Comparing the performance of Japan’s earthquake hazard maps
to uniform and randomized maps

Edward M. Brooks¹, Seth Stein¹,², Bruce D. Spencer³,²

¹Department of Earth and Planetary Sciences, Northwestern University, Evanston, Illinois
²Institute for Policy Research, Northwestern University, Evanston, Illinois
³Department of Statistics, Northwestern University, Evanston, Illinois

INTRODUCTION

The devastating 2011 magnitude 9.1 Tohoku earthquake and the resulting shaking and tsunami were much larger than anticipated in earthquake hazard maps. Because this and all other earthquakes that caused ten or more fatalities in Japan since 1979 occurred in places assigned a relatively low hazard, Geller (2011) 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. Defenders of the maps countered by arguing that these earthquakes are low-probability events allowed by the maps (Hanks et al., 2012), which predict the levels of shaking that should expected with a certain probability over a given time (Cornell, 1968; Field, 2010). Although such maps are used worldwide in making costly policy decisions for earthquake-resistant construction, how well these maps actually perform is unknown. We explore this hotly-contested issue (Kerr, 2011; Stein et al., 2012; Stirling, 2012; Gulkan, 2013; Marzocchi and Jordan, 2014; Wang, 2015) by comparing how well a 510-year-long record of earthquake shaking in Japan (Miyazawa and Mori, 2009) is described by the actual maps, uniform maps, and randomized maps. Surprisingly, as measured by the metric implicit in the maps, i.e. that during the chosen time interval the predicted ground motion should be exceeded only at a specific fraction of the sites, both uniform and randomized maps
do better than the actual maps. However, using the squared misfit between maximum observed shaking and that predicted as a metric, the actual maps do better than uniform or randomized maps. These results indicate that the Japanese maps are not performing as well as expected, that what factors control map performance is complicated, and that learning more about how maps perform and why would be valuable in making more effective policy.

HAZARD MAPS

Japan’s probabilistic seismic hazard maps (Fig. 1) use assumptions about the locations, magnitudes, and probabilities of future earthquakes and the resulting shaking to predict the maximum shaking that should be exceeded only with a certain probability over a given time (Cornell, 1968; Field, 2010). At a point on the map, the probability $p$ that during $t$ years of observations shaking will exceed a value that is expected once in a $T$ year return period is assumed to be

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

This probability is small for $t/T$ small and grows with time (Fig. 2a). For example, shaking with a 475-year return period should have about a 10% chance being exceeded in 50 years, 41% in 250 years, 65% in 500 years, and 88% in 1000 years. Maps are characterized by either their return period (e.g., 475 years) or probability in an observation time (10% in 50 years). Maps are generated for different return periods because greater shaking is anticipated from rarer but larger earthquakes.

Although such maps are used worldwide in making costly policy decisions for earthquake-resistant construction, how well they actually perform is unknown. A map can be assessed by comparing the actual fraction $f$ of sites where shaking exceeded the mapped threshold at that site to $p$. This approach (Ward, 1995) considers many sites to avoid the difficulty that large motions at any given site are rare. For example, a 10%
chance that the maximum shaking at a site during the observation period will be as large or larger than predicted corresponds to a 90% chance that it will be less.

The short time since hazard maps began to be made poses a challenge for assessing how well they work (Beauval et al., 2008; 2010). If during ten years after a map was made large earthquakes produced shaking at 40% of the sites exceeding that predicted, the map may not performing well. However, if in the subsequent 240 years no higher shaking occurred at these sites, the map would be performing as designed. Given this problem, 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; Nekrasova et al., 2014; Wyss et al., 2012; Mak et al., 2014). Although such assessments are not true tests, in that they compare the maps to data that were available when the map was made, they give useful insight into the maps’ performance.
Figure 1: a-d) 2008 version of probabilistic seismic hazard maps for Japan, generated for different return periods (J-SHIS, 2015). e) Map of largest known shaking on the Japan Meteorological Agency (JMA) intensity scale at each grid point for 510 years (Miyazawa and Mori, 2009).
MAP PERFORMANCE

We compared the 2008 version of the Japanese hazard maps to a catalog of shaking data for 1498-2007 (Miyazawa and Mori, 2009), giving the largest known shaking on the Japan Meteorological Agency (JMA) instrumental intensity scale at each grid point in 510 years (Fig. 1e). The observed data and predicted shaking maps cover essentially the same area, but with different resolutions. The predicted shaking maps have a 250 m x 250 m grid and the observed data had been interpolated to 1.7 km x 1.4 km spacing. Because our metrics call for an equal number of predictions and observations, we used ArcGIS to spatially join the two, assigning each observation to the appropriate grid cell.

The probability of exceedance equation predicts the probability for any given observation and return period. Figure 2b shows the predicted probability of exceedance, and thus the expected fraction of sites with maximum shaking above the mapped value, for 510 years of observation for each of the hazard maps in Fig. 1a-d. The predicted probability decreases with longer return period, because progressively rarer levels of shaking are less likely to occur. For example, \( p = 66\% \) of the sites are expected to have shaking higher than that predicted by the map with 475 year return period, whereas only 19\% are expected to be higher than predicted by the map with 2475 year return period.

However, as Fig. 2c shows, only \( f = 27\% \) of the sites plot above the 45° line for the map with 475 year return period. The remaining sites plot below the line, because the map predicted shaking higher than observed (Miyazawa and Mori, 2009). Similar discrepancies appear for the other maps with return periods of 101, 975, and 2475 years, all of which yield \( f < p \). We characterize this effect using a fractional site exceedance metric

\[
M_0(f, p) = |f - p|,
\]
As expected, both $p$ and $f$ decrease for longer return periods (Fig. 2d). Their difference $M0$ also decreases, showing that the map with the longest return period best characterizes the actual exceedance fraction.

A limitation of $M0$ is that a map with exceedances at exactly as many sites as predicted ($M0 = 0$) could still significantly overpredict or underpredict the magnitude of shaking. We thus also consider a squared misfit metric

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

where $x_i$ and $s_i$ are the maximum observed shaking and predicted shaking at each of the $N$ sites. Graphically, $M0$ reflects the fraction of sites plotting above the 45° line in Figure 2c, whereas $M1$ reflects how close to the line sites plot.

For the Japanese data, $M1$ behaves differently from $M0$, in that it increases with return period (Fig. 2d). $M1$ is smallest for the map with 101-year return period (Fig. 1a), consistent with the fact that this map is most visually similar to the data (Fig. 1e). Maps with longer return periods match the data less well, in part because they predict higher shaking than observed along the Japan Trench (e.g., 34°N, 135°E). This makes sense for the 975- and 2475-year maps, because the data span only 510 years, too short for some of the predicted largest shaking to have occurred (Fig. 3).
Figure 2: a) Assumed probability that during a $t$ year long observation period, shaking at a site will exceed a value that is expected on average once in a $T$ year return period. b) Predicted probability of exceedance, and thus the expected fraction of sites with maximum shaking above the mapped value, for data spanning a 510 year observation period and maps of different return period. The predicted probability decreases for longer return periods. Squares denote values for the hazard maps in Fig. 1a-d. c) Comparison of largest observed shaking at sites (Fig. 1e) to predictions of map with 475-year return period (Fig. 1b). d) Actual and predicted fractional exceedance for maps and data in Fig. 1, and corresponding map performance metrics.
Figure 3: Maps of the difference between observed and predicted shaking. The 475-, 975-, and 2475-year maps show a tendency to overpredict shaking, as shown by predominant red coverage.

Although ideally one might expect the map with return period 475 years to best match the 510 years of observation, that fact that it does not reflects the fact that the maps were made by using other data and models to try to predict future earthquake shaking, rather than by fitting the shaking data. In particular, the earthquake magnitudes assumed in the maps were inferred from the fault lengths (Fujiwara et al., 2009), rather than from
past intensity data. The maps were made with knowledge of past earthquakes, but were not tuned by fitting past shaking. Because the hazard map parameters were not chosen to specifically match the past intensity data, comparing the map and data can yield insight.

### Table 1:

<table>
<thead>
<tr>
<th></th>
<th>Return Time</th>
<th>1498-2007</th>
<th>1498-2011</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>M0</td>
<td>M1</td>
</tr>
<tr>
<td>Actual Maps</td>
<td>101 Years</td>
<td>0.40</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>475 Years</td>
<td>0.39</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>975 Years</td>
<td>0.22</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>2475 Years</td>
<td>0.07</td>
<td>0.63</td>
</tr>
<tr>
<td>Uniform Maps</td>
<td>101 Years</td>
<td>0.37</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>475 Years</td>
<td>0.30</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>975 Years</td>
<td>0.12</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>2475 Years</td>
<td>0.003</td>
<td>0.76</td>
</tr>
<tr>
<td>Random Maps</td>
<td>101 Years</td>
<td>0.41</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>475 Years</td>
<td>0.29</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>975 Years</td>
<td>0.11</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>2475 Years</td>
<td>0.03</td>
<td>0.97</td>
</tr>
</tbody>
</table>

Calculated metrics for actual, uniform, and randomized maps. Metrics were recalculated after adding 2011 Tohoku earthquake data to observed maximum shaking data to assess how the fit of the predicted shaking maps changed.

**UNIFORM AND RANDOM MAPS**

We generated uniform hazard maps from each of the four hazard maps by assigning each site the median hazard predicted by that map (Fig. 4). Surprisingly, the uniform maps yield lower values of the exceedance metric $M0$, showing a smaller
difference between the predicted and observed exceedance fractions than for the actual maps.

Figure 4: a) Uniform hazard map, with hazard at all sites set equal to median of corresponding map (Fig. 1c). b) Randomized hazard map, with hazard at sites randomly chosen from values in corresponding map (Fig. 1c). c,d) Performance metrics for applying the actual, uniform and randomized versions of the maps in Fig. 1a-d to data in Fig. 1e.
This effect can be visualized by considering that a uniform map shifts all points sidewise to lie on the vertical median line (Fig. 5). Most points stay either above or below the 45° line, and thus do not change \( f \), the fraction above the line. However, sites in the two triangular regions between the horizontal median line and the 45° line shift from being above to below or vice versa. Because more of these sites are below the 45° line (blue region) than above it (green region), \( f \) increases and \( M0 \) decreases.

![Figure 5: Illustration of how using the median predicted value for all sites can improve a hazard map's performance, as measured by the exceedance metric, if the map overpredicts the observed shaking.](image)

Similar results arise for randomized maps, in which site predictions are chosen at random from the actual predictions (Fig. 4) by giving an index to each point on the map, then shuffling the order of the indices, producing a different prediction at each point.

10,000 randomizations for each map yielded tightly clustered values of \( M0 \) and \( M1 \). The median results for the randomized maps are similar to those for the uniform maps, and thus generally better (lower \( M0 \)) than the actual maps.
However, using the squared misfit metric, the actual maps do better (lower $M1$) than uniform or randomized maps. This occurs because the actual maps better capture the spatial variations in the data than uniform or - even more so - randomized maps.

**INCORPORATING TOHOKU**

We augmented the dataset by adding intensity data from the 2011 Tohoku earthquake, the largest known earthquake in Japan, which occurred after the maps we used were made (Fig. 6a). These data were provided as 2,878 individual intensity measurements from different sites. As with the rest of the data, we used ArcGIS to spatially join this dataset to the prior dataset by selecting the maximum intensity observed.

Adding these data dramatically increases the maximum observed shaking along the east coast from about 35°- 38°N (Fig. 6b). We then repeated the analyses for the actual, uniform, and randomized maps. The exceedance metric $M0$ for each actual map decreased due to the higher shaking values but remained larger than for the uniform and randomized maps. Measured by the squared misfit metric $M1$, the actual maps still outperform uniform or randomized maps. Adding the Tohoku data improves the fit of the actual maps for the 975- and 2475-year return periods, because the predicted shaking for these long return periods is similar to that observed for Tohoku (Fig. 7).
Figure 6: a) Observed shaking in 2011 Tohoku earthquake. b) Historical shaking (1498-2007) map (Fig. 1e) updated with Tohoku data. c,d) Performance metrics for applying uniform and randomized versions of maps in Fig. 1a-d to updated data.
Figure 7: Maps of the difference between observed and predicted shaking with 2011 Tohoku earthquake data added. The increased shaking along the eastern coast reduces the extent of overprediction.

IMPLICATIONS

Our basic finding is that the Japanese hazard maps are not performing as well as might be hoped. Although this possibility was suggested by damaging earthquakes in areas mapped as low hazard, the overall bias seems to be the other way. The mapped
levels of shaking occur at a much lower fraction of sites than predicted, indicating that the maps systematically overpredict shaking and uniform or randomized maps do better from this perspective. However, the actual maps describe the observed shaking better than uniform or randomized maps. This complicated behavior illustrates the value of different metrics, in that $M_0$ is more sensitive to average shaking levels, whereas $M_1$ is more sensitive to spatial variations. It seems that although the Japanese maps are designed to predict shaking levels that should be exceeded a certain fraction of the time, the process by which their parameters are chosen tends to make the mapped shaking more closely resemble the maximum observed.

The observation that the actual maps do worse than uniform or randomized maps by one metric and better by another reflects the fact that a system's performance has multiple aspects. For example, how good a baseball player Babe Ruth was depends on the metric used. In many seasons Ruth led the league in both home runs and in the number of times he struck out. By one metric he did very well, and by another, very poorly.

More generally, how maps perform involves subtle effects. These results are for a particular area, much of which has a high earthquake hazard, and a particular set of maps and data. Although the misfit could be due to downward bias in the historical intensity data (Miyazawa and Mori, 2009), such data are expected to be biased toward higher values (Hough, 2013). The maps could be also biased upward, due to assumptions about the earthquake sources, the ground motion prediction equations, or conversions between the predicted shaking and intensity. Lowering the predicted shaking at all sites by a constant shift improves both $M_0$ and $M_1$ (Fig 8), although the actual misfit is spatially variable, as shown in Figs. 3 and 7. A similar improvement would result from raising the observed intensity values. These results suggest that hazard maps should be evaluated for consistency with what is known about past large earthquakes. Although historic intensity may have biases, hindcasts using them cover much longer time periods than will be practical for forecasts starting from the time a map is made. Situations like this, in which the hindcast does poorly, suggest possible problems that should be investigated.
Figure 8: Change in metrics as a result of applying a uniform shift to the maps' predictions. The 475-, 975-, and 2475-year maps all exhibit improvements for both the fractional exceedance and squared misfit metrics when predictions are decreased by a small amount. The 101-year map has very low predictions and an expected exceedance of 99.4%, which causes the metrics to behave differently from the others when a shift is applied.
Some of the Japanese results would likely apply to other areas, and some not. Presumably the greater the hazard variation within an area, the less likely a uniform or random map is to do better than a detailed map. Many questions need to be explored. In particular, it is important to find out whether better results are best obtained via better choices of parameters in the probabilistic approach (Stein and Friedrich, 2014) or by alternative deterministic approaches (Klugel et al, 2006; Wang, 2011; Peresan and Panza, 2012; Wang and Cobb, 2012).

Most crucially, these results indicate the need to know much more than we do about how well seismic hazard maps actually describe future shaking. Natural hazard forecasts do not be perfect – or even that good - to be useful in making policy (Stein and Stein, 2013; Field, 2015). However, the more we know about how much confidence to place in forecasts, the more effectively they can be used.

ACKNOWLEDGEMENTS

We thank the 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 Masatoshi Miyazawa and Takeo Ishibe for sharing their intensity data. Spencer thanks the Institute for Policy Research at Northwestern University for supporting his research.

REFERENCES


