NORTHWESTERN UNIVERSITY

Assessing the Performance of Earthquake Hazard Maps

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Earth and Planetary Sciences

By

Edward M. Brooks

EVANSTON, ILLINOIS

March 2019

ABSTRACT

Assessing the Performance of Earthquake Hazard Maps

Edward M. Brooks

Despite advancing knowledge about the mechanics of earthquakes, earthquake prediction remains, and will likely remain, an unsolved problem. Hence in order to reduce the risk posed by earthquake shaking, seismologists have developed tools called earthquake hazard maps. Earthquake hazard maps communicate expected future shaking scenarios, and are used by engineers to develop building codes and build safer structures to minimize their chance of failure or collapse in an eventual earthquake. However, 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 forecast shaking.

This thesis defines metrics that measure the effects of over-prediction and under-prediction by quantifying overall performance, making it easy and straightforward to compare the performance of different maps. These metrics consider different aspects of performance, including the probabilistic nature of some maps, and spatial variations of predictions. Although no single metric alone fully characterizes map behavior, using several metrics can provide useful insight for comparing and improving hazard maps.

I use these metrics to explore the performance of earthquake hazard maps made around the world. In Italy, I compare the performance of competing modeling procedures as a demonstration of how maps can be directly compared via the metrics. In Japan, I address criticisms of hazard mapping following the failure of the maps during the 2011 Tohoku earthquake by exploring the performance of simple, synthetic maps that describe uniform or randomized hazard. By the metric implicit in probabilistic seismic hazard assessment, the primary form of map-making worldwide, uniform maps are found to perform better than the original maps. This conclusion motivates additional research addressing over-parameterization, which finds that map performance may be improved by smoothing the predictions from the hazard models. In the United States, I assess the performance of one-year hazard maps made to describe the hazard from induced seismicity. The performance of the 2016 and 2017 maps prove to be better than any other map studied with the metrics I have defined, justifying the procedure for making short-term maps. Finally, I define a Monte Carlo simulation procedure to create artificial realizations of shaking scenarios and expand the utility of the metrics, outlining how they can be used as absolute, rather than relative, measures of performance.

Acknowledgements

There are so many people I would like to thank for all the help and support I have received over during my time at Northwestern. First, I owe many thanks to my graduate committee: Seth Stein, Bruce Spencer, and Donna Jurdy. Their support and guidance has been crucial to me. Additional thanks are owed to Bruce for his mentoring as I completed my master's degree in statistics, and to Seth for encouraging me to explore statistics, as well as any other career or research goal that intrigued me. Seth's staunch advocacy on behalf of his students is something very special.

There are a number of additional people in this department whose help should be acknowledged. I would not have gotten to where I am now without the assistance of Craig Bina, Matt Rossi, and Baudilio Tejerina during my qualifying exams. Thank you to all the more senior grad students who preparing me to pass that exam, in particular Trevor Bollmann and Emily Wolin. I am pleased to be a member of a research group with Leah Salditch and Jamie Neely, and thank them for their friendship and collaboration. That thanks must also be extended to my unofficial research partners, Amir Salaree and Nooshin Saloor. I will remain grateful for their support and tolerance with how frequently I barged into their offices with questions. I am privileged to have been in a cohort with Jamie McFarlin and Everett Lasher, and I thank them both for their companionship. Thanks to Trish Beddows for her guidance and keeping me level-headed when job hunting became too stressful.

Thank you to the Institute for Policy Research for their funding, granting me more time to focus on my studies. Thank you to Antonella Peresan for her guidance in the early stages of my work, and helping arrange my visit to Austria. Thank you to Mark Petersen and Dan McNamara for their outreach, collaboration, and for providing me the opportunity to work with the USGS.

Thank you to every friend who has met me for a beer (or three), who has seen a concert with me, gone with me to board game night, gone curling together, and made life throughout this journey so enjoyable. To my roommates Rachel, Dan, Daniel, Weston, and Deanna, I owe so much for all you did to make me look forward to returning home every night. To Steve, Alex, Matt, and Jack, thanks for the countless laughs. Thanks to Ben and Karen for always being there.

Special thanks must also be given to Daniel Berman for his willingness to help with every problem I encountered, academic and otherwise, even if geologic science was well outside his realm of expertise. I owe so many thanks to Lavanya Sivakumar, whose support has meant so much. Thank you for being there, for helping me plan what comes next, and inspiring me in so many ways.

Finally, on top of all else, I must thank my family. Thank you to the Kaish family for giving me such a great tie to Chicago, for wonderful meals, and for always being around. Thank you, Hannah, for being the best sister I could ask for, for talking and joking and just being fun. And finally, thank you to my parents. Without their encouragement, motivation, and guidance, none of what I have accomplished would have been possible. It means so very much to me. Thank you.

Contents

List of Tables

List of Figures

CHAPTER 1

Thesis Overview

1.1. Introduction

On March 11, 2011, a M_W 9.1 earthquake struck the eastern coast of Japan. This event was substantially larger than seismologists had thought was possible, because the region off Tohoku's shore was thought to have relatively low seismic hazard. Nonetheless, the tsunami that followed the shaking resulted in nearly 20,000 casualties. This event was just one of many in recent years that illustrate a systematic problem facing the seismological community: despite widespread acceptance and usage, we don't understand fully how well earthquake hazard maps perform. Similar surprisingly large shaking events, resulting in billions of dollars in damage and hundreds of thousands of casualties, have happened in Wenchuan, China (2008), Haiti (2010), and Nepal (2015).

The idea behind modern earthquake hazard maps dates back to 1968. Cornell (1968) outlined a procedure for probabilistically describing the shaking at a site within some time span. This method, Probabilistic Seismic Hazard Analysis (PSHA), is now widely used globally for inferring the seismic hazard posed to buildings and other structures, and hence the level of anti-seismic engineering necessary to guarantee safety. The procedure involves gathering as much information as possible about a number of earthquake scenarios, including magnitude, location, rate, and local site effects. Earthquake occurrence over time is widely considered to be a Poissonian random variable, and while this may or may not reflect reality, it is usually assumed in PSHA studies.

By combining all this information with a ground motion model that describes how earthquakes of a given magnitude will yield specific levels of ground shaking as a function of distance, seismologists develop hazard curves, which give the probability of shaking exceeding a certain value within some window of time (frequently fifty years). Probabilistic

hazard maps are an aggregation of these curves over many points, giving at each point on a map a level of shaking thought to have a given probability of being exceeded during the time window (Field, 2010).

By definition, PSHA allows for shaking larger than shown on a hazard map. Hence when a large exceedance occurs, the question is whether it is consistent with the map, or demonstrates an inaccuracy in the map. If an exceedance comes in the first year of a map with a fifty-year lifetime, does it indicate that we will see less shaking in the following forty-nine years? Map-makers, following these large events, tend to assume that the map was flawed in some way, and update the parameters used to make it. But which parameters need fixing? This is a complex question, further compounded by the sometimes subjective decisions made in selecting parameter values (Stein *et al.*, 2015a). Moreover, beyond the subjectivity in parameter selection, uncertainties in these parameters are not clearly communicated to the end-user, resulting in the confusion over how well a map should be expected to work.

This research involves a series of studies addressing the question of how well an earthquake hazard map works. It begins by defining metrics that can be used to quantify different aspects of map behaviors. The primary focus is on the probability of exceedance in the PSHA methodology, but additional attention is given to other aspects of performance, as well as the primary characteristics of the leading alternative to PSHA, Neo-Deterministic Seismic Hazard Analysis (NDSHA). This method reduces the use of probability, and instead seeks to describe the maximum possible shaking. Through this lens, a map's success or failure can be thought of more spatially than probabilistically. The difference between

these metrics, and some of the considerations one must make when assessing a map, are illustrated through an assessment of an Italian hazard map.

Following the development of metrics, this research explores a number of case studies, quantifying the performance of maps for Japan and the United States. Different types of data, including large-scale historic records, crowd-sourced shaking data, and computer simulations, are used to assess their performance. This reflects a growing understanding of both the metrics and how to use them, and lays the framework for how they could be applied them in the future to investigate what type of hazard maps are most useful for anticipating seismic hazard, and how they can be further improved.

1.2. Chapter 2: Metrics for Assessing the Performance of Earthquake Hazard Maps

In Chapter 2, I define metrics to quantify the performance of earthquake hazard maps. At the time of this work, there was no agreement as to how to answer the question of how well hazard maps work, in part because of how complex the models are.

Inspired by a more familiar hazard prediction analogue, weather forecasts, parallels can be drawn that illustrate how to assess how well a forecast predicts what occurs. This leads to a series of metrics, with a primary focus on two: the Fractional Exceedance Metric ($M0$) and the Squared Misfit Metric ($M1$).

I use these metrics in a case study to explore the performance of PSHA and NDSHA maps in Italy. This is the first of many map comparisons, and demonstrates a way of addressing a key question in the hazard mapping community: given multiple earthquake hazard maps, which map is the best?

This chapter has been published as Stein, S., Spencer, B. D., and Brooks, E. M. (2015). Metrics for assessing earthquake-hazard map performance, *Bulletin of the Seismological Society of America*, **105(4)**, 2160 - 2173.

1.3. Chapter 3: Comparing the Performance of Earthquake Hazard Maps to Uniform and Randomized Maps

Chapter 3 is motivated by criticisms levied against the PSHA methodology following the deadly 2011 Tohoku earthquake mentioned in the prior chapter. Tsunami countermeasures were so substantially overwhelmed, due to the same assumptions that made the map of predicted hazard so dramatically wrong. In a paper following the aftermath, Geller (2011) declared "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." If so, a map showing uniform hazard should be preferable to the existing map.

Assessing a uniform map's performance relative to the actual map is simple in practice, using the ideas I discussed in Chapter 2. The hazard maps from Japan in 2008, were compared to historic records of maximal shaking covering centuries before 2008, and then appended with post-Tohoku data. I use "hindcasting," comparing a map to data gathered from before the map was made, for this comparison to overcome the fact that so only a short time has elapsed since the map was made in comparison to its return period. Ultimately, whether a uniform hazard map is preferable to the "true" maps being used is "maybe," depending on the metric used.

This chapter has been published as Brooks, E. M., Stein, S., and Spencer, B. D. (2016). Comparing the performance of Japan's earthquake hazard maps to uniform and randomized maps, *Seismological Research Letters*, **87(1)**, 90-102.

1.4. Chapter 4: The Effects of Smoothing on the Performance of Earthquake Hazard Maps

Chapter 4 follows up the study in Chapter 3, where I explore maps between a uniform map and the original map. Because a uniform map has been completely smoothed across its entire surface area, I explore the performance of maps that have been smoothed over smaller regions. This analysis finds that by one metric, improvements to the performance of the Japanese National Hazard maps come if they are smoothed over $\sim$ 75 km. They run the risk of being over-fitted in their current form.

This chapter is published as Brooks, E. M., Stein, S., and Spencer, B. D. (2017). Investigating the effects of smoothing on the performance of earthquake hazard maps. *International Journal of Earthquake and Impact Engineering*, **2(2)**, 121-134.

1.5. Chapter 5: Earthquake Hazard Map Performance for Natural and Induced Seismicity

In Chapter 5, I apply the methods in Chapter 2 to a new data set and hazard map for the central and eastern United States (CEUS). The map, the 2016 one-percent in one year national seismic hazard map for natural and induced seismicity, was developed by the US Geological Survey. It addresses the recent dramatic increase in seismicity due to hydraulic fracturing ("fracking") and waste water injection associated with unconventional oil and natural gas production. Since 2008, these activities have increased the seismicity

in previously almost-aseismic Oklahoma and the surrounding areas to levels similar to that in California.

The shaking from induced earthquakes depends on human-driven actions, which cannot be treated as stable over time like the recurrence rates of naturally occurring earthquake. Hence the hazard map is designed to be used for a short time span, one year, rather than the traditional fifty years. This allows me to avoid using hindcasting for map assessment. I began this study began in early 2017, after an entire year of seismic shaking data had been gathered by the USGS "Did You Feel It?" (DYFI) system for the full time window over which the map was intended to be used.

The results show that the 2016 USGS map is one of the more successful maps analyzed by the metrics approach. Both on the national scale, and zoomed in on the region surrounding Oklahoma, the metrics suggest that this map performs better than any from prior studies. Furthermore, an in-depth analysis of population, location, and earthquake magnitude all agree that the map is unbiased, and quite successful in forecasting hazard.

This study was published as Brooks, E. M., Stein, S., Spencer, B. D., Salditch, L., Petersen, M. D., and McNamara, D. E. (2018). Assessing earthquake hazard map performance for natural and induced seismicity in the central and eastern United States. *Seismological Research Letters*, **89(1)**, 118-126.

1.6. Chapter 6: Assessing Map Performance Via Shaking Simulations

In the final chapter of this thesis, I explore future paths for better assessing hazard map performance. This chapter examines the performance of the 2017 one-percent in one year national seismic hazard map for natural and induced seismicity, the follow-up to

the prior year's map. The 2017 earthquake record differs substantially from the previous year, featuring substantially fewer events, and hence less shaking. Because the shaking is lessened, the response rate to DYFI also declines. The metrics used to assess map performance indicate a weaker performance compared to the previous year, though still better than the assessments from studies elsewhere discussed previous chapters.

The decrease in performance, and in responses in the DYFI data, prompt questions about how the likelihood that the reduced seismicity and shaking are consistent with the map. Could they, and thus the reduced performance, have occurred purely by chance? Assuming the parameters selected by the USGS for their hazard map accurately describe how the earth behaves, I explore how Monte Carlo simulation can be used to fill in the gaps in responses, explore the effects of uncertainty in the model, and address the question of how likely it is to observe the specific shaking scenario at occurred. This analysis finds that the lower shaking from 2017 is highly unlikely to have occurred by chance, indicating an issue in the selection of map parameters. The likely culprits, the decrease in fluid injection volume and hence seismicity due to regulatory and economic pressures, suggest important considerations to acknowledge and incorporate in future iterations of the map. This approach also lays the groundwork for future improvements in the metrics used to assess map performance by exploring how to assign scores to performance when there is only one map and thus no other maps for comparison, as was the case in Chapters 2 through 4. This improvement, turning the metrics from relative assessments to absolute assessments, is crucial for assessing whether an earthquake hazard map is "good."

Chapter 6 will be submitted shortly to *Seismological Research Letters* as "Assessments of the performance of the 2017 one-year seismic hazard forecast for the central and eastern

United States via simulated earthquake shaking data", by Brooks, E., M., Neely, J. S., Stein, S., Spencer, B. D., and Salditch, L.

CHAPTER 2

Metrics for Assessing the Performance of Earthquake Hazard Maps

2.1. Summary

Recent large earthquakes that did great damage in areas predicted to be relatively safe illustrate the importance of criteria to assess how well earthquake hazard maps used to develop codes for earthquake-resistant construction are actually performing. At present, there is no agreed way of assessing how well a map performed and thus 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 although they significantly under-predict or over-predict shaking, or to be nominally unsuccessful but do well in terms of predicting shaking. This chapter explores some possible metrics that better measure the effects of over-prediction and under-prediction 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 over-predict the recorded shaking in a 2200-year-long historical intensity catalog, illustrating problems in the data (most likely), models, or both.

2.2. Introduction

As Hurricane Irene threatened the U.S. East Coast in August 2011, meteorologist Kerry Emanuel (2011) 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 (type I errors), and sensitivity, lack of false negatives (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."

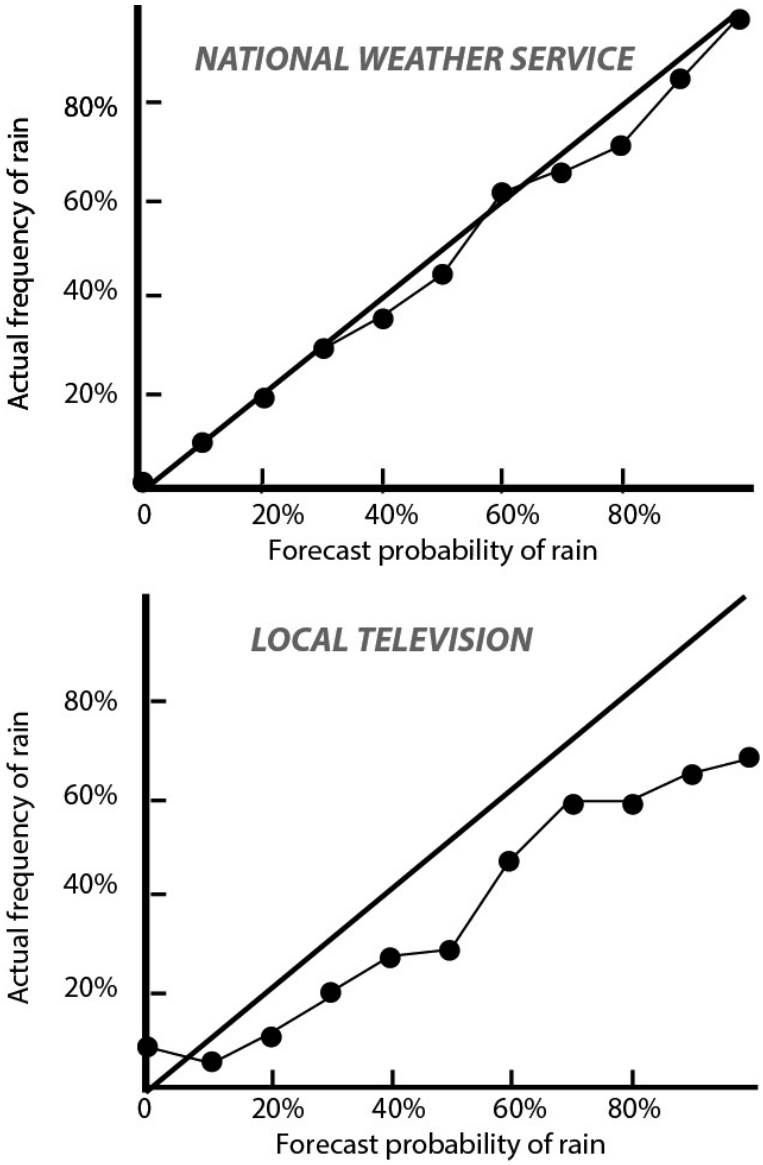


Figure 2.1. Comparison of the predicted probability of rain to that actually observed in National Weather Service and a local television station's forecasts (Silver, 2012).

Figure 2.1 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, whose 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 so 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 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).

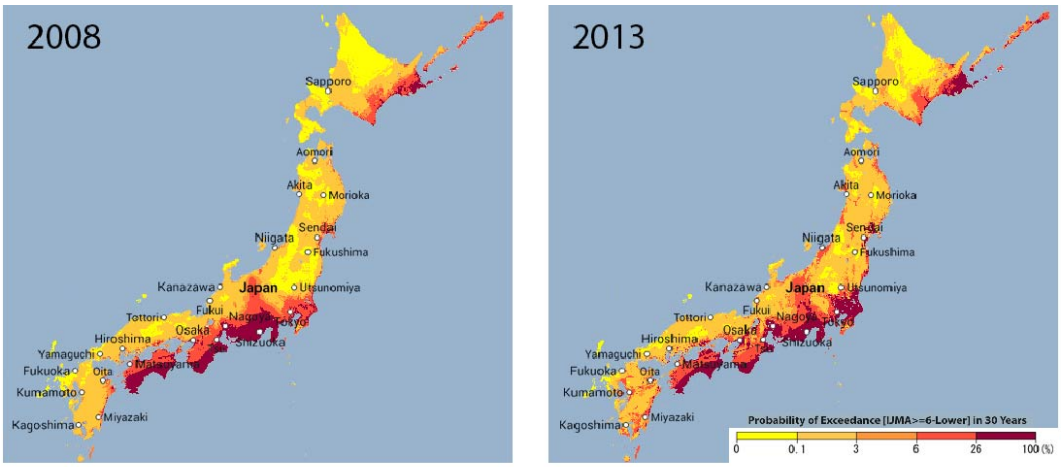
These events have catalyzed discussions among seismologists and earthquake engineers about commonly used earthquake hazard mapping practices (Kerr, 2011; Stirling, 2012; Gulkan, 2013; Iervolinoa, 2013). The underlying question is the extent to which the occurrence of low probability shaking indicates 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 (Figure 2.2) 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 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



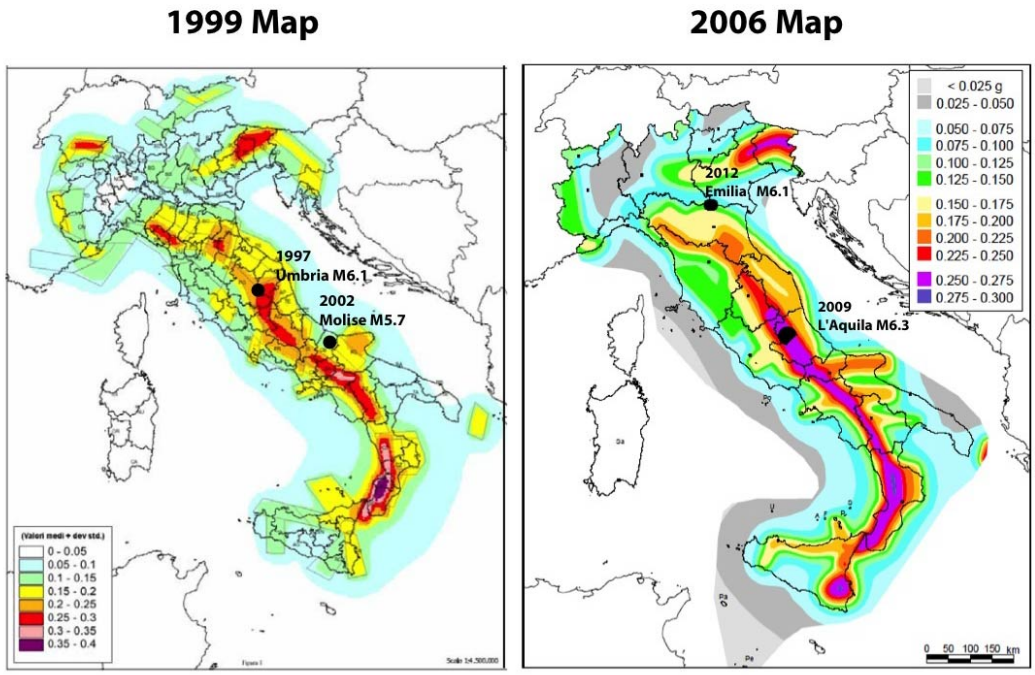


Figure 2.2. (Top) Comparison of Japanese national seismic hazard maps before and after the 2011 Tohoku earthquake. The predicted hazard has been increased both along the east coast, where the 2011 earthquake occurred, and on the west coast (J-SHIS, 2015). (Bottom) Comparison of successive Italian hazard 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 four heads as an unlikely event consistent with your probability model, and so not change it. But, 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., Siliva, 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 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.

In this chapter, I explore some possible metrics to address these issues. Although no single metric can fully characterize map behavior, examining map performance using several metrics can provide useful insight.

2.3. 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; Nekrasova *et al.*, 2014; Mak *et al.*, 2014) and are being developed under auspices of the Global Earthquake Model project.

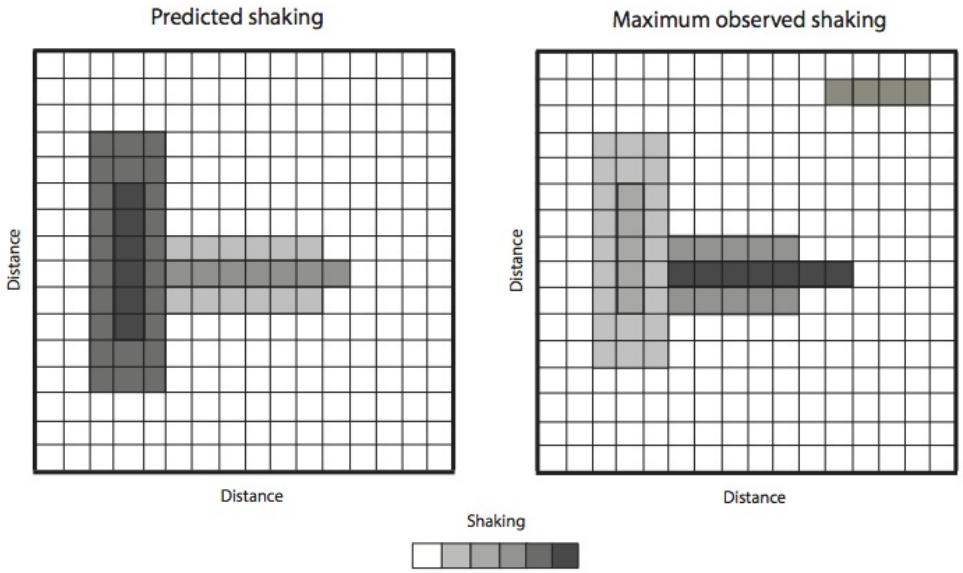


Figure 2.3. Schematic comparison of hazard map prediction to a map of the maximum observed shaking.

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 2.3, where for simplicity one can 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 over-predicted the shaking associated with the north-south striking fault, and under-predicted that associated with the associated 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 (Figure 2.4a) 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 the observation period τ and return period T, used. The Poisson (time-independent) probability p that earthquake shaking at a site will exceed some value in τ years, assuming this occurs on average every T years, is assumed to be

$$p = 1 - e^{-\frac{\tau}{T}}, \tag{2.1}$$

which is approximately τ/T for $\tau \ll T$. This probability is small for τ/T small and grows with time (Figure 5).

For a given return period, higher probabilities occur for longer observation periods. 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. Thus in 50 years 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 (Figure 2.1). In contrast, as discussed shortly, earthquake hazard maps are

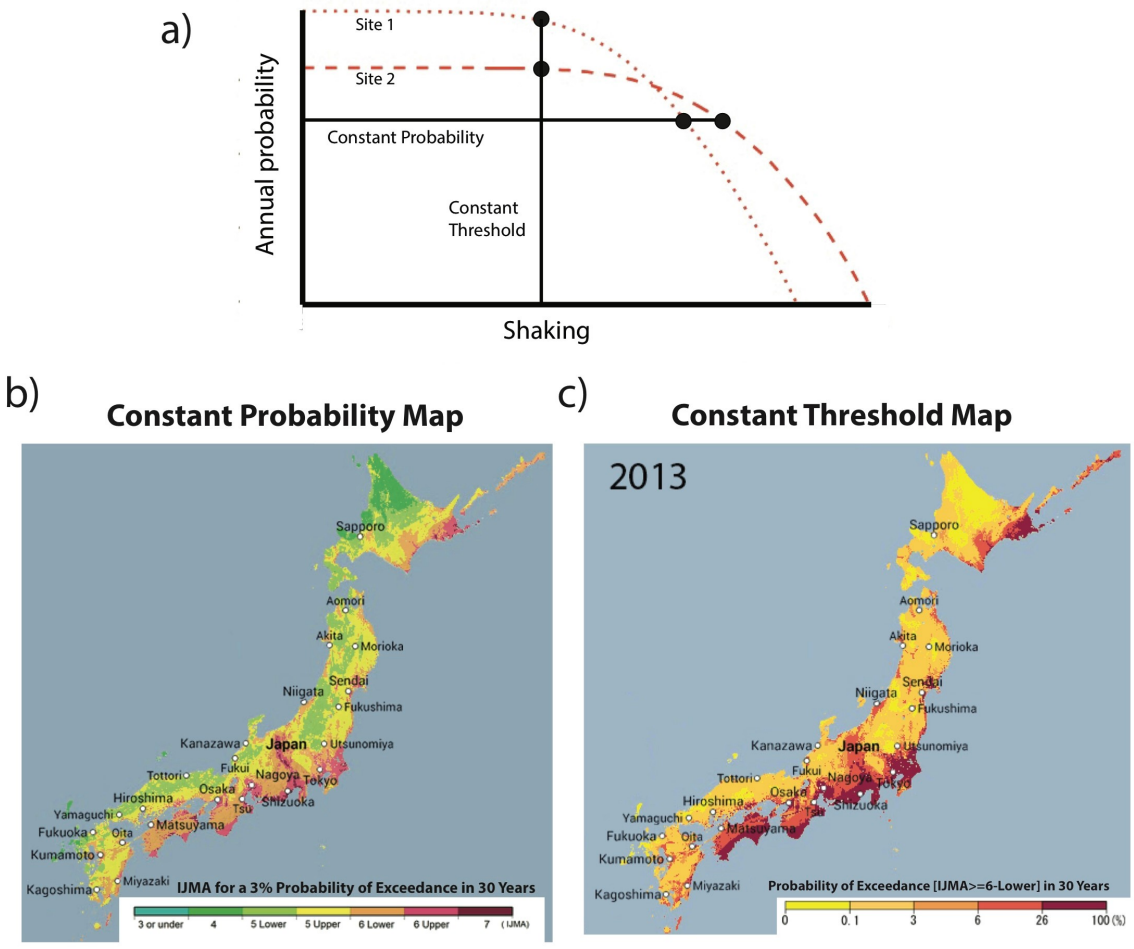


Figure 2.4. (a) Schematic hazard curves for two sites. Constant probability hazard maps like (b) are made by sampling the hazard curves at a fixed probability to predict that the largest shaking in each area will exceed a specific value with a certain probability during a certain time (observation period). 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.

tested by analysis of shaking across many areas to compensate for the short time periods of data available (Ward, 1995).

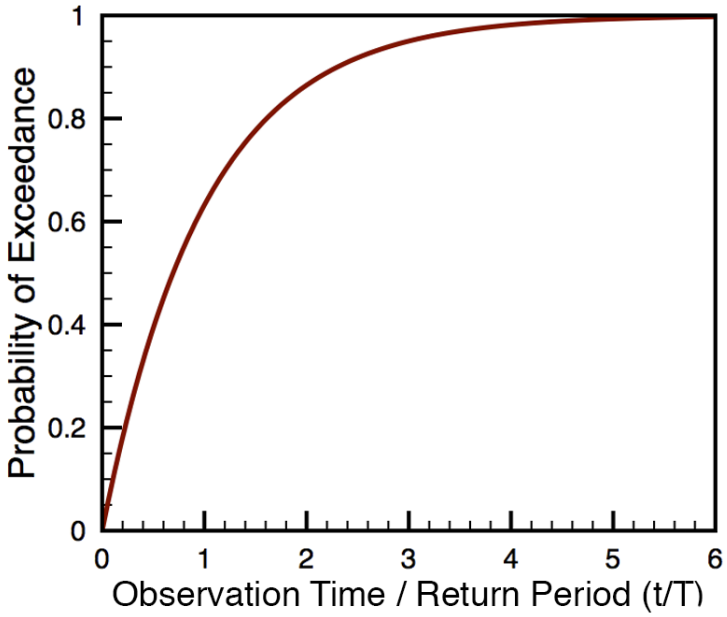


Figure 2.5. Assumed probability p that during a τ years long observation period, shaking at a site will exceed a value that is expected on average once in a T year return period, assuming equation 2.1, $p = 1 - e^{-\frac{\tau}{T}}$.

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 period. Thus the map shows the predicted shaking levels s_i for a given probability $p = P(x_i > s_i)$ for all areas i. For example, Figure 2.4b shows the shaking intensity on the Japan Meteorological Agency scale that is anticipated to have only a 3% chance of exceedance in 30 years. This approach, termed uniform hazard, is used in developing seismic design maps in the U.S. and Europe. An alternative is to present constant threshold hazard maps like that in Figure 2.4c. 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 $p_i = P(x_i > s)$ 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 years. Such maps show how the probability that a structure will be shaken at or above a certain threshold varies across locations.

2.4. Exceedance Metric, $M0$

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 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-year period, corresponding to a 2475-year return period. If the observation period is 250 years, Figure 2.5 shows that there is a 10% chance that the maximum shaking is as large or larger than predicted, and hence a 90% chance that it is less than predicted.

The longer the observation time compared to 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-year period a large earthquake produced shaking exceeding that predicted at 10% of the sites, this situation could imply that the map was not performing well. However, if in the subsequent 200 years 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 whose 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, $E[f]$, 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,

$$f - p = \underbrace{[f - E[f]]}_{\text{random component}} + \underbrace{[E[f] - p]}_{\text{systematic component}}. \tag{2.2}$$

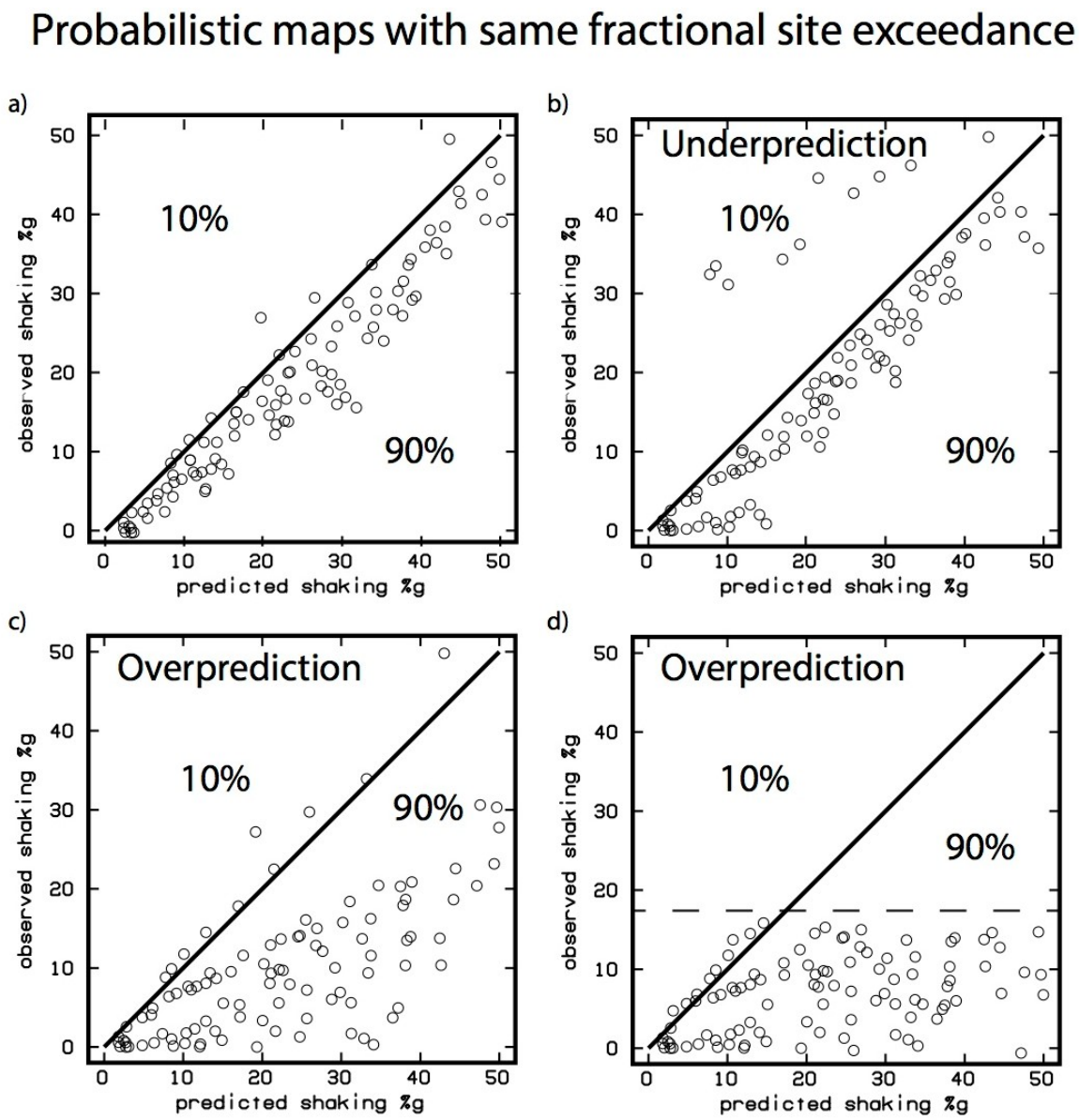


Figure 2.6. Comparison of the shaking predicted in various subregions of hazard maps to the maximum-observed shaking. Each of the four maps satisfies the fractional site exceedance criterion for $p = 0.1$, but (b)-(d) have significant under-predictions or over-predictions.

The systematic component is the difference between the average true probability (which equals $E[f]$) and the average predicted probability p of exceedance. If the map parameters do reasonably well in describing earthquakes in the area, $E[f]$ 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 chapter.

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^o line (Figure 2.6), 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 τ exceeds a shaking value s_i. This estimated probability can be written $p_i = P(x_i > s_i)$. For a sufficiently large number of areas, the fraction f of areas where $x_i > s_i$ should be approximately equal to the average probability for the areas, or $f \approx \bar{p}$, where $\bar{p} = \frac{1}{N} \sum_{i=1}^{N} p_i$. 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 them. This fractional site exceedance metric can be written as

$$M0 = |f - p|, \tag{2.3}$$

where 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 (Figure 2.5). $M0$ ranges from 0 to 1, with the ideal map having $M0 = 0$. If $M0 > 0$, then the map has either positive fractional site exceedance, measured by

$$M0^{+} = \begin{cases} |f - p| & f > p \\ 0 & \text{otherwise,} \end{cases} \tag{2.4}$$

or negative fractional site exceedance, measured by

$$M0^{-} = \begin{cases} |f - p| & f < p \\ 0 & \text{otherwise.} \end{cases} \tag{2.5}$$

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

2.4.1. Limitations of the Exceedance Metric

Although the exceedance metric is reasonable, it only reflects part of what one might want a probabilistic hazard map to do. This issue is illustrated by the results from four hypothetical probabilistic hazard maps (Figure 2.6), 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, or $M0 = 0$. However, some of these maps would be more useful than others.

The map giving rise to the results in Figure 2.6a would be viewed as highly effective, in that the maximum actual shaking plots close to that predicted. The map largely avoided under-prediction, which would have exposed structures built using a building code based on these predictions to greater-than-expected shaking. Similarly, it largely avoided over-prediction, which would have caused structures to be over-designed and thus waste resources.

Mathematically, largely avoiding under-prediction can be posed as saying that in the fN areas where $x_i > s_i$, the excess shaking $x_i - s_i$ should be modest. Similarly, largely avoiding over-prediction means that in the $(1 - f)N$ areas where $x_i < s_i$, the over-predictions should be modest. Maps can do well as measured by the fractional site exceedance metric, but have significant over-predictions or under-predictions.

For example, the map giving rise to the results in Figure 2.6b 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 Figures 2.6c and 2.6d over-predicted the shaking at most sites, although they have zero fractional site exceedance. Figure 2.6c shows a systematic bias toward higher-than-observed values, as could arise from using inaccurate equations to predict ground motion. The map for Figure 2.6d over-predicted the shaking in that the

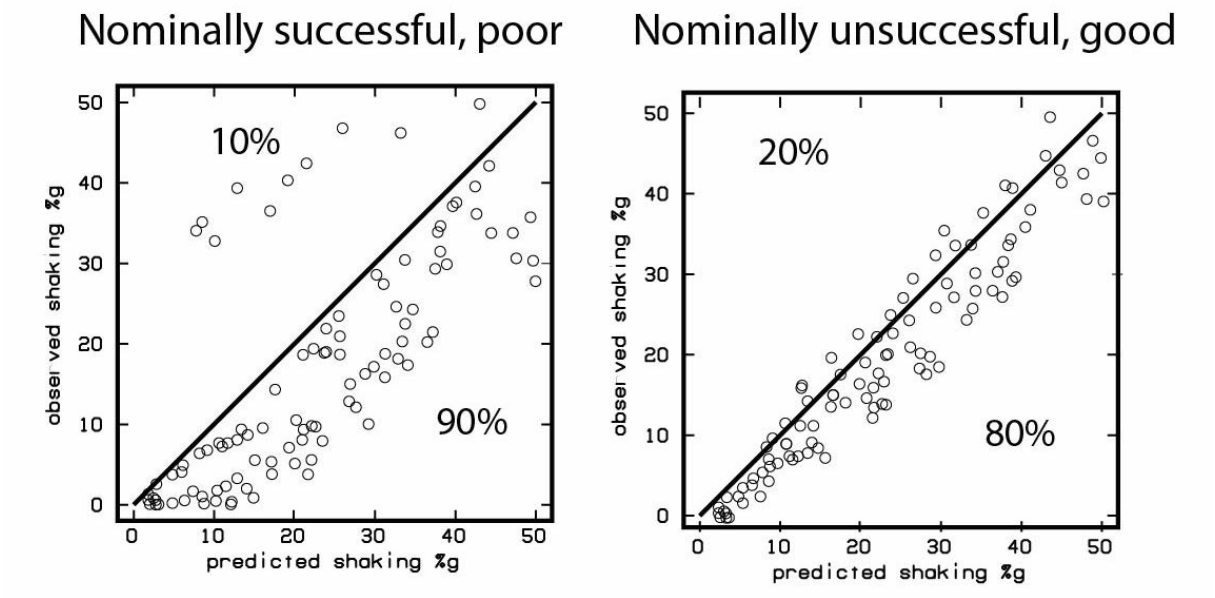


Figure 2.7. Comparison of the results of two hazard maps. The example in (a) is nominally successful, as measured by the fractional exceedance metric, but significantly under-predicts the shaking at many sites and over-predicts that at others. That in (b) is nominally unsuccessful, as measured by the fractional site exceedance metric, but better predicts the shaking at most sites.

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 2.7 for hazard maps in which the predicted shaking threshold for each site should be exceeded with probability 10% in the observation period. The map in Figure 2.7a is nominally very successful as measured by $M0 = 0$, but significantly under-predicts the shaking at many sites and over-predicts 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 $f = 0.2$, and $M0 = 0.1$.

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, i.e. 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.

2.5. Alternative Metrics

Many other metrics could be used to provide additional information for quantifying aspects of the observed vs. predicted graphs in Figures 2.6 and 2.7. As these additional metrics numerically summarize aspects of the graphs, they account for the length of the observation period. Consider four (Figure 2.8) 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 2.6 and 2.7. As these additional metrics numerically summarize aspects of , the hazard maps represented were constructed so that the shaking threshold for each site should be exceeded with probability 10% over the observation period.

One metric is simply the squared misfit to the data

$$M1(s,x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - s_i)^2 \tag{2.6}$$

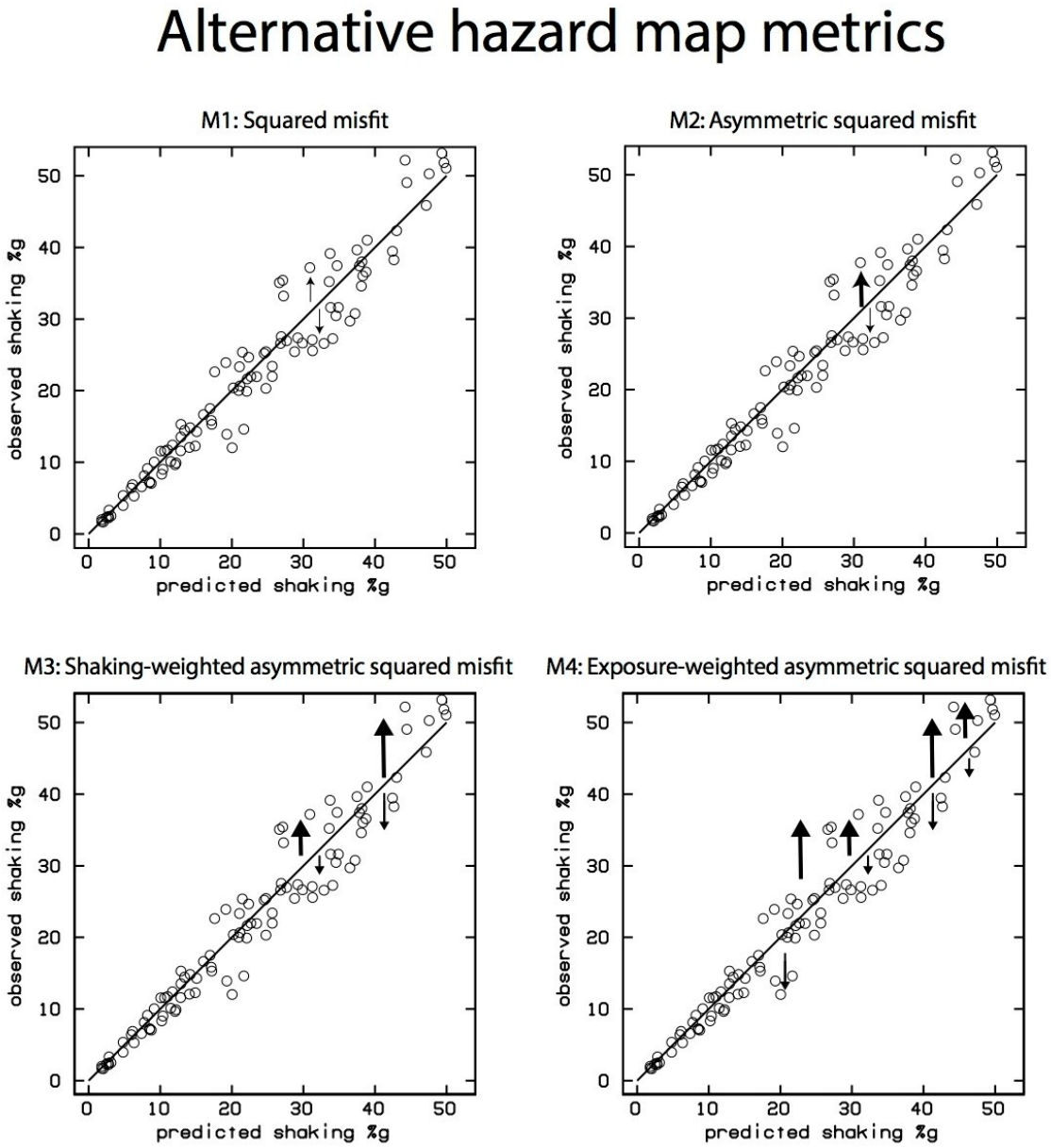


Figure 2.8. Four metrics ($M1$ through $M4$) that provide additional information beyond that from the fractional site exceedance metric.

which measures how well the predicted shaking compares to the highest observed. 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 (a)-(d) in Figure 2.6 have $M1 = 36$, 69, 253, and 370.

Similarly, by this metric, the map in Figure 2.7b ($M1 = 25$) does better than that in Figure 2.7a ($M1 = 135$). 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 one might also use metrics that weight different aspects of the prediction differently. For example, because under-prediction does potentially more harm than over-prediction, one solution could be to weight under-prediction more heavily. One such asymmetric metric is

$$M2(s,x) = \frac{1}{N}\sum_{i=1}^{N} a[(x_i - s_i)^+]^2 + b[(x_i - s_i)^-]^2 \tag{2.7}$$

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

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

$$M3(s,x) = \frac{1}{N}\sum_{i=1}^{N} a(s_i)[(x_i - s_i)^+]^2 + b(s_i)[(x_i - s_i)^-]^2 \tag{2.8}$$

where $a(s_i) > b(s_i) \geq 0$ and both a and b increase with s_i.

Another option is to vary the asymmetric weights a and b so that 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 a metric

$$M4(s,x) = \frac{1}{N}\sum_{i=1}^{N} a(e_i)[(x_i - s_i)^+]^2 + b(e_i)[(x_i - s_i)^-]^2 \tag{2.9}$$

where $a(e_i) > b(e_i) \geq 0$ 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.

2.6. 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, these would be used to evaluate how different maps of an area, made under different assumptions, actually performed. One 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; Nekrasova *et al.*, 2014; Wyss *et al.*, 2012; Mak *et al.*, 2014). Although such "hindcast" assessments are not true tests, in

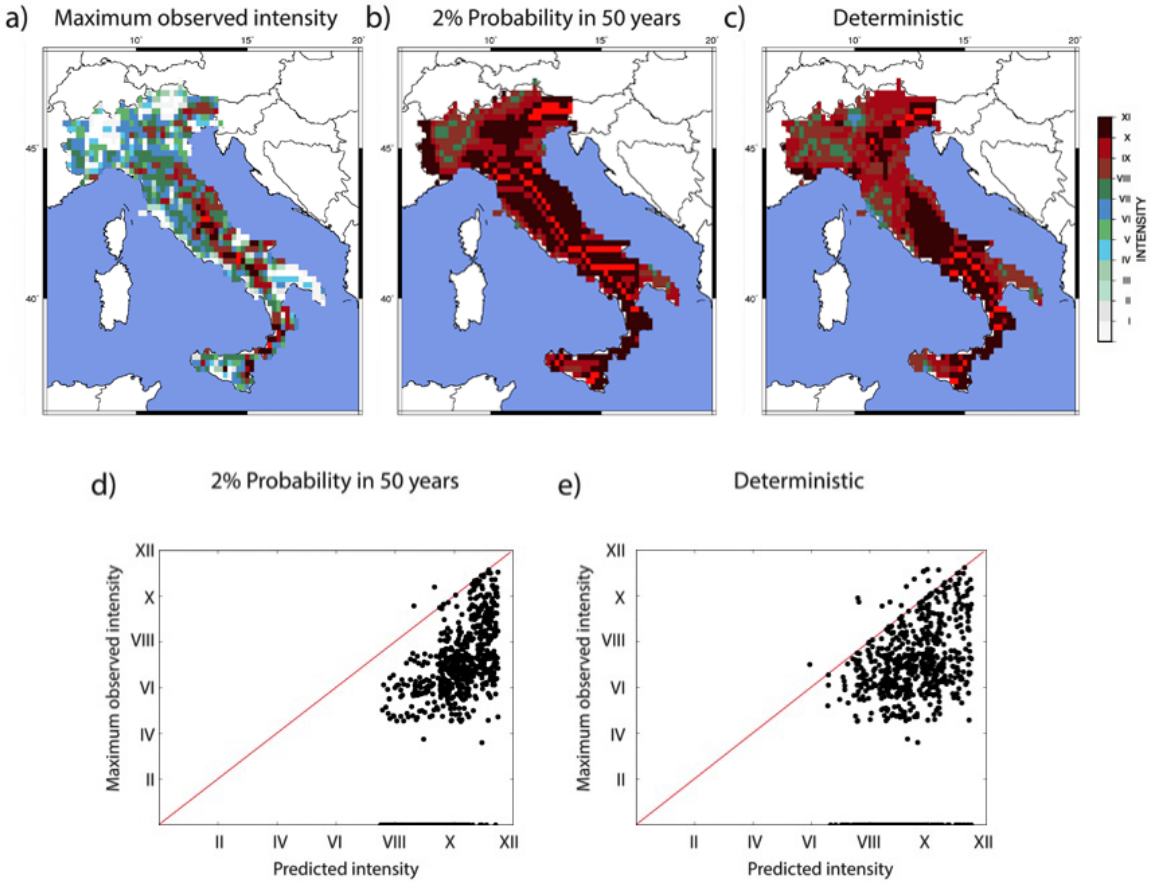


Figure 2.9. Comparison of (a) historical intensity data for Italy to (b) a probabilistic hazard map and (c) a deterministic hazard map, both of which over-predict the observed shaking, as shown in (d) and (e). Several points are moved slightly for clarity.

that they compare the maps to data that were available when the map was made, they give useful insight into the maps' performance.

For example, Figure 2.9a compares historical intensity data for Italy from 217 B.C. to 2002 A.D., developed from a compilation by Gruppo di Lavoro (2004), to a probabilistic map for 2% in 50 years and a deterministic map (Figures 2.9b and 2.9c) (Nekrasova *et al.*, 2014). As seen in Figure 2.5, this ~2200-year observation time and 2475-year return

period correspond to an exceedance probability $p = 58.89\%$. Hence the observed shaking at most sites should exceed that predicted.

However, the probabilistic map has only 2 sites out of 800 for which the observed shaking exceeding exceeds the threshold value, for $f = 0.25\%$. Comparing that with $p = 58.89\%$ there is thus a large negative fractional site exceedance, with $M0 = 0.5864$.

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. In principle, one can use the past performance to crudely calibrate the deterministic map, however. Thus, the empirical probability of exceedance for sites in Italy was 1.62% over 2200 years, corresponding to 2% over 2713 years, or 0.037% over 50 years. A similar approach has been used to calibrate deterministic scenario-based population forecasts (Keyfitz, 1981; Alho and Spencer, 2005). However, as discussed below, there are questions about the data so this example is purely illustrative.

Both hazard maps significantly over-predict the observed shaking, as shown by the $M1$ metric. The deterministic map does better ($M1 = 23.7$) than the probabilistic map ($M1 = 27.2$) because its overall over-prediction 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 years. The poor fit of both maps indicate a problem with the data, maps, or both. The metrics illustrate the problem, but do not indicate its cause.

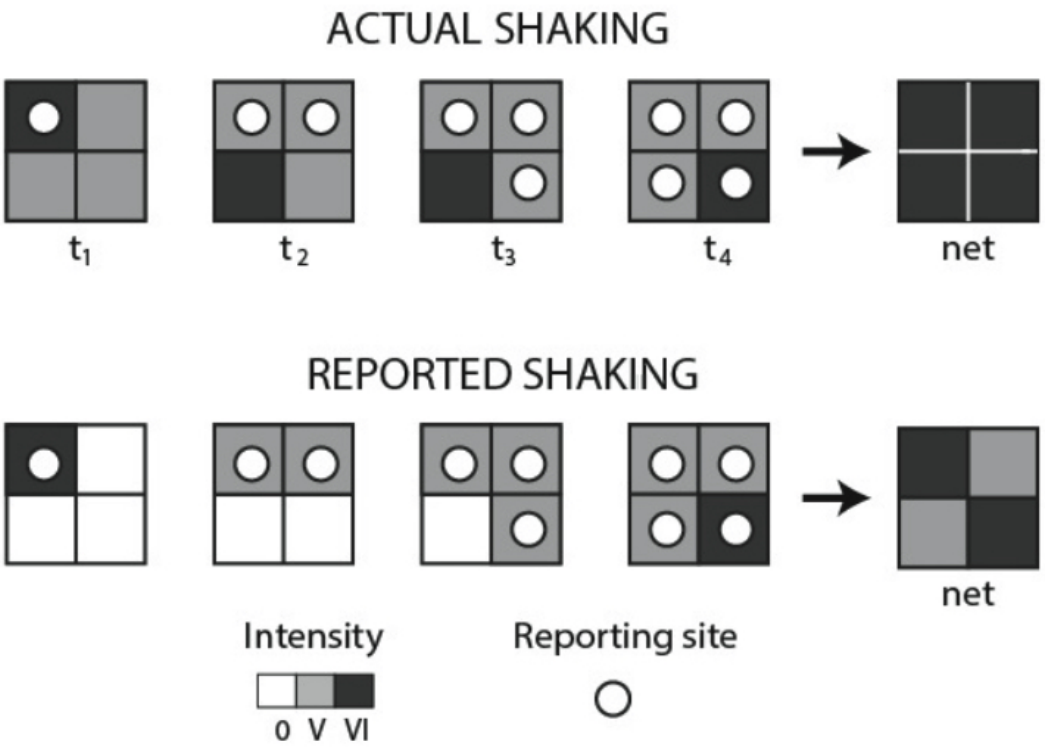


Figure 2.10. Schematic illustration of one way that variations in sampling over time τ could underestimate earthquake shaking. If reports are available only from grid cells including a reporting site (circle), the reported maximum shaking in some cells (lower row) is less than the actual maximum (upper row).

It is possible that some of the assumptions in the hazard map making were biased toward over-predictions. However, it is likely that much of the misfit results from 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 2.10 shows schematically how sampling bias could understate actual shaking, and Hough (2013) shows that sampling bias can also overestimate actual shaking.

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, $\sim$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 to maps produced after the earthquakes occurred. The metrics thus describe how well the maps fit 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.

2.7. The 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 these 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 from Equation 2.2 that for the exceedance metric, $M0$ the difference $f - p$ between observed and forecasted is the sum of the chance component, $f - [E]f$, and the bias $[E]f - p$. To interpret the difference, $f - p$, it's important 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, consider the probability distribution of f, the fraction of sites whose shaking exceeded the specified thresholds, for the Italy data used in Figure 2.9. Take the sites to be a population of interest, rather than a sample from a larger population. Consider only randomness associated with ground motion at each site. Figure 2.9b is a constant probability map, predicting that the probability is 2% that in 50 years shaking at a given site exceeds a threshold value for the site, and thus that in 2200 years 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, i.e., whether the difference is "statistically significant", the known limitations of hypothesis testing notwithstanding (Marzocchi *et al.*, 2012). For each site $i = 1, ..., N$, define $X_i = 1$ if shaking exceeded the threshold and $X_i = 0$ otherwise. Consistent with the model underlying the constant probability map, one can assume each X_i has a Bernoulli distribution with parameter p, i.e., $X_i = 1$ with probability p and $X_i = 0$ with probability $1-p$. If the X_i's 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 2 or smaller or 798 or larger is astronomically small, 1.7×10^{-179}. This probability is vastly smaller than the conventional $\alpha = 0.05$ level of significance, indicating that 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 co-variance 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 $\bar{\rho} = \sum_{i \neq j} \frac{\rho_{ij}}{N(N-1)}$. For example, if each X_i has correlation ρ with exactly k other X_j's and no correlation with all other X_j's, then $\bar{\rho} = \frac{\rho k}{N-1}$. To help interpret the correlation, consider $\rho_{ij} = \rho > 0$ 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 non-exceedance at a given pair of sites equals $(1-p)^2$. But, if the

correlation is $\rho > 0$ then the probability of non-exceedance at the pair of sites increases by $\rho p(1-p)$. Using $p = 0.589$ and, purely for illustrative purposes, taking $\rho = 0.36$, one sees that the probability of non-exceedance 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 chapter. 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 $z = \frac{Nf - Np + c}{\sqrt{Np(1-p)}}$ as if it were Gaussian with mean 0 and variance 1, where the "continuity correction" equals $\frac{1}{2}$ if $f < p$, $-\frac{1}{2}$ if $f > p$, and 0 if $f = p$. With $Np = 471.2$, $Nf = 2$, and $N = 800$, one can calculate $z = -33.7$, which as before (with the binomial model) corresponds to an astronomically small probability. Now, suppose for illustrative purposes that $\bar{\rho} = 0.38$, as discussed in the previous paragraph. To take correlation into account, divide the z-value of -33.7 by $\sqrt{Np(1-p)} = 16.99$ 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 $\alpha = 0.05$. 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 $\alpha = 0.05$. In that case, the difference between Nf and Np would not be "statistically significant" at the $\alpha = 0.05$ 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 that the mean squared deviation between f and p equals the sum of the variance in f and the squared bias in p, that is, $E[f-p]^2 = V(f) + \text{Bias}^2$.

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

$$\hat{p} = (f-p)^2 - V(f). \tag{2.10}$$

For example, for the 2%-in-50-years model with correlation, it is possible to estimate $V(f) = \frac{f(1-f)}{N}[1 + (N-1)\bar{\rho}] = 0.0071$, which does not assume that 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).

2.8. 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

$$SS(s,r,x) = 1 - \frac{M(s,x)}{M(r,x)} \tag{2.11}$$

where 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 negative if they did worse. With this information, it is then possible to 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 that 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, 2012) and can be modeled as the same, including the possibility of an $M_W 9$ earthquake (there is about one every 20 years somewhere on a trench).

The idea that a map including 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 over-parameterization or over-fitting (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 (Figure 2.2). 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 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, 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.

CHAPTER 3

Comparing the Performance of Earthquake Hazard Maps to Uniform and Randomized Maps

3.1. Summary

The devastating 2011 M_w 9.1 Tohoku earthquake and the resulting shaking and tsunami were much larger than anticipated in earthquake hazard maps. Geller (2011) has thus 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. I 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 Japanese national hazard (JNH) maps, uniform maps, and randomized maps. Surprisingly, as measured by the metric implicit in the JNH 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 as a metric the squared misfit between maximum observed shaking and that predicted, the JNH maps do better than uniform or randomized maps. These results indicate that the JNH 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.

3.2. Introduction

Probabilistic seismic hazard maps (Figure 3.1) 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 τ years of observations shaking will exceed a value that is expected once in a T year return period is assumed to be described by Equation 2.1, a Poisson distribution

$$p = 1 - e^{-\frac{\tau}{T}}.$$

This probability is small for τ/T small and grows with observation time τ (Figure 3.2). For example, shaking with a 475-year return period has about a 10% chance being exceeded in 50 years, 41% in 250 years, 65% in 500 years, and 88% in 1000 years.

The assumption that shaking values are described by a Poisson distribution is commonly used for maps in which the earthquake recurrence is assumed to be described by a Poisson process, so the probability of an earthquake of a certain size on a fault is time independent. In the Japanese maps, the probability of earthquake recurrence is modeled on some of the faults as varying with time, whereas that for other faults is modeled as time-independent. The shaking record reflects contributions from many faults, and when the observation period starts and ends is independent of the histories of earthquakes on these faults. Because just the number of exceedances within the observation window are of interest, when within this window the earthquakes occurred has no effect on performance measures. Hence it is straightforward to compare the observed shaking values to those expected from the Poisson distribution.

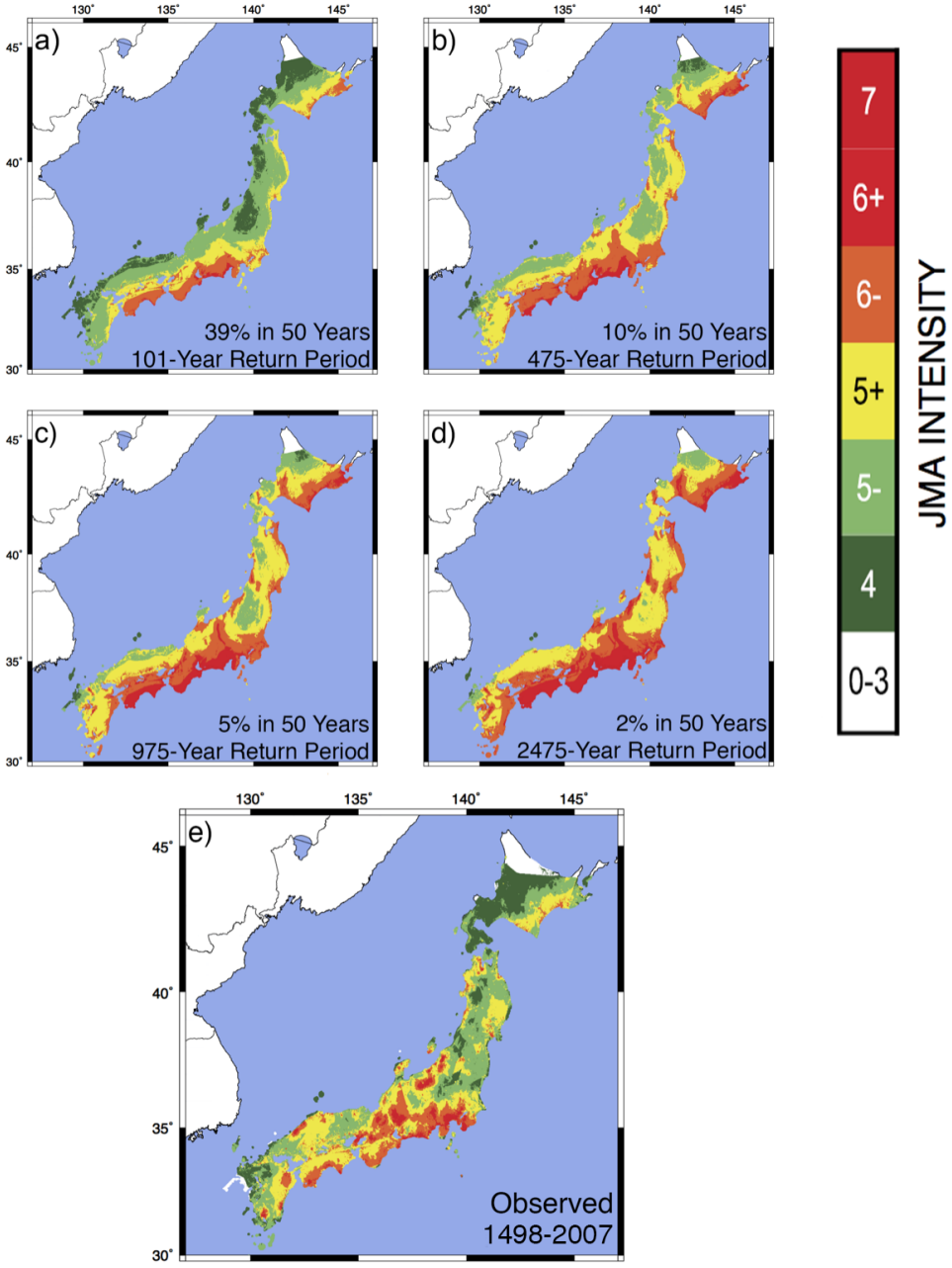


Figure 3.1. (a-d) The 2008 version of probabilistic seismic-hazard maps for Japan, generated for different return periods (J-SHIS, 2015). (e) The largest known shaking on the Japan Meteorological Agency (JMA) intensity scale at each grid point for 510 yrs (Miyazawa and Mori, 2009).

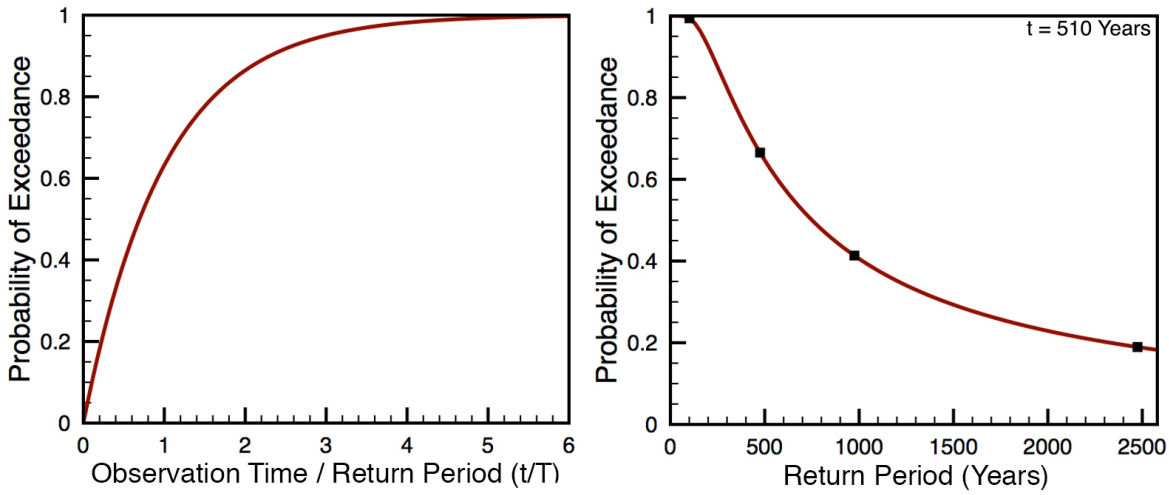


Figure 3.2. (Left) The assumed probability that, during a τ-year-long observation period, shaking at a site will exceed a value that is expected on average once in a T-year-long return period. (Right) Predicted probability of exceedance, and thus the expected fraction of sites with maximum shaking above 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 Figure 3.1.

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. The different maps are forecasts derived from a hazard model whose parameters describe the locations, magnitudes, and probabilities of future earthquakes and the resulting shaking.

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 period 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 10%-in-50 year 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.

3.3. Japanese National Hazard Map Performance

To test hazard map performance, I draw a comparison between the 2008 version of the Japanese National Hazard (JNH) 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 (Figure 3.1). The observed data and JNH maps cover essentially the same area, but with different resolutions. The JNH maps have a 250 m × 250 m grid and the observed data had been interpolated to roughly 1.7 km × 1.4 km ($2^{-6} \times 2^{-6}$ deg^2) spacing (Miyazawa and Mori, 2009). Because the metrics call for an equal number of predictions and observations,

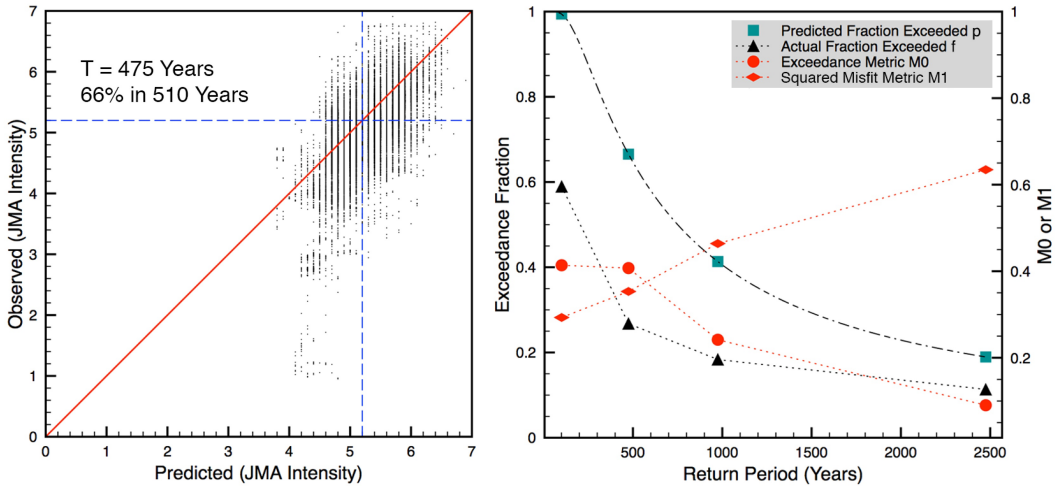


Figure 3.3. (Left) The comparison of largest observed shaking at sites (Figure 3.1e) to predictions of Japanese National Hazard (JNH) map with the 475-year return period (Figure 3.1b). (Right) Actual and predicted fractional exceedance for JNH maps and data in Figure 3.1, and corresponding map performance metrics.

ArcGIS was used to spatially join the two, decimating the JNH data to match the distribution and spacing of the observation data. The effect of site conditions is included in both the predictions and observations, making the two comparable. The observed shaking data are effectively continuous, whereas the JNH maps are discrete to one decimal place, resulting in the discretization seen in the left panel of Figure 3.3.

Although the JNH maps do not state how their performance should be evaluated, assessment of their performance can be completed using two metrics I described in Chapter 2. The first, the fractional exceedance metric, $M0$, is based on the probability of exceedance equation that predicts the probability for any given observation and return period. Figure 3.2 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 JNH maps in Figure 3.1. 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 Figure 3.3 shows, only $f = 27\%$ of the sites plot above the 45^o line (showing a 1:1 observed : predicted ratio) for the JNH 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 JNH maps with return periods of 101, 975, and 2475 years, all of which yield $f < p$. This is the effect characterized by the fractional exceedance metric, $M0$, described in Chapter 2 by Equation 2.3,

$$M0 = |f - p|.$$

As expected, both p and f decrease for longer return periods (Figure 3.3, right panel). Their difference, $M0$, also decreases, showing that the map with the longest return period best characterizes the actual exceedance fraction.

As discussed in Chapter 2, a limitation of $M0$ is that a map with exceedances at exactly as many sites as predicted ($M0 = 0$, Figure 2.6) could still significantly over-predict or under-predict the magnitude of shaking. Thus, it is useful also consider a squared misfit metric, described earlier by Equation 2.6,

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

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^o line, 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 (Figure 3.3). $M1$ is smallest for the map with 101-year return period (Figure 3.1a), consistent with the fact that this map is most visually similar to the data (Figure 3.1e). 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^oN, 135^oE). 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. This difference is mapped in Figure 3.4.

Although, ideally, one might expect the map with return period of 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.*, 2009a), 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.

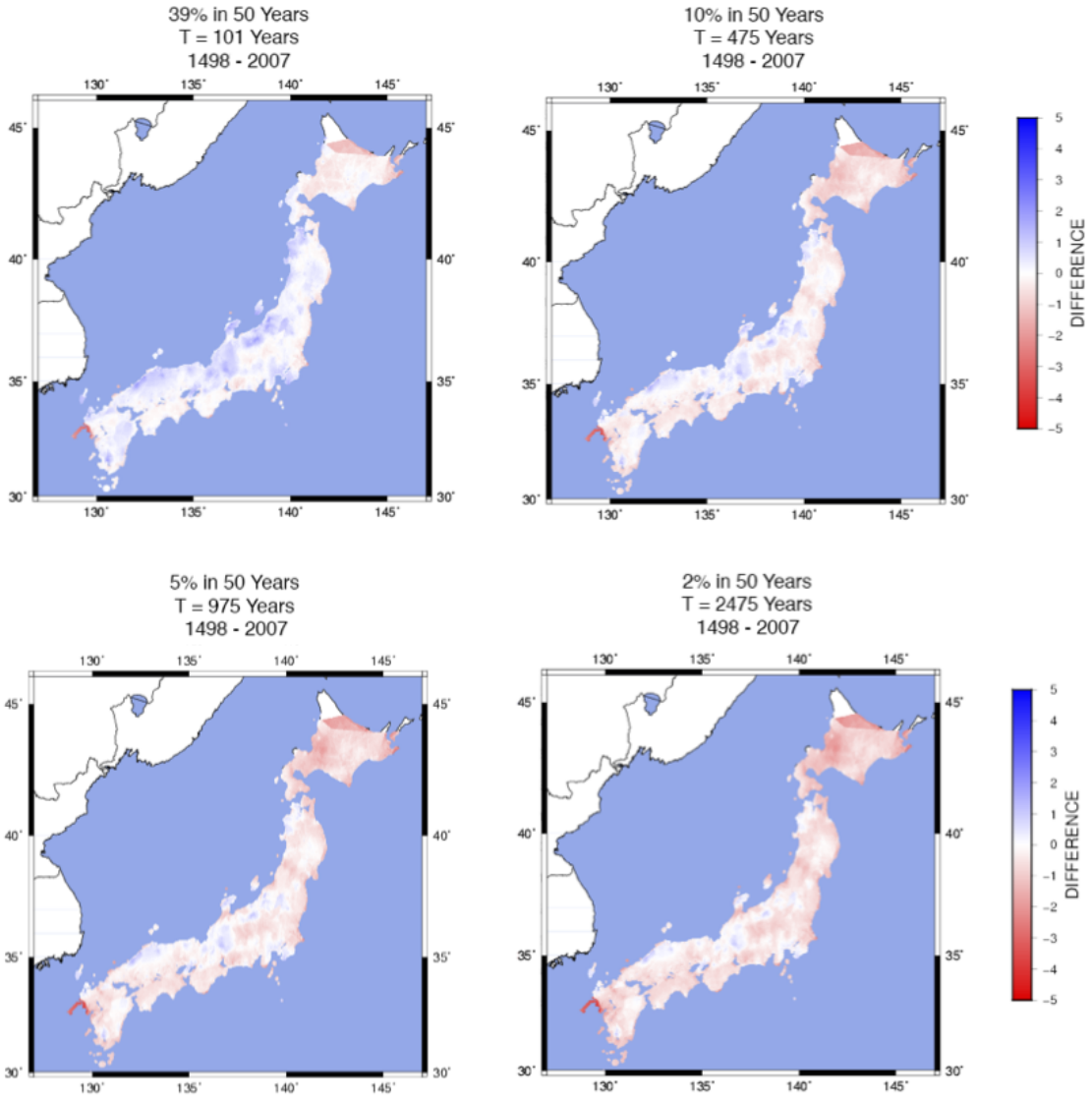


Figure 3.4. The difference between maximum observed and predicted shaking. The 475, 975, and 2475 year JNH maps tend to over-predict shaking, as shown by the predominant red coverage.

3.4. Uniform and Random Maps

Generating uniform hazard maps from each of the four JNH hazard maps is simply done by assigning each site the median hazard predicted by that map (Figure 3.5). 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.

How this effect arises can be visualized by considering that a uniform map shifts all points sideways to lie on the vertical median line, as shown in Figure 3.6. Most points stay either above or below the 45^o 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^o line shift from being above to below or vice versa. Because more of these sites are below the 45^o line (blue region) than above it (red region), f increases and $M0$ decreases.

Similar results arise for randomized maps, also shown in Figure 3.5, in which site predictions are chosen at random from the actual JNH map by giving an index to each point on the JNH map, then shuffling the order of the indices, producing a randomized map with the same median and other statistical properties but a different prediction at each point.

Ten thousand randomizations for each map yielded tightly clustered values of $M0$ and $M1$. For a randomized map, given a set of s predictions and x observations, define f to be $P(x > s)$. This is equivalent to observing a point lies above the schematic 45^o line. Randomizing a map is analogous to taking all (s, x) pairs, fixing x and randomly assigning a new s from the set of all predictions. Solving for $P(x > s)$ given totally

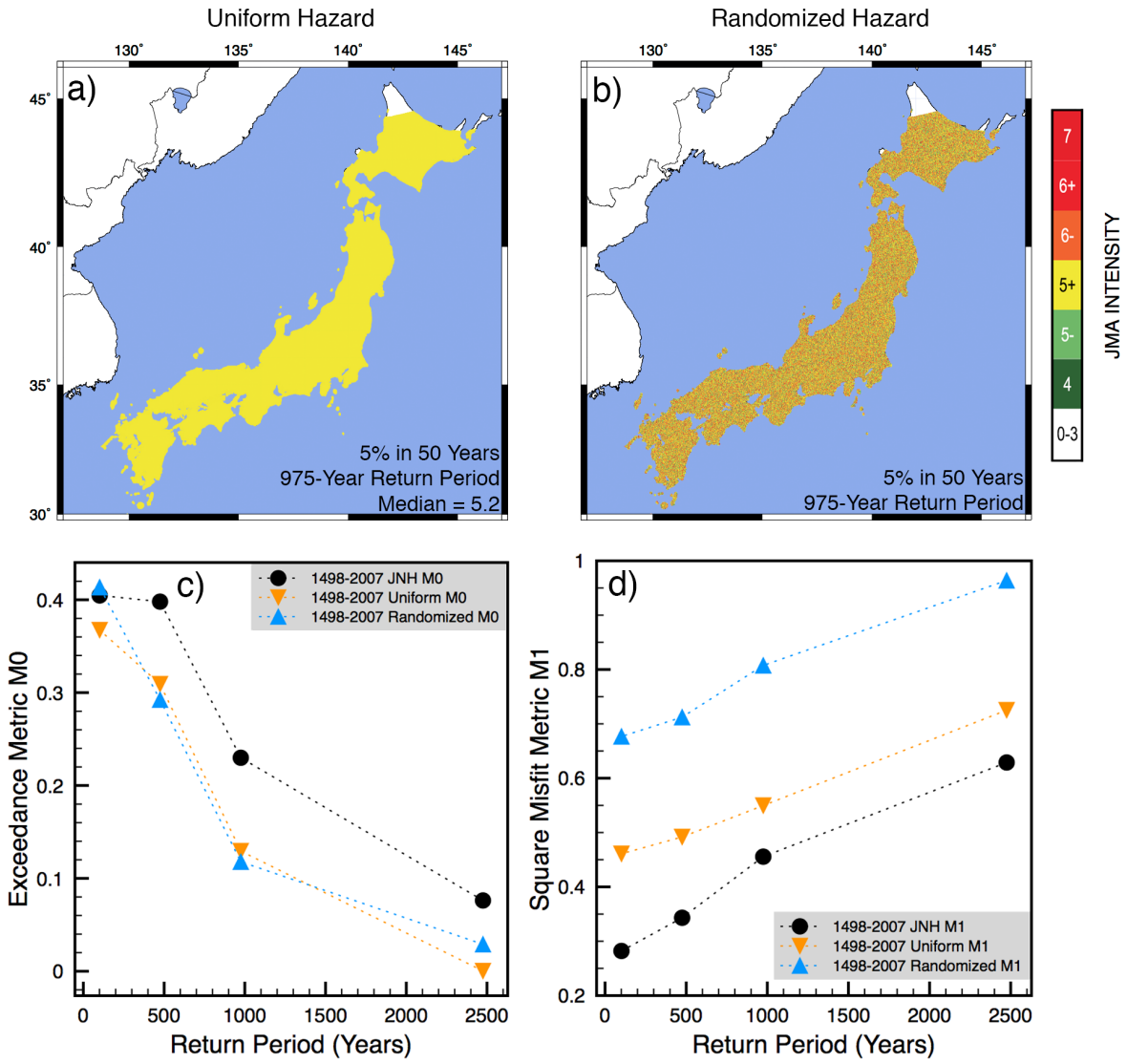


Figure 3.5. (a) The uniform hazard map, with hazard at all sites set equal to the median of the corresponding map from Figure 3.1. (b) Randomized hazard map, with hazard at sites randomly chosen from values in the corresponding JNH map. (c,d) Performance metrics for applying the actual JNH, uniform, and randomized versions of the maps.

undefined distributions is a challenging problem (Thompson, 1933). However, Figure 3.7 demonstrates that the data are roughly normal.

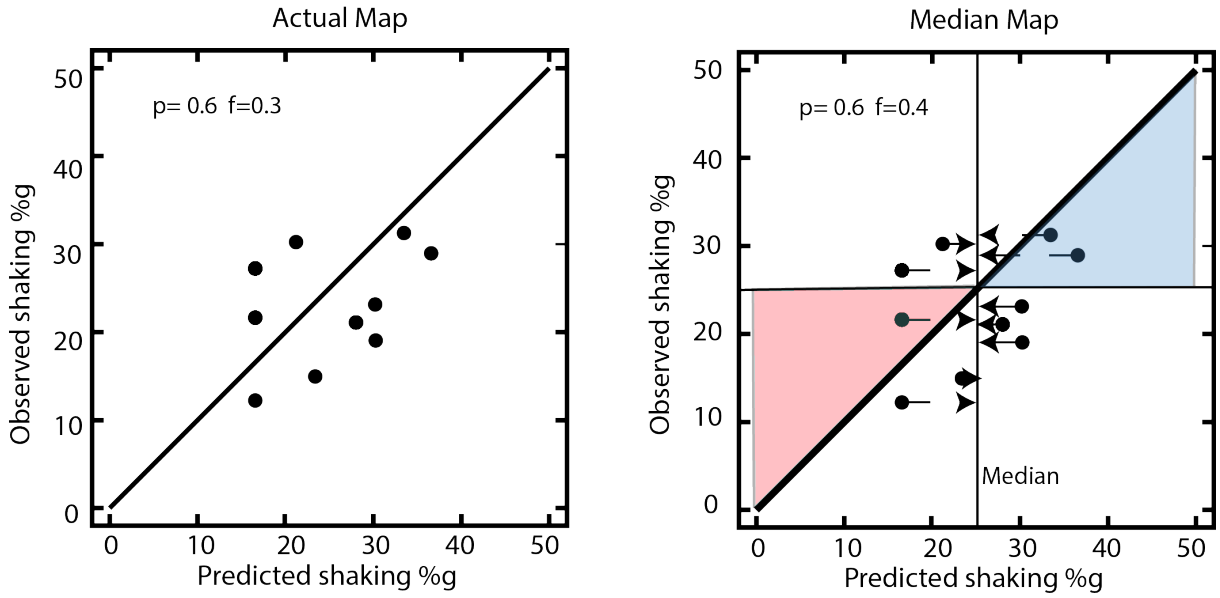


Figure 3.6. Illustration of how using the median predicted value for all sites can improve a hazard map's performance, as measured by the exceedance metric, if the map over-predicts the observed shaking.

This assumption makes calculating $f_{predicted}$ for a randomized map much more straightforward. $P(x > s) = P(x - s > 0)$, and given two normal distributions, this can be calculated as

$$P(x > s) = f_{predicted} = 1 - \Phi\left(\frac{\mu_s - \mu_x}{\sqrt{\sigma_s^2 + \sigma_x^2}}\right) \tag{3.1}$$

where each μ and σ^2 correspond to the predicted or observed data, and $\Phi(z)$ is the cumulative distribution function of the standard normal distribution $N(0, 1)$, so $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{\frac{t^2}{2}} dt$.

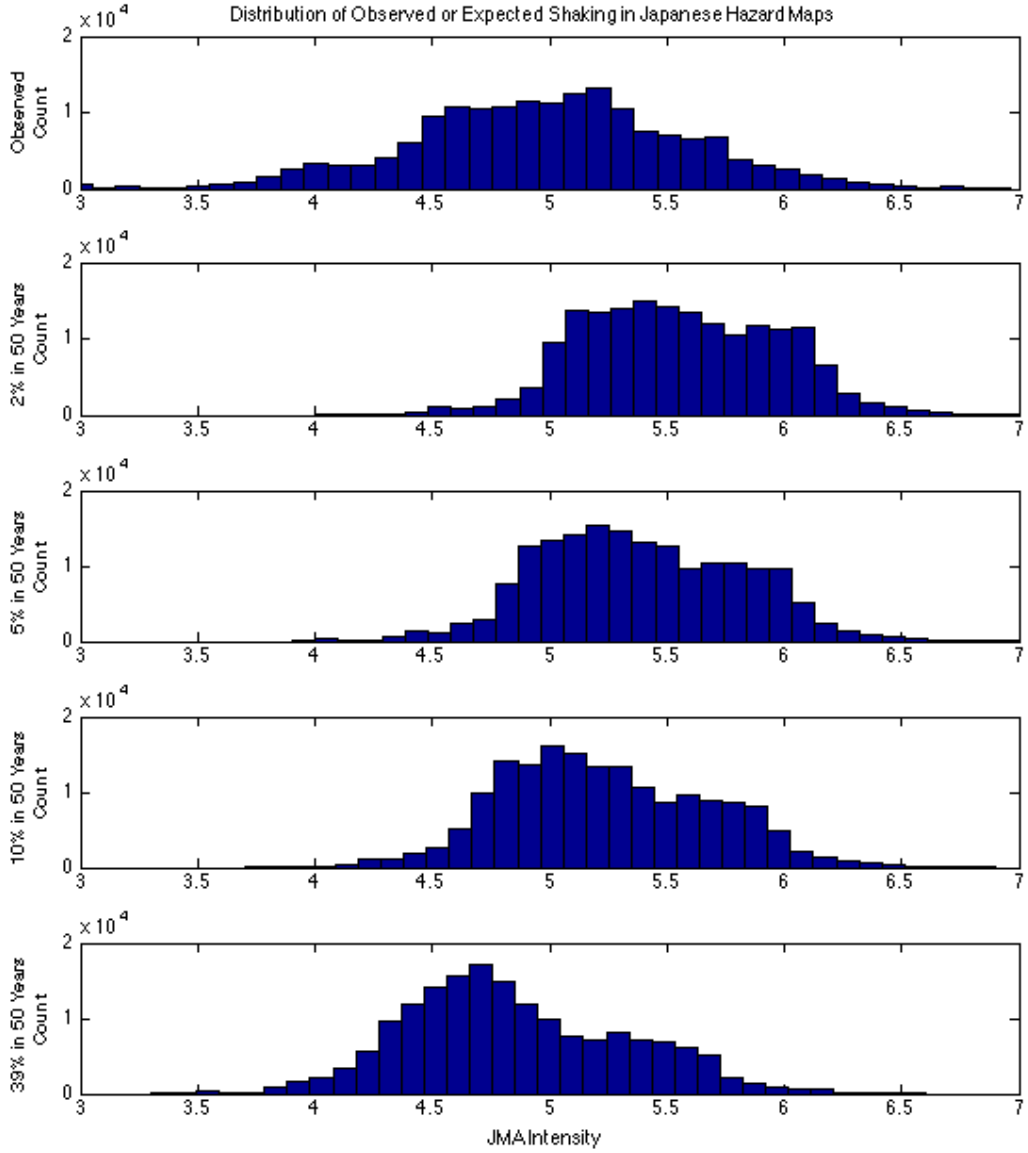


Figure 3.7. Distribution of observation and expectation in each of the Japanese maps from Figure 3.1. Distributions are roughly normal, and are treated as such to predict exceedance for randomized maps.

Table 3.1. Comparison of predicted and actual f for a randomized map, based on the mean and standard deviation of the observed shaking record, and JNH predictions (Figure 3.1). $\mu_x = 4.9612$, $\sigma_s = 0.6595$.

Probability of Exceedance	Return Period	μ_s	σ_s	$f_{predicted}$	$\bar{f}_{observed}$
39% in 50 Years	101 Years	4.8484	0.4266	0.2266	0.2185
10% in 50 Years	475 Years	5.2339	0.4399	0.2968	0.2956
5% in 50 Years	975 Years	5.3843	0.4506	0.3665	0.3733
2% in 50 Years	2475 Years	5.5503	0.4787	0.5551	0.5806

Thus it can be said that the fraction of sites exceeded for a randomly distributed map follows a normal distribution, and there exists a formula to predict the fraction of sites that will exceed observation.

Table 3.1 shows how these predictions compare to the actual, calculated fraction exceeded, and prove the usability of any one randomized map as a point of comparison to the actual JNH maps. $\sigma_{\bar{f}}$ for all maps is substantially smaller than each $\bar{f}$, $\sigma_{\bar{f}} \approx 0.0007$.

The exceeded fraction formula shows that exceedance appears to be a function of the mean observation, where a smaller mean correspond to larger f. The slight deviations between $f_{predicted}$ and $\bar{f}_{observed}$ can be explained by the deviations from true, prefect normalcy seen in Figure 3.7, but the general similarities justify the normalcy assumption.

The median results for the randomized maps are similar to those for the uniform maps, and thus generally better (lower $M0$) than the JNH maps. However, Figure 3.6 shows that using the squared misfit metric, the JNH 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.

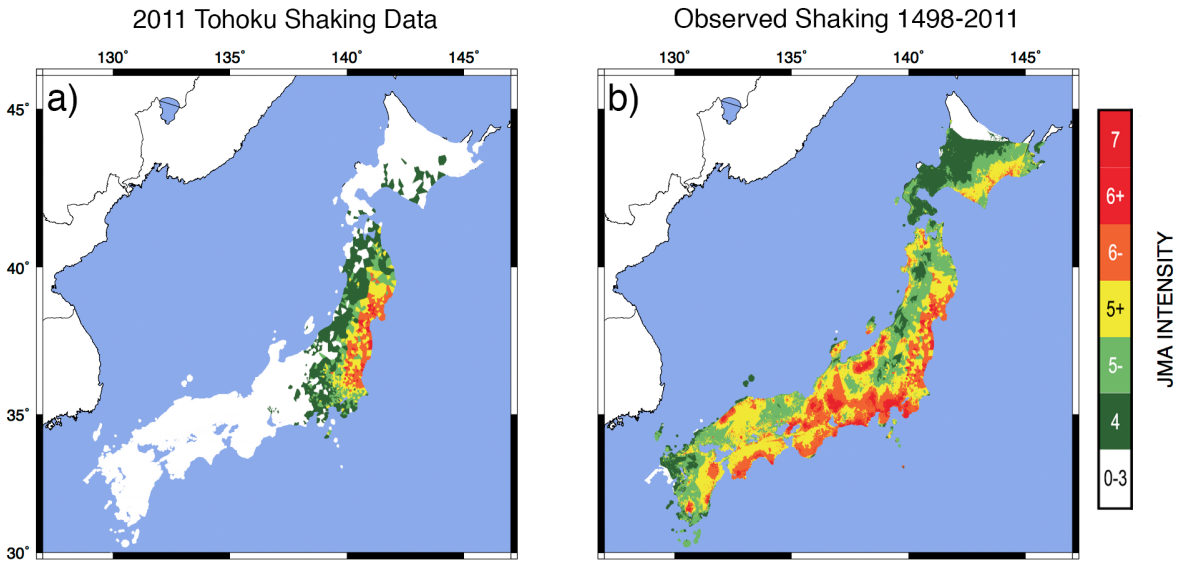


Figure 3.8. (a) Observed shaking in the 2011 Tohoku earthquake. (b) Historical shaking (1498-2007) map (Figure 3.1e) updated with Tohoku data.

3.5. Incorporating Tohoku Data

To expand on the comparison between actual and synthetic maps, the observational data set was expanded with shaking information following the 2011 Tohoku earthquake, effectively increasing the observational window from 1498-2007 to 1498-2011. These data were provided as 2,878 individual intensity measurements from different sites (Figure 3.8a). As with the rest of the data, ArcGIS was used to spatially join this data set to the prior data set, creating two observation data sets, one for 1498-2007, and one of shaking from 2011. Selecting the maximum shaking at each site from these two data sets yielded an updated data set, shown in Figure 3.8b.

Adding these data dramatically increases the maximum observed shaking along the east coast from about 35^o- 38^oN (Figure 3.8b). With this data set, repeated analyses for

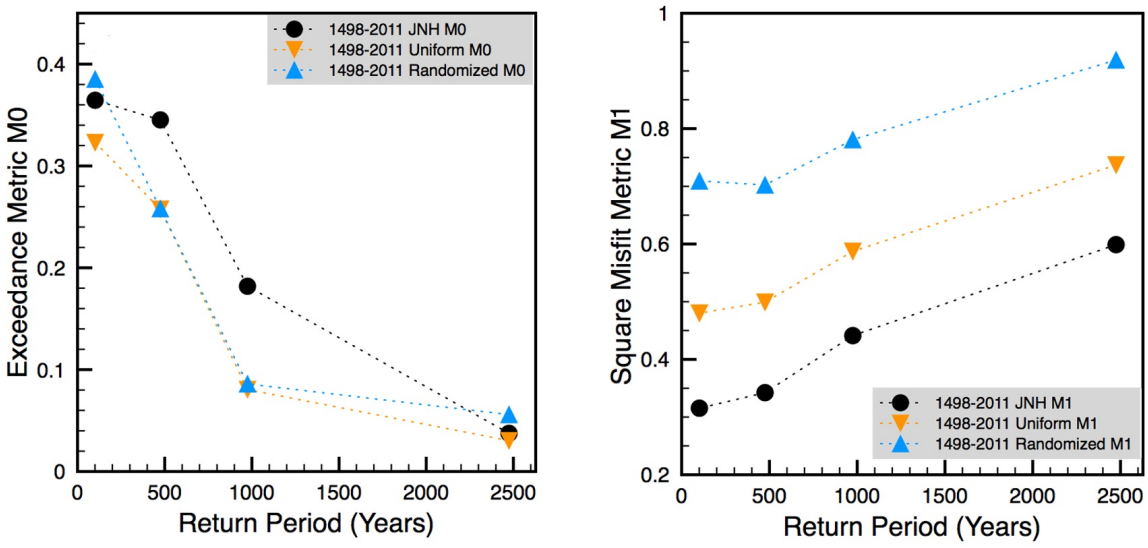


Figure 3.9. Performance metrics for applying uniform and randomized versions of maps to updated data.

the updated JNH, uniform, and randomized maps yield new metric scores (Figure 3.9). The exceedance metric $M0$ for each updated JNH 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 updated JNH maps still outperform uniform or randomized maps. Adding the Tohoku data improves the fit of the JNH 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. Figure 3.10 is an updated version of Figure 3.5, showing the smaller difference between observed and predicted shaking for the longer return-period maps.

3.6. Implications

Table 3.2 summarizes the results from the previous metric calculations. The basic finding is that the Japanese hazard maps are not performing as well as might be hoped.

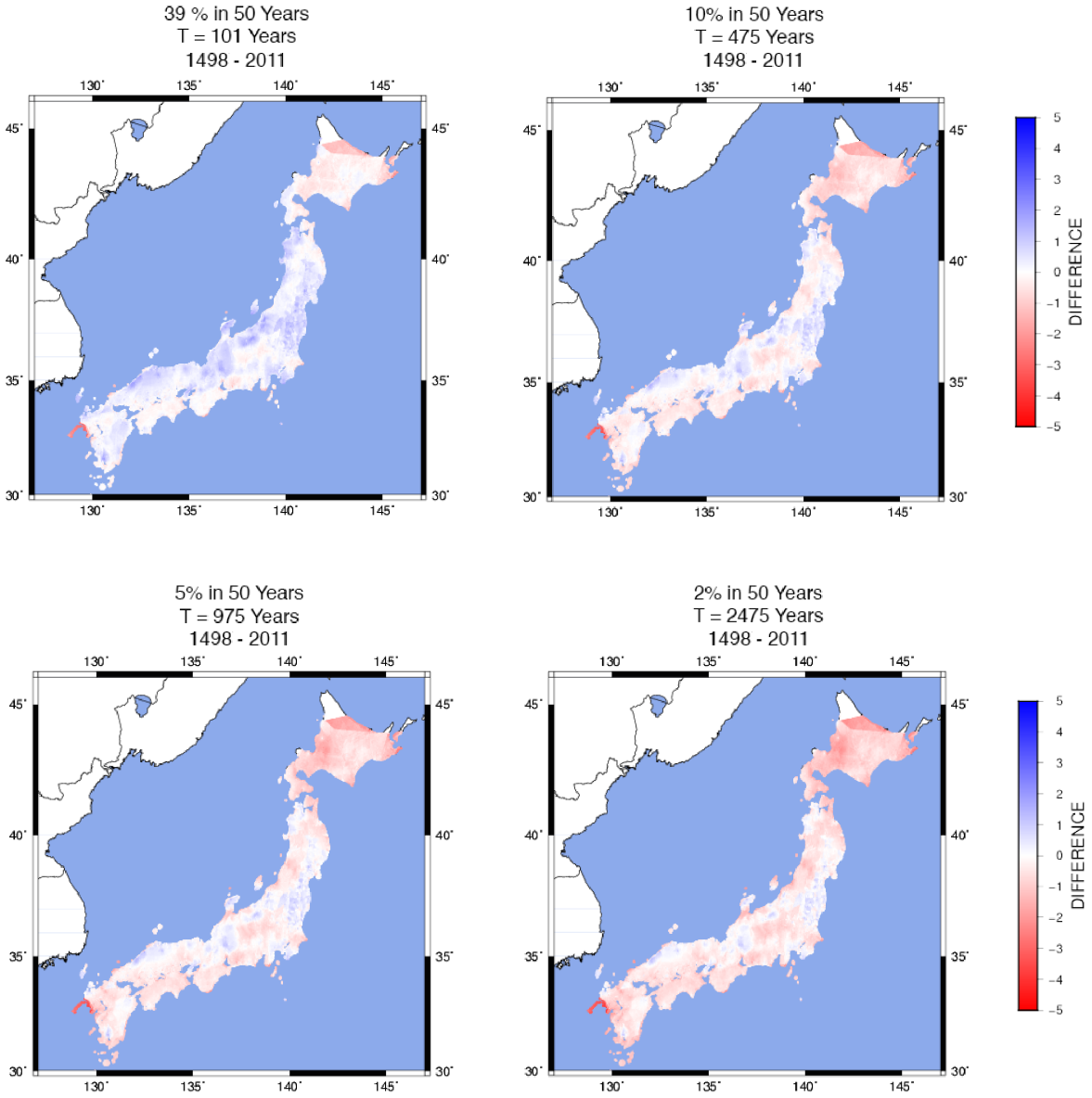


Figure 3.10. The difference between observed and predicted shaking with 2011 Tohoku earthquake data added. The increased shaking along the eastern coast reduces the extent of over-prediction.

Table 3.2. Summary table of metric calculations for JNH, uniform, and randomized hazard maps, with and without Tohoku data for different return periods T.

		1498-2007		1498-2011	
Maps	T (Years)	$M0$	$M1$	$M0$	$M1$
JNH	101	0.40	0.28	0.36	0.32
	475	0.39	0.34	0.34	0.34
	975	0.22	0.46	0.18	0.44
	2475	0.07	0.63	0.03	0.68
Uniform	101	0.37	0.46	0.32	0.48
	475	0.30	0.49	0.25	0.50
	975	0.12	0.55	0.07	0.59
	2475	0.003	0.73	0.03	0.74
Random	101	0.41	0.68	0.38	0.71
	475	0.29	0.71	0.25	0.70
	975	0.11	0.81	0.08	0.78
	2475	0.03	0.97	0.06	0.91

Although this possibility was suggested by damaging earthquakes in areas mapped as low hazard (Geller, 2011), 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 JNH maps systematically over-predict shaking, and uniform or randomized maps do better from this perspective. However, the JNH maps describe the observed shaking better than uniform or randomized maps. This complicated behavior illustrates the value of different metrics, in that $M0$ is more sensitive to average shaking levels, whereas $M1$ is more sensitive to spatial variations. It seems that although the JNH maps are designed to predict shaking levels that should be exceeded at a certain fraction of the sites, the process by

which their parameters are chosen tends to make the mapped shaking more closely resemble the maximum observed. That is to say that while the maps are intended to be judged probabilistically, as done by $M0$, they can be described as successful deterministically, as done by $M1$.

The observation that the JNH 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), the similar histograms for the observed and predicted shaking values (Figure 3.7) argue against a major bias. Moreover, such data are expected to be biased toward higher— not lower— values (Hough, 2013).

Another cause of mismatch could be that the JNH maps are partially time-dependent, in that the probability of earthquake recurrence and hence hazard is modeled on some of the faults as varying with time, whereas that for other faults is modeled as time-independent. However, this should have little effect for evaluating maps for two reasons, as shown schematically in Figure 3.11. First, the predicted hazard at a site is the sum of contributions due to many different faults, which are assumed to be at different stages in their seismic cycles, so the net effect of integrating forward (forecasting) or backwards (hindcasting) will be similar. Second, the longer the data set and return period considered,

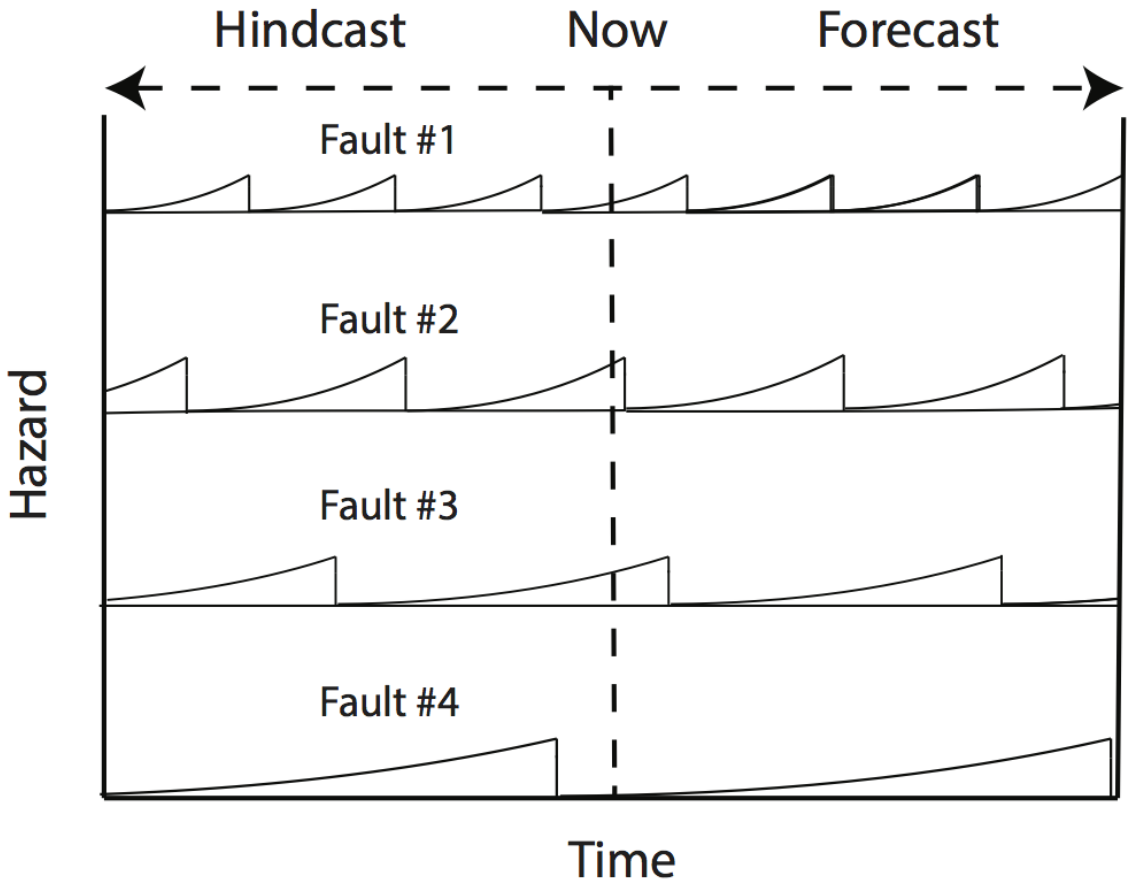


Figure 3.11. The hazard at a site due to four different faults, each of which is presumed to contribute a time-dependent hazard. Because the faults are at different stages in their cycles, their net contribution is similar going forward and backward in time, especially for longer return periods.

the more they average over entire seismic cycles. Hence it is justified, like Miyazawa and Mori (2009), to compare the JNH maps to the 510-year historical data set.

The maps could be also biased upward due to assumptions about the earthquake sources, the ground motion prediction equations, or conversions between the predicted

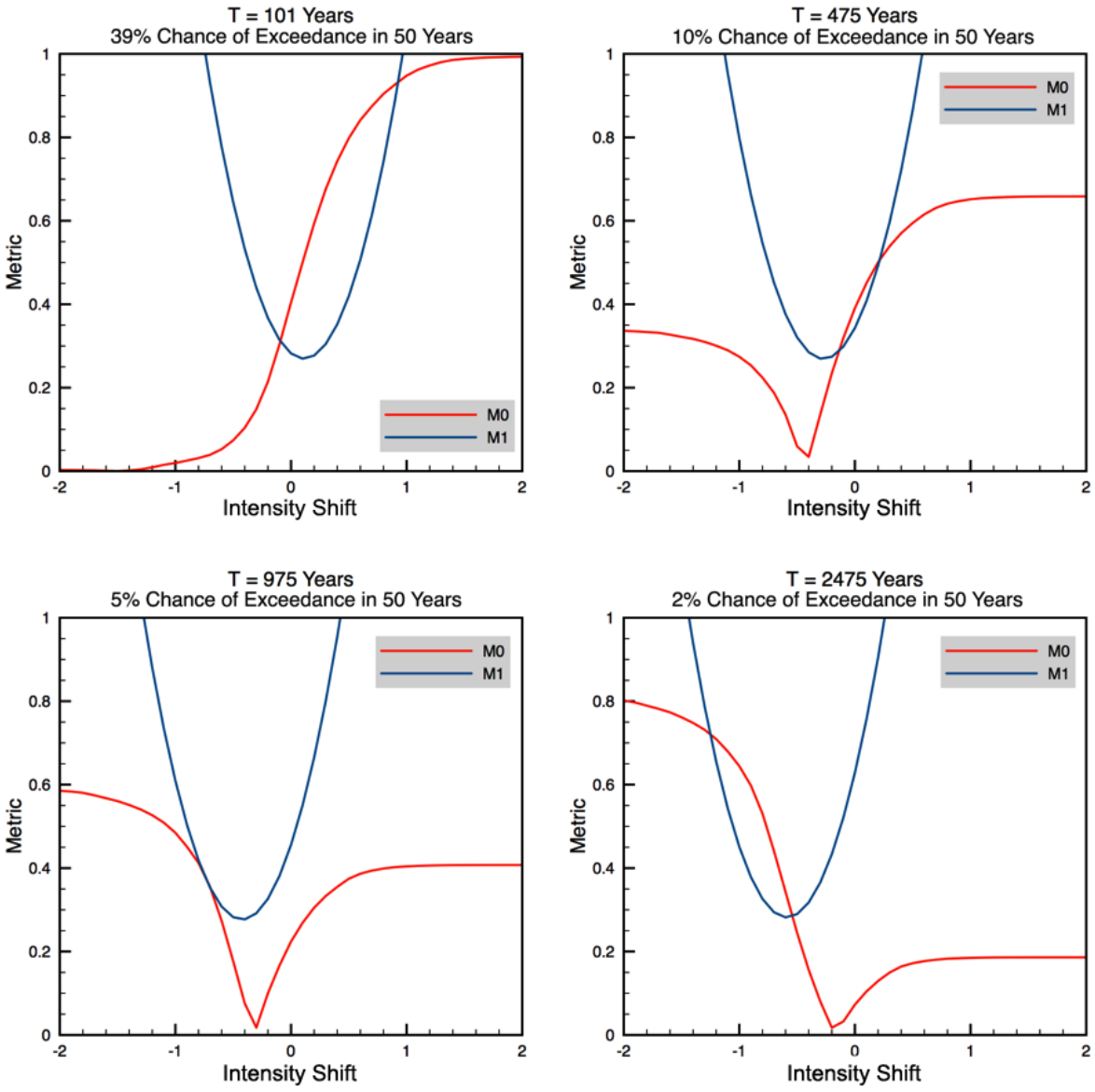


Figure 3.12. Change in metrics as a result of applying a uniform shift to the maps' predictions. The 475, 975, and 2475-year maps all exhibit improvements for both the fractional exceedance and squared misfit metrics when predictions are decreased by a small amount. The 101-year map has very low predictions and a high expected exceedance of 99.4%, which causes the metrics to behave differently from the others when a shift is applied.

shaking and intensity. Lowering the predicted shaking at all sites by a constant shift improves both $M0$ and $M1$, as shown in Figure 3.12, although the actual misfit is spatially variable, as shown in Figures 3.4 and 3.10. 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 data 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.

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. Given its length and quality, the 500-year long Japanese data set is the best known data set for these purposes, but hopefully high-quality historical data sets can be developed for some other areas with long historical records. Among the many questions, is whether better results are best obtained via better choices of parameters in the probabilistic approach (Stein and Friedrich, 2014) or by alternative deterministic approaches (Klügel 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.

CHAPTER 4

The Effects of Smoothing on the Performance of Earthquake Hazard Maps

4.1. Summary

In recent years, it has become clear that the actual performance of earthquake hazard maps often differs from that ideally expected, for reasons that are unclear. As a result, this study explores map behavior to learn more about how they actually perform by taking an empirical approach of asking what maps do, rather than what they should ideally do. Here, I explore whether less detailed 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— that the predicted ground motion should be exceeded only at a specific fraction of the sites— 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, such that a uniform map performs worse than the unsmoothed map. Because the maps were made by using other data and models to try to predict future earthquake shaking, rather than by fitting past shaking data, this result is probably not an artifact of hindcasting rather than forecasting. It suggests that hazard models and the resulting maps can 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. Hence to forecast future hazard, the goal should be not to build the most detailed model, but instead one that is robust or stable in the sense that the forecast is not unduly affected when the Earth does not behave exactly as expected.

4.2. Introduction

This chapter, as was the case in the prior chapters, is motivated by the fact that recent earthquakes that did great damage in areas shown by earthquake hazard maps as relatively safe. This has 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.

Assessing how well maps describe actual shaking 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 probability of exceedance, p. Discussed in prior chapters, this probability $p = 1 - e^{-\frac{\tau}{T}}$ is Poissonian, and is small when τ/T is small, and grows with observation time τ (Figure 3.2). Hence the shaking predicted by a map with a T-year return period should have a 39% chance being exceeded in $\tau = T/2$ years, a 63% chance being exceeded in $\tau = T$ years, and 86% in $\tau = 2T$ years.

It should be reiterated that although such assessments are not true tests, in that they compare the maps to data that were available when the map was made(i.e. forecasts rather than hindcasts), they give useful insight into the maps' performance.

In Chapter 3, I demonstrated that in some cases, one can argue in favor of a uniform hazard map for describing national hazard. By the metric implicit in the PSHA methodology, the fractional exceedance metric $M0$, it was found that the 2008 Japan

National Hazard (JNH) maps performed worse than a map where hazard was set to the median of the JNH map's predictions. Expanding on this, uniform map can be described as 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?

4.3. Smoothing Map Performance

I use the data set from the prior chapter, the observational catalog from Miyazawa and Mori (2009), which gives 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).

I smoothed the JNH maps 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. Given a prediction at some coordinate latitude/longitude, $s_{i,j}$, each smoothed map consists of new measurements at each point

$$s'_{i,j} = \frac{1}{(2D+1)^2} \sum_{j-D}^{j+D} \sum_{i-D}^{i+D} s_{i,j} \tag{4.1}$$

where D is a parameter that describes the size of the smoothing grid.

Iterating over all points on the map using progressively larger values for D yielded maps smoothed to greater degrees. For regions close to the coast, only values on land in Japan were used, disregarding values from the surrounding ocean (null values). 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 ($D = 1$), and each individual cell was $\sim$1.5 km on a side. This smoothing procedure is quite simple, and improved variants that used shapes or Gaussian weights other than squares might do even better. The optimal smoothing algorithm is beyond the scope of this study, rather the goal here is to illustrate how even simplistic smoothing algorithms may yield performance benefits.

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, shown in Figure 4.1. Figure 4.2 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 $D = 100$ (300 km on a side) across. This reinforces an earlier result, in that smoothing over all of Japan produces uniform maps, which was found to 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 $D = 25$ to 50 (75-150 km) smoothing window, and then decreases with further smoothing. This reinforces the earlier result that by this metric uniform maps perform worse than the unsmoothed map. As discussed in the following subsection, the effect of random error on $M1$ is quite small, so the improved fit is significant.

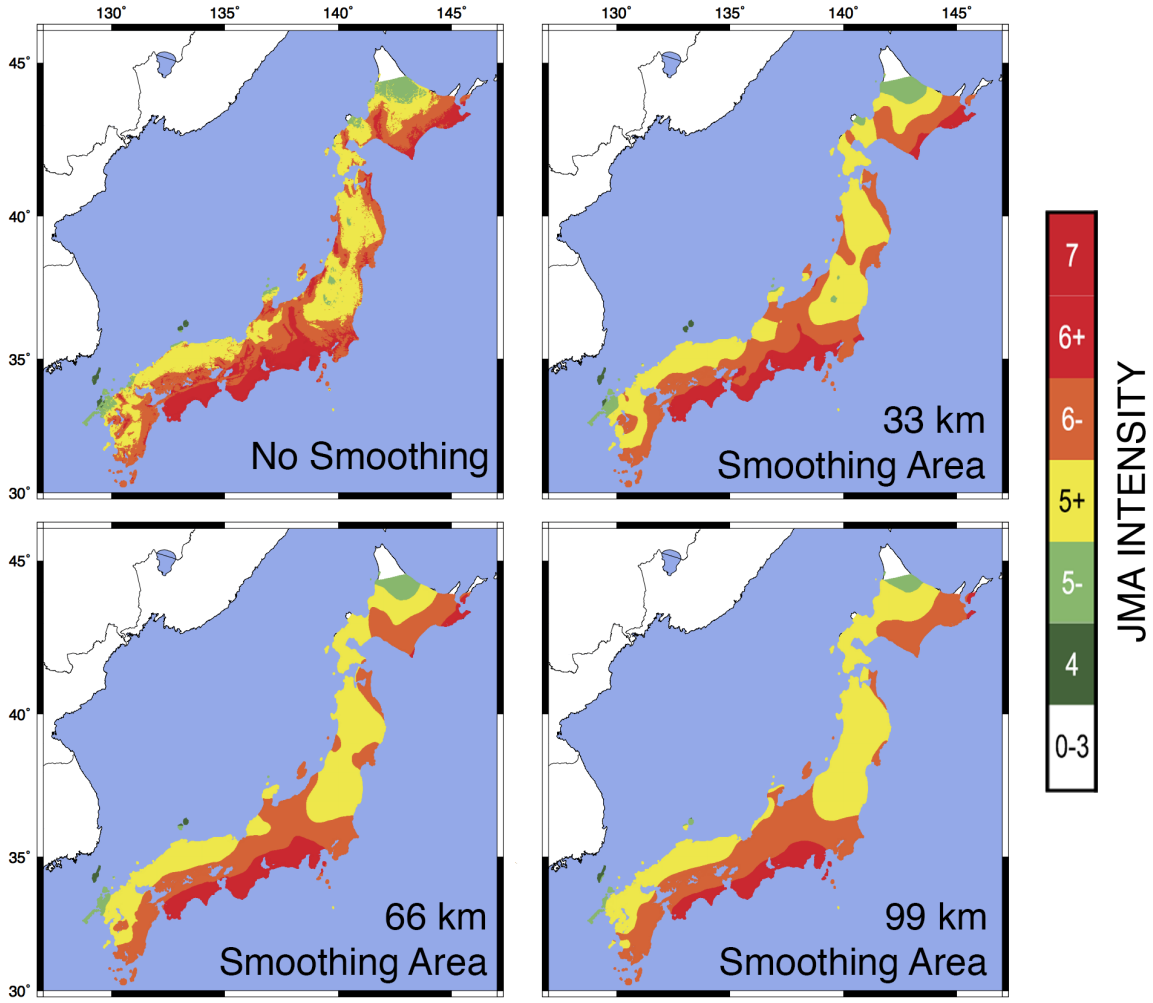


Figure 4.1. Effects of smoothing the JNH map with 475-year return period (a) over progressively larger areas (bd).

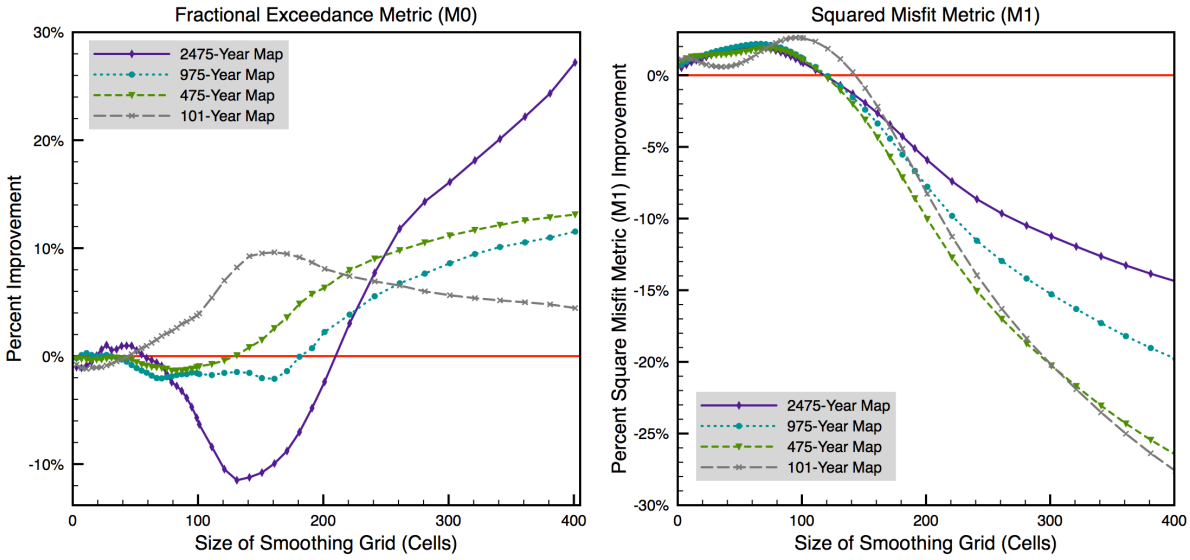


Figure 4.2. Improvement in map performance described by the change in fractional exceedance (a) squared misfit (b) metrics compared to the original map, for different amounts of smoothing.

4.3.1. The Effects of Random Error on $M0$ and $M1$

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 $\frac{f(1-f)}{n}$, with n denoting the equivalent number of statistically independent sites after allowance for spatial correlations. 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.

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

$$E[M0] = \mu\left[1 - 2\Phi\left(-\frac{\mu}{\sigma}\right)\right] + 2\sigma\varphi\left(-\frac{\mu}{\sigma}\right), \tag{4.2}$$

where $\mu = E[f - p]$, $\sigma = \sqrt{\text{var}(f - p)}$, $\Phi(s) = \int_{-\infty}^{s} \varphi(x)dx$, and $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$. 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.

4.3.1.1. 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)(0.73)/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, due to the estimations $\Phi\left(-\frac{\mu}{\sigma}\right) \approx 1$ and $\varphi\left(-\frac{\mu}{\sigma}\right) \approx 0$.

4.3.1.2. Variance of $M1$. Given $x_i - s_i$, which has variance υ^2 and kurtosis β, the variance of $M1$ can be described by

$$V(M1) = \frac{(n-1)^2\upsilon^4}{n^3}\left[\frac{n-1}{n}\beta - \frac{n-3}{n-1}\right] \approx \frac{\upsilon^4}{n}[\beta - 1] \tag{4.3}$$

Consider the 475-year return period. Denote the deviations $x_i - s_i$ by d_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 again the equivalent number of independent sites is 500. Then υ^2 is estimated by 0.2695 and β, the kurtosis,

is estimated by $\frac{0.4557}{0.2695^2} = 6.274$. Equation 4.3 hence estimates $V(M1)$ by 0.000764 and the standard error of the $M1$ statistic by 0.028. The estimate of $M1$ was 0.34 (Table 3.2), and so the coefficient of variation or relative standard error was 8.1%.

4.3.1.3. Variance of Change in $M1$ Due to Smoothing. Consider the apparent improvement in $M1$ due to smoothing, again for the 475-year return period for a 66 km smoothing area (bottom left of Figure 4.1). Denote the unsmoothed predictions by s_i and the smoothed predictions by s'_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 $M1_{unsmoothed}$ and $M1_{smoothed}$. The variance of the change in $M1$ due to smoothing $V(M1_{unsmoothed} - M1_{smoothed}) = V(M1_{unsmoothed}) + V(M1_{smoothed}) - 2\rho\sqrt{V(M1_{unsmoothed})V(M1_{smoothed})}$, with ρ denoting the correlation between $M1_{smoothed}$ and $M1_{unsmoothed}$. To estimate this, use the sample moments from the prior subsection. 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 averages of δ_i^2 and δ_i^4 were found in the above subsection, $\bar{\delta}^2 = 0.2965$ and $\bar{\delta}^4 = 0.4557$. $\bar{\delta'}^2 = 0.2653$ and $\bar{\delta'}^4 = 0.4837$. Applying Equation 4.3 to the smoothed data yields $V(M1_{smoothed}) = 0.000827$. ρ is approximated using $\rho \approx \frac{\sum_i \delta_i^2 \delta_i'^2}{\sqrt{\sum_i \delta_i^4 \sum_i \delta_i'^4}}$. This yields $\rho = 0.9664$. Thus, $V(M1_{unsmoothed} - M1_{smoothed}) = 0.00006$, which gives a standard error of 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).

4.4. Implications of Smoothing

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 variability.

Hazard maps are not expected to perform perfectly. Aspects of future earthquake behavior will differ from those of past earthquakes, the details of which are only partly known. Some of the assumed details of future earthquake behavior 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 trade-off (Parker, 1977) states that more detailed a model, 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.

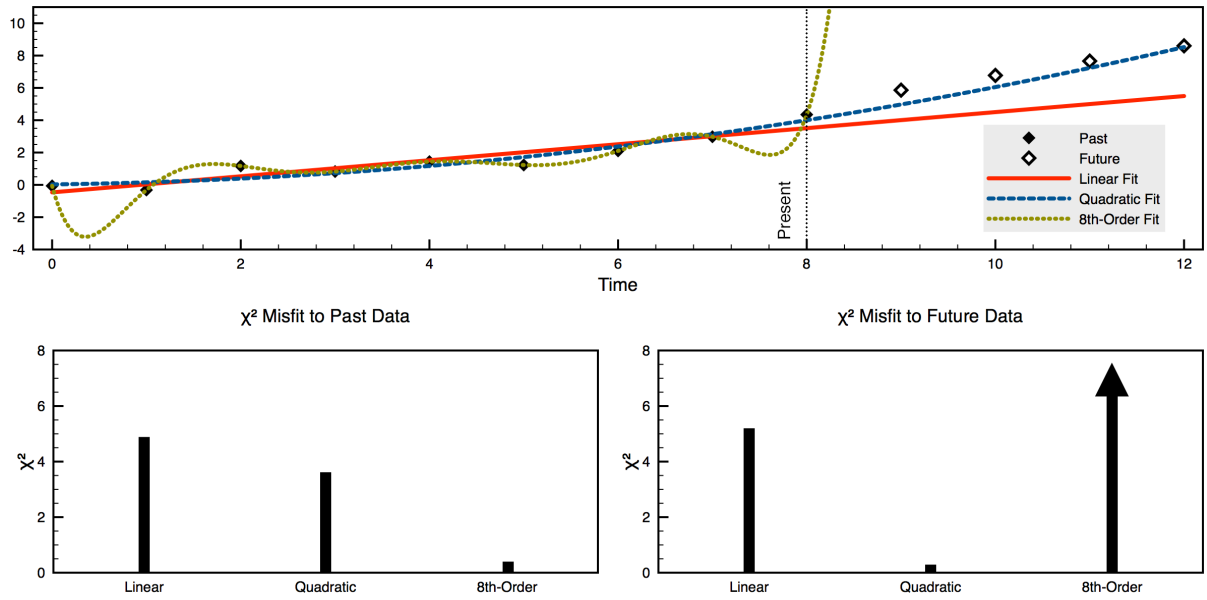


Figure 4.3. Example of the effect of over-parameterization on forecasting. A high order polynomial fits past data better than linear or quadratic models, but this more detailed model predicts the future worse than the simpler models.

An analogous phenomenon is recognized in other applications and termed "over-fitting" or "over-parameterization." 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 4.3 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 8th 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).

These findings showing that an improved fit results 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 data set used. Second, this 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 Chapter 3, neither problem appears large enough to invalidate this 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.*, 2009b), 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 over-fitting past data leads to poorer future predictions, suggests that similar effects could arise for earthquake hazard maps elsewhere. This approach

involved smoothing maps resulting from a 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 this 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 is robust or stable in the sense that the forecast is not unduly affected when the Earth does not behave exactly as expected.

CHAPTER 5

Earthquake Hazard Map Performance for Natural and Induced Seismicity

5.1. Summary

Seismicity in the central United States has dramatically increased since 2008 due to the injection of waste water produced by oil and gas extraction. In response, the USGS created a one-year probabilistic hazard model and map for 2016 to describe the increased hazard posed to the central and eastern United States. Using the intensity of shaking reported to the "Did You Feel It?" system during 2016, I assess the performance of this model using a metric that compares the fraction of sites at which the maximum shaking exceeded the mapped value to the fraction expected. These fractions are similar for both the central and eastern United States as a whole, and for the region within it with the highest amount of seismicity, Oklahoma and its surrounding area. The greatest mismatch is observed in northern Texas, with hazard over-stated, presumably because lower oil and gas prices and regulatory action reduced the water injection volume relative to the previous year. I also assess the model using a misfit metric that compares the spatial patterns of predicted and maximum observed shaking. This hazard map performs better by both metrics than other hazard maps studied in prior chapters. These results imply that such hazard maps can be valuable tools for policy makers and regulators in managing the seismic risks associated with unconventional oil and gas production.

5.2. Introduction to Shaking in the Central and Eastern United States

Since 2008, seismicity in the central United States has increased dramatically, largely due to the injection of waste water produced by unconventional oil and gas extraction (Ellsworth, 2013; Keranen *et al.*, 2013, 2014; Kim, 2013; Hough, 2014; Rubinstein and

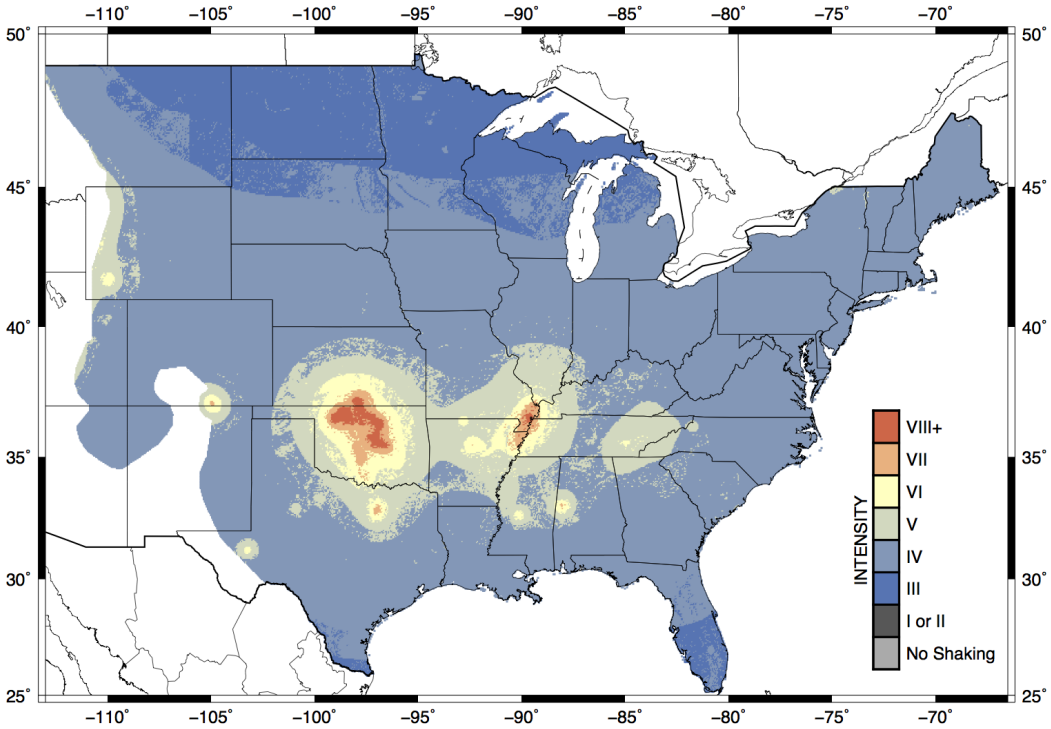


Figure 5.1. 2016 one-percent in one-year national seismic hazard map, showing the hazard for the central and eastern United States from induced and natural earthquakes (Petersen *et al.*, 2016a).

Mahani, 2015; Weingarten *et al.*, 2015). This increased seismic activity poses a higher hazard than historically experienced in areas that are generally unprepared for the resulting levels of shaking (Liu *et al.*, 2014; Ellsworth *et al.*, 2015).

The increased likelihood of damage necessitated reassessment of the seismic hazard in the area. Accordingly, the USGS produced a new seismic hazard map for the central and eastern U.S. (Petersen *et al.*, 2016a, b), including the effects of both induced and

natural seismicity (Figure 5.1). The largest change to the prior map (the 2014 U.S. national seismic hazard map), which did not incorporate induced earthquake effects, was the significantly increased hazard predicted in the area covering southern Kansas, Oklahoma, and northeast Texas (Petersen *et al.*, 2015). Induced seismicity was incorporated into the hazard map by defining zones where earthquakes do not appear natural, indicated by a noticeable increase in seismicity near injection wells, both spatially and temporally. Petersen *et al.* (2016a) defines separate logic trees for seismicity inside and outside these zones, which differ largely in the parameters used to describe catalog duration, smoothing distance, maximum magnitude, and ground motion models. Seismicity rates are inferred from injection rates from the prior year, which are assumed to be unchanged for 2016. The updated national map shows between 5-12% probability of shaking at or above MMI VI in this area for the one-year time window in 2016, similar to the predicted hazard from natural seismicity in historically much more active regions like California (Petersen *et al.*, 2016b).

The new model is a one-year forecast, showing the level of shaking that should have a 1% chance of exceedance at any point on the map during the year. The model used in making this map assumed that earthquake rates would remain relatively stationary and could be used to forecast shaking during 2016. This approach includes the effects of non-tectonic earthquakes, in contrast to the 2014 model that excluded non-tectonic earthquakes.

Such one-year models are potentially valuable for policy makers and regulators dealing with the complex question of how to address the hazard due to induced earthquakes. To this end, this chapter investigates how well the model forecasted the shaking that

actually occurred in 2016, and quantifies the performance of the map using two metrics that summarize different aspects of the map's performance.

As is the case in prior studies, the performance of earthquake hazard maps is assessed using two metrics to numerically compare a map's predictions to records of shaking. The first, the fractional exceedance metric $M0$, is described by $M0 = |f - p|$, where f and p are the fraction and predicted fraction of sites that are exceeded for some observational time period τ. p is typically treated as a Poisson variable, so $p = 1 - e^{-\frac{\tau}{T}}$. While this metric is implicit in PSHA methodology, it is also binary ("above" or "below"), so the squared misfit metric, $M1 = \frac{1}{N}\sum_{i=1}^{N}(x_i - s_i)^2$, where x_i and s_i are each site's observed and predicted shaking, can provide an alternative view on map performance. For the purposes of this study, $p = P(x_i > s_i) = 1\%$.

5.3. Comparison to Observed Shaking

To assess the performance of the 2016 model, I use a record of observed shaking captured after the map was made. The best and most extensive data available are from the "Did You Feel It?" (DYFI) database (Wald *et al.*, 1999; Atkinson and Wald, 2007). DYFI is an online tool allowing anyone who experiences ground motion to report it. Responses are compiled and geocoded by zip code to characterize the shaking distribution from an earthquake. After a year of data is collected, the USGS compiles maps of the annual maximum shaking at sites reported to the DYFI system, gridded at a 10 km resolution (Quitoriano *et al.*, 2017). Despite possible issues of quality and data completeness, DYFI is considerably more complete than available instrumental data and proves to be one of the most thorough and robust data sets available (Wald *et al.*, 2012).

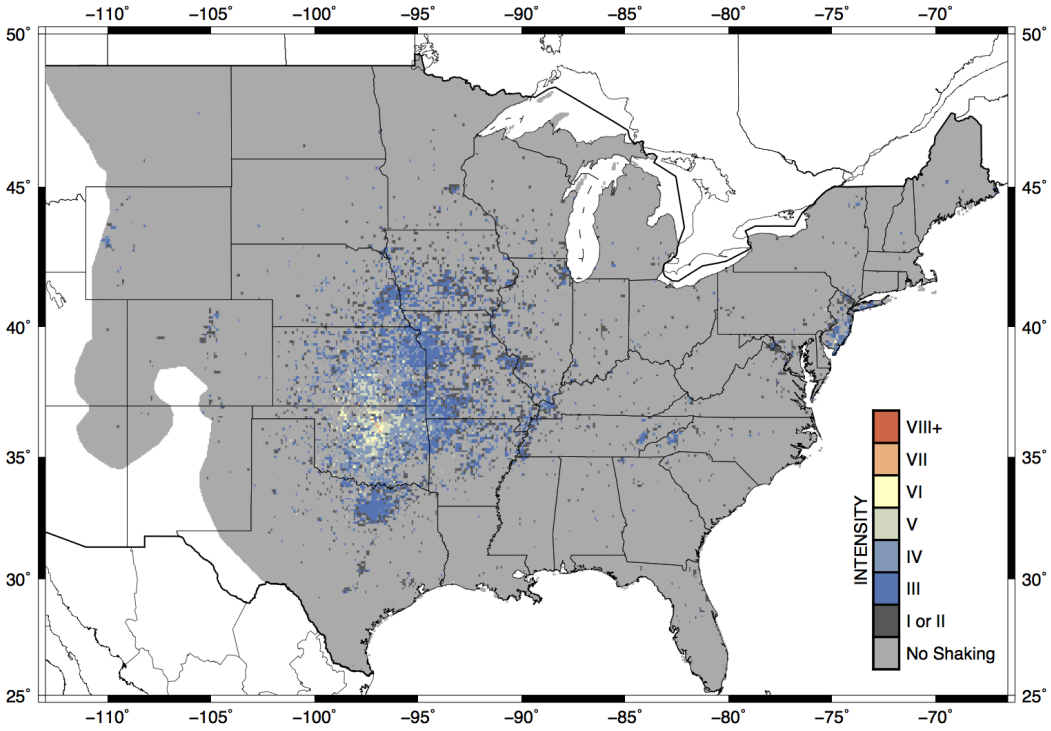


Figure 5.2. Maximum reported shaking recorded from "Did You Feel It?" (DYFI) in 2016 for the central and eastern United States.

The 2016 maximum DYFI response map (Figure 5.2) shows large areas of the central and eastern United States with no responses, with clusters of responses in the seismically most active regions. Absence of response can occur either because no shaking was felt, or because no one responded to DYFI after an earthquake, perhaps due to low population or "earthquake fatigue" following numerous events (Mak and Schorlemmer, 2016b). Assuming that all regions without a response did not experience shaking is unrealistic, especially given the low populations in portions of the study area. Furthermore, such treatment would incorrectly imply that the map severely over-predicts shaking. Instead,

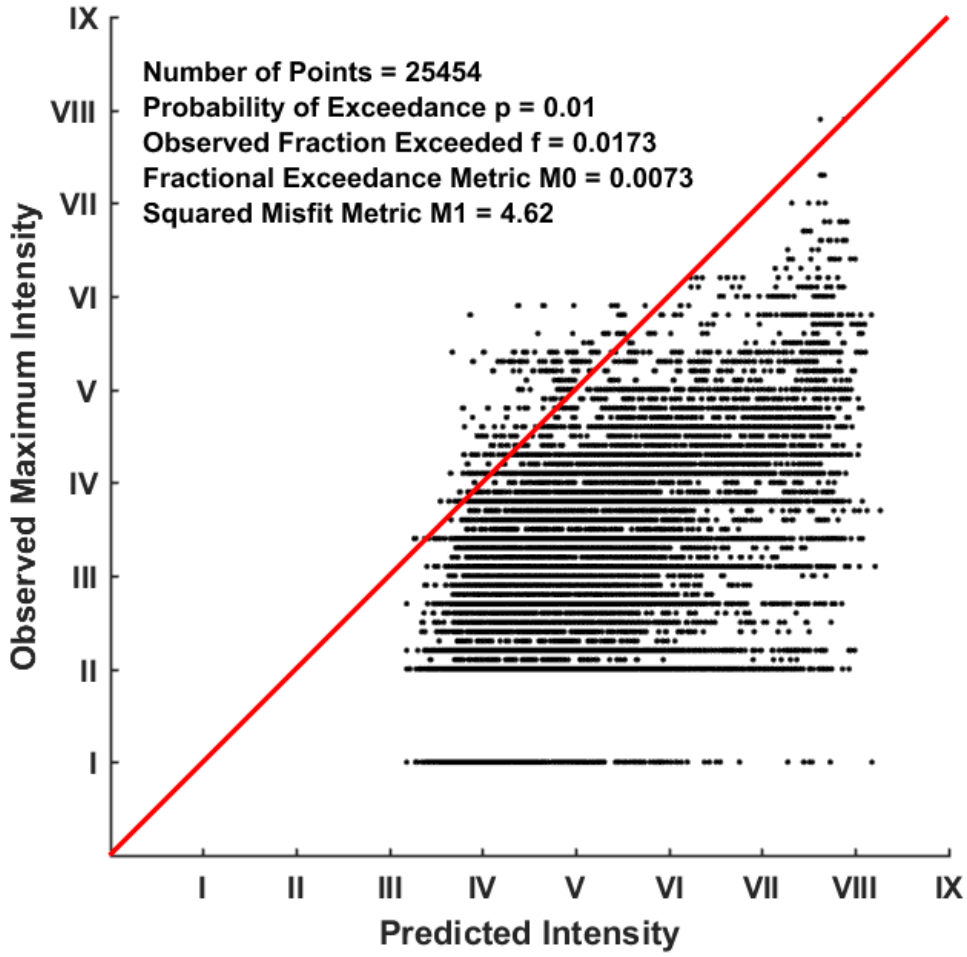


Figure 5.3. Comparison plot of 2016 map predictions and 2016 DYFI observations for all points with a DYFI response in the central and eastern United States.

it is better to treat regions lacking response as null or missing values and exclude them in evaluating the metrics.

Across the entire central and eastern United States, roughly 10% (25,454 out of 236,578 points) of the map has a DYFI response. All these points are used, while recognizing that spatial correlation among the shaking at these points vastly reduces the effective

sample below the nominal 25,454 (See Sections 2.7 and 4.3.1). The comparison between the predicted and observed maximum shaking is plotted in Figure 3. The fractional exceedance metric, $M0$, compares the fraction of points f above the diagonal line— where the largest observed shaking exceeds prediction— to the fraction p expected. About 1% of all sites should be above this line, and the actual fraction is 1.73%, leading to a fractional exceedance $M0 = 0.0073$. The squared misfit metric $M1$ is 4.62, reflecting the visual similarity between the hazard map predictions and the map of maximum observed shaking (Figures 5.1 and 5.2).

The difference between 1% and 1.73% site exceedances results from few hundred more exceedances than expected. To see how large a mismatch this is, consider how much of an increase in predicted shaking would make $p = f = 0.01$, and thus $M0 = 0$. This would occur if the average predicted shaking were 0.24 MMI units, or 5% higher, than that predicted. This would decrease the number of exceedances observed to exactly that predicted.

5.4. The Greater Oklahoma Area

When considering these results, it is important to note that the data are sparse on a national scale. Thus, it is also valuable to examine the most seismically active portion of the mapped area. Figure 5.4 shows the predicted and maximum observed shaking for this "greater Oklahoma" area. Here, data completeness is improved relative to the entire region, with 45% (10,160 out of 22,560) of sites having DYFI responses.

Figure 5.5 shows the metrics for this smaller area. The fractional exceedance metric $M0 = 0.0069$ shows that the fraction of sites in this area experiencing higher than expected

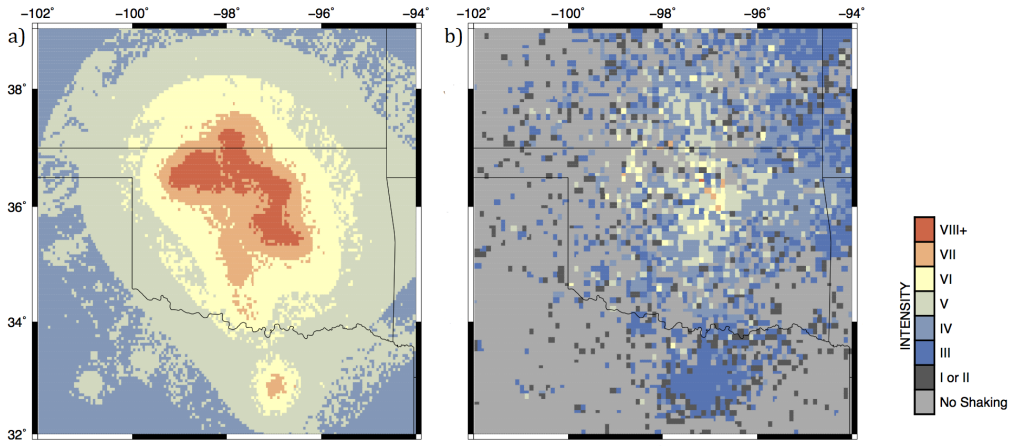


Figure 5.4. a) 2016 One-year seismic hazard model for induced and natural earthquakes for the greater Oklahoma region (Petersen *et al.*, 2016). b) Maximum reported shaking recorded from DYFI in 2016 for this region.

shaking is about the same as for the entire area, $M0 = 0.0073$. The squared misfit metric increases slightly, from $M1 = 4.62$ to 5.01.

The only notable difference between observation and prediction occurs for the Dallas area, where the map over-predicts the amount of shaking. Despite a maximum shaking forecasted of intensity VII, enough for moderate damage, the highest shaking widely reported is intensity III. This difference explains the increase in the squared misfit metric relative to the map as a whole, because a larger percentage of the local map, which covers roughly 15% of the greater Oklahoma region both in area and number of DYFI reports, is misfit.

The mismatch in Dallas could be most easily explained in two different ways:

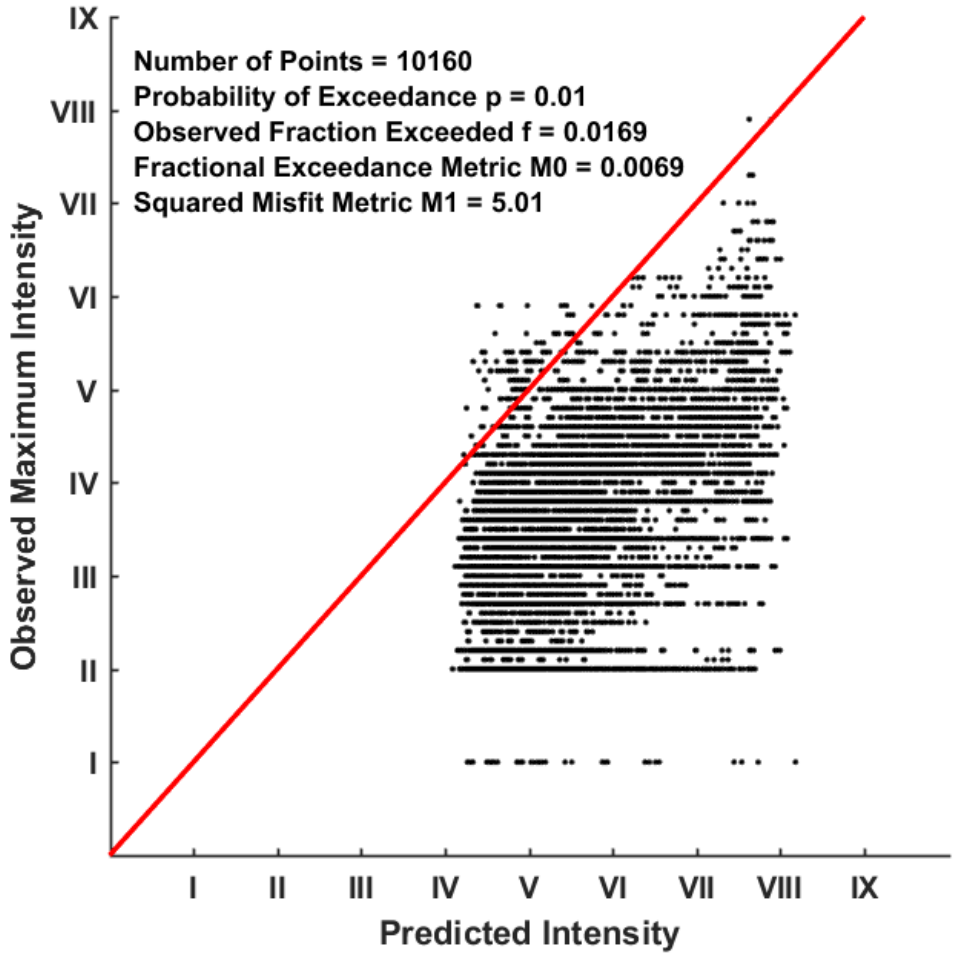


Figure 5.5. Comparison of 2016 map predictions and DYFI observations for all points with a DYFI response in the greater Oklahoma region.

(1) The DYFI data reflect the actual shaking that the map over-predicted.

(2) Fewer people in Dallas responded to DYFI, leading to under-reporting of the maximum shaking.

However, investigation into the DYFI data refutes the latter explanation. As shown in Figure 5.6, there appears to be a strong relationship between the number of reports

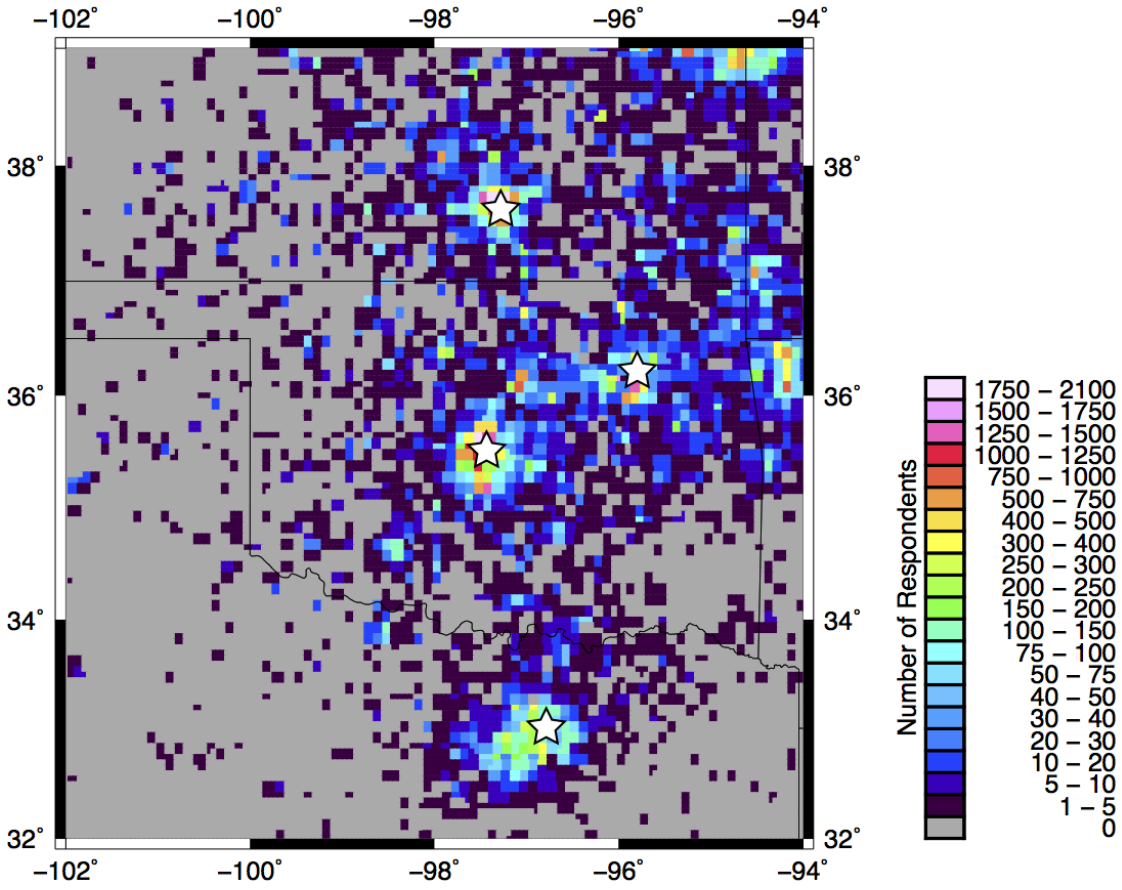


Figure 5.6. Number of respondents who reported shaking when the maximum intensity DYFI event occurred. Stars denote major population centers.

contributing to the maximum observed shaking and population, but not the intensity of the maximum shaking. The DYFI system seems to be well known enough to yield good reporting from a large population, even without intense shaking (Mak and Schorlemmer, 2016b).

It thus appears that the mismatch between observation and prediction in the Dallas region reflects a decrease in seismicity. The 2016 model assumed a constant level of human

activity, i.e. waste water injection rates remaining unchanged. However, unconventional oil development is tied closely to economic factors (Campbell and Laherrre, 1998; Murray, 2016). Perhaps due to changing oil prices, or in response to seismicity assumed to be associated with waste water injection, injection rates in northern Texas diminished in 2016, rather than staying stable (Hornbach *et al.*, 2016; Kuchment, 2017). As a result, the model over-predicted the shaking in 2016. A similar but smaller decrease in seismicity is also occurring in Oklahoma (Murray, 2016).

5.5. Supplementing Missing Data

The fact that DYFI only has responses from about half of the greater Oklahoma area prompts the question of how the data set can be supplemented for a more thorough picture. Null responses do not necessarily imply no shaking has occurred, and gaps between regions with reports of high shaking where reports of shaking are low or missing are likely to reflect low population rather than low shaking. One approach is to set non-reporting regions to intensity I, which is "not felt" (Boatwright and Phillips, 2017). However, setting null points to I is similar to setting them to 0, in that it also unfairly penalizes the map.

An alternative is to use models of expected shaking following known earthquakes to fill in sites without reported DYFI intensities. The USGS ShakeMap program predicts ground shaking following an earthquake, taking into account its magnitude, location, and geologic setting (Wald *et al.*, 2005). Although it is a model, rather than direct observations like DYFI, ShakeMap provides reasonably accurate data augmentation in null regions and regions where the reported shaking is surprisingly low given their locations (Wald *et al.*, 2012).

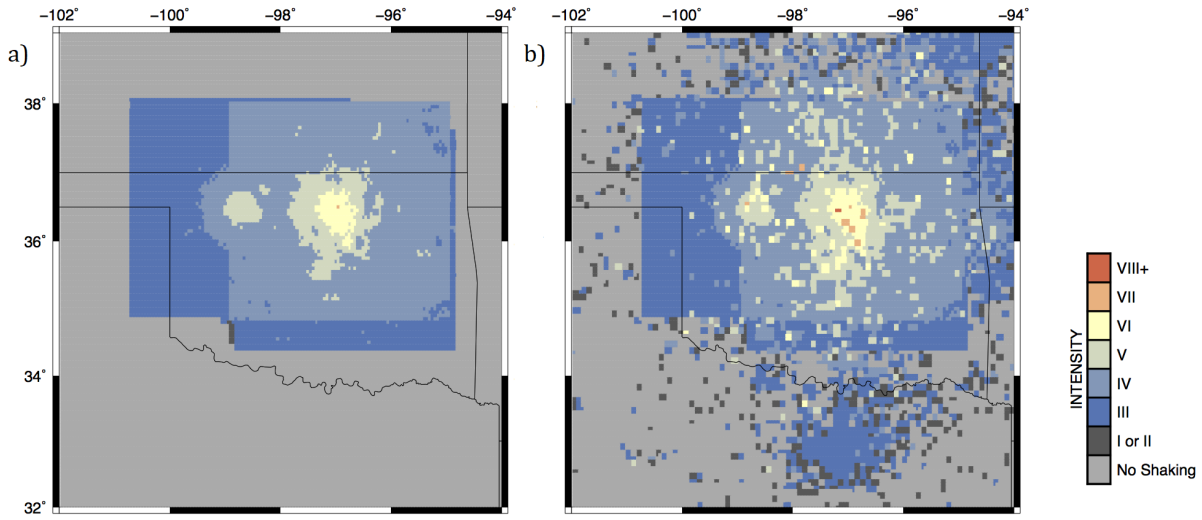


Figure 5.7. a) Maximum shaking predicted by ShakeMap from 21 earthquakes greater than magnitude 4 in Oklahoma in 2016. b) Result of combining ShakeMap predictions with existent DYFI data.

In 2016, 21 earthquakes with magnitude 4.0 or greater occurred in the greater Oklahoma region. 4.0 was selected as the minimum magnitude to reduce the difficulty of assembling a data set, given that the distribution of earthquakes across the region was sufficiently spread out such that smaller events would not produce higher shaking than from the larger events. Figure 5.7a shows the highest shaking modeled from the 21 earthquakes. ShakeMap predicts no exceedances relative to the hazard map, but the squared misfit is reasonably close to the original reported value. Hence, there is no extreme bias toward high or low values in the ShakeMap predictions, and treat the latter as minimum estimates of the maximum shaking at points without data.

Figure 5.7b shows the result of combining the ShakeMap predictions and DYFI data. In this combined data set, almost 60% of the sites have a shaking value (13,427 sites out of 22,560). Figure 5.8 shows the metrics based on this combined data set. While the number

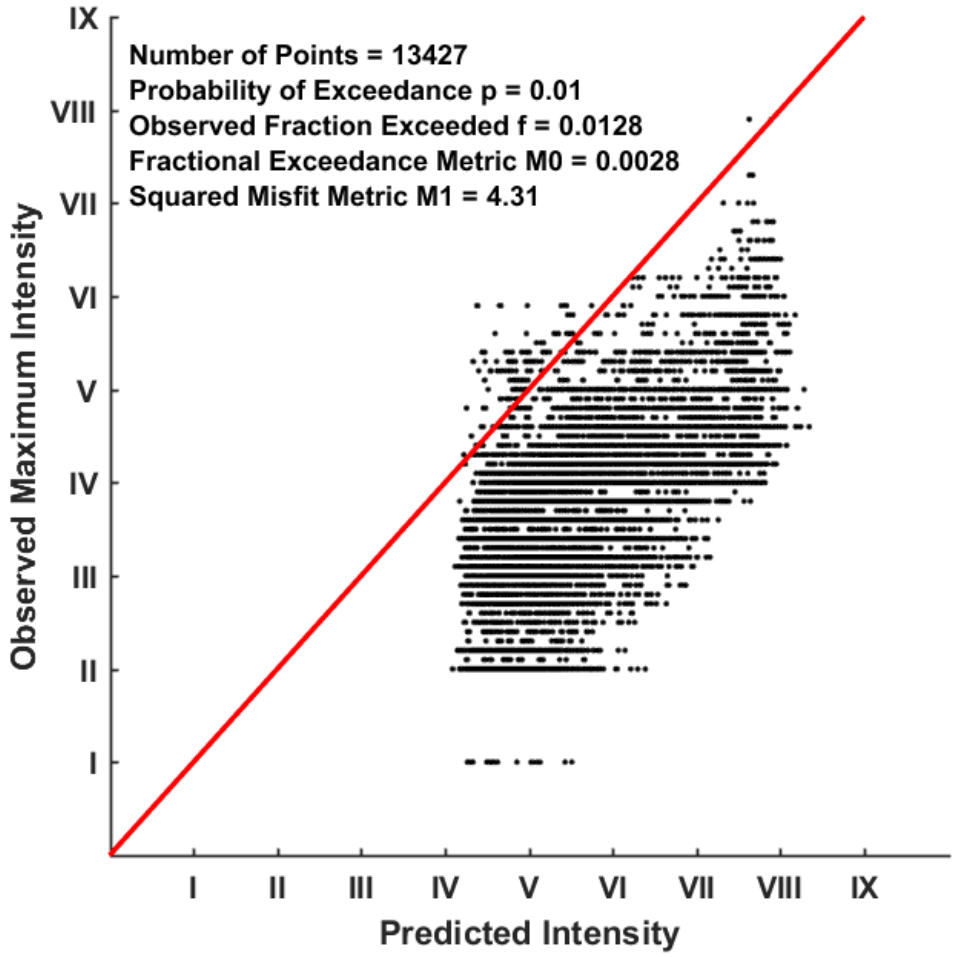


Figure 5.8. Plot comparing 2016 map predictions to combined DYFI observations and ShakeMap predictions for the greater Oklahoma region.

of exceedances remains constant, the number of sites increases, decreasing the fraction of sites that exceed the predicted shaking. As a result, the fraction of sites exceeding the map predictions is 1.28%, yielding $M0 = 0.0028$. The squared misfit decreases to $M1 = 4.31$. These reductions suggest that the missing data, and anomalously low reports in areas of high shaking made the map appear to perform less well than it actually did.

5.6. Trends in the Data

In addition to calculating metrics, comparing the maximum observed and predicted shaking (Figures 5.3, 5.5, and 5.8) can highlight trends in the data and show how the map performance varies. The original DYFI data, in addition to maximum intensity and number of respondents, also indicate the magnitude of the earthquake that caused the felt shaking. Figure 5.9 shows the predicted-observed plot for the greater Oklahoma region with this magnitude data added to the original DYFI data. Although the database of shaking covers a wide range of magnitudes, the shaking reports are almost entirely dominated by the M_W 5.8 Pawnee earthquake, the largest recorded in the state of Oklahoma (Yeck *et al.*, 2017).

A few exceedances come from small events with $M < 3.8$. Of the 172 exceedances, 168 came from the Pawnee earthquake, and 4 came from other sources. The DYFI records only note the magnitude, number of respondents, and number of events that drove responses, so which small event prompted each of these exceedances is unknown.

Three earthquakes in Oklahoma in 2016 had $M \geq 5.0$. Because each had a different magnitude, the maximum DYFI reports associated with each magnitude can be given a known epicenter, and an epicentral distance from the site to each can be calculated. Figure 5.10 shows the predicted-observed plot as a function of distance for the sites where epicentral distance is known (96% of all DYFI observations). Both the predicted and observed shaking decrease with distance, as expected. Exceedances arise primarily where the map predicts intensity IV-V shaking, and come from a range of distances with no clear bias favoring one distance. A single exceedance for predicted intensity VIII appears to have occurred very close to the source event.

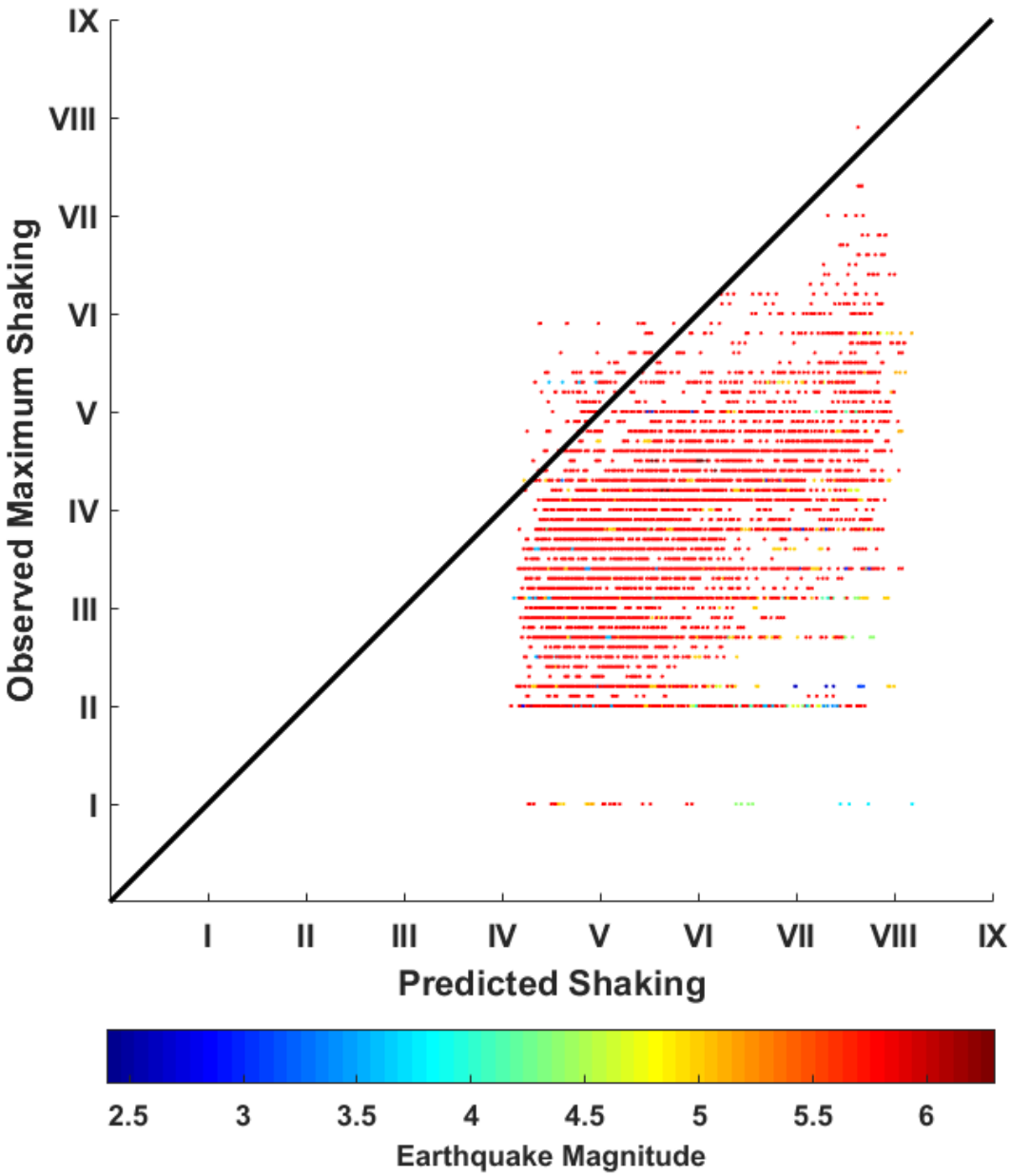


Figure 5.9. Plot comparing 2016 national seismic hazard map predictions to maximum DYFI observations for points with DYFI responses in the greater Oklahoma region, further broken down by magnitude.

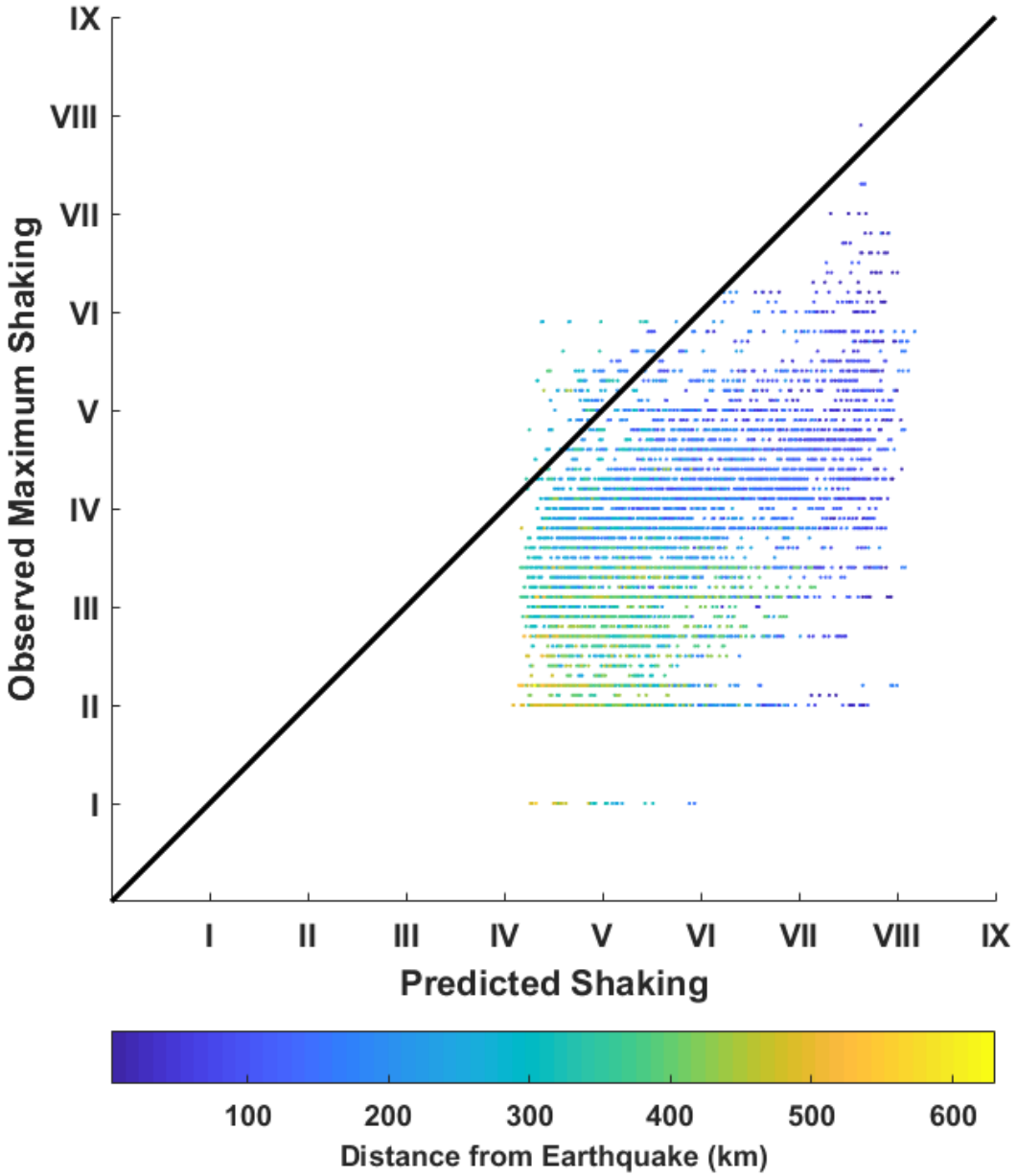


Figure 5.10. Plot comparing 2016 national seismic hazard map predictions to maximum DYFI observations for points with DYFI responses in the greater Oklahoma region, further broken down by distance from the earthquake epicenter.

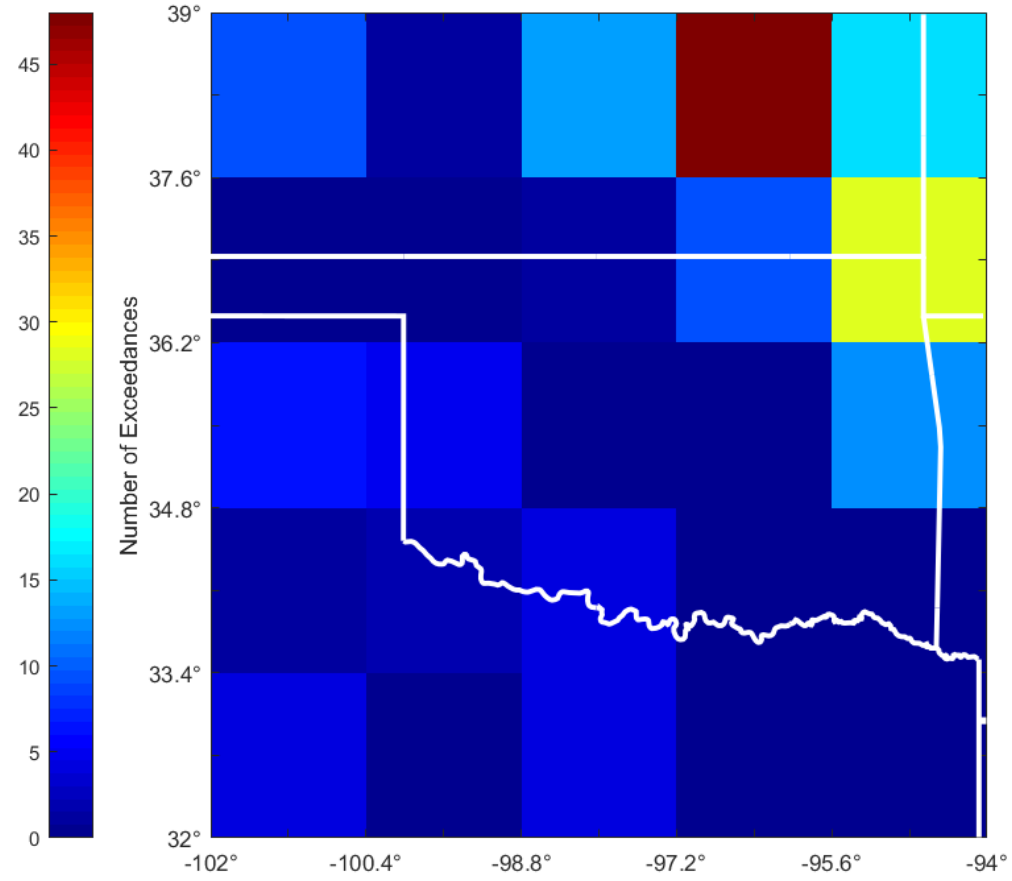


Figure 5.11. Count of exceedances in site groups. Nine of twenty-five site groups have no exceedances, showing the limit of such a fine-scale breakdown.

Finally, an analysis was completed the spatial distribution of exceedances. Sites in the greater Oklahoma region were grouped into a coarse $1.2^o \times 1.4^o$ grid, such that the expected number of exceedances in each grid square was at least five. At this small scale, exceedances cluster (Figure 5.11). Many regions have zero exceedances, leading to an incalculable fractional exceedance metric. The lack of exceedances in the south shows the

over-prediction in Dallas and north Texas. The high number in the northeast portion of the map shows where predictions were lower than the observed shaking.

5.7. Conclusions

The 2016 one-year national seismic hazard map for the central and eastern United States performed very well. It predicted the observed shaking well, both as measured by the fractional exceedance metric, and spatially, as measured by the misfit metric. Because hazard map assessment is a relatively new enterprise and only a few cases have so far been assessed there is currently no threshold defined for a "good" score on $M0$ and $M1$ metrics. This, and the related question of how well a model could realistically be expected to describe observations remain questions for future work. However, the $M0$ fractional exceedance scores for the 2016 map are far lower than those for maps in the prior chapters. The results of those studies are summarized in Table 5.1. By this view, $M0 = 0.0073$ on the national level and 0.0069 for the greater Oklahoma region indicate strong performance. The reduction to an even smaller $M0 = 0.0028$ when supplementing with ShakeMap data further reinforces that with additional information and a more thorough coverage of data, that the hazard map succeeds in what is trying to do. Furthermore, that this map succeeds without using hindcasting is more evidence of the strength of its performance.

As noted earlier, a 5% increase in the average predicted shaking for the national map would yield a perfect match between predicted and observed fractional exceedances. Such a small difference could easily occur by chance due to which earthquakes occur in the short time period sampled (Vanneste *et al.*, 2017). The $M1$ squared misfit metric also demonstrates strong spatial (and hence visual) similarity between the predicted and

Table 5.1. Comparison of metric scores from studies in prior chapters. T is return period, τ is catalog length. While $M1$ is unitless, it is not fair to directly compare scores derived from different intensity units, and thus $M1_{Japan}$ is excluded, as it was calculated using JMA intensity, not MMI.

					Metric	
Year	Country	T (years)	τ (years)	Hindcast	$M0$	$M1$
2004	Italy	2475	2220	Yes	0.59	27.2
2008	Japan	475	510	Yes	0.39	—
2016	United States	100	1	No	0.0073	4.62

observed shaking maps. A map with a score of $M0 = 0$ may not be perfect, there can easily be regions of over-prediction balanced by areas of under-prediction. However, both metrics combined suggest strong performance, both in terms of fulfilling PSHA objectives and spatial accuracy. The model benefited from the fact that the 2016 seismicity rates across this region were generally similar in Oklahoma to those observed during 2015.

The largest misfit occurred in northeastern Texas, where shaking was substantially over-predicted. This appears to reflect the limitations of the map's assumption that earthquake rates would remain relatively stationary, which would not be the case if water injection rates change due to regulatory or economic forces. While this change highlights a limitation of the model, it indicates the value of making hazard maps for such short timescales in areas where induced seismicity is a major factor, because economic and regulatory factors can change waste water injection rates rapidly (Petersen *et al.*, 2017). This situation differs from natural seismicity hazard maps, where any time-dependent (earthquake cycle) effects occur on longer timescales.

Independently assessing successive one-year maps offers the prospect of improving the models used to generate them, in that the factors contributing to the map's performance

(spatial variability, magnitude, ground motion prediction model, etc.) can be evaluated. Similarly, as more is learned about the mechanisms of induced seismicity, this information can be included in the modeling. If successive models continue to perform well— and even improve— they can be valuable tools for policy makers in managing the seismic risks associated with unconventional oil and gas productions.

CHAPTER 6

Assessing Map Performance Via Shaking Simulations

6.1. Summary

As a result of waste water injection from non-conventional oil and gas production, the central and eastern United States experienced a dramatic increase in seismicity. To better characterize the resulting hazard, the U.S. Geological Survey began producing one-year seismic hazard maps intended to capture both natural and induced seismicity as of 2016. In its first year, I found that the map performed very well, demonstrating both a good match between the observed and expected number of exceedances, and between observed and predicted shaking. I repeat this analysis for the 2017 map, using "Did You Feel It?" (DYFI) data to explore the map's performance in different regions of the country. I find that the 2017 model performed well, but not as well as the previous year's model. I explore the likelihood of observing the performance seen in 2017, by simulating earthquake shaking realizations using the assumptions of the 2017 hazard model, including a- and b-values, locations of induced earthquakes, and ground motion models. These simulations indicate a low likelihood of this decrease in performance happening by chance if the assumptions in the hazard model were appropriate. Hence, it is likely that the map's performance reflects a reduction in waste water injection rates, possibly due to regulatory and economic pressures. Future maps could benefit from better modeling how seismic rates may change year-to-year and improved ground motion models.

6.2. Introduction

Increases in non-conventional oil and gas production in the central and eastern United States (CEUS) since 2008 have resulted in significantly increased seismicity, most notably in Oklahoma and the surrounding regions (Horton, 2012; Ellsworth, 2013; Keranen *et al.*, 2013, 2014). This region historically has not experienced high shaking and is generally unprepared for the increased seismicity (Liu *et al.*, 2014; Ellsworth *et al.*, 2015).

The increased seismicity necessitated reassessment of the resulting hazard. For this purpose, the USGS produced a series of hazard maps intended to be used for one year, which focus on the hazard that results from human activity, namely waste-water injection (Petersen *et al.*, 2016, 2017, 2018a). Developing maps for one year of usage, versus a longer window like 50 years as in other maps (Petersen *et al.*, 2015), allows responses to the changes that may happen in human activities, a non-steady state variable.

The USGS's first one-year map (Petersen *et al.*, 2016) accounted for the induced seismicity by defining zones where earthquakes do not appear natural, indicated by a noticeable increase in seismicity near injection wells, both spatially and temporally. They defined separate logic trees for seismicity inside and outside these zones, which differ largely in the parameters used to describe catalog duration, smoothing distance, maximum magnitude, and ground motion models. Seismicity rates are inferred from injection rates from the prior two years.

The one-year model has an advantage for assessing the resulting performance of the map; the time necessary to gather data is not so long that one must resort to historic data instead. While "hindcasting," using historic data to assess hazard maps where catalogs of subsequent shaking don't exist, is useful (as discussed in Chapters 2-4), gathering data

generated entirely after the map was made is preferable for assessing its performance as it is a true test of a model's forecast.

In the prior chapter, I found that the 2016 model performed better than previous maps studied using "Did You Feel It?" (DYFI) data to compare seismic intensity observations to the model's predictions. Both within the entire CEUS, and in the area surrounding Oklahoma where induced seismicity is most prevalent, the data were in good accord with the model's predictions. I thus concluded that the 2016 model was a very good model. Subsequently, other studies have looked at the 2016 model and found general agreement between observation and prediction using DYFI and instrumental data (White *et al.*, 2018; Mousavi and Beroza, 2018).

For the following year, another model for 2017 was developed (Figure 6.1). The 2017 model employs the same logic trees and ground motion models (GMMs) as the 2016 model, but uses an updated earthquake catalog for the additional year of seismicity observed (Petersen *et al.*, 2017). A second year of seismic intensity records from DYFI allowed assessment of the 2017 model's performance using the metrics employed in the previous chapters.

The first metric, the fractional exceedance metric $M0$, is $M0(f,p) = |f - p|$, where p is the predicted fraction of sites where the highest shaking is expected to exceed the model's predictions, and f is the observed fraction of sites where this actually occurs. The probability p is derived from the fact that probabilistic hazard models seek to predict a level of shaking that should be exceeded only with a certain probability in some time window (Cornell, 1968; Field, 2010). At any point on the map, the probability of exceedance is given by an exponential distribution $p = 1 - e^{-\frac{\tau}{T}}$. For one year of observation,

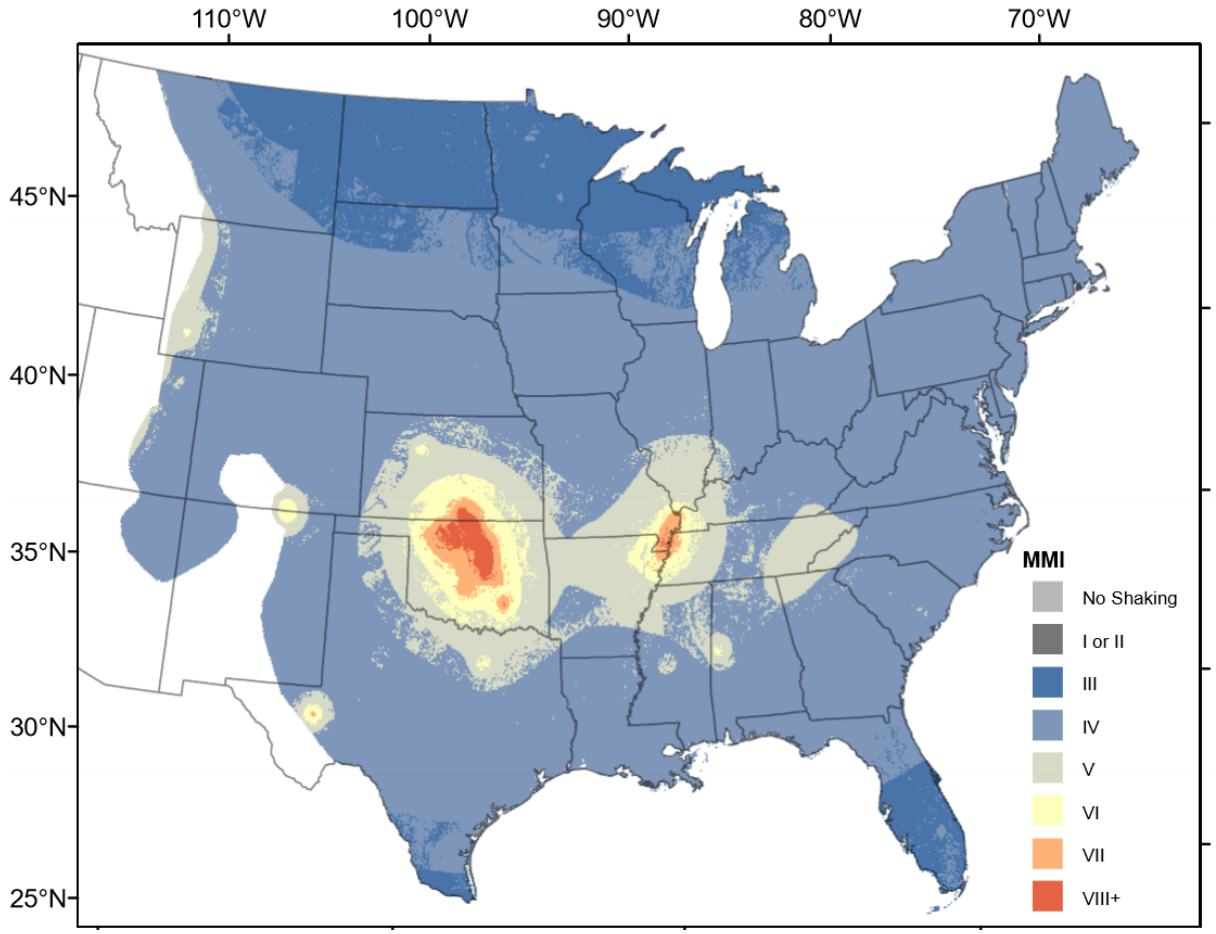


Figure 6.1. 2017 One-Year Seismic Hazard Forecast for the CEUS (Petersen *et al.*, 2017). Shaking levels are communicated in Modified Mercalli Intensity units (MMI).

$\tau = 1$, and the model assumes a return period of $T = 100$ years. Hence, the probability of exceedance p for the model is roughly $p = 0.01$.

The fractional exceedance metric is implicit in probabilistic seismic hazard analysis (PSHA). This metric is binary, and only considers whether an observation is over or under the map's prediction. Thus, an alternative is the squared misfit metric $M1$: $M1(s, x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - s_i)^2$, in which x_i and s_i are the maximum observed shaking and predicted

shaking at each of the sites $i = 1, 2, \ldots, N$. While not the goal of PSHA, this metric captures other important aspects of map performance, notably the spatial match between prediction and observation. Because a map can be successful by $M0$, but less useful overall as a map, it is better to consider both metrics to get a clearer understanding of map behavior. For both metrics, a perfect match between prediction and observation will yield a score of 0, hence higher scores reflect relatively weaker performance.

6.3. Seismicity in 2017

The 2017 model emphasizes the hazard most strongly in Oklahoma and surrounding states, mainly Texas and Kansas. This emphasis was also present in 2016. To assess the map's performance, I use the shaking record from the DYFI database (Wald *et al.*, 1999; Atkinson and Wald, 2007). A number of studies have compared DYFI data to predicted seismic hazard in recent years (Mak and Schorlemmer, 2016a; Cremen *et al.*, 2017; White *et al.*, 2018), and noted DYFI's utility for broad areal coverage (Atkinson and Wald, 2007; Hough, 2012, Mak and Schorlemmer, 2016b). Reports for individual events are geocoded by zip code, and annual summaries of maximum observed intensities are compiled on a 10 km grid (Quitoriano *et al.*, 2017). The DYFI database, specifically the annual maximum data, is one of the most thorough and robust seismic intensity data sets available, providing the most observations over the largest area.

Figure 6.2 shows the maximum shaking reported to DYFI for 2017. The map shows 17,391 sites on the one-year map where at least one report was made. Though the data are sparse, there is a match, broadly speaking, between the expected shaking and the intensity in the reports made, including the highest shaking congregated within Oklahoma. The

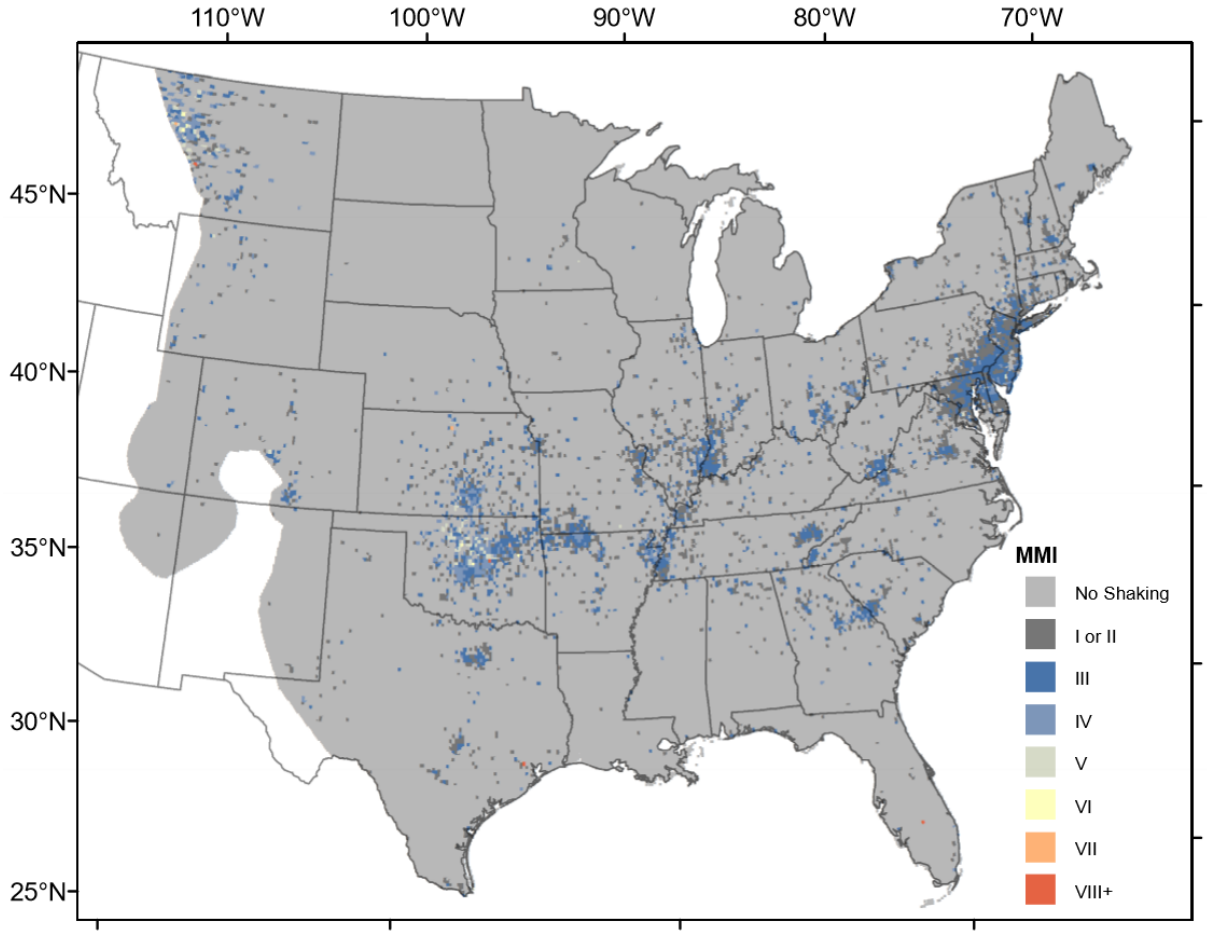


Figure 6.2. Maximum "Did You Feel It?" (DYFI) responses in 2017 for the CEUS. Gray regions indicate an absence of DYFI responses, but do not necessarily imply a lack of shaking.

map also features many reports to the east in the Pennsylvania/Delaware/Jersey tri-state area, and to the west in Montana. These reports are geographically consistent with the location of the largest earthquakes observed in the CEUS in 2017, shown in Figure 6.3.

Figure 6.3 shows that despite the expected high seismicity in the greater Oklahoma area, the largest earthquake in the CEUS in 2017 was a M 5.8 event near Lincoln, Montana (McMahon *et al.*, 2017). Similarly, the east coast experienced a M 4.1 earthquake in

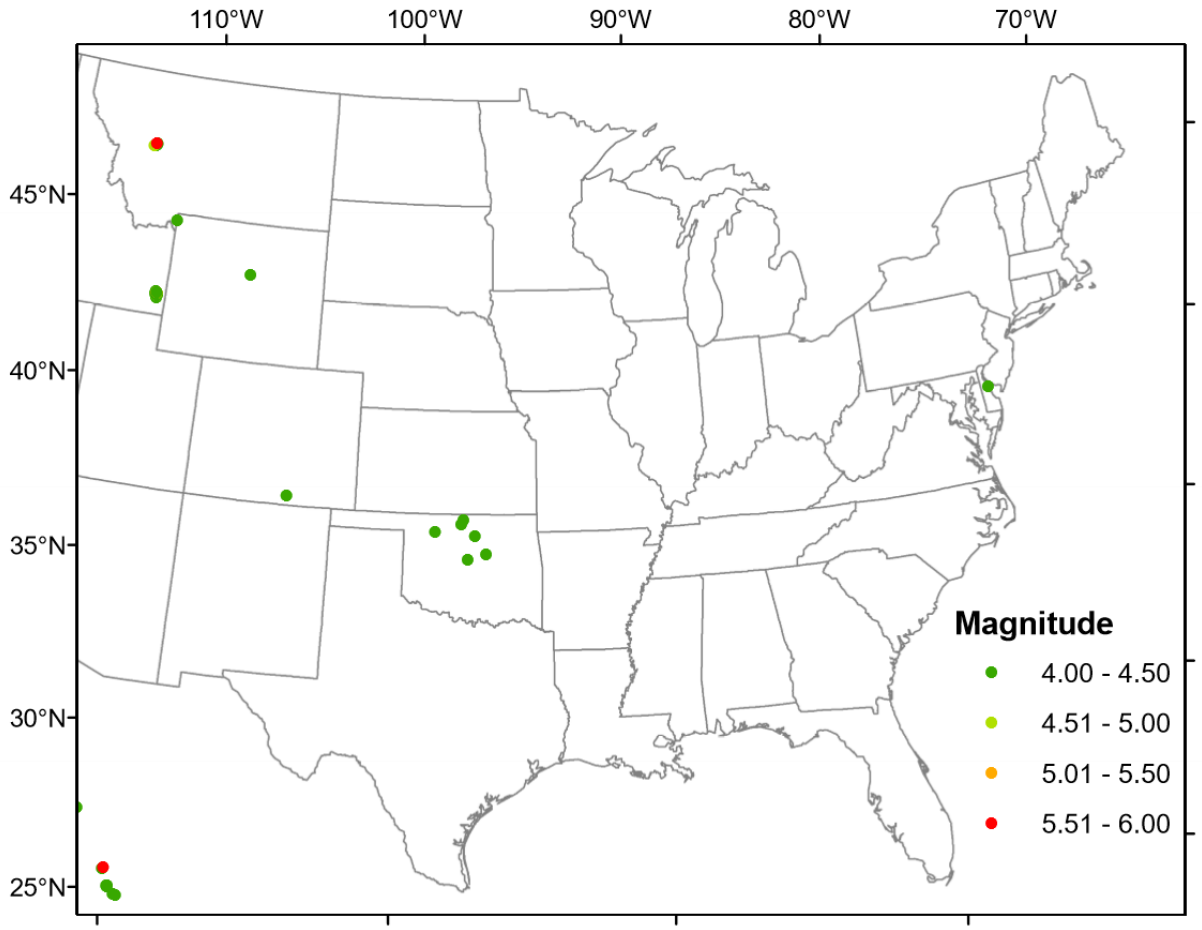


Figure 6.3. Occurrence of M4+ earthquakes near the central and eastern United States in 2017. Earthquakes in Oklahoma, Montana, and Delaware were the primary generators of DYFI responses.

December 2017, located in Dover, Delaware, where seismicity is expected to be low. Conversely, Oklahoma experienced only a handful of large seismic events, especially in comparison to the previous year, when it experienced a number of high-shaking events, including the largest recorded in its history, the M 5.8 Pawnee event (Yeck *et al.*, 2017). While the distribution of these 2017 events differs from the previous year's, where large events occurred in Oklahoma and small events were located elsewhere, they provide an

opportunity to see how well a map performs when a number of "black swans," rare and unexpected scenarios, occur (Stein *et al.*, 2012).

6.4. DYFI and Map Performance

Figure 6.4 illustrates the map's overall performance. From the $N = 17391 responses$, we see that 501 sites reported shaking exceeding that predicted. This corresponds to $f = 0.0288$, hence for $p = 0.01$, $M0 = 0.0188$. Furthermore, $M1 = 5.39$, a relatively low score reflecting a reasonable spatial correlation between the maximum shaking at sites that responded to DYFI and the predicted maximum shaking there. The prior chapter found that for the 2016 model, the CEUS had $M0 = 0.0073$ and $M1 = 4.62$. While the previous year's lower scores indicate a slightly better performance overall, in general they are similar. Although this is the first study focusing on successive iterations of maps, prior studies (Chapters 2 through 4) have found maps that yield both $M0$ and $M1$ orders of magnitude higher.

For areas where predicted intensity is high (e.g. intensity VII+), the range of observed shaking spans intensity II to VIII. A similar trend is present for low-intensity predictions (i.e. intensity III to V). However, between these two regions (e.g. intensity V to VII), observed shaking levels are predominantly lower. As a result, most exceedances come from sites where shaking was predicted to be relatively low. Most of those exceedances cluster around the 45^o line that marks a perfect match between observation and prediction, suggesting that most exceedances are small. Roughly two dozen points are exceedances where intensity is VII+. Further insight into spatial variations in map performance as a

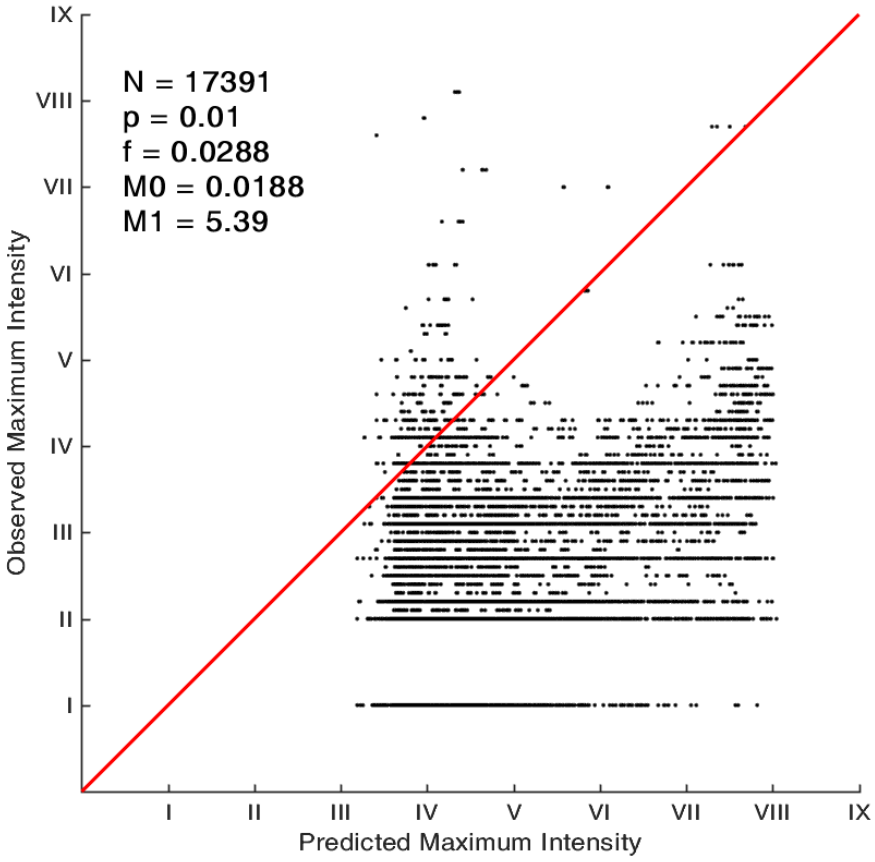


Figure 6.4. Comparison of predictions from the 2017 hazard map to maximum reported intensity for the entire central and eastern United States.

function of the DYFI data can be had by subdividing Figure 6.1 into smaller geographic regions.

Figure 6.5 (top row) shows the predicted and observed shaking, and metric score for the greater Oklahoma area, the region where induced seismicity in the US is the highest. The greater Oklahoma region contains the majority of locations where observed shaking is well below model predictions. There are only eight exceedances from 3,410 observations, which leads to an observed exceedance fraction $f = 0.0023$ and $M0 = 0.0077$, similar

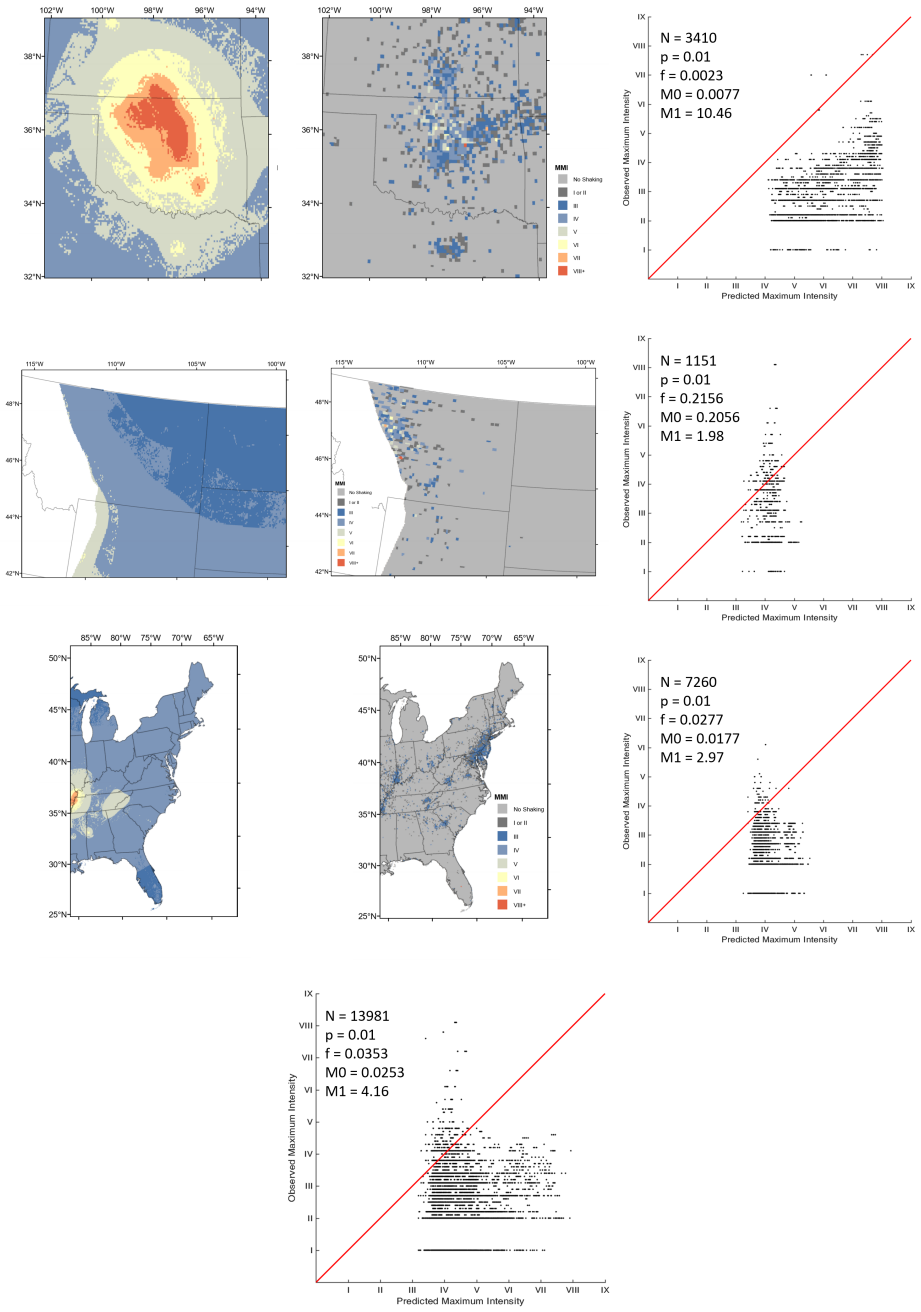


Figure 6.5. Predicted (left column) and observed shaking maps (middle column), and predicted-observed plots for the greater Oklahoma area (top row), Montana (second row), and the east coast of the US (third row). Fourth row shows observed-predicted plot for the entire CUES except the greater Oklahoma area.

to the 2016 model score. The 2016 model under-predicted shaking and the Oklahoma region experienced more exceedances than expected. In 2017, the opposite occurred, and shaking was over-predicted. With so few exceedances in the area, the significance of a lower $M0$ score is lessened (discussed in section 5.6). In contrast to the similarities in $M0$ between 2016 and 2017, 2017's greater Oklahoma $M1 = 10.46$, substantially higher than for the entire CEUS in 2017 and the CEUS and greater Oklahoma area $M1$ scores for 2016 (in 2016, $M1_{CEUS} = 4.62$, and $M1_{OK} = 5.01$).

The effects of the low shaking reported in Oklahoma can be highlighted by examining the CEUS metrics excluding the greater Oklahoma area. The bottom row of Figure 6.5 shows observed versus predicted shaking for the opposite of the top row, i.e. the entire CEUS without the box in the top of Figure 6.5. The map lacks the exceedances that occurred at higher predicted maximum intensities, because the areas with the highest predicted shaking due to induced seismicity are removed from the map. Hence, there are far more exceedances from areas with lower predicted intensities, largely the Montana and Delaware earthquakes. Oklahoma shaking was heavily over-predicted, but the map as a whole under-predicted shaking, so removing Oklahoma from the data set yields a larger fraction of site exceedances. Hence $M0$ increases to 0.0253, a decrease in map performance. However, the large mismatch in the intensity of the shaking predicted in Oklahoma is reflected by the squared misfit $M1$, double the CEUS $M1$ score. Thus, removing the greater Oklahoma area from the data improves the map performance by the squared misfit metric, so $M1 = 4.16$.

Additional subdivisions illustrate other strengths and weaknesses of map performance. Consider responses to the Montana earthquake in the northwest, shown in the second row

of Figure 6.5. Because this was an unexpectedly large earthquake for the area, it led to a high number of exceedances. Over 20% of sites reported shaking exceeding the predicted levels, so $M0 = 0.2063$. However, this number may be artificially inflated by a lack of distant responses in areas where shaking may not have exceeded predictions, perhaps due to low population. The few reports from distant areas that do describe shaking are close to the predicted values. As a result of this strong match, this region of the map scores $M1 = 1.98$. This is notably lower than any other $M1$ scores generated for 2017, indicating that the map generally succeeds at matching predictions to observations.

Finally, the third row of Figure 6.5 shows the eastern half of the CEUS map, where the seismicity is largely non-induced. The region tends to be very aseismic, though there is a history of events happening along the coast in the past (Hough, 2012; Wolin *et al.*, 2012). This portion of the map contains 40% of the data reported in the entire CEUS but has $M0 = 0.0177$, similar to that of the entire region. The data lack large exceedances, and while there are some instances of over-predicted shaking in Tennessee and the New Madrid region, broadly speaking there is a good visual match between predictions and observations. Hence, a low squared misfit arises, with $M1 = 2.97$.

The metric calculations indicate that data fit the 2017 map's predictions reasonably well, although not as closely as the 2016 data fit that map. This map performs better than all previous studies conducted of hazard maps using the metrics approach, with one exception: the 2016 one-year map. This difference opens questions of how much variability to expect in performance from year to year, and whether the poorer performance in 2017 is likely to have arisen by chance or instead reflects a meaningful change in seismicity. Furthermore, the decrease in responses, which may be a function of lower

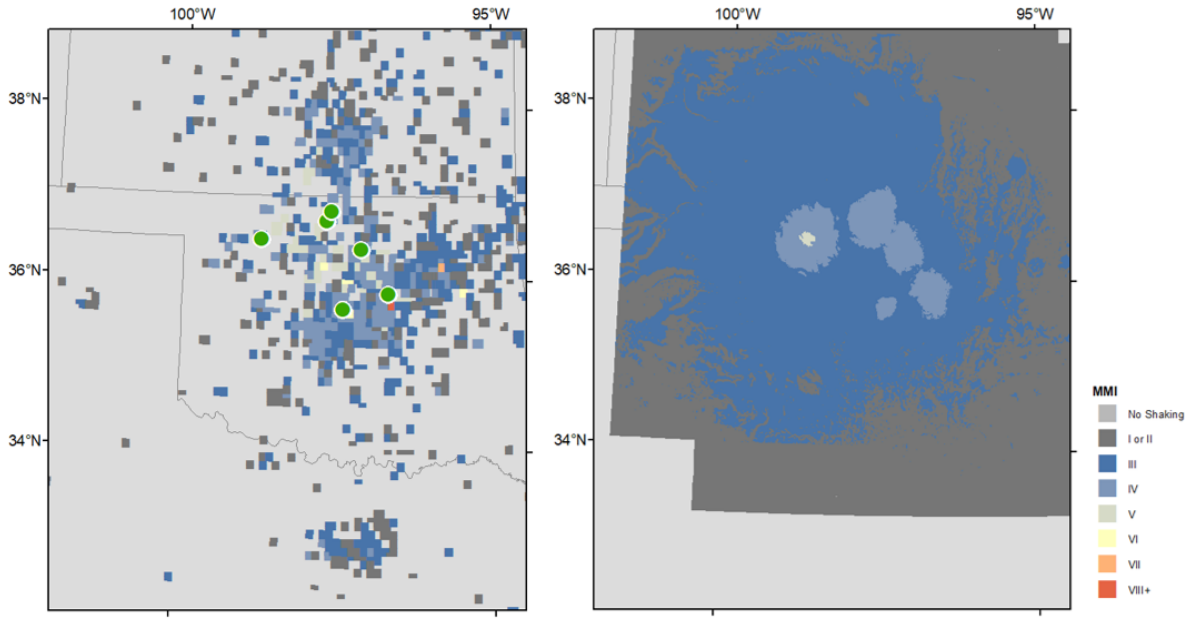


Figure 6.6. Comparison between shaking reported to DYFI (left) and predictions from ShakeMap (right). Earthquakes in Oklahoma M 4+ are marked with green circles on the left. They are excluded on the right to not obscure contour changes, but are located at the center of each local maxima in the ShakeMap predictions.

shaking, low population at the epicenters of events, or "earthquake fatigue" driving down response rates— or a combination of all three— raises questions regarding data quality and completeness (Mak and Schorlemmer, 2016b).

Focusing specifically on Oklahoma, between 2016 and 2017 the number of responses catalogued in the maximum shaking felt per year in the DYFI system dropped by nearly two thirds. 15% of the region had shaking reported in 2017, down from 40% in 2016. Figure 6.6 shows the absence of responses by comparing the DYFI map to what ShakeMap predicted the intensity of shaking would be (Wald *et al.*, 2005). I found in the study of

the previous year's map that, in absence of information, ShakeMap is a suitable approximation of a lower bound for reported shaking. While there are some regions where DYFI shaking exceeds ShakeMap predictions, there are many more areas devoid of responses that ShakeMap suggests should have felt shaking, particularly towards the western part of Oklahoma. This is likely due to issues in population density compounding issues with decreased response rate. This decrease in reports is also accompanied by a decrease in the intensity of shaking reported. The median maximum annual shaking reported in 2017 is 2.9, down half an intensity unit from the median in 2016, 3.4.

6.5. Simulating Shaking

To address the questions of whether low responses are a function of low shaking levels or low response rates, I use Monte Carlo simulation to characterize the variability of possible shaking histories for the greater Oklahoma area in 2017. This approach is similar to that of Vanneste *et al.* (2018). The simulations can address the likelihood of the observed decrease in shaking occurring by chance, explore data incompleteness issues, and give insight into how the metrics used for assessment describe the map's performance. For simplicity, the simulations use Oklahoma as the region of interest, and so did not need to account for multiple zones of induced seismicity across CEUS, or the effect of natural seismicity in areas like New Madrid region. Consider the four random processes that define the maximum shaking experienced in a year:

(1) Where the earthquakes occur

(2) How many earthquakes occur

(3) The magnitude of an earthquake

(4) Uncertainties in the ground motion models that describe shaking

The first three processes control the distribution of earthquakes in the region, and can be described simply. Petersen *et al.* (2016) define regions where waste water injection is expected to lead to induced seismicity. Consider this area to be uniformly susceptible to induced earthquakes.

In the models, the rate of induced seismicity is defined for an upcoming year as a weighted average of the past two years, with weights of .8 and .2 for the most recent year and the previous year, respectively (Petersen *et al.*, 2014; Petersen *et al.*, 2017). For 2017, given a minimum magnitude of completeness of 2.7, we observe 162 independent (e.g. declustered, with no aftershocks) earthquakes from 2016 and 152 earthquakes in 2015 for the Oklahoma induced zone (Petersen *et al.*, 2018b). Hence the expected number of earthquakes $\lambda = 160$. With this λ, it is simple to model the number of possible earthquakes that can occur as a Poisson random variable.

Earthquake magnitudes are assigned based on the Gutenberg-Richter relationship, assuming $b = 1$, as the 2017 model does (Petersen *et al.*, 2017). Zhuang and Touati (2015) define a method for assigning magnitudes, given a b-value and a minimum magnitude for completeness, using inverse transform sampling. An event's magnitude, m, is randomly generated, assuming that

$$m = -\frac{1}{b \ln 10} \ln U + m_0, \tag{6.1}$$

where b is the model's b-value ($b = 1$), U is a value obtained randomly from a uniform distribution on $[0, 1]$, and the minimum magnitude for completeness is $m_0 = 2.7$.

For a more accurate characterization of the 2017 earthquake record, this process of simulating earthquake occurrence is repeated a second time, using $\lambda = 4$, to include earthquakes that happened outside the defined box of induced seismicity.

Simulating these first three random processes generates many realizations of seismicity in Oklahoma in 2017. To describe the resulting ground shaking, I use the GMMs used by Petersen *et al.* (2017). Nine different models with varying weights (Petersen *et al.*, 2014) are used and then aggregated to describe ground shaking (Frankel *et al.*, 1996; Toro, 2002; Silva *et al.*, 2002; Campbell, 2003; Tavakoli and Prezeshk, 2005; Atkinson and Boore, 2006; Atkinson, 2008; Prezeshk *et al.*, 2011; Atkinson, 2015). Because the resulting shaking is given as PGA, and the model being assessed predicts shaking as MMI, the same techniques used in Petersen *et al.* (2017) to convert PGA to MMI are employed (Worden *et al.*, 2012).

Each ground motion model, as well as the conversion from PGA to MMI, has an error term in the form of a Gaussian random variable. The error terms for ground motion are treated as uncorrelated at each site, and can be treated as representing the uncertainty in each model, as well as the influence of site effects, directionality, or other modifiers to shaking. The PGA to MMI conversion error is correlated, and assumed to be equal at all sites. Using the GMMs, the shaking from each earthquake in a given realization is calculated. After calculation for all events' shaking, the maximum "observed" shaking at each site, gridded on a 10 km scale, is selected and used to calculate the metrics for map performance.

This procedure generated 10,000 simulations to explore the full range of outcomes, and call this model which allows for full variance of all parameters "unconstrained."

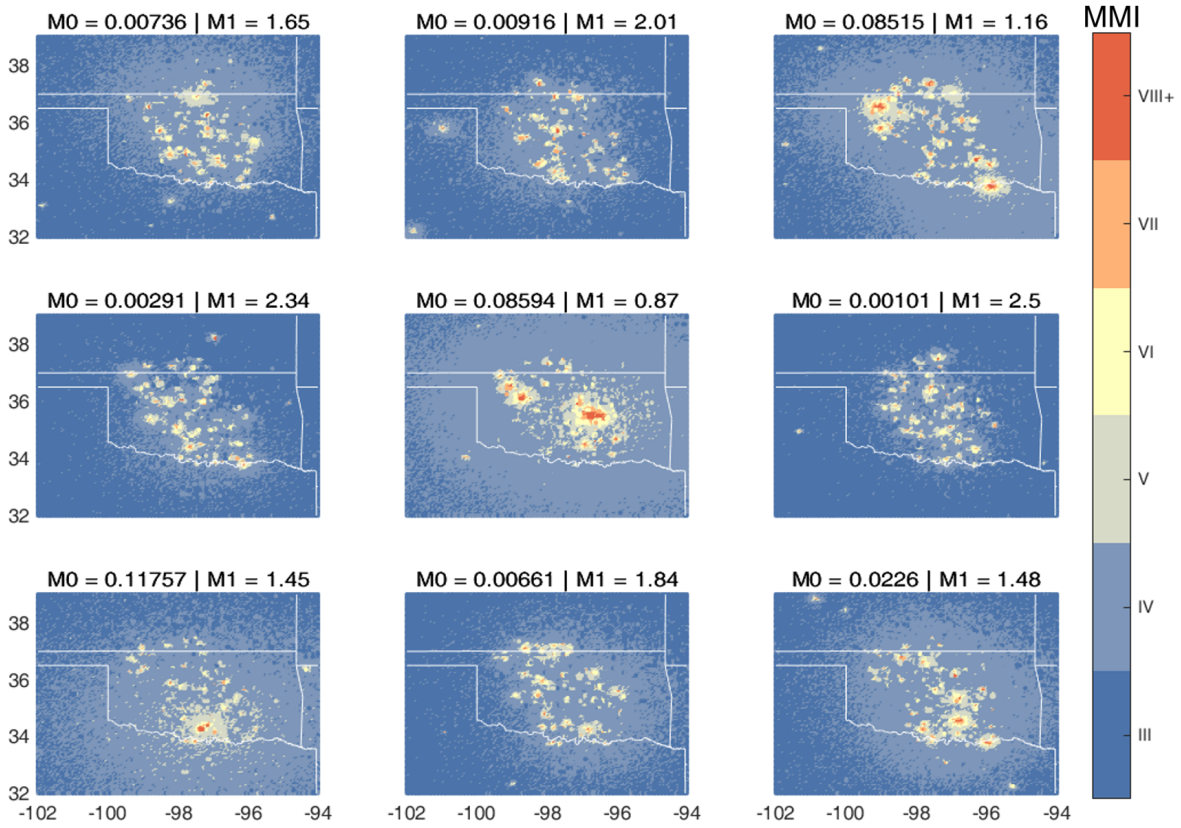


Figure 6.7. Simulation outputs, illustrating shaking and metric results, for nine different realizations of 2017. The shaking model is fully unconstrained, with randomness in earthquake count, location, magnitude, and in the GMMs. Intensity is reported in MMI.

Figure 6.7 shows nine of these realizations, and the metric calculations associated with each. Intensity is calculated at each site within the greater Oklahoma region. Hence, the metrics here show performance that would arise with a 100% response rate.

The metrics calculated for the 10,000 "unconstrained" simulations— where uncertainties in earthquake count, location, magnitude, ground motion models, and PGA to MMI conversion are allowed to vary— are shown in Figure 6.8. The results show generally low $M0$ scores, reflecting a tendency for $f \approx p$. A tail drops off for larger $M0$ scores,

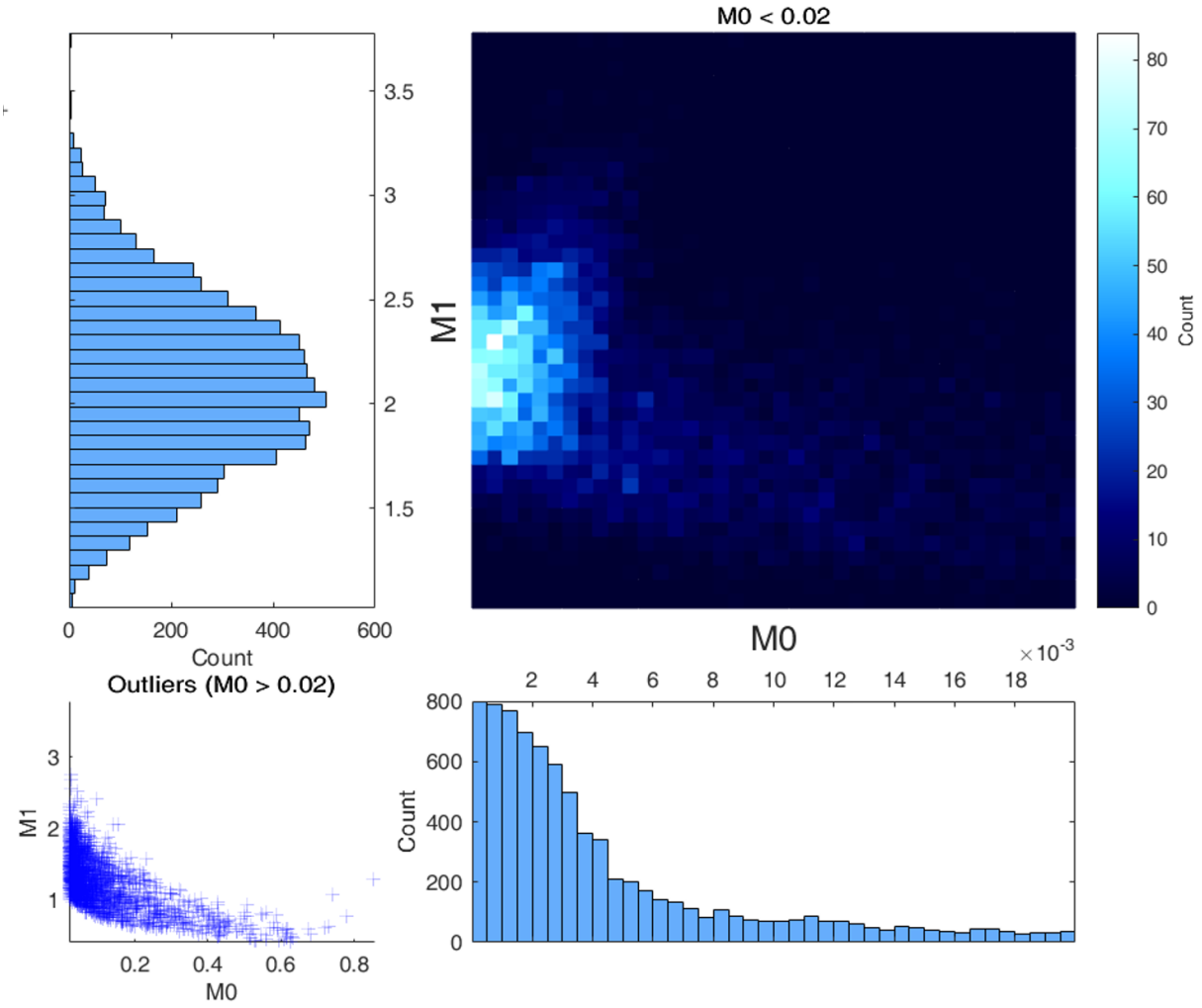


Figure 6.8. Heat map of distribution of metric scores for 10,000 realizations of 2017 seismicity. The x-axis is the fractional exceedance metric, $M0$, and the y-axis is the squared misfit metric, $M1$. Each axis has a histogram of each metric's distribution, independent of the other. Outliers, defined by $M0 > 0.02$, are plotted in the bottom left.

showing the small possibility of achieving a very large score, indicative of scenarios where $f >> p$. Counts of $M0 > 0.02$ decrease substantially, and are shown by an extra plot beneath the heat map. Though the values reach as high as $M0 = 0.8$, most simulations have $M0 \leq 0.02$. 80% of the simulations fall below this cutoff, 95% of the simulations

result in metrics where $M0 \leq 0.15$. Generally speaking, this distribution agrees with the 2017 DYFI result ($M0 = 0.0077$). The squared misfit metric $M1$ is roughly characterized by a normal curve centered around a mean $M1 = 2.09$, a very low score compared to the 2017 DYFI data ($M1 = 10.46$). The standard deviation for this curve is 0.41. There is an inverse correlation between the fraction of sites that exceed predicted shaking and $M1$. This can be observed by the two trend lines that grow out of the dense grouping of points centered around $(M0, M1) = (0, 2.09)$, which is where $f = p = 0.01$. The sharp upward trend terminates at $M0 = 0.01$, the point where no exceedances are observed ($f = 0$). The shallow downward trend continues for all $M0$ values in the heat map and outlier plot. Hence, as f increases, the squared misfit decreases. While only the fractional exceedance metric is implicit in the definition this hazard map, this result suggests it should be possible to minimize both metrics for a given set of predictions and expected number of exceedances.

The metrics obtained with DYFI data for 2017 ($M0 = 0.0077$ and $M1 = 10.46$) do not fall within the simulation's distribution due to the high squared misfit metric, $M1$, though they are in reasonable accord with the distribution of the fractional exceedance metric, $M0$. However, the discrepancy between the metrics for DYFI and the simulations may result from incompleteness in the DYFI data. Some of the largest earthquakes in Oklahoma in 2017 have few DYFI responses (Figure 6.6), and some of the neighboring responses report MMI II- intensities, which seem far too low for their proximity to some events' epicenters. It appears there is an issue with the response rate of the DYFI data, resulting in many inconsistent and missing responses.

It is not a fair comparison to compare the metrics for the DYFI data, given the low response rate, to those for simulated data, where data coverage is perfect. In Chapter 5, it was shown through the addition of ShakeMap data that an inverse relationship exists between response rate and $M1$. Thus, while both DYFI and simulation can be used to assess performance individually, it is unclear whether a comparison between the two is useful for comparing aspects of map performance, or the relationship between $M1$ and the number of responses.

Hence, to better address data discrepancies and generate a more consistent comparison of map performance, the same simulation procedure is repeated, but with fixed earthquake count, location, and magnitude to the values for declustered events that occurred in 2017, considering only the uncertainties in ground motion and PGA to MMI conversion. By repeating this analysis with these fixed parameters, a comparable data set of observations to contextualize the results of the simulation is made. This approach is analogous to using ShakeMap data, and approximates having DYFI data with a 100% response rate for the earthquakes observed in 2017, but has the added benefit of also incorporating the uncertainty in the ground motion models.

Figure 6.9 shows nine of the 5,000 shaking simulations calculated with the observed earthquake catalog in Oklahoma in 2017. Half the number of simulations are made to reduce computation time, as substantially smaller variance in the output is expected when there are many fewer input variables. Whereas the original simulations model is termed "unconstrained," these new simulations are considered to be a "constrained" simulation model. The variance in shaking output between simulations appears to be much smaller because they simulate the same events so ground motion varies less between simulations.

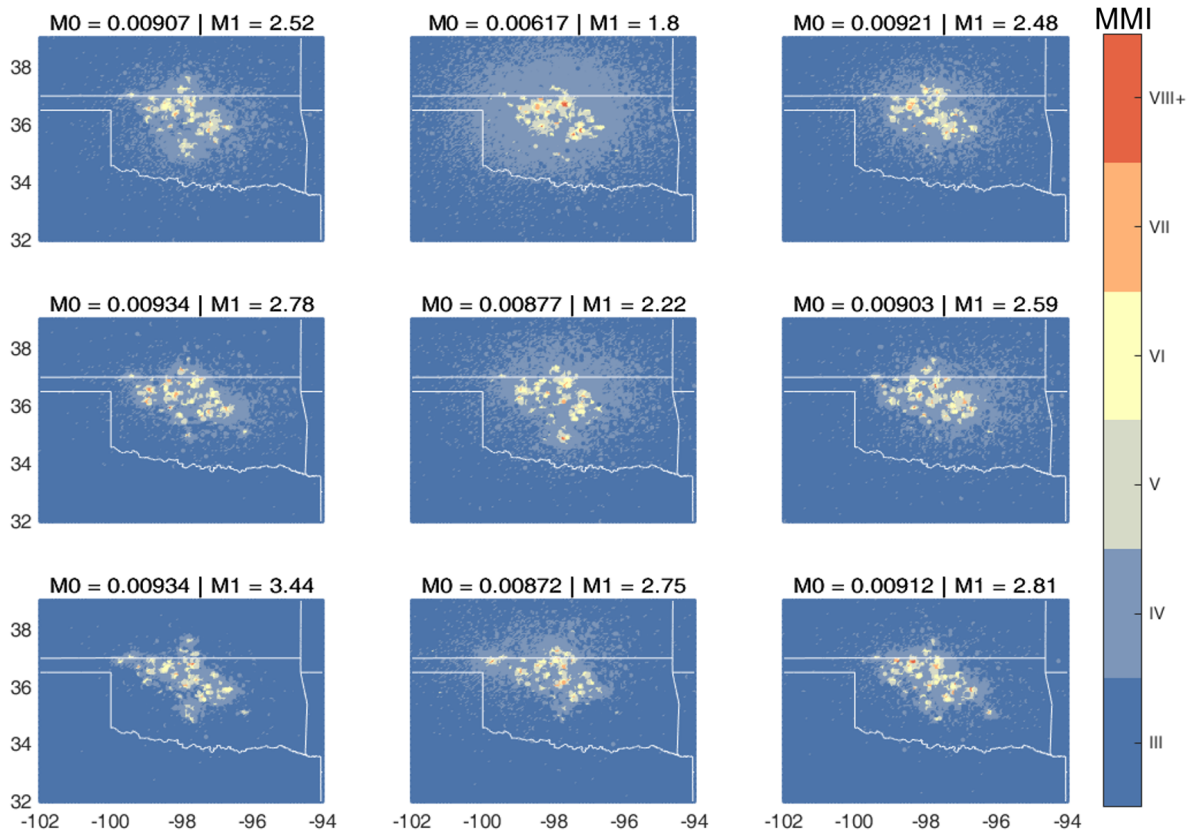


Figure 6.9. Sample of simulation outputs, illustrating shaking and metric variability for the observed earthquake record in 2017. These constrained simulations have randomness only due to the GMMs. Intensity is reported in MMI.

Figure 6.10 shows the variability in the metrics for the constrained simulations, which is due only to GMM uncertainty. $M0$ is tightly clustered between 0.008 and 0.01, differing strongly from the distribution of $M0$'s for the unconstrained model. 97% of the simulations have $M0 \leq 0.02$. $M1$ in the constrained simulation has a slightly larger standard deviation than the unconstrained model, with a wider normal curve centered at the mean $M1 = 2.33$. The standard deviation of the $M1$ distribution is 0.48. The

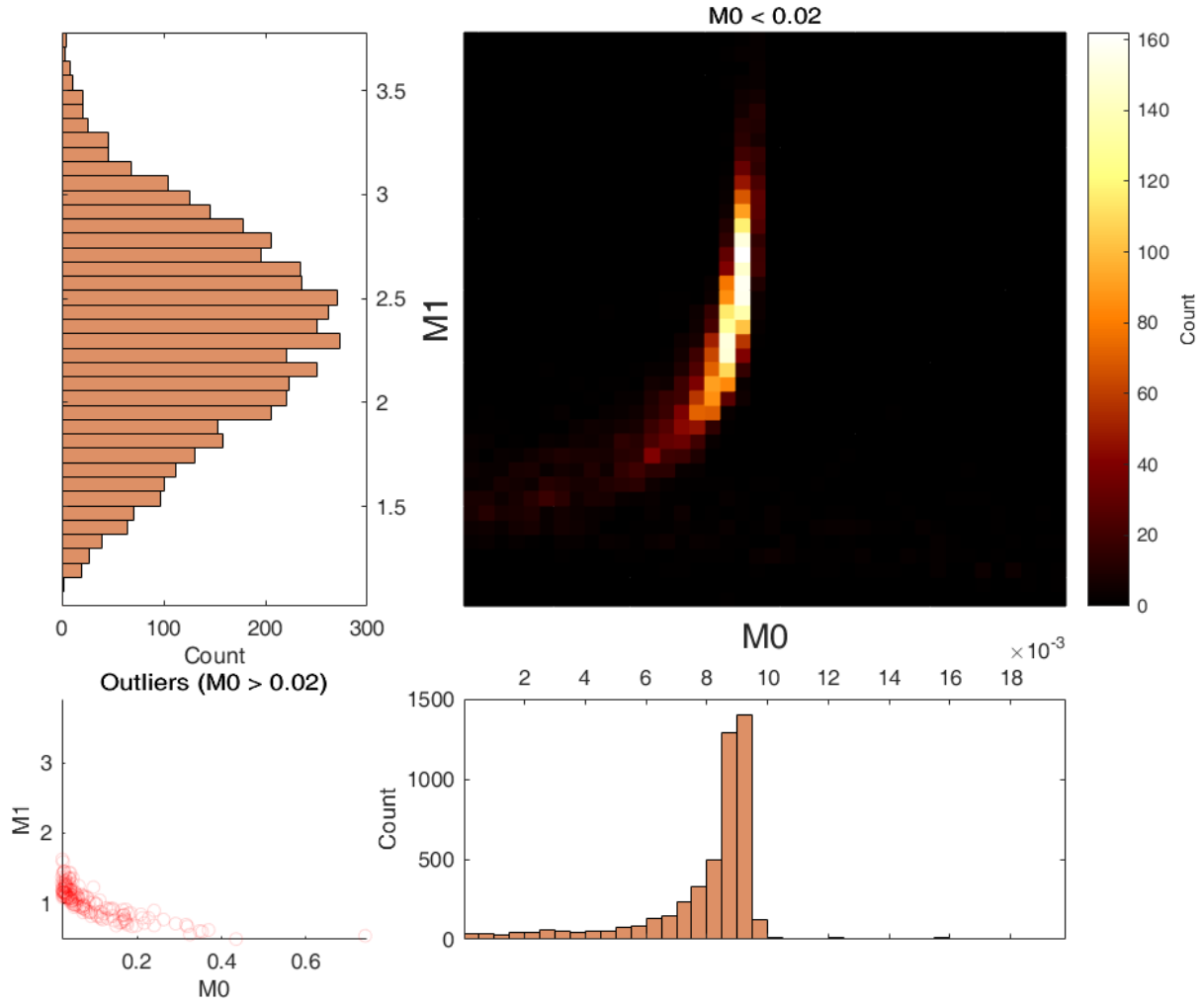


Figure 6.10. Heat map of distribution of metric scores from for 5,000 realizations of 2017. The x-axis is the fractional exceedance metric, $M0$, and the y-axis is the squared misfit metric, $M1$. Each axis shows a histogram of each metric's distribution, independent of the other. Outliers, defined by $M0 > 0.02$, are plotted in the bottom left.

outliers, $M0 > 0.02$, are far fewer in number, because the lower shaking produces fewer exceedances.

The results from both sets of simulations are superimposed in Figure 6.11. Counts are normalized to show relative frequency, illustrating the likelihood of getting a specific

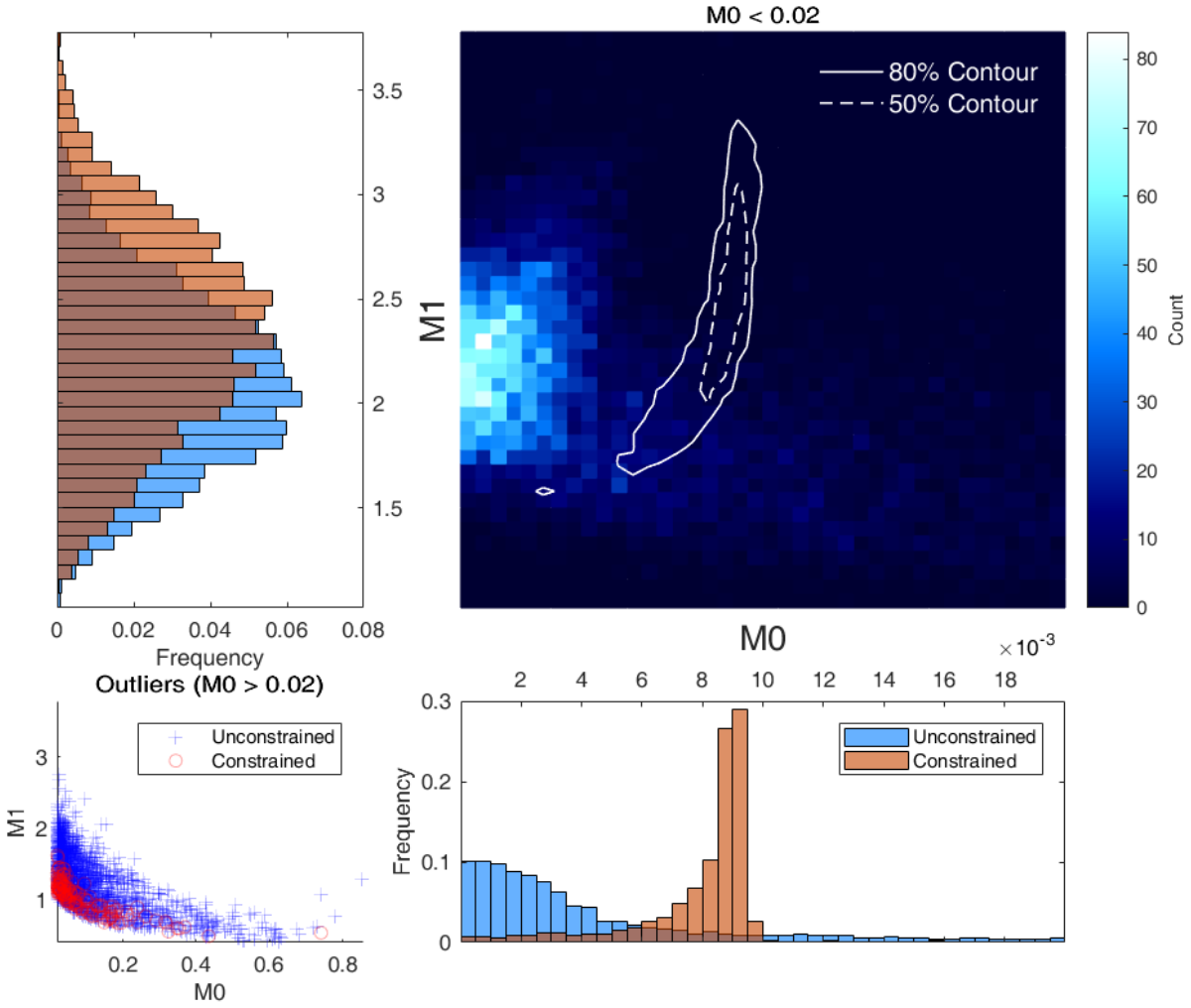


Figure 6.11. Superposition of unconstrained (background heat map, blue histogram) and constrained simulations (white contour lines, orange histogram). Contour lines represent where 50% and 80% of the data reside. Histograms shows relative frequencies in the distribution of each metric. Outliers ($M0 > 0.02$) are plotted on the bottom left.

metric result from the unconstrained and constrained simulations, and comparing the distributions of the two. Through this comparison, we examine the likelihood that the higher misfit observed in 2017 could be attributed to either bad luck, or a flawed assumption in the underlying model (Stein *et al.*, 2011). A model that can accurately forecast

the shaking for a given year would be expected to have a metric score that falls within the range of metric scores achieved from a fully unconstrained simulation. In essence, we would expect what we observe to be one of the scenarios predicted by the unconstrained model. However, the contours of the constrained model, representing 50% and 80% bounds on the data, are displaced upward for both metrics, giving higher scores, indicating weaker performance than what might be expected from a random scenario from the unconstrained simulation. Though it is possible, based on the distribution of the unconstrained simulation, to have an "expected" earthquake scenario result in poor map performance (as indicated by the simulations that yielded very high metric scores, indicative of even weaker performance), there is little overlap between the constrained (contours) and the unconstrained (heat map) simulation results. Hence, from this comparison, it appears that the poorer performance of the 2017 map arises not from bad luck, but a flawed assumption. Some physical aspect of shaking is not accurately described by the model for the map, otherwise one might expect to see the constrained simulations overlap with the unconstrained output. Specifically, the tight clustering of the fractional exceedance metric around the predicted number of exceedances ($p = 0.01$) seems to suggest that the earthquakes that occurred in 2017 were insufficient to generate shaking large enough to cause exceedances. This lack of large shaking can also explain the upward shift in the squared misfit metric, $M1$. There were too large events in 2017. Larger events would generate lower scores for both $M0$ and $M1$. This conclusion is reinforced by the issues in the DYFI data, which similarly suffers due to the lack of large shaking events necessary to generate a broad response of quality reports.

6.6. Conclusions

Comparison of shaking observed in 2017 to that predicted by the 2017 hazard model shows an over-prediction of shaking. The shaking record for 2017 contained so little shaking that it generated essentially no exceedances. This is in stark contrast with 2016, which had many exceedances throughout the Greater Oklahoma Area, due to numerous large events, specifically three M 5+ episodes, including the M 5.8 Pawnee earthquake. These large events, and moderate to large shaking episodes in general, dominate the maximum shaking record.

The greatest mismatch between prediction and observation for the 2016 model was in northern Texas (Chapter 5). Substantial shaking in Dallas did not occur as predicted, and maximum DYFI reports in the area were linked to distant larger earthquakes in Oklahoma. Waste water injection has been found to be linked strongly with Texan seismicity (Hornbach *et al.*, 2016), and though earthquakes may persist after waste water injection has halted, rates decline following the closure of an injection site (Ogwari *et al.*, 2018). I believe the change in seismicity that follows a reduction in waste water injection rates cause an increase in metric scores.

A similar effect leading to a decrease in large events may be going on in Oklahoma. Regulatory efforts have capped waste water injection rates in Oklahoma, leading to a gradual decline in the number of larger earthquakes. Furthermore, oil prices and earthquake rates are correlated (Roach, 2018), so the sharp decline in prices since 2014 (Prest, 2018) may influence rates. Combined economic and regulatory pressures thus led to the decrease in the maximum shaking observed in Oklahoma.

It appears the parameters used to predict seismicity in the 2017 hazard model did not fully account for these changes. B-values in induced zones may appear higher amid the earthquake swarms that occur, which may be occurring presently (Goertz-Allman and Wiemer, 2012). An increase of the b-value in this setting would decrease the likelihood of observing higher magnitude events, a possible explanation for what was observed in 2017. Despite this, the b-value for the 2017 model, as well as prior years, was set to $b = 1$. Further gains in performance can come from improving GMMs. The ground motion models used for the 2017 map are derived for scenarios that may not apply to the Oklahoma region, including non-induced seismicity and the tectonic setting of the western United States. More localized ground motion models may reduce the very large uncertainty that contributes to the misfit in model performance (Novakovic *et al.*, 2018; Moschetti *et al.*, 2018).

While there is room for improvement in the hazard model, the resulting map is still useful as a whole. The metric scores, as calculated with the DYFI data, tend to be similar to the performance of the previous year's model, with only slightly worse fractional exceedance metrics than for the 2016 model. All regions except Oklahoma, where seismic rates are assumed to be better known and stable, have consistently small squared misfit metric values. As a whole, these results are better than many maps we have analyzed by this approach (Chapters 2 through 4). Hence, despite weaker performance compared to the 2016 map, I believe the 2017 model is a good map. This conclusion is reinforced by the results of the seismicity simulation, which shows performance weaker than the previous year's map, but stronger than performance from our other studies. While the mismatch in simulated metric distributions appears to reflect assumptions that could be improved,

the model's performance is still better, with a much smaller discrepancy between observed and predicted shaking, than that of many of other maps for natural seismicity assessed previously. For the purposes of mitigating risk and anticipating shaking in the future, the 2017 model can still inform users about the hazards posed by waste water injection and other seismically inducing activities.

Beyond assessing the performance of this specific map, these results have implications for the general issue of how to assess and improve earthquake hazard maps' performance. The simulations approach is useful for filling in gaps in data and exploring the uncertainty in a map's predictions. Furthermore, the metrics defined in Chapter 2 were intended to be used as a comparative tool to assess the performance of different maps. Through work on many maps, a general understanding of what constitutes high and low metric scores may become apparent, though it is harder to assess a map's performance with no basis for a comparison. Simulation is a tool to address this problem, allowing for comparison of many different shaking realizations. The simulations allow better understanding of map performance if there are no comparisons to be made, and for a better understanding of the likelihood of observing a single outcome (Vanneste *et al.*, 2018). This is important for better using metrics to assess map performance in that it will let researchers move from measuring relative performance to measuring absolute performance. Further research on these analyses can push how the metrics are used, so rather than asking if a map is "better," we can ask if a map is "good." These advances will help researchers better understand the performance of maps and thus how to better use them for earthquake hazard mitigation.

References

[1] Albarello, D., and D'Amico, V. (2008). Testing probabilistic seismic hazard estimates by comparison with observations: an example in Italy, *Geophys. J. Int.*, **175**, 1088-1094.

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

[3] Atkinson, G. M., and Boore, D. M. (2006). Earthquake ground-motion prediction equations for eastern North America. *Bull. Seismol. Soc. Am.*, **96(6)**, 2181-2205.

[4] Atkinson, G. M., and Wald, D. J. (2007). "Did You Feel It?" intensity data: A surprisingly good measure of earthquake ground motion. *Seismol. Res. Lett.*, **78(3)**, 362-368.

[5] Atkinson, G. M. (2008). Ground-motion prediction equations for eastern north America from a referenced empirical approach: implications for epistemic uncertainty. *Bull. Seismol. Soc. Am.*, **98(3)**, 1304-1318.

[6] Atkinson, G. M. (2015). Ground-motion prediction equation for small-to-moderate events at short hypocentral distances, with application to induced-seismicity hazards. *Bull. Seismol. Soc. Am.*, **105(2A)**, 981-992.

[7] 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.*, **98**, 509-520.

[8] 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.*, **100**, 3329-3331.

[9] Boatwright, J., and Phillips, E. (2017). Exploiting the demographics of "Did You Feel It?" responses to estimate the felt area of moderate earthquakes in California. *Seismol. Res. Lett.*, **88(2A)**, 335-341.

[10] Box, G. E. P. (1979). Robustness in the strategy of scientific model building. *Robustness in Statistics*, **1**, 201-236.

[11] Campbell, C. J., and Laherrère, J. H. (1998). The end of cheap oil. *Scientific American*, **278(3)**, 60-5.

[12] Campbell, K. W. (2003). Prediction of strong ground motion using the hybrid empirical method and its use in the development of ground-motion (attenuation) relations in eastern North America. *Bull. Seismol. Soc. Am.*, **93(3)**, 1012-1033.

[13] Cao, T., Petersen, M. D., and Reichle, M. S. (1996). Seismic hazard estimate from background seismicity in southern California. *Bull. Seismol. Soc. Am.*, **86(5)**, 1372-1381.

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

[15] Cremen, G., Gupta, A., and Baker, J. W. (2017). Evaluation of ground motion intensities from induced earthquakes using "Did You Feel It?" data. In *16th World Conf. on Earthquake Engineering*.

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

[17] Ellsworth, W. L. (2013). Injection-induced earthquakes. *Science*, **341(6142)**, 1225942.

[18] Ellsworth, W. L., Llenos, A. L., McGarr, A. F., Michael, A. J., Rubinstein, J. L., Mueller, C. S., Petersen, M.D. and Calais, E. (2015). Increasing seismicity in the US midcontinent: Implications for earthquake hazard. *The Leading Edge*, **34(6)**, 618-626.

[19] Field, E. (2010). Probabilistic seismic hazard analysis: A primer, http://www.opensha.org/ (last accessed May, 2018).

[20] Field, E. (2015). All models are wrong, but some are useful. *Seismol. Res. Lett.*, **86**, 291-293.

[21] Frankel, A., Mueller, C., Barnhard, T., Perkins, D., Leyendecker, E., Dickman, N., Hanson, S., and Hopper, M. (1996). *National seismic-hazard maps: documentation June 1996*, (pp. 96-532). Reston, VA: US Geological Survey.

[22] Frankel, A., Harmsen, S., Mueller, C., Calais, E., and Haase, J. (2010). Documentation for initial seismic hazard maps for Haiti, *Open-File Report*, 2010-1067, U.S. Government Printing Office, Washington, D.C.

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

[24] Fujiwara, H., Kawai, S., Aoi, S., Morikawa, N., Senna, S., Kudo, N., Ooi, M., Hao, K. X., Wakamatsu, K., Ishikawa, Y., and Okumura, T. (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.

[25] 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.*, **80**, 458-464.

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

[27] Goertz-Allmann, B. P., and Wiemer, S. (2012). Geomechanical modeling of induced seismicity source parameters and implications for seismic hazard assessment. *Geophysics*, **78(1)**, KS25-KS39.

[28] Gruppo di Lavoro (2004). *Catalogo parametrico dei terremoti italiani, versione 2004 (CPTI04)*. INGV, Bologna.

[29] Gulkan, P. (2013). A dispassionate view of seismic-hazard assessment, *Seism. Res. Lett.*, **84**, 413-416.

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

[31] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition.*, Springer, New York.

[32] Hornbach, M. J., Jones, M., Scales, M., DeShon, H. R., Magnani, M. B., Frohlich, C., Stump, B., Hayward, C., and Layton, M. (2016). Ellenburger waste water injection and seismicity in North Texas. *Phys. of the Earth and Plan. Int.*, **261**, 54-68.

[33] Horton, S. (2012). Disposal of hydrofracking waste fluid by injection into subsurface aquifers triggers earthquake swarm in central Arkansas with potential for damaging earthquake. *Seismol. Res. Lett.*, **83(2)**, 250-260.

[34] Hough, S. E. (2012). Initial assessment of the intensity distribution of the 2011 Mw 5.8 Mineral, Virginia, earthquake. *Seismol. Res. Lett.*, **83(4)**, 649-657.

[35] 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.

[36] Hough, S. E. (2014). Shaking from injectioninduced earthquakes in the central and eastern United States. *Bull. Seismol. Soc. Am.*, **104(5)**, 2619-2626.

[37] Iervolinoa, I. (2013). Probabilities and fallacies: why hazard maps cannot be validated by individual earthquakes, *Earthquake Spectra*, **29(3)**, 1125-1136.

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

[39] Kagan, Y. Y., and Jackson, D. D. (2013). Tohoku earthquake: a surprise, *Bull. Seismol. Soc. Am.* **103**, 1181-1194.

[40] Keranen, K. M., Savage, H. M., Abers, G. A., and Cochran, E. S. (2013). Potentially induced earthquakes in Oklahoma, USA: Links between wastewater injection and the 2011 Mw 5.7 earthquake sequence. *Geology*, **41(6)**, 699-702.

[41] Keranen, K. M., Weingarten, M., Abers, G. A., Bekins, B. A., and Ge, S. (2014). Sharp increase in central Oklahoma seismicity since 2008 induced by massive wastewater injection. *Science*, **345(6195)**, 448-451.

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

[43] Keyfitz, N. (1981). The limits of population forecasting, *Pop. and Dev. Rev.*, **7**, 579-59.

[44] Kim, W. Y. (2013). Induced seismicity associated with fluid injection into a deep well in Youngstown, Ohio. *Jour. of Geophys. Res.: Solid Earth*, **118(7)**, 3506-3518.

[45] Klügel, J.-U., Mualchin, L., and Panza, G. F. (2006) A scenario-based procedure for seismic risk analysis. *Engineering Geology*, **88**, 1-22.

[46] Kossobokov, V., and Nekrasova, A. (2012). Global Seismic Hazard Assessment Program maps are erroneous, *Seismic instruments*, **48**, 162-170.

[47] Kruskal, W. (1988). Miracles and statistics: the casual assumption of independence, *J. Am. Stat. Assoc.*, **83**, 929-940.

[48] Kuchment, A. (2017) Are earthquakes gone from our area for good? Dallas News. https://www.dallasnews.com/business/energy/2017/03/01/earthquakes-gone-area-good-scientists-try-solve-mystery Last accessed: May 5, 2017.

[49] Langenbruch, C., Weingarten, M., and Zoback, M. D. (2018). Physics-based forecasting of man-made earthquake hazards in Oklahoma and Kansas. *Nature comms.*, **9(1)**, 3946.

[50] Liu, M., Luo, G., Wang, H., and Stein, S. (2014). Long aftershock sequences in North China and Central US: implications for hazard assessment in mid-continents. *Earthquake Science*, **27(1)**, 27-35.

[51] Mak, S., Clements, R. A., and Schorlemmer, D. (2014). The Statistical Power of Testing Probabilistic Seismic-Hazard Assessments, *Seismol. Res. Lett.*, **85**, 781-783.

[52] Mak, S., and Schorlemmer, D. (2016a). A comparison between the forecast by the United States National Seismic Hazard Maps with recent ground-motion records. *Bull. Seismol. Soc. Am.*, **106(4)**, 1817-1831.

[53] Mak, S., and Schorlemmer, D. (2016b). What Makes People Respond to "Did You Feel It?"?. *Seismol. Res. Lett.*, **87(1)**, 119-131.

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

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

[56] Marzocchi, W., and Jordan, T. H. (2014). Testing for ontological errors in probabilistic forecasting models of natural systems. *Proc. Natl. Acad. Sci. U.S.A.*, **111(33)**, 11973-11978.

[57] McMahon, N. D., Stickney, M., Aster, R. C., Yeck, W., Martens, H. R., and Benz, H. (2017). Spatiotemporal analysis of the foreshock-mainshock-aftershock sequence of the 6 July 2017 M 5.8 Lincoln, Montana earthquake. In *AGU Fall Meeting Abstracts*.

[58] Minoura, K., Imamura, F., Sugawa, D., Kono, Y., and Iwashita, T. (2001). The 869 Jogan tsunami deposit and recurrence interval of large-scale tsunami on the Pacific coast of Northeast Japan, *J. Natural Disaster Sci.*, **23**, 83-88.

[59] Miyazawa, M., and Mori, J. (2009). Test of seismic hazard map from 500 years of recorded intensity data in Japan, *Bull. Seismol. Soc. Am.*, **99**, 3140-3149.

[60] Montilla, J. A., Hamdache, P. M., and Casado, C. L. (2003). Seismic hazard in Northern Algeria using spatially smoothed seismicity: Results for peak ground acceleration. *Tectonophysics*, **372(1)**, 105-119.

[61] Moschetti, M. P., Thompson, E. M., Powers, P. M., Hoover, S. M., and McNamara, D. E. (2018). Ground motions from induced earthquakes in Oklahoma and Kansas. *Seismol. Res. Lett.*

[62] Mostafa Mousavi, S., and Beroza, G. C. (2018). Evaluating the 2016 one-year seismic hazard model for the central and eastern United States using instrumental ground-motion data. *Seismol. Res. Lett.*, **89(3)**, 1185-1196.

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

[64] Murphy, A.H. (1993). What is a good forecast? an essay on the nature of goodness in weather forecasting, *Weather and Forecasting*, **8**, 281-293.

[65] Murray, K. E. (2016). Seismic moment versus water: A study of market forces and regulatory actions. *Geol. Soc. Am. Abstr. Progr.*, **48(7)**.

[66] 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, *Nat. Haz.*, **70**, 629-641.

[67] Novakovic, M., Atkinson, G. M., and Assatourians, K. (2018). Empirically calibrated ground-motion prediction equation for Oklahoma. *Seismol. Res. Lett.*, **108(5A)**, 2444-2461.

[68] Ogwari, P. O., DeShon, H. R., and Hornbach, M. J. (2018). The Dallas-Fort Worth airport earthquake sequence: seismicity beyond injection period. *Jour. of Geophy. Res.: Solid Earth*, **123(1)**, 553-563.

[69] Parker, R. L. (1997). Understanding inverse theory. *An. Rev. of Earth and Plan. Sci.*, **5**, 35-64.

[70] Peresan, A., and Panza, G. F. (2012). Improving earthquake hazard assessments in Italy: An alternative to "Texas sharpshooting." *Eos, Transactions, American Geophysical Union*, **93**, 538.

[71] Petersen, M. D., Moschetti, M. P., Powers, P. M., Mueller, C. S., Haller, K. M., Frankel, A. D., Zeng, Y., Rezaeian, S., Harmsen, S. C., Boyd, O. S. and Field, N. (2015). The 2014 United States national seismic hazard model. *Earthquake Spectra*, **31(S1)**, S1-S30.

[72] Petersen, M. D., Mueller, C. S., Moschetti, M. P., Hoover, S. M., Llenos, A. L., Ellsworth, W. L., Michael, A. J., Rubinstein, J. L., McGarr, A. F., and Rukstales, K. F. (2016a). *2016 one-year seismic hazard forecast for the Central and Eastern United States from induced and natural earthquakes* (No. 2016-1035). US Geological Survey.

[73] Petersen, M. D., Mueller, C. S., Moschetti, M. P., Hoover, S. M., Llenos, A. L., Ellsworth, W. L., Michael, A. J., Rubinstein, J. L., McGarr, A. F., and Rukstales, K. F. (2016b). Seismic hazard forecast for 2016 including induced and natural earthquakes in the central and eastern United States. *Seismol. Res. Lett.*, **87**, 1327-1341.

[74] Petersen, M.D., Mueller, C. S., Moschetti, M. P., Hoover, S. M., Shumway, A. M., McNamara, D. E., Williams, R. A., Llenos, A. L., Ellsworth, W. L., Michael, A. J., and Rubinstein, J. L. (2017). 2017 one-year seismic hazard forecast for the central and eastern United States from induced and natural earthquakes. *Seismol. Res. Lett.*, **88(3)**, 772-783.

[75] Petersen, M. D., Mueller, C. S., Moschetti, M. P., Hoover, S. M., Rukstales, K. S., McNamara, D. E., and Llenos, A. L. (2018a). 2018 one-year seismic hazard forecast for the central and eastern United States from induced and natural earthquakes. *Seismol. Res. Lett.*, **89(3)**, 1049-1061.

[76] Petersen, M. D., Mueller, C. S., Moschetti, M. P., Hoover, S. M., Rukstales, K. S., McNamara, D. E., Williams, R. A., Shumway, A. M., Powers, P. M., Earle, P. S., Llenos, A. L., Michael, A. J., Rubinstein, J. L., Norbeck, J. H., and Cochran, E. S. (2018b). Data release for 2018 one-year seismic hazard forecast for the central and eastern United States from induced and natural earthquakes, U.S. Geological Survey data release, https://doi.org/10.5066/F7Cf9PC4.

[77] Prest, B. C. (2018). Explanations for the 2014 oil price decline: supply or demand? *Energy Economics*.

[78] Prezeshk, S., Zandieh, A., and Tavakoli, B. (2011). Hybrid empirical ground-motion prediction equations for eastern North America using NGA models and updated seismological parameters. *Bull. Seismol. Soc. Am.*, **101(4)**, 1859-1870.

[79] Quitoriano, V., Thompson, E. M., Smoczyk, G. M., and Wald, D. J. (2017). Access to "Did You Feel It?" data for induced earthquake studies. Paper presented in *2017 Seismological Society of America Annual Conference*; April 20, 2017; Denver, CO.

[80] Reyners, M. (2011). Lessons from the destructive M_W 6.3 Christchurch, New Zealand, earthquake, *Seismol. Res. Lett.*, **82**, 371-372.

[81] Roach, T. (2018). Oklahoma earthquakes and the price of oil. *Energy Policy*, **121**, 365-373.

[82] Rubinstein, J. L., and Mahani, A. B. (2015). Myths and facts on wastewater injection, hydraulic fracturing, enhanced oil recovery, and induced seismicity. *Seismol. Res. Lett.*, **86(4)**, 1060-1067.

[83] Sagiya, T. (2011). Integrate all available data, *Nature*, **473**, 146147.

[84] Silva, W., Gregor, N., and Darragh, R. (2002). Development of regional hard rock attenuation relations for central and eastern North America. *Pacific Engineering and Analysis*, El Cerrito, CA.

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

[86] Stein, S., Geller, R., and Liu, M. (2011). Bad assumptions or bad luck: Why earthquake hazard maps need objective testing. *Seismol. Res. Lett.*, **82(5)**, 623-626.

[87] Stein, S., Geller, R. J., and Liu, M. (2012). Why earthquake hazard maps often fail and what to do about it, *Tectonophysics*, **562/563**, 1-25.

[88] Stein, S., Geller, R. J., and Liu, M. (2013). Reply to comment by Arthur Frankel on "Why earthquake hazard maps often fail and what to do about it", *Tectonophysics*, **592**, 207-209.

[89] Stein, S., and Stein, J. L. (2013). How good do natural hazard assessments need to be? *GSA Today*, **23(4)**.

[90] Stein, S., and Friedrich, A. (2014). How much can we clear the crystal ball? *Astronomy and Geophysics*, **55**, 2.11-2.17.

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

[92] Stirling, M.W., and Petersen, M. D. (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.

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

[94] Stirling, M. W. (2012). Earthquake hazard maps and objective testing: the hazard mapper's point of view, *Seismol. Res. Lett.*, **83**, 231-232.

[95] Stephenson, D. (2000). Use of the "Odds Ratio" for diagnosing forecast skill, *Weather and Forecasting*, **15**, 221-232.

[96] Stucchi, M., Albini, P., Mirto, C., and Rebez, A. (2004). Assessing the completeness of Italian historical earthquake data, *Annals of Geophys.*, **47**, 659-673.

[97] 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. *Geophys. Jour. Int.*, **198(3)**, 1554-1571.

[98] Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 285-294.

[99] Toro, G. R. (2002). Modification of the Toro *et al.* (1997) attenuation equations for large magnitudes and short distances. *Risk Engineering Technical Report.*

[100] Vanneste, K., Stein, S., Camelbeeck, T., and Vleminckx, B. (2018). Insights into earthquake hazard map performance from shaking history simulations, *Scientific Reports*, **8(1)**, 1855.

[101] Wald, D. J., Quitoriano, V., Dengler, L. A., and Dewey, J. W. (1999). Utilization of the internet for rapid community intensity maps. *Seismol. Res. Lett.*, **70(6)**, 680-697.

[102] Wald, D. J., Worden, B. C., Quitoriano, V., and Pankow, K. L. (2005). *ShakeMap manual: technical manual, user's guide, and software guide*, (No. 12-A1).

[103] Wald, D. J., Quitoriano, V., Worden, C. B., Hopper, M., and Dewey, J. W. (2012). USGS "Did You Feel It?" internet-based macroseismic intensity maps. *Annals of Geophysics*, **54(6)**.

[104] Wang, Z. (2011). Seismic hazard assessment: issues and alternatives, *Pure. Appl. Geophys.*, **168**, 11-25.

[105] Wang, Z. and Cobb, J. (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*, GSA, Boulder, CO.

[106] Wang, Z. (2015). Predicting or forecasting earthquakes and the resulting ground motion hazards: a dilemma for earth scientists, *Seismol. Res. Lett.*, **86**, 1-5.

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

[108] Weingarten, M., Ge, S., Godt, J. W., Bekins, B. A., and Rubinstein, J. L. (2015). High-rate injection is associated with the increase in US mid-continent seismicity. *Science*, **348(6241)**, 1336-1340.

[109] White, I. J., Liu, T., Luco, N., and Liel, A. B. (2018). Considerations in comparing the US Geological Survey one-year induced-seismicity hazard models with "Did You Feel It?" and instrumental data. *Seismol. Res. Lett.*, **89(1)**, 127-137.

[110] Wolin, E., Stein, S., Pazzaglia, F., Meltzer, A., Kafka, A., and Berti, C. (2012). Mineral, Virginia, earthquake illustrates seismicity of a passive-aggressive margin. *Geophys. Res. Lett.*, **39(2)**.

[111] Worden, C. B., Gerstenberger, M. C., Rhoades, D. A., and Wald, D. J. (2012). Probabilistic relationships between groundmotion parameters and modified Mercalli intensity in California. *Bull. Seismol. Soc. Am.*, **102(1)**, 204-221.

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

[113] Yeck, W. L., Hayes, G. P., McNamara, D. E., Rubinstein, J. L., Barnhart, W. D., Earle, P. S., and Benz, H. M. (2017). Oklahoma experiences largest earthquake during ongoing regional wastewater injection hazard mitigation efforts. *Geophys. Res. Lett.*, **44(2)**, 711-717.

[114] Zhuang, J., and Touati, S. (2015). Stochastic simulation of earthquake catalogs. *Community Online Resource for Statistical Seismicity Analysis.*