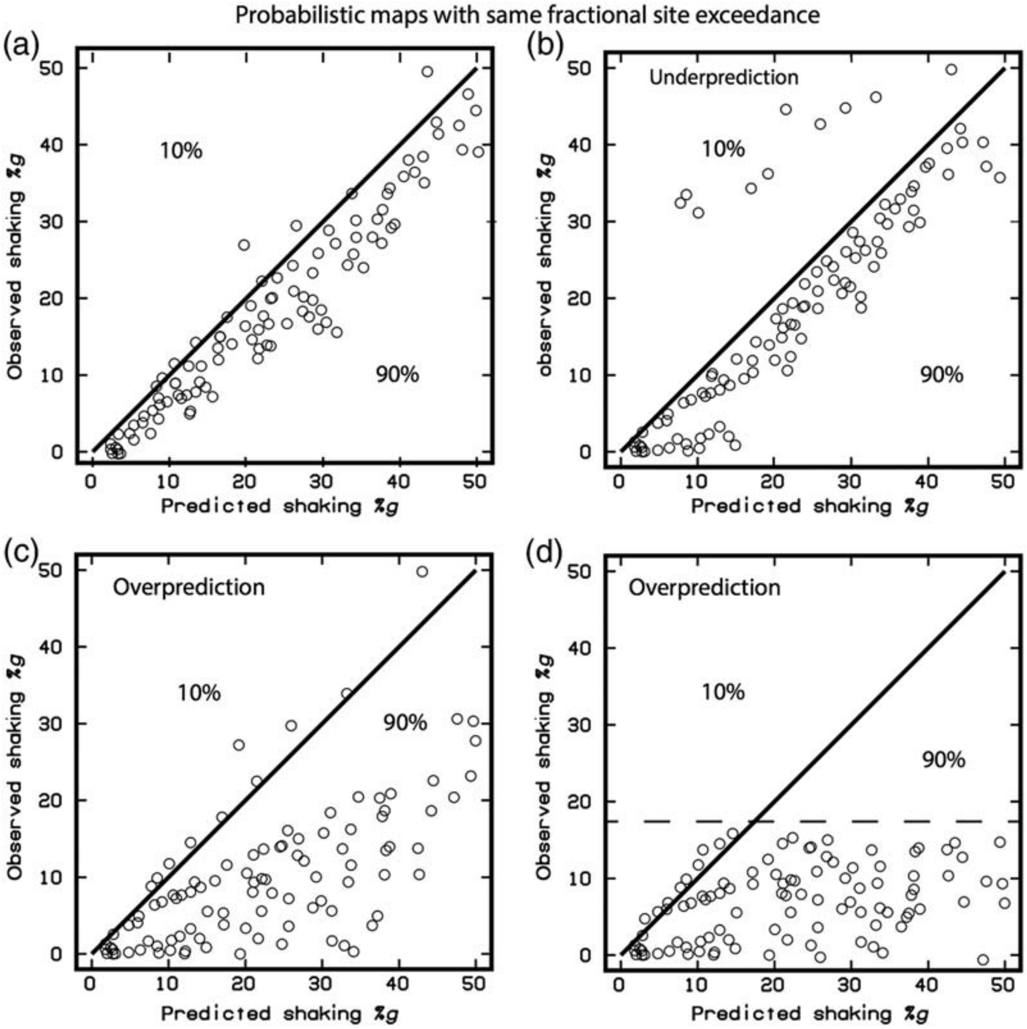

Figure 6. Comparison of the shaking predicted in various subregions of hazard maps to the maximum-observed shaking. Each of the four maps satisfies the fractional site exceedance criterion for $p = 0.1$ but (b)–(d) have significant underpredictions or overpredictions.

or M0 = 0. However, some of these maps would be more useful than others.

The map giving rise to the results in Figure 6a would be viewed as highly effective in that the maximum actual shaking plots close to the predicted shaking. The map largely avoided underprediction, which would have exposed structures built using a building code based on these predictions to greater-than-expected shaking. Similarly, it largely avoided overprediction, which would have caused structures to be overdesigned and thus waste resources.

Mathematically, largely avoiding underprediction can be posed as saying that in the fN areas in which $x_i \geq s_i$, the excess shaking $x_i - s_i$ should be modest. Similarly, largely avoiding overprediction means that in the $(1 - f)N$ areas in which $x_i < s_i$, the overpredictions should be modest. Maps can do well as measured by the fractional site exceedance metric but still have significant overpredictions or underpredictions.

For example, the map giving rise to the results in Figure 6b exposed some areas to much greater shaking than predicted. This situation could reflect faults that were unrecognized or more active than assumed. Hence, although the map satisfies the fractional site exceedance metric that it was designed to achieve, we would not view this map as very effective.

Conversely, the maps in Figure 6c,d overpredicted the shaking at most sites, although they have zero fractional site exceedance. Figure 6c shows a systematic bias toward higher-than-observed values, as could arise from using inaccurate equations to predict ground motion. The map for Figure 6d overpredicted the shaking in that the actual shaking was everywhere less than a threshold value (dashed line), as could arise from overestimating the magnitude of the largest earthquakes that occurred.

Hence, the fractional site exceedance metric M0 measures only part of what we would like a map to do, as illustrated in Figure 7 for hazard maps in which the predicted shaking threshold for each site should be exceeded with 10% probability in the observation period. The map in Figure 7a is nominally very successful as measured by M0 = 0 but significantly underpredicts the shaking at many sites and

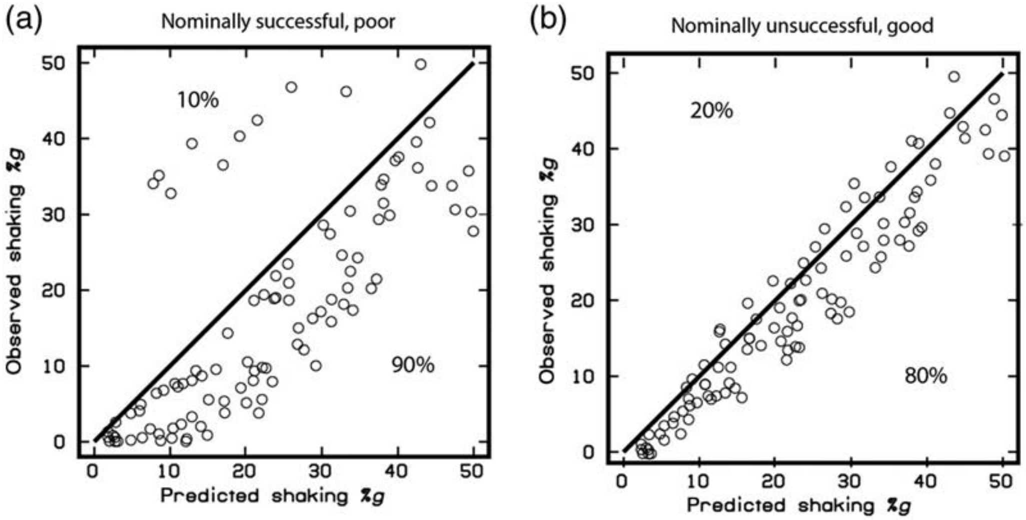

Figure 7. Comparison of the results of two hazard maps. That in (a) is nominally successful, as measured by the fractional exceedance metric, but significantly underpredicts the shaking at many sites and overpredicts that at others. That in (b) is nominally unsuccessful, as measured by the fractional site exceedance metric, but better predicts the shaking at most sites.

overpredicts it at others. Conversely, the map in Figure 7b is nominally unsuccessful as measured by M0, because ground shaking at 20% of the sites exceeds that predicted, so $f = 0.2$. However, it does a reasonable job of predicting the shaking at most sites. Thus, in many ways, the nominally unsuccessful map is better than the nominally successful one.

In this formulation, a map would be considered to be doing poorly if M0 is much greater than 0, that is, if the observed and predicted fractions of exceedances differ enough. This situation could arise from a single very large event causing shaking much larger than anticipated over a large portion of a map but will generally reflect what occurs (or does not occur) in many events in many places over time, as for the Italian maps discussed later.

Alternative Metrics

Many other metrics could be used to provide additional information for quantifying aspects of the observed versus predicted graphs in Figures 6 and 7. Because these additional metrics numerically summarize aspects of the graphs, they account for the length of the observation period. We consider four metrics (M1–M4; Fig. 8) that compare the maximum observed shaking x_i in each of the map's N subregions over some time interval to the map's predicted shaking s_i. Like those in Figures 6 and 7, the hazard maps represented here were constructed so that the shaking threshold for each site should be exceeded with 10% probability over the observation period.

One metric (M1) is simply the squared misfit to the data,

$$\mathrm{M1}(s,x) = \sum_{i=1}^{N}(x_i - s_i)^2/N,$$

which measures how well the predicted shaking compares with the highest observed shaking. Given the probabilistic nature of the ground-motion prediction, scatter above and below the predicted value is expected (Beauval *et al.*, 2010). Even so, smaller overall deviations correspond to better-performing maps. Hence, maps in Figure 6a–d have M1 = 36, 69, 253, and 370.

Similarly, by this metric, the map in Figure 7b (M1 = 25) does better than that in Figure 7a (M1 = 135). Hence from a purely seismological view, M1 seems an appropriate metric that tells more than M0 about how well a map performed.

However, a hazard map's goal is societal—to guide mitigation policies and thus reduce losses in earthquakes. Hence, we might also use metrics that weight different aspects of the prediction differently. For example, because underprediction does potentially more harm than overprediction, we could weight underprediction more heavily. One such asymmetric metric (M2) is the asymmetric squared misfit,

$$\mathrm{M2}(s,x) = \frac{1}{N}\sum_{i=1}^{N} a[(x_i - s_i)^+]^2 + b[(x_i - s_i)^-]^2,$$

in which $(x_i - s_i)^+ = \max(x_i - s_i, 0)$ and $(x_i - s_i)^- = \max(s_i - x_i, 0)$ and $a \geq b \geq 0$.

A refinement would be to vary the asymmetric weights a and b so that they are larger for the areas predicted to be the most hazardous, such that the map is judged most on how it does there. In this metric (M3, a shaking-weighted asymmetric squared misfit),

$$\mathrm{M3}(s,x) = \frac{1}{N}\sum_{i=1}^{N} a(s_i)[(x_i - s_i)^+]^2 + b(s_i)[(x_i - s_i)^-]^2,$$

in which $a(s_i) \geq b(s_i) \geq 0$ and both a and b increase with s_i.

Another option (M4, the exposure-weighted asymmetric squared misfit) varies the asymmetric weights a and b so that

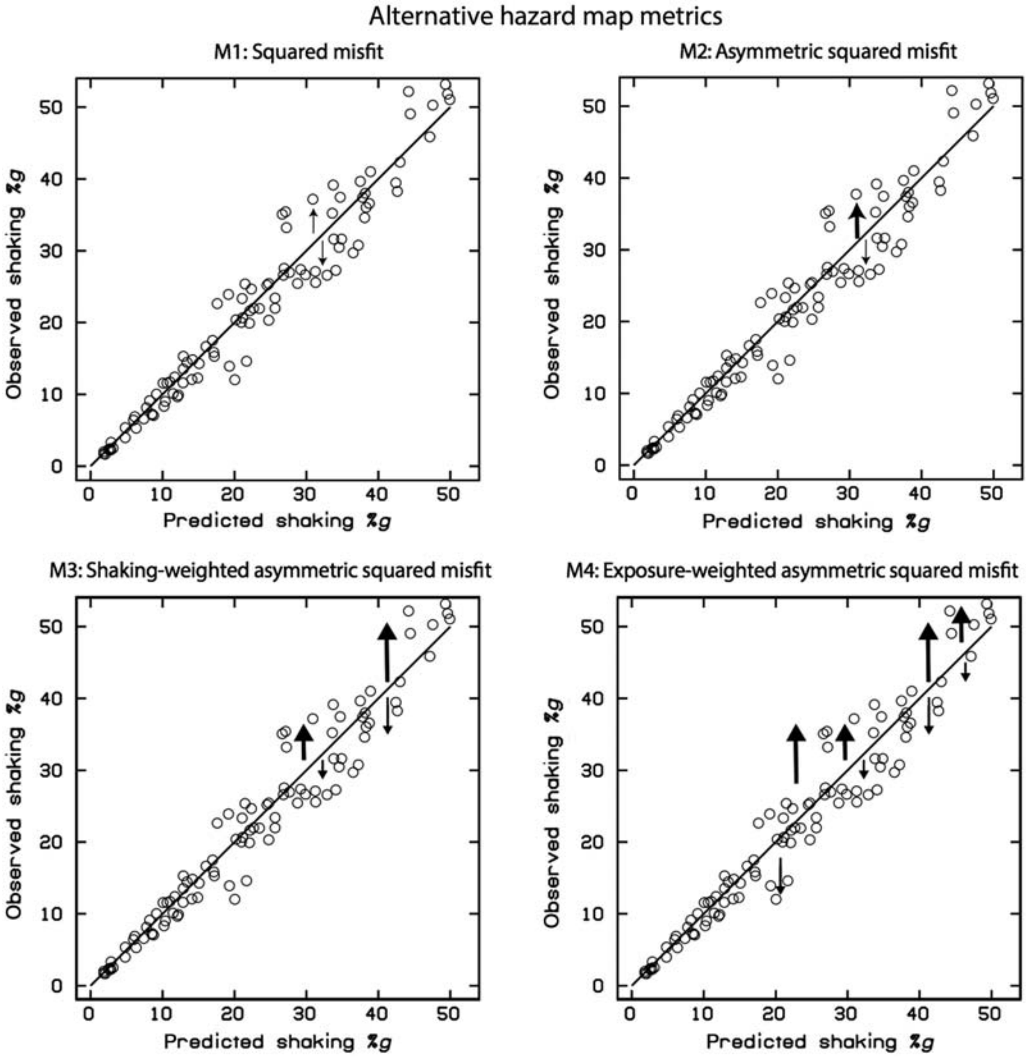

Figure 8. Four metrics (M1–M4) that provide additional information beyond that from the fractional site exceedance metric.

they are larger for areas with the largest exposure of people and/or property, such that the map is judged most on how it does there. Defining e_i as a measure of exposure in the ith region yields the metric

$$\mathrm{M4}(s,x) = \frac{1}{N}\sum_{i=1}^{N} a(e_i)[(x_i - s_i)^+]^2 + b(e_i)[(x_i - s_i)^-]^2,$$

in which $a(e_i) \geq b(e_i) \geq 0$ and both a and b increase with e_i.

Although these metrics are discussed in terms of probabilistic hazard maps, they can also be applied to deterministic maps.

Example

The examples here illustrate some of the many metrics that could be used to provide more information about how well an earthquake-hazard map performs than is provided by the implicit fractional site exceedance metric. Ideally, we would use them to evaluate how different maps of an area, made under different assumptions, actually performed. We would then be in a position to compare the results of the different maps and identify which aspects require improvement.

However, the short time since hazard maps began to be made poses a challenge for assessing how well they work. Hence, various studies examine how well maps describe past shaking (Stirling and Peterson, 2006; Albarello and D'Amico, 2008; Stirling and Gerstenberger, 2010; Kossobokov and Nekrasova, 2012; Wyss *et al.*, 2012; Mak *et al.*, 2014; Nekrasova *et al.*, 2014). Although such hindcast assessments are not true tests, in that they compare the maps with data that were available when the map was made, they give useful insight into the maps' performance.

For example, Figure 9a compares historical intensity data for Italy from 217 B.C. to A.D. 2002, developed from a compilation by Gruppo di Lavoro (2004), with a probabilistic map for 2% in 50 yrs and a deterministic map (Fig. 9b and 9c, respectively) (Nekrasova *et al.*, 2014). As seen in Figure 5, this ~2200 yr observation time and 2475 yr return period cor-

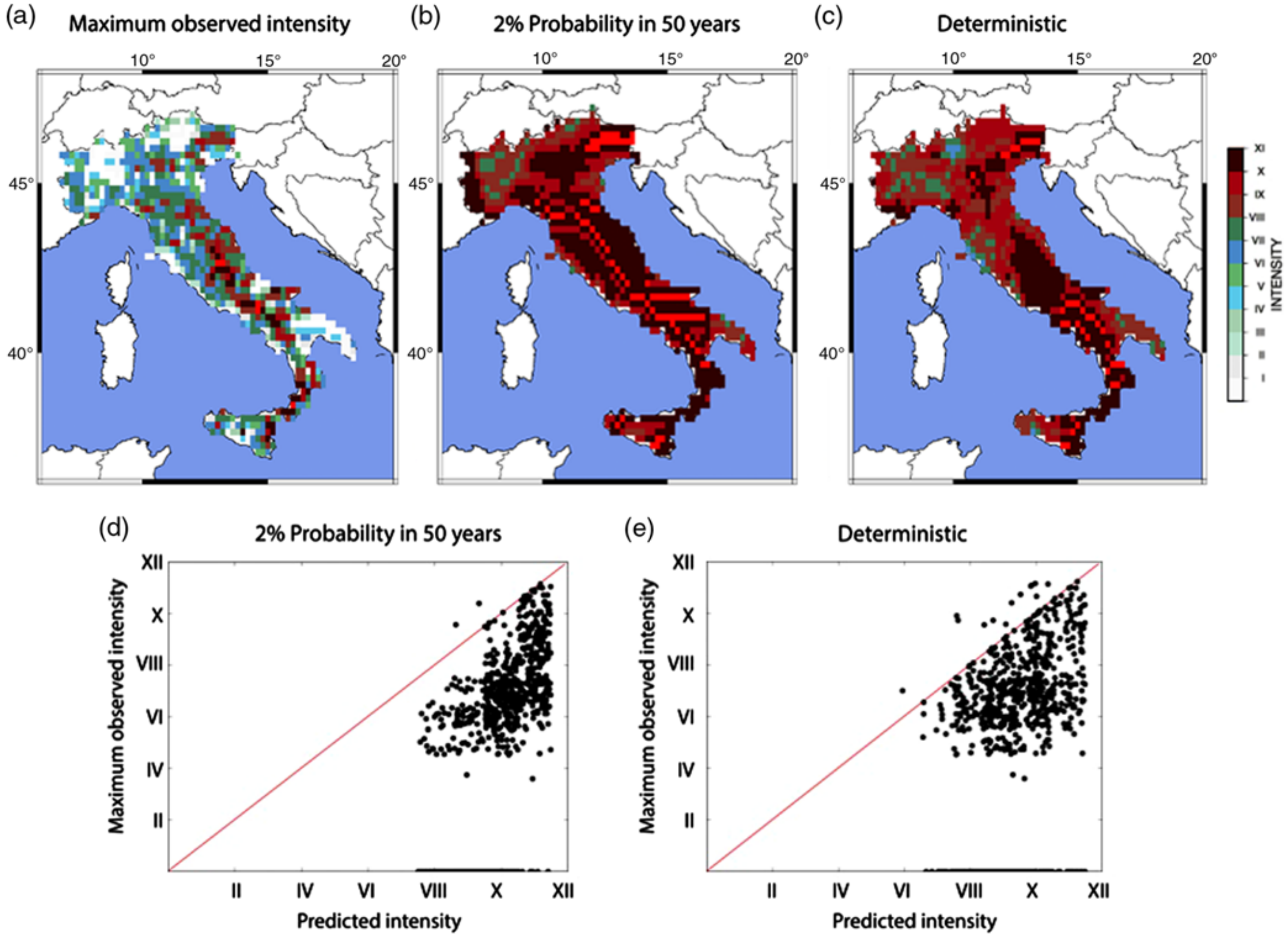

Figure 9. Comparison of (a) historical intensity data for Italy to (b) a probabilistic hazard map and (c) a deterministic hazard map, both of which overpredict the observed shaking, as shown in (d) and (e). Several points are moved slightly for clarity. The color version of this figure is available only in the electronic edition.

respond to an exceedance probability $p = 58.89\%$, hence the observed shaking at most sites should exceed that predicted.

However, the probabilistic map has only two sites out of 800 for which the observed shaking exceeds the threshold value, for $f = 0.25\%$. Comparing that with $p = 58.89\%$, we find large negative fractional site exceedance, with $\mathrm{M0} = 0.5864$.

For the deterministic map, the predicted threshold of ground motion was exceeded at 13 of the 800 sites, so $f = 1.62\%$. The fractional exceedance metric for the deterministic map cannot be computed, because the map does not provide a stated probability of exceedance over a time period. However, in principle, we can use the past performance to crudely calibrate the deterministic map. Thus, the empirical probability of exceedance for sites in Italy was 1.62% over 2200 yrs, corresponding to 2% over 2713 yrs, or 0.037% over 50 yrs. A similar approach has been used to calibrate deterministic scenario-based population forecasts (Keyfitz, 1981; Alho and Spencer, 2005). Because there are questions about the data, as discussed below, this example is purely illustrative.

Both hazard maps significantly overpredict the observed shaking, as shown by the M1 metric. The deterministic map does better ($\mathrm{M1} = 23.7$) than the probabilistic map ($\mathrm{M1} = 27.2$), because its overall overprediction is somewhat less.

The large misfit between the data and probabilistic map shown by M0 is unlikely to have occurred purely by chance, given the length of the historical catalog, which is comparable to the map's return period of 2475 yrs. The poor fit of both maps indicates a problem with the data, maps, or both. The metrics illustrate the problem but do not indicate its cause.

It is possible that some of the assumptions used when making the hazard map were biased toward overpredictions. However, it is likely that much of the misfit results from the catalog being biased to too-low values. The historical catalog is thought to be incomplete (Stucchi *et al.*, 2004) and may underestimate the largest actual shaking in areas due to a space–time sampling bias and/or difficulties with the historically inferred intensities. Figure 10 schematically shows how sampling bias could understate actual shaking, and Hough (2013) shows that sampling bias can also overestimate actual shaking.

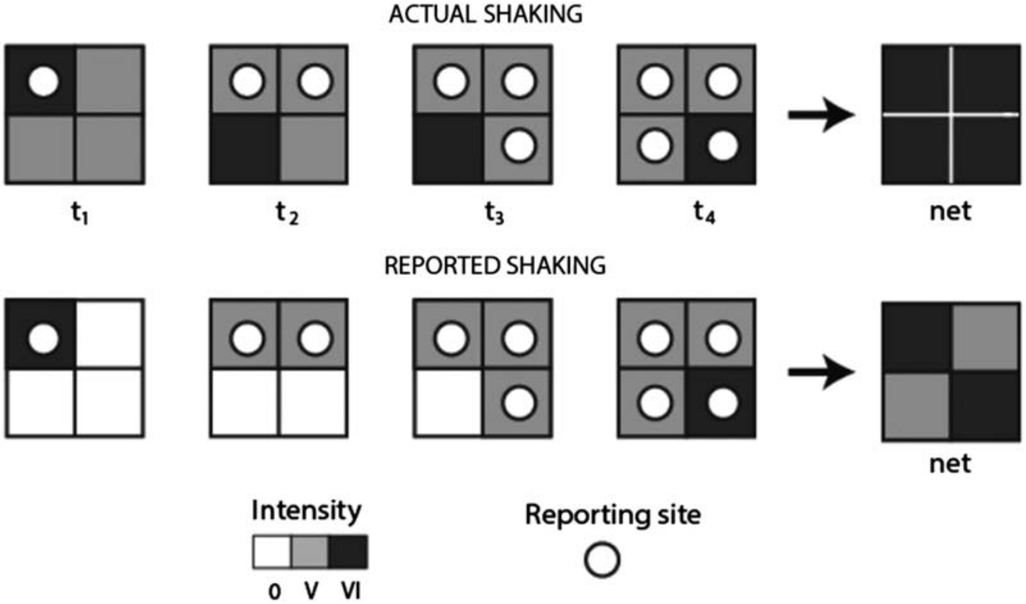

Figure 10. Schematic illustration of one way that variations in sampling over time t could underestimate earthquake shaking. If reports are available only from grid cells including a reporting site (circle), the reported maximum shaking in some cells (lower row) is less than the actual maximum (upper row).

This example also illustrates other complexities. The historical intensity data have a long enough observation time for reliable comparison with the 2% map. However, they have the difficulty that regions can have no reported shaking, either because no shaking large enough to be reported occurred or because such shaking occurred but is not reflected in the historical record. When the sites with no reported shaking are omitted, M1 values for the probabilistic map drop from 27.2 to 10.4, and M1 values for the deterministic map drop from 23.7 to 7.2. The difference in M1 values between the probabilistic and deterministic maps stays about the same (~3). Because f is so small relative to p for the probabilistic map, the M0 value just barely changes, decreasing from 0.5864 to 0.5857. These issues would not arise for instrumentally recorded data for which low values can be distinguished from missing values (no data).

Another complexity is that hazard maps predict average effects over some area for a uniform site response, whereas actual ground shaking includes local site effects. Hence, ideally, site effects would be included or removed if the structure were adequately known. Otherwise, nearby sites could be averaged to reduce the effect of variations on a spatial scale smaller than can be modeled with available information.

Most crucially, this analysis compared a set of observations with maps produced after the earthquakes occurred. The metrics thus describe how well the maps fit the data that were used in making them. Such retrospective analysis has been the norm to date, given that hazard mapping is a relatively new technology compared to the earthquake record. Prospective testing will be needed to see how well maps predicted future shaking. By examining how well a map described what happened (or happens) over its entire area, metrics like those discussed here have the benefit of requiring a much shorter time span of data than would be required to assess how the map performed at individual sites.

Effect of Random Error and Bias on Metrics

Although metrics measure how well the predicted shaking matches that observed, assessing their statistical properties requires also assuming and applying a probability model to the data underlying the metrics.

The situation is analogous to deciding if a diet is working. Using your weight as a metric shows changes over time, but deciding whether the changes could have occurred purely by chance or are significant requires assuming and applying a probability model for the scale's weight measurements. The probability model involves the properties of the scale: different scales all measure weight but with different precision and accuracy. Hence statistical significance depends on the model assumed to describe the data.

Recall that for the exceedance metric M0, the difference $f - p$ between observed and forecasted is the sum of the chance component $f - Ef$ and the bias $Ef - p$. To interpret the difference $f - p$, we want to know how large the chance component might be and then to assess whether the bias appears to be appreciable. Statistical significance tests often are used for this purpose in analogous applications.

Understanding the effect of chance and biases on numerical values of metrics requires considering the sources of randomness and bias. Are the sites the whole population or a sample? How was the sample chosen? How accurate are the measurements of shaking, and what is the joint probability distribution of shaking?

One also needs to consider how the map was developed. To the extent that past shaking data were used in developing the hazard curves underlying the map, the numerical values of the metrics applied to past data may not reflect their numerical values when applied to future events. This is a potential problem, because the forecasts' purpose is to predict the future, not the past. Cross-validation methods may be useful, but the limited number of sites and their correlations over space and time may pose difficulties.

For illustrative purposes, we consider the probability distribution of f, the fraction of sites for which shaking exceeded the specified thresholds, for the Italy data used in Figure 9. We take the sites to be a population of interest rather than a sample from a larger population. We consider only randomness associated with ground motion at each site, and for clarity of exposition we use a simple model. Figure 9b is a constant probability map, predicting that the probability is 2% that in 50 yrs shaking at a given site exceeds a threshold value for the site, and thus that in 2200 yrs the probability of exceedance is $p = 58.89\%$.

It is of interest to test whether the difference between the observed number of exceedances and the expected number is greater than what would be likely to occur by chance when the model is correct, that is, whether the difference is statistically significant (the known limitations of hypothesis testing notwithstanding) (Marzocchi *et al.*, 2012). For each site $i = 1, \ldots, N$, define $X_i = 1$ if shaking exceeded the threshold and $X_i = 0$ otherwise. Consistent with the model underlying the constant probability map, we assume each X_i has a Bernoulli distribution with parameter p, that is, $X_i = 1$ with probability p and $X_i = 0$ with probability $1 - p$. If the X_is are mutually independent, then the total number of exceedances, $Nf = \sum_{i=1}^{N} X_i$, has a binomial distribution with parameters N and p.

For $N = 800$ sites, the data show two exceedances, so $Nf = 2$ and thus $f = 0.0025$. In contrast, for a binomial model with parameters 800 and 0.5889 (the probability specified for the map), the expected number of exceedances is $Np = 471.2$, and the probability that the observed count Nf is either ≤ 2 or ≥ 798 is astronomically small (1.7×10^{-179}). This probability is vastly smaller than the conventional 0.05 level of significance, indicating the discrepancy between Nf and Np, or equivalently between f and p, is statistically significant. If the assumed model is correct, there is almost no chance that the observed number of exceedances would be so far from the expected number. Either an incredibly unlikely small amount of exceedance occurred just by chance or there are problems with the model or data, as previously discussed.

Another possibility is that the model's assumption of independence across sites could be wrong, so exceedances at different sites are correlated. Although, as discussed earlier, this correlation does not bias the metric, it would affect significance tests because it affects the amount of chance variability in the number of exceedances. If the average correlation is positive, the observations carry less information, so the evidence against $p = 0.5889$ is weaker.

To see this, note that under the Bernoulli model, the covariance of X_i and X_j for sites i and j equals $\rho_{ij} p(1 - p)$, with the correlation ρ_{ij} reflecting the spatial correlation. The average correlation across different sites is $\bar{\rho} = \sum_{i \neq j} \rho_{ij} / [N(N - 1)]$. For example, if each X_i has correlation ρ with exactly k other X_js and no correlation with all other X_js, then $\bar{\rho} = \rho k / (N - 1)$. To help interpret the correlation, consider $\rho_{ij} = \rho \geq 0$ for all distinct sites i and j. This implies $\bar{\rho} = \rho$. If there is independence or, more generally, if $\rho = 0$, the probability of nonexceedance at a given pair of sites equals $(1 - p)^2$. But, if the correlation is $\rho > 0$, then the probability of nonexceedance at the pair of sites increases by $\rho p(1 - p)$. Using $p = 0.589$ and, purely for illustrative purposes, taking $\rho = 0.36$, we see the probability of nonexceedance at the pair of sites increases from 0.169, the probability under independence, to 0.256, a relatively large increase (52%).

In general, the variance of Nf is $Np(1 - p)[1 + (N - 1)\bar{\rho}]$. The term in square brackets is an inflation factor for the binomial variance when $\bar{\rho} > 0$. Empirical estimation of $\bar{\rho}$ is beyond the scope of this article. Once $\bar{\rho}$ has been specified, however, the significance calculations can easily accommodate spatial correlation if the Gaussian approximation to the binomial distribution is used. Under the independence assumption, a simple approximation to the binomial distribution of Nf is based on treating $z = (Nf - Np + c)/\sqrt{Np(1 - p)}$ as if it were Gaussian with mean 0 and variance 1, in which the continuity correction c equals 0.5 if $f < p$, -0.5 if $f > p$, and 0 if $f = p$. With $Np = 471.2$, $Nf = 2$, and $N = 800$, we calculate $z = -33.7$, which, as before (with the binomial model), corresponds to an astronomically small probability. Now, suppose for illustrative purposes that $\bar{\rho} = 0.36$, as discussed in the previous paragraph. To take correlation into account, we divide the z-value of -33.7 by $16.99 = \sqrt{1 + (N - 1)\bar{\rho}}$ to get an adjusted z-value of -1.98. This corresponds to a two-tailed probability of 0.047, which is still smaller than the conventional significance level of 0.05. If the correlation parameter $\bar{\rho}$ were even larger, say 0.37, the adjusted z-value would increase, and the associated two-tailed probability would exceed 0.05. In that case, the difference between Nf and Np would not be statistically significant at the 0.05 level. It is clear that an assumption of independence can make a huge difference in these calculations (Kruskal, 1988).

Starting with the decomposition of $f - p$ given earlier, squaring both sides, and taking expected values, shows the mean-squared deviation between f and p equals the sum of the variance in f and the squared bias in p; that is, $E(f - p)^2 = V(f) + Bias^2$.

When the variance $V(f)$ is not too large, we may use the following estimator of the squared bias in the specification of p:

$$\text{Estimator of squared bias of } p = (f - p)^2 - V(f).$$

For example, for the 2%-in-50-yrs model with correlation, we can estimate $V(f)$ by $f(1-f)[1+(N-1)\bar{\rho}]/N$, or 0.0071, which does not assume the specification of p is correct. The estimate of squared bias is 0.337. The ratio of the square root of 0.337 to p is 0.99. According to this analysis, then, based on illustrative assumptions that may not capture reality, the estimate of p is almost all systematic error (bias).

Map Comparison and Updating

The metrics discussed here can also be used to compare the maximum shaking observed over the years in regions within a hazard map to that predicted by the map and by some null hypotheses. This could be done via the skill score, a method used to assess forecasts including weather forecasts:

$$SS(s,r,x) = 1 - M(s,x)/M(r,x),$$

in which M is any of the metrics, x is the maximum shaking, s is the map prediction, and r is the prediction of a reference map produced using a reference model (referred to as a null hypothesis). The skill score would be positive if the map's predictions did better than those of the map made with the null hypothesis and would be negative if they did worse. We could then assess how well maps have done after a certain time, and whether successive generations of maps do better.

One simple null hypothesis is that of regionally uniformly distributed seismicity or hazard. Geller (2011) suggests the Japanese hazard map in use prior to the Tohoku earthquake is performing worse than such a null hypothesis. Another null hypothesis is to start with the assumption that all oceanic trenches have similar b-value curves (Kagan and Jackson, 2013) and can be modeled as the same, including the possibility of an M_{w} 9 earthquake (there is about one every 20 yrs somewhere on a trench).

The idea that a map that includes the full detail of what is known about an area's geology and earthquake history may not perform as well as assuming seismicity or hazard are uniform at first seems unlikely. However, it is not inconceivable. An analogy could be describing a function of time composed of a linear term plus a random component. A detailed polynomial fit to the past data describes them better than a simple linear fit but can be a worse predictor of the future than the linear trend. This effect is known as overparameterization or overfitting (Silver, 2012). A way to investigate this possibility would be to smooth hazard maps over progressively larger footprints. There may be an optimal level of smoothing that produces better performing maps because, on a large scale, regional differences are clearly important.

Metrics for hazard maps can also be useful in dealing with the complex question of when and how to update a map. A common response to unexpected earthquakes or shaking is to remake a hazard map to show higher hazard in the affected areas (Fig. 2). The revised map (e.g., Frankel *et al.*, 2010) would have better described the effects of past earthquakes and is anticipated to better represent the effects of future earthquakes. Maps are also remade when additional information, such as newly discovered faults or improved ground-motion prediction models, are recognized or become available.

Although remaking maps given new information makes sense, it is done without explicit assessment of how well the existing map has performed to date or with explicit criteria for when a map should be remade. Similarly, this process provides no explicit way to quantify what improvements are hoped for from the new map. These issues can be explored using metrics like those here. Statistical models, including Bayesian models, could be used to simultaneously provide appropriate updating as new data become available and to smooth the maps. Specification of such models will involve an interesting blending of modern statistical modeling with advancing seismological knowledge.

In summary, we believe that metrics like those discussed here can help seismologists assess how well earthquake-hazard maps actually perform, compare maps produced under various assumptions and choices of parameters, and develop improved maps.

Data and Resources

The Japanese hazard maps are from http://www.j-shis.bosai.go.jp/map/?lang=en (last accessed January 2014). The historical intensity data for Italy from 217 B.C. to A.D. 2002 from a compilation by Gruppo di Lavoro (2004) (http://emidius.mi.ingv.it/CPTI04; last accessed January 2014), and the digital values for the hazard maps from Nekrasova *et al.* (2014) were provided by A. Peresean.

Acknowledgments

We thank the U.S. Geological Survey (USGS) John Wesley Powell Center for Analysis and Synthesis for hosting a working group under auspices of the Global Earthquake Model project, whose stimulating discussions inspired this work, and Paola Albini for helpful comments. S. Stein thanks the Alexander von Humboldt Foundation for supporting his stay at the Georg-August-Universität Göttingen and Ludwig-Maximilians Universität München. B. D. Spencer thanks the Institute for Policy Research at Northwestern University for supporting his research.

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,** 1088–1094.

Alho, J. M., and B. D. Spencer (2005). *Statistical Demography and Forecasting*, Springer, New York, New York.

Beauval, C., P.-Y. Bard, and J. Douglas (2010). Comment on "Test of seismic hazard map from 500 years of recorded intensity data in Japan" by Masatoshi Miyazawa and Jim Mori, *Bull. Seismol. Soc. Am.* **100,** 3329–3331.

Beauval, C., P.-Y. Bard, S. Hainzl, and P. Guéguen (2008). Can strong motion observations be used to constrain probabilistic seismic hazard estimates? *Bull. Seismol. Soc. Am.* **98,** 509–520.

Cornell, C. A. (1968). Engineering seismic risk analysis, *Bull. Seismol. Soc. Am.* **58,** 1583–1606.

Field, E. (2010). *Probabilistic seismic hazard analysis: A primer*, http://www.opensha.org/ (last accessed May 2014).

Frankel, A. (2013). Comment on “Why earthquake hazard maps often fail and what to do about it,” by S. Stein, R. J. Geller, and M. Liu, *Tectonophysics* **592,** 200–206.

Frankel, A., S. Harmsen, C. Mueller, E. Calais, and J. Haase (2010). Documentation for initial seismic hazard maps for Haiti, *Open-File Rept. 2010-1067*.

Geller, R. J. (2011). Shake-up time for Japanese seismology, *Nature* **472,** 407–409.

Gruppo di Lavoro (2004). *Catalogo parametrico dei terremoti italiani, versione 2004 (CPTI04)*, Istituto Nazionale di Geofisica e Vulcanologia (INGV), Bologna, Italy, http://emidius.mi.ingv.it/CPTI04 (last accessed January 2014).

Gulkan, P (2013). A dispassionate view of seismic-hazard assessment, *Seismol. Res. Lett.* **84,** 413–416.

Hanks, T. C., G. C. Beroza, and S. Toda (2012). Have recent earthquakes exposed flaws in our misunderstandings of probabilistic seismic hazard analysis? *Seismol. Res. Lett.* **83,** 759–764.

Hough, S. E. (2013). Spatial variability of “Did You Feel it?” intensity data: Insights into sampling biases in historical earthquake intensity distributions, *Bull. Seismol. Soc. Am.* **103,** 2767–2781, doi: 10.1785/0120120285.

Iervolinoa, I. (2013). Probabilities and fallacies: Why hazard maps cannot be validated by individual earthquakes, *Earthq. Spectra* **29,** no. 3, 1125–1136.

Kagan, Y. Y., and D. D. Jackson (2013). Tohoku earthquake: A surprise, *Bull. Seismol. Soc. Am.* **103,** 1181–1194.

Kerr, R. A. (2011). Seismic crystal ball proving mostly cloudy around the world, *Science* **332,** 912–913.

Keyfitz, N. (1981). The limits of population forecasting, *Popul. Dev. Rev.* **7,** 579–593.

Klügel, J.-U., L. Mualchin, and G. F. Panza (2006). A scenario-based procedure for seismic risk analysis, *Eng. Geol.* **88,** 1–22.

Kossobokov, V. G., and A. K. Nekrasova (2012). Global Seismic Hazard Assessment Program maps are erroneous, *Seismic Instruments* **48,** 162–170.

Kruskal, W. (1988). Miracles and statistics: The casual assumption of independence, *J. Am. Stat. Assoc.* **83,** 929–940.

Mak, S., R. A. Clements, and D. Schorlemmer (2014). The statistical power of testing probabilistic seismic-hazard assessments, *Seismol. Res. Lett.* **85,** 781–783.

Manaker, D. M., E. Calais, A. M. Freed, S. T. Ali, P. Przybylski, G. Mattioli, P. Jansma, C. Prepetit, and J. B. De Chabalie (2008). Interseismic plate coupling and strain partitioning in the northeastern Caribbean, *Geophys. J. Int.* **174,** 889–903.

Marzocchi, W., J. D. Zechar, and T. H. Jordan (2012). Bayesian forecast evaluation and ensemble earthquake forecasting, *Bull. Seismol. Soc. Am.* **102,** 2574–2584.

Minoura, K., F. Imamura, D. Sugawa, Y. Kono, and T. Iwashita (2001). The 869 Jogan tsunami deposit and recurrence interval of large-scale tsunami on the Pacific coast of northeast Japan, *J. Nat. Disast. Sci.* **23,** 83–88.

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,** 3140–3149.

Mucciarelli, M., D. Albarello, and V. D’Amico (2008). Comparison of probabilistic seismic hazard estimates in Italy, *Bull. Seismol. Soc. Am.* **98,** 2652–2664.

Murphy, A. H. (1993). What is a good forecast? An essay on the nature of goodness in weather forecasting, *Weather Forecast.* **8,** 281–293.

Nekrasova, A., V. Kossobokov, A. Peresan, and A. Magrin (2014). The comparison of the NDSHA, PSHA seismic hazard maps and real seismicity for the Italian territory, *Nat. Hazards* **70,** 629–641.

Peresan, A., and G. F. Panza (2012). Improving earthquake hazard assessments in Italy: An alternative to “Texas sharpshooting”, *Eos Trans. AGU* **93,** 538.

Rice, J. A. (2007). *Mathematical Statistics and Data Analysis*, Third Ed., Duxbury Press, Belmont, California.

Sagiya, T. (2011). Integrate all available data, *Nature* **473,** 146–147.

Silver, N. (2012). *The Signal and the Noise*, Penguin, New York, New York.

Sivia, D. S. (2006). *Data Analysis: A Bayesian Tutorial*, Second Ed.,Oxford University Press, Oxford, United Kingdom.

Stein, S., and A. Friedrich (2014). How much can we clear the crystal ball? *Astron. Geophys.* **55,** 2.11–2.17.

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.

Stein, S., R. J. Geller, and M. Liu (2013). Reply to “Comment on ‘Why earthquake hazard maps often fail and what to do about it’ by Arthur Frankel”, *Tectonophysics* **592,** 207–209.

Stein, S., B. D. Spencer, and E. Brooks (2015). Bayes and BOGSAT: Issues in when and how to revise earthquake hazard maps, *Seismol. Res. Lett.* **86,** 6–10.

Stephenson, D. (2000). Use of the “odds ratio” for diagnosing forecast skill, *Weather Forecast.* **15,** 221–232.

Stirling, M. W. (2012). Earthquake hazard maps and objective testing: The hazard mapper’s point of view, *Seismol. Res. Lett.* **83,** 231–232.

Stirling, M. W., and M. Gerstenberger (2010). Ground motion-based testing of seismic hazard models in New Zealand, *Bull. Seismol. Soc. Am.* **100,** 1407–1414.

Stirling, M. W., 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,** 1978–1994.

Stucchi, M., P. Albini, C. Mirto, and A. Rebez (2004). Assessing the completeness of Italian historical earthquake data, *Ann. Geophys.* **47,** 659–673.

Wang, Z. (2011). Seismic hazard assessment: Issues and alternatives, *Pure Appl. Geophys.* **168,** 11–25.

Wang, Z., and J. Cobb (2012). A critique of probabilistic versus deterministic seismic hazard analysis with special reference to the New Madrid seismic zone, in *Recent Advances in North American Paleoseismology and Neotectonics East of the Rockies*, Geological Society of America, Boulder, Colorado.

Ward, S. (1995). Area-based tests of long-term seismic hazard predictions, *Bull. Seismol. Soc. Am.* **85,** 1285–1298.

Wyss, M., A. Nekraskova, and V. Kossobokov (2012). Errors in expected human losses due to incorrect seismic hazard estimates, *Nat. Hazards* **62,** 927–935.

Department of Earth and Planetary Sciences
Institute for Policy Research
Northwestern University
Evanston, Illinois 60208
(S.S., E.M.B.)

Department of Statistics
Institute for Policy Research
Northwestern University
Evanston, Illinois 60208
(B.D.S.)

Manuscript received 7 June 2014;
Published Online 21 July 2015