Bulletin of the Seismological Society of America, Vol. 96, No. 5, pp. 1624–1633, October 2006, doi: 10.1785/0120050195

Methodologies for Earthquake Hazard Assessment: Model Uncertainty and the WGCEP-2002 Forecast

by Morgan T. Page and J. M. Carlson

Abstract Model uncertainty is prevalent in probabilistic seismic hazard analysis (PSHA) because the underlying statistical signatures for hazard are unknown. Although methods for incorporating parameter uncertainty of a particular model in PSHA are well understood, methods for incorporating model uncertainty are more difficult to implement because of the high degree of dependence between different earthquake-recurrence models. We show that the method used by the 2002 Working Group on California Earthquake Probabilities (WGCEP-2002) to combine the probability distributions given by multiple earthquake-recurrence models has several adverse effects on their results. In particular, WGCEP-2002 uses a linear combination of the models that ignores model dependence and leads to large uncertainty in the final hazard estimate. Furthermore, model weights were chosen based on data, which has the potential to systematically bias the final probability distribution. The weighting scheme used in the Working Group report also produces results that depend on an arbitrary ordering of models. In addition to analyzing current statistical problems, we present alternative methods for rigorously incorporating model uncertainty into PSHA.

Introduction

The goal of probabilistic seismic hazard analysis (PSHA) is to provide a quantitative estimate of the likelihood of exceeding a given threshold of earthquake-caused ground motions in a specific region during a given time period (Senior Seismic Hazard Analysis Committee [SSHAC], 1997). PSHA is characterized by deep uncertainty, for not only is there parameter uncertainty regarding the values of various model inputs needed to estimate hazard, there is also model uncertainty. This type of uncertainty relates to the statistical signatures for hazard, that is, how best to represent the earthquake renewal process in a recurrence model. Although methods for incorporating parameter uncertainty are widely used, model uncertainty is less well understood (Aposolakis, 1995). Nevertheless, it is prevalent in PSHA and must be handled properly.

The 2002 Working Group on California Earthquake Probabilities (WGCEP-2002, Working Group, or WG02) differed from previous reports in that an attempt to quantify and incorporate model uncertainty was made. Unlike previous consensus reports in 1988, 1990, and 1995 (WGCEP, 1990a, b, 1995), in which a single model was agreed upon, the WG02 report (WGCEP, 2003) used multiple models to generate the 2002 forecast. Model uncertainty was incorporated by taking a linear combination of the probability distributions given by several different models. Model uncertainty comprises a large portion of the total uncertainty in the WG02 forecast.

In this article, we refer periodically to the "true hazard" of an earthquake in a given region. This notion is not completely straightforward, as discussed by Freedman and Stark (2003). On one hand, either the earthquake will happen or it will not, so in a sense the true hazard of a certain event is 0 or 1. However, it will most likely never be possible to estimate hazard with complete confidence. Because of irreducible (aleatory) uncertainty in earthquake forecasting, the probability of an earthquake occurring must be expressed as just that, a probability. For the purposes of this article, it is helpful to think of true hazard as the probability of an earthquake occurring that we could deduce with only irreducible uncertainty, that is, the best estimate we could make in light of uncertainties that are stochastic in character. For more information on aleatory uncertainties see the SSHAC (1997) report.

We use the word "model" in this article in the sense of an earthquake-recurrence model. Note that model uncertainty may be present in other areas of PSHA as well, but these are not the focus of this work. The role of ground-motion model uncertainty has been analyzed by F. Scherbaum *et al.* in light of the recent PEGASOS project (unpublished manuscript, 2006). They find that incorporating multiple ground-motion models with a logic tree approach leads to an overestimation of the total epistemic uncertainty.

The goal of this article is to provide a careful critique

of the WG02 report based on the underlying probabilistic methods that were used. Using simple examples, we illustrate several key issues and assumptions that are problematic and lead to biases in the final results. In addition, we discuss various methods for avoiding these issues in the future. Precisely incorporating model uncertainty and dependence in PSHA will allow for the most precise formulation of hazard that the data allow.

This article is divided into three sections. First, we give a brief summary of the various earthquake-recurrence models used in the WG02 report and the current methodology that combines these models into a single probabilistic forecast. Second, we diagnose several unintended biases inherent in this type of PSHA formulation. In particular, any dependence between the earthquake-recurrence models is ignored in a weighted average of the individual model probability distributions. We also show that the model weights, which are based in part on data availability, have the potential to systematically skew the final prediction. Model weights also vary from fault to fault, and this is formulated in such a way that the arbitrary order of the models themselves in the methodology affects the result. Finally, we present alternative methods for incorporating model uncertainty and model dependence. We discuss Bayesian methods, which provide a framework to update hazard estimates as more data become available. Also, copulas (dependence models) provide a means to combine multiple probability distributions into a single distribution in a way that incorporates dependence. We conclude with a simple example using copulas in a Bayesian framework.

The WG02 Report

WG02 concluded that the probability of one or more earthquakes exceeding moment magnitude 6.7 in the San Francisco Bay Region for the 30-year period from 2002 to 2031 is 62% (with 95% confidence bounds of 27% and 87%). In addition to the regional estimate, the Working Group calculated probabilities for fault segments, rupture sources, fault systems, and the background. Below we give a summary of their methodology relevant to our analysis.

The Working Group used five models to estimate earthquake probability in their report on earthquake hazards in the San Francisco Bay Region. These models are statistical models that attempt to estimate the probability of earthquake occurrence. The first of these models, the Poisson model, assumes that earthquakes randomly occur with a time-independent probability. This model has only one parameter, λ, the average rate of earthquake occurrence (seismicity). The Poisson probability density for an event as a function of time t since the last event is given by

$$f_{\text{Pois}}(t) = \lambda\, e^{-\lambda t}. \tag{1}$$

The second model, the empirical model, is also a one-parameter Poisson-type model, but with a different value of λ. Whereas in the Poison model background seismicity is taken to be the long-term, historical rate of earthquake occurrence, in the empirical model post-1906 seismicity is used to estimate the corresponding λ. Because recent seismicity in the San Francisco Bay Region is lower than the historical rate, the empirical model predicts lower probabilities than the Poisson model.

The third model, the Brownian passage time (BPT) model (Matthews *et al.*, 2002; Kagan and Knopoff, 1987), uses two parameters to compute the probability of an event: λ and the aperiodicity of events α. The probability density for an event is given by

$$f_{\text{BPT}}(t) = \sqrt{\frac{1}{2\pi\lambda\alpha^2 t^3}}\, e^{-\lambda(t-1/\lambda)^2/2t\alpha^2}. \tag{2}$$

Note that the probability density is zero at $t = 0$, as a new earthquake is thought unlikely until stress reaccumulates on the fault segment. The aperiodicity α is estimated from data (Ellsworth *et al.*, 1999). When $\alpha = 0$, earthquake occurrence is periodic, and for large values of α this model behaves similarly to a Poisson process.

The fourth model, the BPT-step model, is a BPT model that also incorporates the effects of stress interactions from events on nearby faults. This interaction is incorporated into the model by a "clock change," that is, changing the value of t in equation (2). WG02 used this model to incorporate stress changes from two events: the 1906 San Francisco earthquake and the 1989 Loma Prieta earthquake.

The final model is the time-predictable model (Shimazaki and Nakata, 1980). This model uses the slip in the last earthquake, coupled with the slip rate of the fault segment, to calculate the expected time of the next rupture. The fault is expected to rupture once all of the strain released in the last earthquake has reaccumulated on the fault. In the WG02 report this model is used only on the San Andreas fault segment.

Each of these models employs different data to estimate earthquake hazard, and each of them has different assumptions as to what parameters drive the hazard. Consequently, each of these models gives a different prediction for the level of earthquake hazard in the San Francisco Bay Region. The probability distributions given by each model for one or more regional earthquakes with moment magnitude 6.7 or greater are shown in Figure 1a. These histograms were generated by using 3000 Monte Carlo iterations for each model. Each Monte Carlo iteration samples from possible values for all model inputs, such as seismicity or aperiodicity. The weights for the parameter values are estimated from data or determined from expert opinion. Without this parameter uncertainty, each model would give a point prediction for the probability. However, as there are indeed uncertainties in the inputs for these models, each of these models has its own associated uncertainty. The spread of each distribution in Figure 1a is a result of parameter uncertainty within each

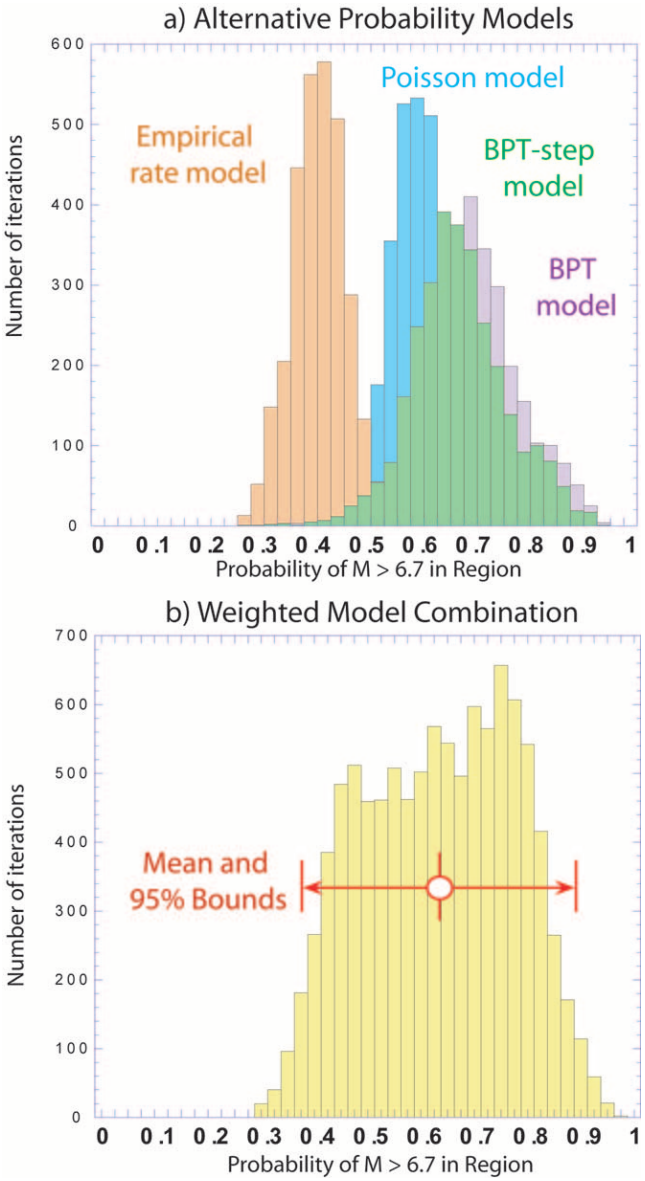

Figure 1. Adapted from the WG02 report (WGCEP, 2003). Distributions of the regional probability of a M 6.7 earthquake calculated by using various probability models. (a) Overlapping histograms show probability calculated in 3000 iterations using each of four models separately. The shape and width of each distribution reflects epistemic uncertainty in the choice of underlying models and parameters. (b) Corresponding distribution calculated in 10,000 iterations using the weighted combination of models shown in Figure 4. The broad shape of this distribution reflects the combination of distinct behaviors of the alternate models. Additional mass near $P = 0.8$ corresponds to realizations that employ the time-predictable model on the San Andreas fault, not shown in (a).

model. Thus, the figure is showing probability distributions for regional probability.

The issue at hand, which arises broadly in hazard analysis, is the following. Given all these different models and different predictions, what is the best estimate for earthquake hazard and uncertainty for the San Francisco Bay Region? Each model gives a different probability distribution. The problem reduces to combining the probability distributions from different models into a single probability distribution function that best represents current knowledge about earthquake hazard in the San Francisco Bay Region.

To arrive at one estimate of earthquake hazard, the Working Group performed what is essentially a weighted averaging of the models. That is, they carried out a Monte Carlo sampling of all five models, so that the final answer is a function of not just the mean prediction of a given model, but depends on the entire probability distribution given by that model. The top level of the logic tree, that is, the first level randomly sampled in a given Monte Carlo run, determines which model will be used. The models themselves are allowed to differ from fault to fault during a given Monte Carlo run. We still refer to this methodology as an "averaging," since neglecting correlations of the models between faults, the result of the weighted Monte Carlo sampling is a weighted average of the probability distributions produced by each model.

The final result of what is, to first order, a weighted average of the individual model probability distribution functions is shown in Figure 1b. The combined probability distribution has an average of 62%, which is similar to the mean result of the 1999 Working Group report (WGCEP, 1999). However, the uncertainty in the combined result, as illustrated by the large variance in Figure 1b, is considerably larger. Although new data have been added, they have not improved the final uncertainty; the cost of adding new data has been the incorporation of additional models, and these have increased the uncertainty.

Current Statistical Biases in PSHA

The first step in providing a rigorous methodology to incorporate model uncertainty in PSHA is a critique of previous analyses. The three problems with current methodology that we discuss here are the linear combination of models, choosing model weights based upon data availability, and arbitrary ordering of models. Below we present an analysis that illustrates key issues and points toward opportunities to improve hazard estimates.

Linear Combination of Models

Does averaging make sense for the earthquake-hazard models? There is a fundamental problem assigning probability weights to different models. Averaging can be justified if one and only one model describes the underlying mechanism generating hazard. In this case, the weights used in

the average would reflect the probability that a given model is this one "true model" (Morgan and Henrion, 1990). However, this condition—that one model is "correct" and the other models are not—is, in fact, a very strict condition, and in most cases where hazard is generated by the interaction of many factors in an unknown way, it will not be satisfied. What is more likely is that each model captures some aspect of the underlying process but is not in and of itself a complete description, so that the probability that a given model is correct is near zero. As the models are not collectively exhaustive, the probability weights given to the models will not sum to one (Morgan and Henrion, 1990; Winkler, 1995). Nonetheless averaging of probabilities produced by different models is practiced not just in the field of earthquake-hazard estimation, but in climate-change studies as well (Lempert *et al.*, 2004).

It can be argued that the weights assigned to earthquake recurrence models are not meant to be probabilities, but they simply judge the "relative merit" of the individual models. While there is some debate on this subject (see, for example, Abrahamson and Bommer [2005], McGuire *et al.* [2005], and Musson [2005]), the Working Group, in their analysis, treats the weights as probabilities in their aggregation of the models. Here we posit that averaging over the model weights cannot be valid, because the weights themselves are neither exclusive nor exhaustive probabilities.

Consider our case of earthquake-hazard estimation. One model, for example, the time-predictable model, attempts to quantify the hazard based on the slip in the last earthquake. Another model, the BPT-step model, makes predictions based on the stress shadows from previous earthquakes. Averaging the predictions given by these two models is not valid because it is unlikely that earthquake hazard is a function of slip in the last earthquake but not stress shadows, or vice versa (Bier, 1995). To put it another way, the models are not mutually exclusive. One would expect that both of these factors are important in estimating earthquake hazard. The best approach would use all available information to arrive at a prediction and aggregate the models in a statistically sound way.

To further demonstrate exactly why averaging models can be incorrect, consider the following example. Suppose we are trying to assess hazard for a given region. This region has "risk factor A" and "risk factor B," which could be any type of input used in a hazard model to make a prediction. For example, these "risk factors" could be time since the last earthquake, recent seismicity, strain data, etc. Suppose also that we have two models for earthquake occurrence on hand: model A, which makes predictions for earthquakes based on risk factor A, and model B, which makes predictions based on risk factor B. Based on the historical catalog and/or physical reasoning, model A gives an event probability of α for regions having risk factor A, and thus for the region in question. Similarly, model B assigns an event probability of β to the region. Now, as neither of these models is a perfect predictor, α is not equal to β. As is shown in Figure 2a, α and

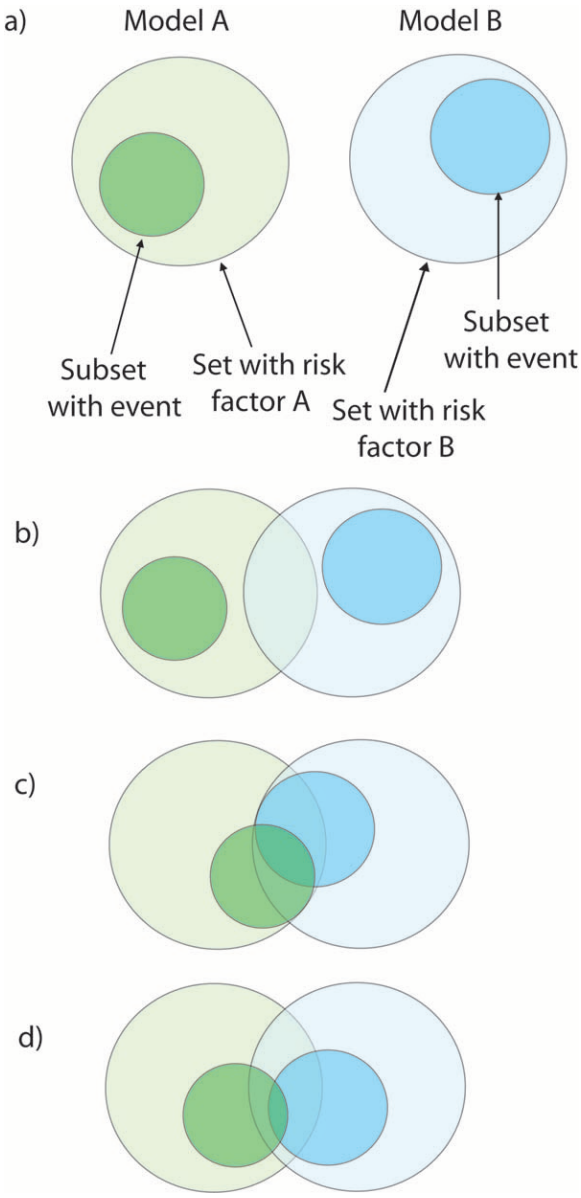

Figure 2. (a) Model A calculates hazard for the entire set shown having risk factor A. The smaller circle shows those members of the entire set that will in actuality have the hazardous event. Model A gives a probability of α, which is equal to the percentage of the entire set that is in the smaller circle. Similarly, model B calculates a probability of β, which is proportional to the number of elements with risk factor B that are in the small blue circle. What is the event probability for an element that has both risk factor A and risk factor B? It could be 0%, as shown in b, or nearly 100%, as shown in c, depending on how the risk factors interact. (d) In some situations it might be wise to predict that the risk factors are independent, giving a probability of $1 - (1 - \alpha)(1 - \beta)$. In each of these cases, the true hazard cannot be obtained by any average of α and β, and thus, it is not correct to average the results of model A and model B.

β are proportional to the size of the subset with the event divided by the size of the set with the risk factor. In the light of conflicting estimates, what is our best guess for the probability of an event? Should we average α and β according to expert opinion in the validity of these models? What can we say mathematically about the possible values for the true hazard?

In fact, without knowing how risk factors A and B interact, we can place no constraint on what the true hazard is, even if both model A and model B describe how risk factors A and B affect earthquake hazard perfectly. The Venn diagrams in Figure 2 illustrate this point. It is perfectly consistent with the information given that the probability of earthquake occurrence in this region is anywhere between 0% and 100%, regardless of the values of α and β. Risk factors A and B could interact as shown in Figure 2b, so that the region has no probability of an event, or as shown in Figure 2c, in which an earthquake is nearly guaranteed. If risk factors A and B are completely independent, the event probability is then given by $1 - (1 - \alpha)(1 - \beta)$, which is greater than both α and β. Thus, many of the plausible values for event probability can not be obtained by any average.

To make this example more concrete, consider the case where model A in our example is the Poisson model and model B is the empirical model from the Working Group report. In this case, risk factor A is historical seismicity and risk factor B is recent (post-1906) seismicity. If all we know about the system is historical seismicity, then we can apply the Poisson model and achieve an unbiased (although not very precise) hazard estimate. Similarly, we can imagine that the empirical model would be a "correct" model in the sense that systemwide it gives unbiased estimates, and that if all we know about the system is recent seismicity, it will give a best estimate with this information. Using both models, however, is problematic unless we understand how recent seismicity and historical seismicity interact. In this case if $\alpha = \beta$, we would expect the actual probability to indeed be the α. However, $\alpha > \beta$ in the case of the San Francisco Bay Region. That is, the Poisson model gives a higher probability than the empirical model because historical seismicity is higher than recent seismicity. It is at least plausible that true hazard in this case could be greater than α if systems that currently have a dearth in seismicity relative to the historical average have a higher hazard, perhaps because a great deal of stress has accumulated during the recent lull in seismicity. This interpretation is somewhat controversial, and we are not advocating this as fact, but simply presenting a plausible example in which the true hazard cannot be obtained by any average. It is possible for parameters in different models to be dependent in this way, and this possibility is ignored by the linear combination of models used in the WG02 report.

Choosing Model Weights Based on Data Availability

Seven faults that are believed capable of generating an earthquake of magnitude 6.7 or greater are incorporated into the WG02 hazard estimate. Each of the five probability models is assigned a weight for these faults, and these model weights are sampled in the logic tree. Because the models are not mutually exclusive, there is no simple interpretation for the relative weights assigned to the different models for each fault. At first glance, it appears that they reflect the probability that a particular model is the "correct model," although we have shown that this is not as straightforward as it sounds, as more than one model can be "correct" in the sense of being unbiased. We could consider the more "correct" model to be that which most accurately describes earthquake generation, or, alternatively, the model which is the most precise (that is, resolves the most parameters). However, even this problematic interpretation cannot be used, for the model weights differ from fault to fault, and we surely do not expect the process of earthquake generation (with the exception of the model parameters themselves) to differ from fault to fault. The model weights are a function of the fault in question because weights were based on the "relative amount and quality of geologic data" available for each fault. However, weighting based on availability of data can lead to systematic errors. Certainly the data available do not determine which model is the more correct model, as the physics of earthquakes is not a function of what it is possible to measure. Hence, it is incorrect to weight models based on this.

To see how weighting based on data quality can lead to systematic errors, consider two faults, fault A and fault B. Suppose we also have two models of the earthquake probability, the Poisson model and a recurrence model that gives low probabilities after large earthquakes and higher probabilities after a seismic lull. Suppose further that fault A has been active historically, so that data exist from past earthquakes. We might be apt to give the Poisson model a low weight for fault A, as the recent seismicity and abundance of data seems to make the recurrence model a better fit. After all, we should be able to make better predictions than the Poisson model (which while it is unbiased has low precision) for fault A because we have such an abundance of data. Now, suppose fault B has been seismically inactive during the historical record. We have few data, making it difficult to apply the recurrence model in this case, which requires information about past earthquakes. So we might be apt to give the Poisson model a larger weight. But notice what we have done. Compare the model predictions for the two faults as shown in Figure 3. The recurrence model would give a low probability for fault A because there were recent earthquakes and a high probability for fault B because there were not. In each case, we weighted the model that gave the lowest prediction the most. In this example, giving model weights based on data availability systematically skewed the hazard estimates to the low side.

Arbitrary Ordering of Models

Because the model weights in the WG02 report varied from fault to fault, it was not possible to have the same

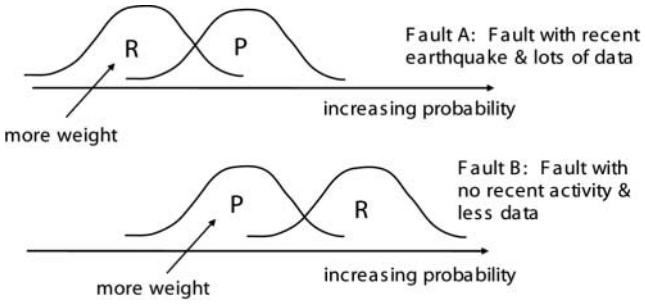

Figure 3. Choosing model weights based on data availability can lead to systematic bias, as shown in this example. Fault A has recent activity, and a recurrence model (R) gives a low probability for future rupture. Fault B has no recent earthquakes, and thus the recurrence model gives a high probability. The Poisson model (P) gives the same probability estimate for both faults. However, by weighting the models based on the amount of data available, we systematically choose the model giving the lowest estimate in each case.

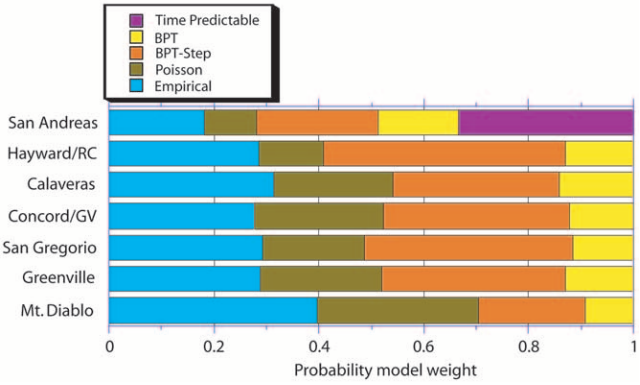

Figure 4. Adapted from the WG02 report (WGCEP, 2003). Division of weight assigned to each probability model for each fault, as determined by expert opinion. The time-predictable model was applied only to the San Andreas fault.

model in effect on each fault for a given Monte Carlo iteration. To mitigate this problem, the models were organized in the order shown in Figure 4. Then, for a given Monte Carlo iteration, a single random number between 0 and 1 determined which model would be in effect on each fault. The method can be seen graphically from Figure 4: a given random number determines the horizontal position on the graph. A vertical line drawn at that position specifies which model is employed on each fault. For example, if the random number is 0.6, the Mt. Diablo fault uses the Poisson model, the San Andreas fault uses the BPT model, and the remaining faults use the BPT-step model. This method has several problematic outcomes: certain models can interact, while others, such as the empirical model and the BPT model, will never both be used in the same iteration. Perhaps the most troubling result of this method is that the ordering of the models changes the result. That is, the Working Group positioned first the empirical model, followed by the Poisson model, the BPT-step model, and the BPT model, and arranged the time-predictable model last, as shown in Figure 4. However, a different ordering of these five models would lead to a different, albeit only slightly different, result for the combined hazard. This is troubling since the ordering is arbitrary. According to the WG02 report, an independent sampling of model weights on each fault would have resulted in the same mean regional probability, but a smaller variance.

Note that the problems associated with choosing weights based on data availability and the arbitrary ordering of models could be avoided by making the model weights constant from fault to fault. This would result in large parameter uncertainty on some faults for some models. The probability distributions of unknown parameters (for example, for the time-predictable model) would have to be assessed in this implementation.

Solutions

The simplest way to incorporate model uncertainty is to reparameterize model uncertainty as parameter uncertainty. Morgan and Henrion (1990) suggest forming a metamodel, which reduces to various individual models in special cases. This is possible if the models are sufficiently similar, as with WG02's BPT and BPT-step model, where model uncertainty could be recast as parameter uncertainty regarding the size of the stress-step interaction.

However, to combine probability distributions from sufficiently different models, a careful consideration of model dependence is needed. Here, in brief we discuss Bayesian methods and copulas (dependence models) and present a simple example using this machinery to aggregate different models based on the level of dependence between them.

Bayesian Methods

A straightforward formulation for the model-dependence problem is offered by Bayesian statistics (Gelman *et al.*, 1995). If model A gives a distribution f_A, and model B gives a distribution f_B, then Bayes' Rule gives the posterior probability distribution

$$f_{\text{Post}}(f_A, f_B) \propto f(p)L(f_A, f_B|p). \quad (3)$$

Here, $f(p)$ is the prior probability distribution and $L(f_A, f_B|p)$ is the likelihood that both model A and model B will give distributions f_A and f_B, respectively, if the actual probability is p (Bier, 1995). The Bayesian formulation provides a natural way to incorporate expert information as prior knowledge; alternatively, one could use a Poisson distribution as a prior. As more information is acquired for a given region or fault, the prior is updated, and a new distribution is obtained. Naturally, assessing the likelihood function is the real challenge (Clemen and Winkler, 1999). In the case of mul-

tiple models, as shown above, the likelihood function incorporates all notions of bias and dependence.

The advantage to the Bayesian approach is that the same framework can easily be used in regions of high or low seismicity, regardless of the amount of data available. A Poisson prior is in many ways ideal as it gives the level of hazard one would assume having no information other than seismicity. This is exactly the function of a prior in Bayesian analysis—to assess the state of knowledge before looking at other sources of information. This prior can then be updated as new information is acquired for a particular fault. For example, updating a Poisson prior with a likelihood function based on the additional information of time since the last earthquake would multiply the probability based on historical seismicity alone by the likelihood that the time since the last earthquake is consistent with that historical seismicity, according to a recurrence model of the analyst's choice.

Two potential pitfalls here must be avoided. First, to avoid bias it is necessary to incorporate all information that is known. For example, suppose for fault A we know the time since the last earthquake. On fault B we know there has been no earthquake in the past several hundred years. In addition to updating our prior for fault A with this new information, the prior for fault B must be updated as well. In this example, we have a lower limit on the time since the last earthquake for fault B, and this information must be incorporated into the hazard estimate. Should Bayesian methodology be used, failure to account for all information can lead to systematic bias. For example, it would be incorrect to lower hazard estimates on faults that ruptured in recent history without also raising estimates for faults with no recent activity. The Poisson model can have no bias overall, that is, it can be correctly normalized over the entire system of faults. We want to use more information than the Poisson model alone, since the Poisson model has poor precision. However, when we update the Poisson prior, we must be careful to do it in such a way that total hazard systemwide remains unchanged. Otherwise, we risk sacrificing low bias for precision, which is a poor trade.

A second pitfall must be avoided in the Bayesian formulation. In general, it is possible to update a prior with several types of data. The prior can be updated with some data, yielding a posterior distribution, which then becomes the "new prior" that can be updates via Bayes' Rule with additional data. However, this is only possible when the data are independent. If the data are dependent, the dependence must be included in a single likelihood function as shown in equation (3). The prior and the likelihood function as well must be independent to multiply them as is done in Bayes' Rule.

Copulas

Copulas are dependence models that are ideally suited to the task of combining distributions. They are often used to combine knowledge from different experts into a single-probability distribution (Clemen *et al.*, 2000; Jouini and Clemen, 2002). The expert-aggregation problem is similar to that of the model-aggregation problem. As with model information, expert knowledge is partially dependent, because experts share knowledge. It is not a matter of which expert is right and which is wrong; rather, each expert gives additional information (presumably a function of the knowledge that differs between experts). As this is very similar to the case of model uncertainty, copulas could be used to combine multiple probability distributions from individual models into a single probability distribution.

Copulas are functions that combine univariate marginal distributions into multivariate distribution functions. A key theorem here is Sklar's theorem (Sklar, 1959), which states that given an n-dimensional distribution function $H_n(x_1,\ldots, x_n)$ with marginal distributions $h_1(x_1),\ldots, h_n(x_n)$, there exists a copula C such that

$$H_n(x_1,\ldots, x_n) = C(h_1(x_1),\ldots, h_n(x_n)). \quad (4)$$

Furthermore, if $h_1(x_1),\ldots, h_n(x_n)$ are continuous, C is unique.

Copulas thus combine the information from the marginal distributions (in this case, the individual model probability distributions) into a single distribution H_n. Consider two models that make a prediction based on different sets of parameters. If true hazard is not a function of one set of parameters or the other, but rather an albeit complicated function of both sets of parameters, then each model is giving information that should determine the final hazard. The final probability distribution for hazard should be a function of each model output. The copula is the function that combines the two probability distributions into one distribution.

From Sklar's theorem one can see that any multivariate distribution defines a copula. The choice of copula is a function of the dependence structure of the marginals, but not a function of the marginals themselves. Choosing the copula that correctly describes the dependence between two models is the most important part of this formulation. Clemen and Reilly (1999) discussed how to use expert opinion to this end. If the models are exchangeable in terms of their dependence, an Archimedian copula is ideal, as it treats the marginal distributions symmetrically (Jouini and Clemen, 2002). For more flexibility, the multivariate normal copula can be used (Clemen and Reilly, 1999). It encodes dependence by using pairwise correlation coefficients, for example, Spearman's ρ or Kendall's τ (Lehmann, 1966). This can be ideal in situations where there are few data, for the statistical measures of dependence ρ or τ can be assessed with expert opinion.

A Simple Example

Jouini and Clemen (2002) presented a simple method using copulas in a Bayesian framework to combine multiple probability distributions from experts. We follow their method here to combine two distributions from the WG02

report: the probability distributions given by the Poisson and empirical models.

Jouini and Clemen (2002) used a uniform prior distribution for the application of Bayes' Rule (equation 3). To calculate the likelihood function, they combined the individual expert probability distributions $f_i(\theta)$ using a family of copulas described by Frank (1979). This family of copulas is indexed by a single parameter, Kendall's τ (Kendall, 1938). For two independent and identically distributed pairs of random variables, Kendall's τ is defined as the probability of concordance minus the probability of discordance. The copula family of Frank (1979) can capture the full range of positive dependence from $\tau = 0$ (independence) to $\tau = 1$ (perfect positive dependence). Furthermore, Jouini and Clemen assume for simplicity that given a random median M_i given by the ith marginal, the marginal distributions are only dependent through their estimation errors $\theta - M_i$.

Under these assumptions, the posterior probability distribution is proportional to

$$c(1 - F_1(\theta), \ldots, 1 - F_n(\theta))f_1(\theta) \ldots f_n(\theta), \qquad (5)$$

where c is the copula density function, and F_i is the cumulative probability distribution generated from the ith marginal f_i.

An example of this copula-based method is applied to the Poisson and empirical probability distributions from the WG02 report and shown in Figure 5. We fit the individual

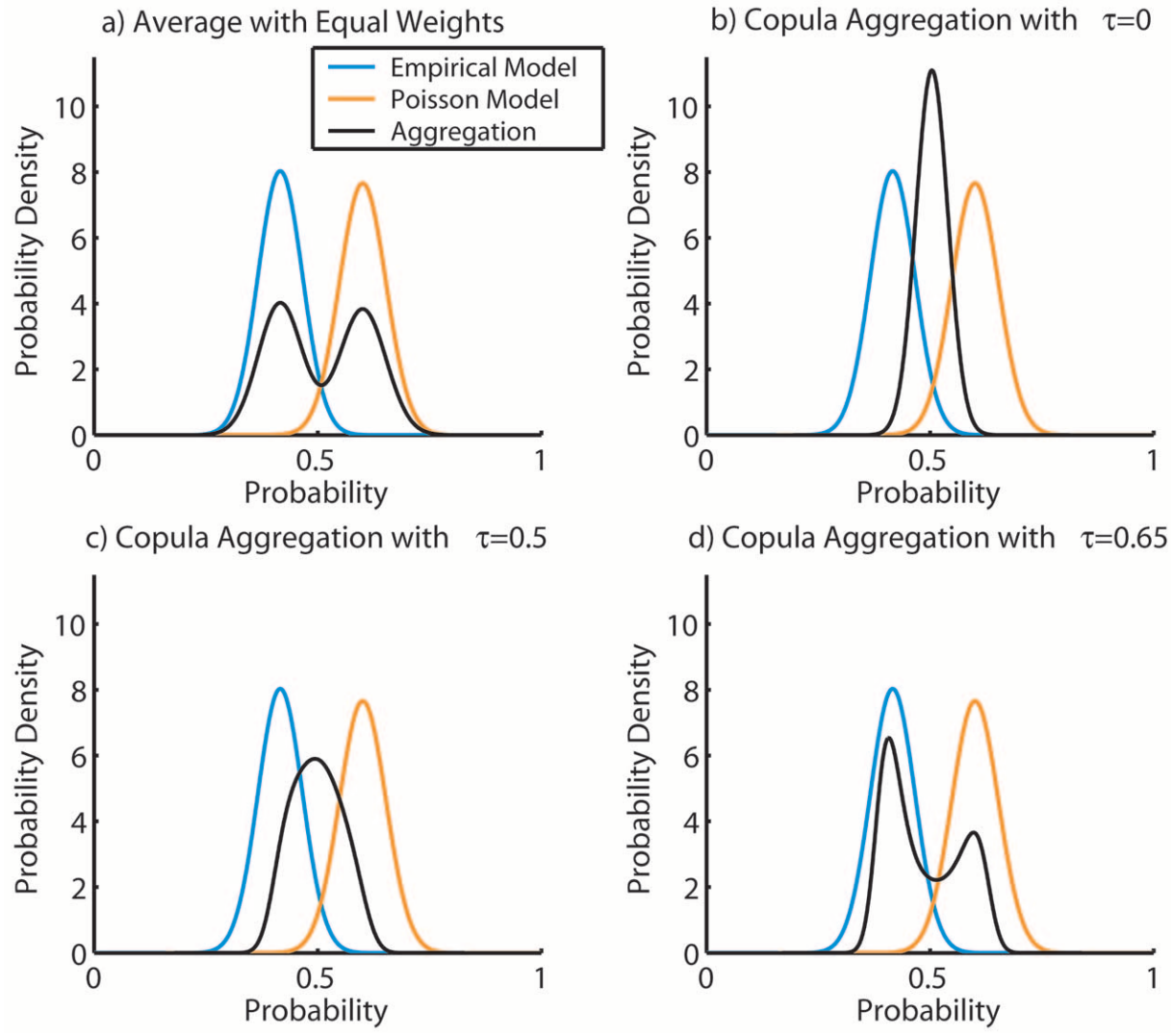

Figure 5. Combining the Poisson and empirical models with an equally weighted linear combination (a) leads to quite different results than a using a copula-based aggregation method (b)–(d). If the two models are independent, Kendall's τ equals zero and our copula reduces to the independence copula. This aggregation has the least variance (b). If we assume more dependence between the input models, then there is less total information. As we expect, in that case the combined model has higher variance, as shown (c) and (d).

probability distributions from Figure 1a with Gaussian distributions. Recall that each statistical model produces a distribution of values for the probability of an earthquake of magnitude 6.7 of greater. The width of the Gaussian that we fit is determined by the parameter uncertainty of the individual models. In Figure 5a, we show the regional probability distributions given by the Poisson and empirical models, along with an equally weighted average of the two in black. This is analogous to the WG02 methodology. Figure 5b, c, and d shows examples of the copula-based aggregation for three different levels of dependence. Figure 5b shows an aggregation in which the two models are assumed to be independent ($\tau = 0$). In this case, Frank's copula reduces to the independence copula $C = f_1 f_2$. If the two models are independent, then together they provide the most information, and the variance in the combined distribution is as small as possible. Higher levels of dependence result in more variance, as more weight is added to regions where the two distributions differ. Large values of Kendall's τ result in a bimodal combined distribution (Fig. 5d). In the case of two Gaussian marginals with equal variance, the combined distribution would be symmetric. In general, Frank's copulas give more weight to the marginal with less variance, and thus the empirical model is given more weight in the aggregated distributions in this example.

The copula-based aggregations shown in Figure 5b, c, and d do not assume that the Poisson and empirical models are mutually exclusive. Rather, they treat both models as marginal distributions. The Poisson model is a marginal distribution for historical seismicity and the empirical model is a marginal distribution for post-1906 seismicity. The copula is the function that combines the two marginals into the bivariate probability density.

The preceding approach can easily be generalized to incorporate more than two models. For each pair of models, one pairwise correlation coefficient is needed to completely define the copula from this particular family. Expert opinion could easily be used to this end. The main drawback of this formulation is that the answer is highly dependent on the type of dependence structure between the models, and thus on the copula chosen. Archimedian copulas, which treat the dependence between the marginal distributions symmetrically, are certainly the easiest to implement. The copulas we use here from Frank (1979) are members of this class. Modeling complex dependence structures, however, requires a more sophisticated analysis. To this end, Clemen and Reilly (1999) discuss methods using the copula underlying the multivariate normal distribution. In addition, MacKenzie (1994) develops a class of copulas with even more flexibility.

Treating the model weights as probabilities (as a linear combination does) is problematic because the weights are not mutually exclusive or collectively exhaustive. Copulas provide a way to combine multiple models without abandoning probabilism.

Conclusion

We have identified several problems with WG02's formulation of model uncertainty. In particular, using a linear combination of different models ignores model dependence and results in large uncertainties in their results. Although the true epistemic uncertainty may indeed be large, the actual amount of uncertainty is ultimately a function of the dependence structure between the models, which was not assessed. In addition, we find that choosing model weights based on data availability can lead to systematic bias. Eliminating bias is paramount because hazard estimates are used to assess insurance rates, building codes, and public policy. We seek a hazard formulation that is both correct in the sense that there is no statistical bias and as good as possible in the sense that it is as precise as the data allow. To do this, a proper formulation of model uncertainty and dependence is needed. We have presented here one possible solution that combines Bayesian methods with copulas to model dependence structures. These methods must be implemented with careful consideration of the type of dependence between different models. If implemented properly, these methods may reduce the total uncertainty in hazard estimates, as well as eliminate sources of bias in existing methodology.

Acknowledgments

We thank Ned Field, Ralph Archuleta, and Eric Dunham for many useful conversations. We also thank our anonymous reviewer, as well as Mark Petersen and Norm Abrahamson, for comments that improved this work. M.T.P. acknowledges the support of a Broida fellowship and Eugene Cota-Robles fellowship from University of California at Santa Barbara (UCSB), as well as a LEAPS fellowship as part of an NSF GK-12 grant to UCSB. In addition, this work was supported by the James S. McDonnell Foundation (Grant 21002070), the William. M. Keck Foundation, NSF Grant PHY99-07949, NSF Grant DMR-9813752, the David and Lucile Packard Foundation, and USGS NEHRP Grant 06HQGR0046.

References

Abrahamson, N. A., and J. J. Bommer (2005). Probability and uncertainty in seismic hazard analysis, *Earthquake Spectra* **21,** no. 2, 603–607.

Aposolakis, G. (1995). A commentary on model uncertainty, in *Model Uncertainty: Its Characterization and Quantification*, Center for Reliability Engineering, University of Maryland, College Park, Maryland, 13–22.

Bier, V. M. (1995). Some illustrative examples of model uncertainty, in *Model Uncertainty: Its Characterization and Quantification*. Center for Reliability Engineering, University of Maryland, College Park, Maryland, 93–100.

Clemen, R. T., and T. Reilly (1999). Correlations and copulas for decision and risk analysis, *Manage. Sci.* **45,** 208–224.

Clemen, R. T., and R. L. Winkler (1999). Combining probability distributions from experts in risk analysis, *Risk Anal.* **19,** 197–203.

Clemen, R. T., G. W. Fischer, and R. L. Winkler (2000). Assessing dependence: some experimental results, *Manage. Sci.* **46,** 1100–1115.

Ellsworth, W. L., M. V. Matthews, R. M. Nadeau, S. P. Nishenko, P. A. Reasenber, and R. W. Simpson (1999). A physically-based earthquake recurrence model for estimation of long-term earthquake probabilities, *U.S. Geol. Surv. Tech. Rep. OFR 99-522*.

Frank, M. J. (1979). On the simultaneous associativity of $F(x, y)$ and $x + y - F(x, y)$, *Aeq. Math.* **19,** 194–226.

Freedman, D. A., and P. B. Stark (2003). What is the chance of an earthquake?, *NATO Sci. Ser. IV: Earth Environ. Sci.* **32,** 201–213.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubinand (1995). *Bayesian Data Analysis*, Chapman and Hall, Boca Raton, Florida.

Jouini, M. N., and R. T. Clemen (2002). Copula models for aggregating expert opinions, *Oper. Res.* **44,** 444–457.

Kagan, Y. Y., and L. Knopoff (1987). Random stress and earthquake statistics: time dependence, *Geophys. J. R. Astr. Soc.* **88,** 723–731.

Kendall, M. G. (1938). A new measure of rank correlation, *Biometrika* **30,** 81–93.

Lehmann, E. L. (1966). Some concepts of dependence, *Ann. Math. Statist.* **37,** 1137–1153.

Lempert, R., N. Nakicenovic, D. Sarewitz, and M. Schlesinger (2004). Characterizing climate-change uncertainties for decision-makers, *Climatic Change* **65,** 1–9.

MacKenzie, G. (1994). Approximately maximum-entropy multivariate distributions with specified marginals and pairwise correlations, *Ph.D. Thesis*, University of Oregon.

Matthews, M. V., W. L. Ellsworth, and P. A. Reasenberg (2002). A Brownian model for recurrent earthquakes, *Bull. Seism. Soc. Am.* **92,** 2233–2250.

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

Morgan, M. G., and M. Henrion (1990). *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*, Cambridge University Press, Cambridge, 67–68.

Musson, R. M. W. (2005). Against fractiles, *Earthquake Spectra* **21,** no. 3, 887–891.

Senior Seismic Hazard Analysis Committee (SSHAC) (1997). Recommendations for probabilistic seismic hazard analysis: guidance on uncertainty and use of experts, Tech. Rep. NUREG/CR-6372, Lawrence Livermore National Laboratory.

Shimazaki, K., and T. Nakata (1980). Time-predictable recurrence model for large earthquakes, *Geophys. Res. Lett.* **7,** 279–282.

Sklar, A. (1959). Fonctions de répartition à n dimentions et leurs marges, *Pub. Inst. Stat. Univ. Paris* **8,** 229–231.

Winkler, R. L. (1995). Model uncertainty: probabilities for models? in *Model Uncertainty: Its Characterization and Quantification*, Center for Reliability Engineering, University of Maryland, College Park, Maryland, 109–118.

Working Group on California Earthquake Probabilities (WGCEP) (1990a). Probabilities of large earthquakes occurring in California on the San Andreas fault, *U.S. Geol. Surv. Tech. Rep. OFR 88-398*.

Working Group on California Earthquake Probabilities (WGCEP) (1990b). Probabilities of large earthquakes in the San Francisco Bay Region, California, *U.S. Geol. Surv. Circular 1053*, 51.

Working Group on California Earthquake Probabilities (WGCEP) (1995). Seismic hazards in southern California: probable earthquakes, 1994–2024, *Bull. Seism. Soc. Am.* **85,** 379–439.

Working Group on California Earthquake Probabilities (WGCEP) (1999). Earthquake probabilities in the San Francisco Bay Region: 2000–2030—a summary of findings, *U.S. Geol. Surv. Tech. Rep. OFR 99-517*.

Working Group on California Earthquake Probabilities (WGCEP) (2003). Earthquake probabilities in the San Francisco Bay Region: 2002–2031, *U.S. Geol. Surv. Tech. Rep. OFR 03-214*.

Department of Physics
University of California at Santa Barbara
Santa Barbara, California 93106-9530
pagem@physics.ucsb.edu
carlson@physics.ucsb.edu

Manuscript received 28 September 2005.