Investigating the effects of smoothing on the performance of earthquake hazard maps

Edward M. Brooks* and Seth Stein

Department of Earth and Planetary Sciences,
Institute for Policy Research,
Northwestern University,
2145 Sheridan Road, Evanston IL,
60208-3130, USA
Email: eddie@earth.northwestern.edu
Email: seth@earth.northwestern.edu
*Corresponding author

Bruce D. Spencer

Department of Statistics and Institute for Policy Research,
Northwestern University,
Evanston, IL, 60208-4070, USA
Email: bspencer@northwestern.edu

Abstract: We explore whether less detailed probabilistic hazard maps might perform better by assessing how smoothing Japan's national earthquake hazard maps affects their fit to a 510-year record of shaking. As measured by the fractional exceedance metric implicit in such probabilistic hazard maps, simple smoothing over progressively larger areas improves the maps' performance such that in the limit a uniform map performs best. However, using the squared misfit between maximum observed shaking and that predicted as a metric, map performance improves up to a ~75–150 km smoothing window, and then decreases with further smoothing. This result suggests that the probabilistic hazard models and the resulting maps may be over-parameterized, in that including too high a level of detail to describe past and future earthquakes may lower the maps' ability to predict future shaking.

Keywords: earthquake hazard maps; probabilistic seismic hazard analysis; smoothing; metrics; Japan; earthquake; hindcasting; parameterization.

Reference to this paper should be made as follows: Brooks, E.M., Stein, S. and Spencer, B.D. (2017) 'Investigating the effects of smoothing on the performance of earthquake hazard maps', *Int. J. Earthquake and Impact Engineering*, Vol. 2, No. 2, pp.121–134.

Biographical notes: Edward M. Brooks is an Earth and Planetary Sciences PhD student at Northwestern University. His research is focused on assessing the performance of earthquake hazard maps and testing statistical models to describe historic seismicity rates. He is also currently pursuing a Master's degree in Statistics.

Seth Stein is a Deering Professor of Geological Sciences at Northwestern. He graduated from MIT in 1975 (BS) and Caltech (PhD) in 1978. His research interests are in plate tectonics, earthquake seismology, earthquake hazards and space geodesy. He has been awarded the James B. Macelwane Medal of the American Geophysical Union, the George Woollard Award of the Geological Society of America, the Stephan Mueller Medal of the European Geosciences Union, the Price Medal of the Royal Astronomical Society, and a Humboldt Foundation Research Award.

Bruce D. Spencer is a Professor of Statistics at Northwestern University. He graduated from Yale in 1979 with his PhD. He works actively with government agencies on major statistical programs and conducts related research in sampling theory and methods and in demographic estimates. For many years he has worked with the Census Bureau on how to estimate population and how to evaluate the accuracy of their estimates. He is a Faculty Fellow of Northwestern's Institute for Policy Research and a Fellow of the American Statistical Association.

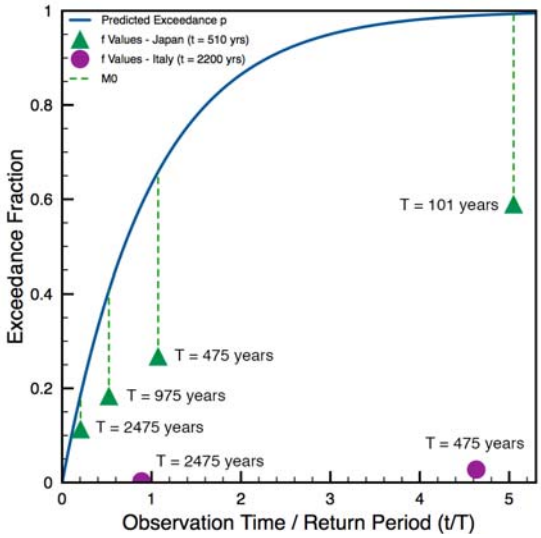
1 Introduction

Recent earthquakes that did great damage in areas shown by earthquake hazard maps as relatively safe have generated interest in the question of how well these maps forecast future shaking (Kerr, 2011; Reyners, 2011; Stein et al., 2011, 2012; Peresan and Panza, 2012; Stirling, 2012; Gulkan, 2013; Marzocchi and Jordan, 2014; Wang, 2015). These discussions have brought home the fact that although the maps are designed to achieve certain goals, we know little about how well they actually perform.

Commonly used probabilistic seismic hazard maps seek to predict the maximum shaking that should be exceeded only with a certain probability over a given time (Cornell, 1968; Field, 2010). At all points 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 described by an exponential distribution, $p = 1 - \exp(-t / T)$. This probability is small for t / T small and grows with observation time t [Figure 1(a)]. Hence the shaking predicted by a map with a T-year return period should have a 39% chance being exceeded in $t = T / 2$ years, a 63% chance being exceeded in $t = T$ years, and 86% in $t = 2T$ years.

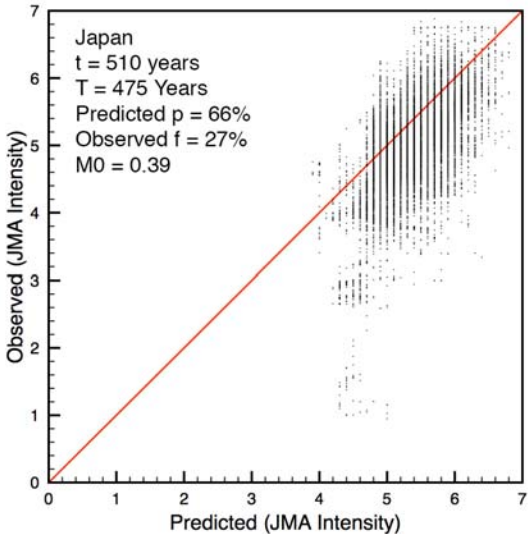
Shaking higher than shown on a probabilistic map often occurs in large earthquakes. Such shaking does not invalidate the map, so long as the fraction of sites at which this occurs is consistent with the map parameters. An alternative deterministic approach seeks to predict maximum values of shaking that will not be exceeded in a specified time period. Although the deterministic approach is not used in the maps we assess here, its predictions can also be compared to observations. It is worth noting that probabilistic and deterministic approaches are not incompatible; each seeks to predict different aspects of future shaking.

Figure 1(a) Probability of exceedance as a function of observation time divided by return period (see online version for colours)



Notes: The observed fraction of exceedances for probabilistic Japanese and Italian hazard maps (triangles and circles) are well below the predicted fraction (solid line) (Gruppo di Lavoro, 2004; Japanese Seismic Hazard Information Station, 2015). Green dashed lines indicate the difference between predicted and observed fractions, defined as the M0 metric.

Figure 1(b) Comparison between predicted and observed exceedances at individual sites for the 475-year Japanese map (see online version for colours)



Assessing how well maps describe actual shaking, relative to this ideal criterion, is challenging. Because the maps forecast the shaking expected over periods of hundreds or thousands of years, the short time period since they began to be made makes assessing

how well they perform difficult (Beauval et al., 2008, 2010). Hence maps can be assessed by comparing the fraction of sites where shaking exceeded the mapped threshold at that site to p. This approach, introduced by Ward (1995) and used in many subsequent analyses (e.g., Albarello and D'Amico, 2008; Fujiwara et al., 2009b; Stirling and Gerstenberger, 2010; Tasan et al., 2014; Nekrasova et al., 2014) considers many sites to avoid the difficulty that large motions at any given site are rare. Given this problem, various studies examine how well maps describe past shaking (Stirling and Petersen, 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.

2 Prior results

We have previously found (Brooks et al., 2016) that some hazard maps behave quite differently from the ideal. Figure 1(b) compares the largest known shaking at points within Japan in 510 years to that predicted by the Japanese National Hazard (JNH) map with a 475-year return period. We characterise the results by showing for each site the level of shaking (which we refer to as the predicted shaking) that has probability p of being exceeded during the return period. Although p = 66% of the sites are expected to have shaking higher than that predicted by the map with 475-year return period, only f = 27% of the sites plot above the 45° line that shows a 1:1 observed: predicted ratio. Similar discrepancies arise for the Japanese maps with other return periods, as shown by the triangles in Figure 1(a), all of which are below the expected exceedance curve. Similar but larger discrepancies arise in a similar analysis of 2200 years of data for Italy (Stein et al., 2015a). These discrepancies could reflect problems with the data, the maps, or both.

We use two metrics to quantify map performance (Stein et al., 2015a). The fractional exceedance metric.

$$M0(f, p) = | f - p |,$$

measures the magnitude of the difference between p, the expected number of sites at which the observed shaking should exceed that predicted by the map, and f, the actual number of such sites. Because this metric does not consider the magnitude of the difference between the predicted and actual shaking, we also assess maps with a squared misfit metric.

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

which compares x_i and s_i, the maximum observed and predicted shaking at each of the N sites. The two metrics characterise different aspects of map performance. In our previous paper, we found that $M0$ is sensitive to how well the map predicts average shaking levels, in that uniformly decreasing the predictions or increasing the observed shaking levels (to adjust for possible map or data biases) reduces the values of $M0$. $M1$ is more sensitive to how well a map predicts the spatial variations in shaking. Visually comparing maps of predicted and observed shaking amounts to using $M1$.

The $M0$ values show the maps' non-ideal behaviour. For the Japanese data and maps (Figure 1) $p > f$, so fewer sites than expected have shaking above that predicted. Moreover, although we might expect the map with return period of 475 years to best match the 510 years of observation (i.e., be closest to the curve), maps with longer return periods actually do better (f closer to p, hence lower $M0$). As discussed in Stein et al. (2015a) and this paper's Appendix, the effect of random error on $M0$ appears to be far too small to produce the large differences between p and f.

Given that some maps are behaving quite differently from how they should ideally (even after allowing for random error), we are exploring map behaviour to learn more about how they actually perform. We take an empirical approach of asking what maps actually do, rather than what they should ideally do. We similarly explore possible changes in the maps that could make their behaviour better match that expected.

Our earlier paper (Brooks et al., 2016) considered Geller's (2011) proposal 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", in which case maps less detailed than present ones would be preferable. We examined how well a 510-year-long record of earthquake shaking in Japan is described by the current JNH maps compared to uniform and randomised versions of these maps. We found that, as measured by the $M0$ metric, both uniform and randomised maps do better than the actual maps. However, using the squared misfit ($M1$) metric, the JNH maps do better than uniform or randomised maps. Similarly, by this metric, the 475-year map works better with 510 years of data than maps with longer return periods. Although $M1$ (unlike $M0$) does not explicitly depend on return period, $M1$ might be expected to be smallest (showing better fit) for maps with return period close to the data length.

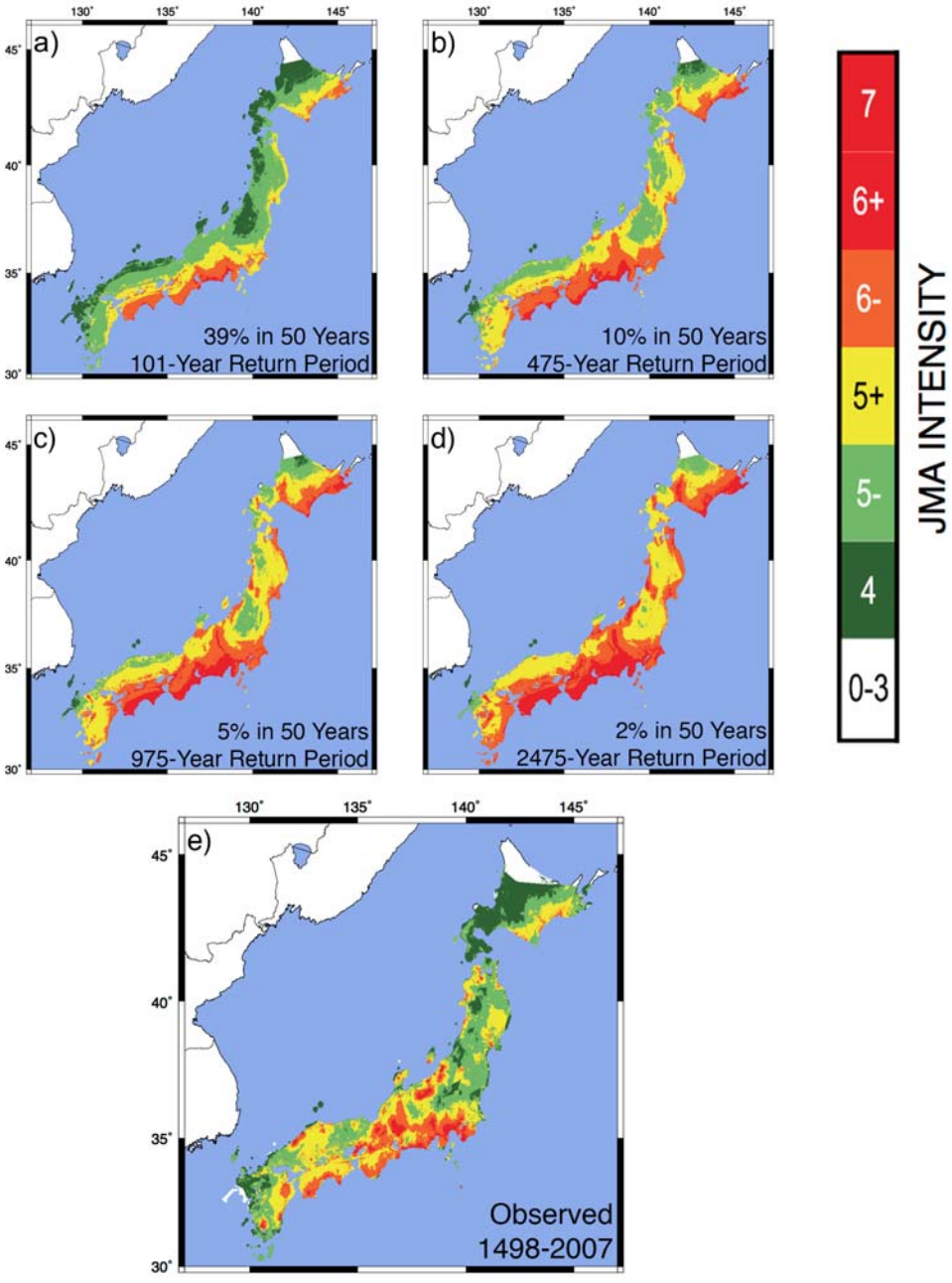
A uniform map is one smoothed (averaged) over the entire country, with all spatial details removed. Hence these results lead to the question of what the effect of smoothing over a smaller area may be. Is there some level of smoothing that preserves an intermediate level of detail that better describes the shaking?.

3 Smoothed map performance

As in our previous paper, we compared a catalog (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 (1498–2007) to four JNH maps for different return periods (J-SHIS, 2015) (Figure 2). The effect of site conditions is included in the maps so their predictions should be comparable to observations.

The JNH maps were smoothed by placing a square composed of cells over each point on the map, averaging the predictions within the square, and assigning that value to the central cell. Iterating over all points on the map using progressively larger squares yielded maps smoothed to greater degrees. For regions close to the coast we used only values on land in Japan, disregarding values from the surrounding ocean. This procedure preserves the number of points in each map, so successive iterations can be compared to the observed history of shaking via the two metrics. The smallest smoothing square was 3×3, and each individual cell was ~1.5 km on a side. Our smoothing procedure is quite simple, and improved variants that used shapes other than squares or rectangles might do even better.

Figure 2 (a–d) Probabilistic seismic hazard maps for Japan, generated for different return periods in 2008, (e) Largest known shaking on the JMA intensity scale in 510 years (see online version for colours)



Source: The Japanese hazard maps are from http://www.j-shis.bosai.go.jp/map/?lang=en (last accessed February 2015). The catalog of historic intensity data from Miyazawa and Mori (2009) was provided by M. Miyazawa

Smoothing over a small area preserves many details of the hazard maps, suppressing only the sharpest high and low hazard features. Progressively larger smoothing areas suppress more of the details (Figure 3). Figure 4 shows plots of the change in map performance as a function of smoothing area, for each of the four maps using both metrics.

The fractional exceedance metric ($M0$) generally improves as the smoothing area increases. Fluctuations are present for smaller smoothing areas, but performance increases steadily for smoothing areas above 200 cells (300 km on a side) across. This reinforces our earlier result, in that smoothing over all of Japan produces uniform maps, which we found perform better than the JNH maps as measured by $M0$.

In contrast, as measured by the squared misfit metric ($M1$), map performance improves somewhat up to a 50–100 cell (75–150 km) smoothing window, and then decreases with further smoothing. This reinforces our earlier result that by this metric uniform maps perform worse than the unsmoothed map. As discussed in the Appendix, the effect of random error on $M1$ is quite small, so the improved fit is significant.

Figure 3 Effects of smoothing the JNH map with 475-year return period (a) over progressively larger areas (b–d) (see online version for colours)

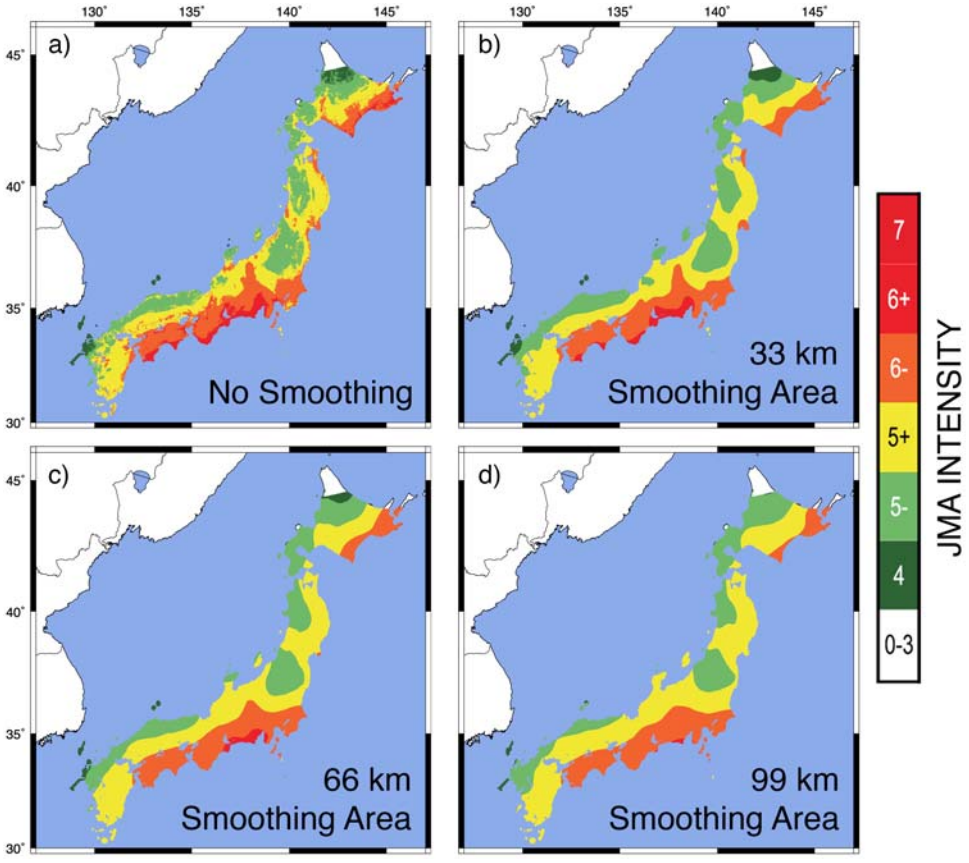
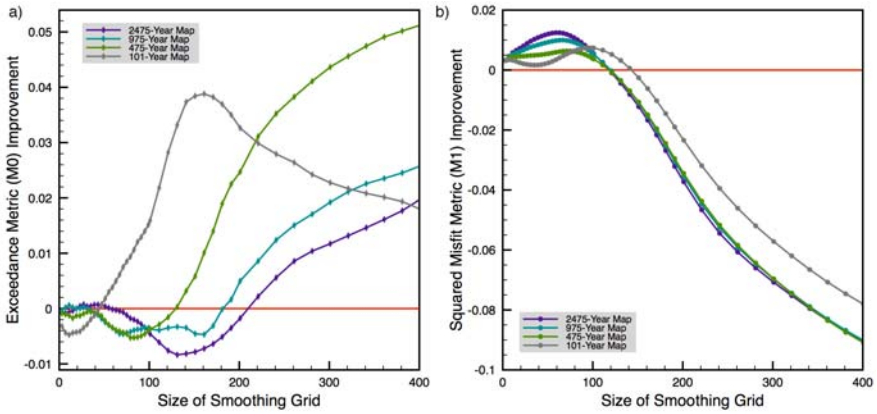


Figure 4 Improvement in map performance described by the change in fractional exceedance (a) squared misfit (b) metrics compared to the original map, for different amounts of smoothing (see online version for colours)



Note: Each cell of the grid is roughly 1.5 km on a side.

We repeated these comparisons for an updated map that incorporated shaking from the 2011 Tohoku event. Brooks et al. (2016) noted that adding these high shaking values improved the JNH maps' performance as measured by both metrics, but their performance relative to uniform and randomised maps remained the same. Similarly, we found that the effects of smoothing on performance remained essentially the same.

4 Implications

These results suggest that including too high a level of detail to describe past or future earthquakes may lower hazard maps' ability to predict future shaking. Such an effect seems plausible given the variability in space and time of earthquake recurrence, so previous earthquakes do not completely show what will happen in the future. Longer records including paleoseismic data, complemented with inferences from geological and geodetic data about faults, are naturally better. However, even a very long record is unlikely to fully capture the natural variability and uncertainty.

We would not expect a hazard map to perform perfectly. Aspects of future earthquake behaviour will differ from those of past earthquakes, the details of which are only partly known. Some of the assumed details of future earthquake behaviour will differ from what actually occurs. Hazard maps require a wide range of assumptions about earthquake source locations, recurrence, and magnitudes, along with models of the resulting ground motion.

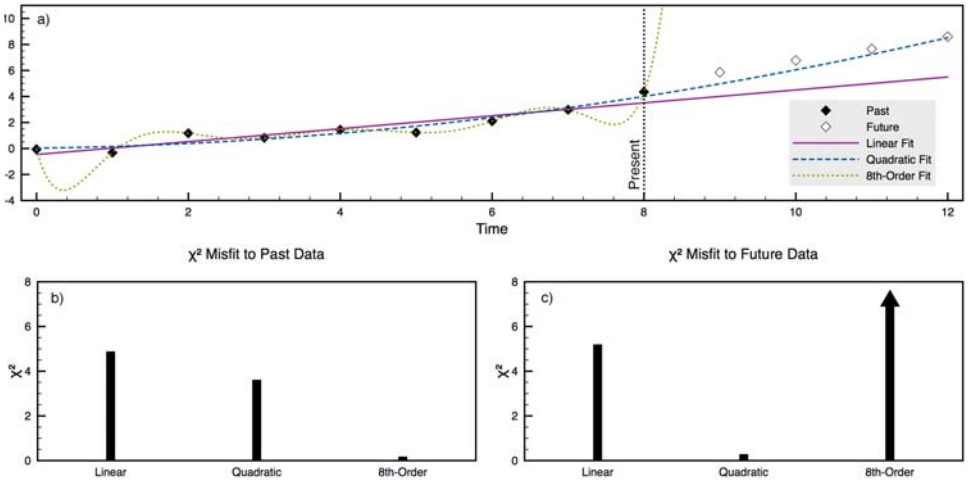
The classic resolution-stability tradeoff (Parker, 1977) tells us that more detailed a model is, the more sensitive it is to uncertainty, and thus the more likely it is to perform worse when assumptions fail. For example, prescribing a detailed rupture scenario will make a map's prediction for the future better if the earth does what is expected, but can make it worse than a simpler model if the earth fails to do what was expected – as in the Tohoku earthquake. Similarly a time-dependent rupture forecast will make a map better than a simple time-independent model if the earth does what is expected, but can make it

worse otherwise. Hence the challenge is to seek an optimal level of detail that incorporates and reflects uncertainties in the model and possible outcomes.

An analogous phenomenon is recognised in other applications and termed 'overfitting'or 'overparameterisation'. For example, given a set of observations at k distinct points in time, one can perfectly fit them with a curve based on k parameters, such as a polynomial of degree $k - 1$. However, a perfect fit to past data need not yield a good fit to future data. A variety of methods are available to trade off closeness of fit to observed data against the complexity of the model, including cross-validation and the Akaike information criterion (AIC) among others (Hastie et al., 2009). Figure 5 shows an example of using a model derived from past data to predict the future evolution of a function. A linear model fits the past data and predicts the future reasonably well, and a quadratic does both even better. However, an eight order polynomial that fits the past data perfectly does a poor job of predicting the future. The more detailed model seems better because it matches the past so well, but imposing that level of detail makes the model predict the future worse.

This situation is common in both geophysical and other forecasting applications. Hence to forecast the future, the goal should be not to build the most detailed model, but instead one that is robust or stable in the sense that small changes in the uncertain model parameters do not dramatically change the model's forecasts (Parker, 1977; Box, 1979).

Figure 5 Example of the effect of overparameterisation on forecasting (see online version for colours)



Notes: A high order polynomial fits past data better than linear or quadratic models, but this more detailed model predicts the future worse than the simpler models.

Our results showing an improved fit resulting from smoothing do, however, have other possible interpretations. First, the fact that the smoother models fit better could result from some features of the historical shaking dataset used. Second, our approach involves comparing a time-dependent hazard model to past data (hindcasting) rather than the more desirable comparison with future data (forecasting). As discussed in our earlier paper comparing these maps and data, we do not believe either problem is large enough to invalidate our approach. Most crucially, the maps were made by using other data and

models to try to predict future earthquake shaking, rather than by fitting past shaking data. In particular, the earthquake magnitudes assumed in the maps were inferred from the fault lengths (Fujiwara et al., 2009a), rather than from past intensity data. Because the hazard map parameters were not chosen to specifically match the past intensity data, comparing the map and data is a useful comparison.

These results are for a particular area, much of which has a high earthquake hazard, and a particular set of maps and data. However, these results, combined with the fact that in many applications overfitting past data leads to poorer future predictions, suggest that similar effects could arise for earthquake hazard maps elsewhere. Our approach involved smoothing maps resulting from a probabilistic hazard model. Hence it has similarities to the way certain hazard map input parameters are smoothed, which uses less detailed models to produce maps that should be more stable. For example, seismicity catalogs are often smoothed to compute seismicity rates (e.g., Cao et al., 1996; Montilla et al., 2003). Essentially our approach smooths the net effect of all inputs. Whether for inputs or outputs, it appears that smoothing may be valuable. It worthwhile exploring to find an appropriate level of model complexity to forecast future hazard (Field, 2015) in a way that is robust or stable in the sense that the forecast is not unduly affected when the earth does not behave exactly as expected. Whether to change a map after an earthquake yielding shaking larger than anticipated depends on whether one regards the high shaking as a low-probability event consistent with the map, or – as is often done – as indicating deficiencies in the map (Stein et al., 2015b).

References

Albarello, D. and D'Amico, V. (2008) 'Testing probabilistic seismic hazard estimates by comparison with observations: an example in Italy', *Geophys J. Int.*, Vol. 175, No. 3, pp.1088–1094.

Beauval, C., Bard, P-Y. and Douglas, J. (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.*, Vol. 100, No. 6, pp.3329–3331.

Beauval, C., Bard, P-Y., Hainzl, S. and Guéguen, P. (2008) 'Can strong motion observations be used to constrain probabilistic seismic hazard estimates?, *Bull. Seismol. Soc. Am.*, Vol. 98, No. 2, pp.509–520.

Box, G.E.P. (1979) 'Robustness in the strategy of scientific model building', *Robustness in Statistics*, Vol. 1, pp.201–236.

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

Cao, T., Petersen, M.D. and Reichle, M.S. (1996) 'Seismic hazard estimate from background seismicity in southern California', *Bull. Seismol. Soc. Am.*, Vol. 86, No. 5, pp.1372–1381.

Cornell, C.A. (1968) 'Engineering seismic risk analysis', *Bulletin of the Seismological Society of America*, Vol. 58, No. 5, pp.1583–1606.

Efron, B. and Tibshirani, R.J. (1993) *An Introduction to the Bootstrap*, Chapman & Hall, New York.

Field, E. (2010) *Probabilistic Seismic Hazard Analysis: A Primer* [online] http://www.opensha.org/ (accessed 1 May 2017).

Field, E.H. (2015) 'All models are wrong, but some are useful', *Seismol. Res. Lett.*, Vol. 86, No. 2A, pp.291–293.

Fujiwara, H. et al. (2009a) '*Technical Reports on National Seismic Hazard Maps for Japan*', Technical Note of the National Research Institute for Earth Science and Disaster Prevention, No. 336.

Fujiwara, H., Morikawa, N., Ishikawa, Y., Okumura, T., Miyakoshi, J.I., Nojima, N. and Fukushima, Y. (2009b) '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.*, Vol. 80, No. 3, pp.458–464.

Geller, R.J. (2011) 'Shake-up time for Japanese seismology', *Nature*, Vol. 472, No. 7344, pp.407–409.

Gruppo di Lavoro (2004) *Catalogo Parametrico dei Terremoti Italiani, Versione 2004 (CPTI04)*, Istituto Nazionale di Geofisica e Vulcanologia (INGV) [online] http://www.emidius.mi.ingv.it/CPTI04 (last accessed January 2014).

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

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer, New York.

Japanese Seismic Hazard Information Station (J-SHIS) (2015) [online] http://www.j-shis.bosai.go.jp/map/JSHIS2/download.html?lang=en (last accessed May 2015).

Kerr, R.A. (2011) 'Seismic crystal ball proving mostly cloudy around the world', *Science*, Vol. 332, No. 6032, pp.912–913.

Kossobokov, V.G. and Nekrasova, A.K. (2012) 'Global seismic hazard assessment program maps are erroneous', *Seismic Instruments*, Vol. 48, pp.162–170.

Mak, S., Clements, R.A. and Schorlemmer, D. (2014) 'The statistical power of testing probabilistic seismic-hazard assessments', *Seismol. Res. Lett.*, Vol. 85, No. 4, pp.781–783.

Marzocchi, W. and Jordan, T.H. (2014) 'Testing for ontological errors in probabilistic forecasting models of natural systems', *Proc. Natl. Acad. Sci.*, USA, Vol. 111, No. 33, pp.1973–11978.

Miyazawa, M. and Mori, J. (2009) 'Test of seismic hazard map from 500 years of recorded intensity data in Japan', *Bull. Seismol. Soc. Am.*, Vol. 99, No. 6, pp.3140–3149.

Montilla, J.A.P., Hamdache, M. and Casado, C.L. (2003) 'Seismic hazard in Northern Algeria using spatially smoothed seismicity', *Results for Peak Ground Acceleration Tectonophysics*, Vol. 372, No. 1, pp.105–119.

Nekrasova, A., Kossobokov, V., Peresan, A. and Magrin, A. (2014) 'The comparison of the NDSHA, PSHA seismic hazard maps and real seismicity for the Italian territory', *Natural Hazards*, Vol. 70, No. 1, pp.629–641.

Parker, R.L. (1997) 'Understanding inverse theory', *Annual Review of Earth and Planetary Sciences*, Vol. 5, No. 1, pp.35–64.

Peresan, A. and Panza, G.F. (2012) 'Improving earthquake hazard assessments in Italy: an alternative to 'Texas sharpshooting', *Eos., Transactions, American Geophysical Union*, Vol. 93, No. 51, p.538.

Reyners, M. (2011) 'Lessons from the destructive Mw 6.3 Christchurch, New Zealand, earthquake', *Seismol. Res. Lett.*, Vol. 82, No. 3, pp.371–372.

Stein, S., Geller, R.J. and Liu, M. (2012) 'Why earthquake hazard maps often fail and what to do about it', *Tectonophysics*, Vol. 562, No. 1, pp.1–25.

Stein, S., Geller, R.J. and Liu, M. (2011) 'Bad assumptions or bad luck: why earthquake hazard maps need objective testing', *Seismol. Res. Lett.*, Vol. 82, No. 5, pp.623–626.

Stein, S., Spencer, B.D. and Brooks, E.M. (2015a) 'Metrics for assessing earthquake hazard map performance', *Bull. Seismol. Soc. Am.*, Vol. 105, No. 4, pp.2160–2173, DOI: 10.1785/0120140164.

Stein, S., Spencer, B.D. and Brooks, E.M. (2015b) 'Bayes and BOGSAT: issues in when and how to revise earthquake hazard maps', *Seismol. Res. Lett.*, Vol. 86, No. 1, pp.6–10.

Stirling, M.W. (2012) 'Earthquake hazard maps and objective testing: the hazard mapper's point of view', *Seismol. Res. Lett.*, Vol. 83, No. 2, pp.231–232.

Stirling, M.W. and Gerstenberger, M. (2010) 'Ground motion-based testing of seismic hazard models in New Zealand', *Bull. Seismol. Soc. Am.*, Vol. 100, No. 4, pp.1407–1414.

Stirling, M.W. and Petersen, M. (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.*, Vol. 96, No. 6, pp.1978–1994.

Tasan, H., Beauval, C., Helmstetter, A., Sandikkaya, A. and Guéguen, P. (2014) 'Testing probabilistic seismic hazard estimates against accelerometric data in two countries: France and Turkey', *Geophysical Journal International*, Vol. 198, No. 3, pp.1554–1571.

Wang, Z. (2015) 'Predicting or forecasting earthquakes and the resulting ground motion hazards: a dilemma for earth scientists', *Seismol. Res. Lett.*, Vol. 86, No. 1, pp.1–5.

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

Wyss, M., Nekraskova, A. and Kossobokov, V. (2012) 'Errors in expected human losses due to incorrect seismic hazard estimates', *Natural Hazards*, Vol. 62, No. 3, pp.927–935.

Appendix

Effects of correlation and random error on M0 and M1 metrics

Both the predicted and observed shaking are spatially correlated. High shaking levels from large earthquakes are strongly correlated between nearby sites, as are low shaking levels at nearby sites where no close strong earthquakes took place during the study period). The difference $f - p$ between observed and forecasted exceedance equals the sum of a chance component, $f - Ef$, and a systematic or bias component, $Ef - p$. The expected square of the chance component equals the variance of f, say $V(f)$ The variance can be large when spatial correlation is high and the number of sites is moderate or small [Stein et al., (2015a), pp.2170–2172]. Spatial correlations do not affect the expected values of p and f, and thus the $M0$ metric. However, they affect the variance $V(f)$. If the variance is known, then an estimator of the squared bias is provided by the larger of $(f - p)^2 - V(f)$ and zero.

The effect of random error on the metrics $M0$ and $M1$ depends on the stochastic model assumed to describe the deviations $x_i - s_i$. If the predictions are taken to be fixed and not to depend on the observed shaking values, then the variance of the empirical fraction of exceedances f may be estimated by $f(1 - f) / n$, with n denoting the equivalent number of statistically independent sites after allowance for spatial correlations (Stein et al., 2015a). This model is overly simple, however, because at least some of the same observations that are used to develop the earthquake hazard maps are also used to compute the deviations. How to model spatial correlations and set a realistic value of n is an interesting and challenging problem that is beyond the scope of this paper.

For large enough values of n (depending on how far the expected value of f is from 0 or 1), f will have an approximately Gaussian distribution. If the distribution of f were exactly Gaussian, the expected value of $M0$ would be

$$E(M0) = \mu[1 - 2\Phi(-\mu / \sigma)] + 2\sigma\varphi(-\mu / \sigma),$$

with $\mu = E(f - p), V(f - p) = \sigma^2, \Phi(s) = \int_{-\infty}^{s} \varphi(x)dx$, and $\varphi(x) = (2\pi)^{-1/2} \exp(-x^2 / 2)$. If the bias μ is large relative to the standard deviation σ then $E(M0) \approx | \mu |$. On the other hand, if $E(f) = p$, so that $\mu = 0$, then $E(M0) = 2\sigma\varphi(0) \approx 0.8\sigma$, which tends to zero as n increases.

Example 1. Variance of M0: For the 475-year return period, the observed value of f was 0.27 compared to the specified probability of exceedance p of 0.66. For illustrative purposes, suppose the equivalent number of independent sites is 500. Then the estimated variance of the observed exceedance is (0.27)(.73) / 500 or 0.0004, and the estimated standard error is the square root of that, or 0.02. This is quite small relative to the value of $M0$ of 0.39. In addition, the bias in $M0$ is negligible, since we estimate $\Phi(-\mu / \sigma) \approx 1$ and $\varphi(-\mu / \sigma) \approx 0$.

The variance of $M1$ is given by:

$$V(M1) = \frac{(n-1)^2 v^4}{n^3}\left[\frac{n-1}{n}\beta - \frac{n-3}{n-1}\right]$$
$$\approx \frac{v^4}{n}[\beta - 1],$$

with v^2 the variance and β the kurtosis of the deviations $x_i - s_i$.

Example 2. Variance of M1: Consider the 475-year return period. Denote the deviations by $d_i = x_i - s_i$ The average deviation across the sites is $\bar{d} = -0.2722$. The average of $(d_i - \bar{d})^2$ is 0.2695 and the average of $(d_i - \bar{d})^4$ is 0.4557. For illustrative purposes, suppose the equivalent number of independent sites is 500. Then v^2 is estimated by 0.2695 and β is estimated by 0.4557 / 0.2695^2 or 6.274. We estimate $V(M1)$ by 0.000764 and we estimate the standard error of the $M1$ statistic by 0.028. The estimate of $M1$ was 0.34 (Brooks et al., 2016), and so the coefficient of variation or relative standard error was 8.1%.

Example 3. Variance of change in M1 due to smoothing: Consider the apparent improvement in $M1$ due to smoothing, again for the 475-year return period. Denote the unsmoothed predictions by x_i and the smoothed predictions by x_i'. Denote the corresponding deviations by $d_i = x_i - s_i$ and $d_i' = x_i' - s_i$. The corresponding values of $M1$ will be denoted by $M1_{\text{unsmoothed}}$ and $M1_{\text{smoothed}}$ The variance of the change in $M1$ due to smoothing, or $V(M1_{\text{unsmoothed}} - M1_{\text{smoothed}})$, equals $V(M1_{\text{unsmoothed}}) + V(M1_{\text{smoothed}}) - 2\rho\sqrt{V(M1_{\text{unsmoothed}})}\sqrt{V(M1_{\text{smoothed}})}$, with ρ denoting the correlation between $M1_{\text{unsmoothed}}$ and $M1_{\text{smoothed}}$. To estimate this, we use sample moments as in Example 2. The average deviation for the smoothed predictions across the sites is $\bar{d}' = 0.2687$. Define $\delta_i = d_i - \bar{d}$ and $\delta_i' = d_i' - \bar{d}'$. The average of δ_i^2 and the average of δ_i^4 are as in Example 2. The average of $\delta_i'^2$ is 0.2653 and the average of $\delta_i'^4$ is 0.4837. The same kind of calculations as carried out in Example 2 now yield the estimate of 0.000827 for $V(M1_{\text{smoothed}})$. To estimate the correlation, ρ we use $\sum_i \delta_i^2 \delta_i'^2 \Big/ \sqrt{\sum_i \delta_i^4 \sum_i \delta_i'^4}$. The

average of $\delta_i^2\delta_i'^2(d_i-\bar{d})^2(d_i'-\bar{d}')^2$ is 0.4537, which leads to an estimated correlation of 0.9664. The variance of the change in $M1$ is thus estimated by 0.00006 = 0.000764 + $0.000827-2(0.9664)\sqrt{0.000764}\sqrt{0.000827}$. The estimated standard error of the change in $M1$ is 0.0074, which is relatively small. It is important to note that this variance calculation is subject to the various limitations identified above. In addition, the variance as calculated does not take into account the randomness due to searching for the optimal smoothing. One way to carry out more realistic variance calculations would be to first model the spatial correlation structure and then to use an appropriate bootstrap procedure (Efron and Tibshirani, 1993).