Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN
GEOFISICA
Ciclo XXVI

Settore Concorsuale di afferenza: 04/A4
Settore Scientifico disciplinare: Geo10 – GEOFISICA DELLA TERRA SOLIDA

EARTHQUAKE FORECASTING AND SEISMIC HAZARD ANALYSIS: SOME INSIGHTS ON THE TESTING PHASE AND THE MODELING

Taroni Matteo

Coordinatore Dottorato

Prof. Michele Dragoni

Relatore

Dott. Warner Marzocchi

Anno 2014

Index

Introduction

> *"Il calcolo delle probabilità, sorto da umili origini, attira ogni giorno più l'interesse dei matematici e dei cultori delle scienze fisiche e sociali. Per quale ragione questa disciplina, di cui l'antichità classica non avrebbe forse nemmeno inteso lo spirito, ha preso uno sviluppo così rigoglioso nell'epoca che vede il trionfo dei metodi sperimentali? Il ravvicinamento non è fortuito. Stretti e molteplici sono i legami tra il calcolo delle probabilità e le scienze di osservazione."*
>
> "The calculus of probability, born from humble origins, attracts every day more the interest of mathematics and researchers of physical and social sciences. Why this subject, of which maybe the ancient Greek thinking doesn't have understood the wit, is growing so high during the age of experimental methods? This approaching is not made by chance. There are various and strong connections between the calculus of probability and the observational sciences."

This is the first paragraph of the introduction of the Guido Castelnuovo's book "Calcolo delle probabilità" that he wrote in 1918.
Why starting a thesis in geophysics using the words of a book on calculus of probability? The calculus of probability and statistics became very important in seismology since the '50s when Gutenberg and Richter formulated their famous law about the (statistical) distribution of the magnitudes of earthquakes. In the following decades there has been a substantial increase of the use of probabilistic and statistics methods in seismology.
Nowadays one of the main goals of modern seismology is to forecast the future seismic activity in one region. When deterministic predictions are not possible, we can only rely on probabilities. Paraphrasing Decker, concerning potentially active volcanoes, "*geologists can find out what did happen, geophysicists can determine what is happening, but to find out what might happen requires statistics*."
De facto, probabilities are the basic component of probabilistic seismic hazard analysis (PSHA). The scope of the PSHA is to calculate the probability to exceed a certain level of PGA (Peak Ground Acceleration) in a defined space-time domain. A trustworthy PSHA has to be based on reliable seismological information and models, and a proper probabilistic framework to combine all

of them. The evaluation of the reliability of models and the definition of a proper probabilistic framework to combine all information and uncertainties are the main targets of this thesis. The study covers both general and specific aspects of these issues.

Using a metaphor that was used at the beginning of the past century relatively to mathematics, the tree of the modern statistical seismology applications grows very rapidly increasing the number of branches in the last years; nevertheless, we note that this increase was not balanced by careful evaluation of the state of the roots of this discipline. A tall tree without robust roots may collapse. Some evidence of the lack of robust roots can be found in the recent critics to the probabilistic assessment of the seismic hazard made by several authors and that are not yet seriously tackled.

In the first chapter we analyse the results of the world forecasting experiment run by the *Collaboratory for the Study of Earthquake Predictability* (CSEP), that is an international cooperation of researchers conducting numerical seismicity forecast experiments. After many decades of *ex post facto* earthquake predictions and individual case studies exploring precursory seismicity patterns, earthquake scientists now broadly view prospective forecast experiments, such as those that CSEP conducts, as the only 'true' test of a model.
Models that forecast earthquake rates in one specific space-time domain are an important part of the PSHA procedure. We take the opportunity of this experiment to contribute to the definition of a more robust and reliable statistical procedure to evaluate earthquake forecasting models. In this chapter we first present the models and the target earthquakes to be forecast. Then we explain the consistency and comparison tests that are used in CSEP experiments to evaluate the performance of the models.
Our contribution consists of:

i) Putting forward a new method to evaluate models that does not use the classical log-likelihood as measure of the performance. This is important because the log-likelihood tends to privilege

models that never fail with respect to model that may fail once but they may have a much better overall performances over the rest of earthquakes.

ii) Providing some clues about how to rank the models' performance. This is very important to understand which model is better and to quantify these differences.

iii) Introducing a methodology to create ensemble forecasting models. We show that models, when properly combined, are almost always better performing that any single model. This modelling is very important from practical purposes because it offers a model that can be used in any kind of PSHA and it accounts for the so-called epistemic uncertainty.

In the second chapter we discuss in depth one of the basic features of PSHA: the declustering of the seismicity rates. We first introduce the Cornell-McGuire method for PSHA and we present the different motivations that stand behind the need of declustering seismic catalogs. Our contribution to this topic consists is summarized here:

i) We show that declustering seismic catalogs unavoidably brings to an underestimation of PSHA.

ii) Using a theorem of the modern probability (the Le Cam's theorem) we show that the declustering is not necessary to obtain a Poissonian behaviour of the exceedances that is usually considered fundamental to transform exceedance rates in exceedance probabilities in the PSHA framework.

iii) We present a method to correct PSHA for declustering, building a more realistic PSHA.

In the last chapter we explore the methods that are commonly used to take into account the epistemic uncertainty in PSHA. The most widely used method is by far the logic tree that stands at the basis of the most advanced seismic hazard maps. In the first part of the chapter we illustrate the probabilistic structure of the logic tree, and then we show that this structure is not adequate to describe the epistemic uncertainty. This means that then output of the logic tree is often misinterpreted. We then propose a new probabilistic framework based on the ensemble modelling that properly accounts for epistemic uncertainties in PSHA. Noteworthy, we show that in most of practical applications the logic tree is not implemented in its standard (and proper) way, but de facto as a tool to create an ensemble model.

I want to thank Warner Marzocchi, who worked more than he should for this thesis, Jeremy Zechar, Pamela Roselli and the pollajo's guys who helped me in many different ways.

Assessing annual global M6+ seismicity forecasts

Taroni M.[1], Zechar J.D.[2,3], Marzocchi W.[1]

[1] Istituto Nazionale di Geofisica e Vulcanologia, Rome, Italy

[2] Swiss Seismological Service, ETH Zurich, Switzerland

[3] Department of Earth Sciences, University of Southern California, USA

Abstract

We consider a seismicity forecast experiment conducted during the last four years. At the beginning of each year, 3 models make a 1-year forecast of the distribution of large earthquakes everywhere on Earth. The forecasts are generated and the observations are collected in the Collaboratory for the Study of Earthquake Predictability (CSEP). We apply CSEP likelihood measures of consistency and comparison to see how well the forecasts match the observations, and we compare results from some intuitive reference models. These results illustrate some undesirable properties of the consistency tests: the tests can be extremely sensitive to only a few earthquakes, and yet insensitive to seemingly obvious flaws—a naïve hypothesis that large earthquakes are equally likely everywhere is not always rejected. The results also suggest that one should check the assumptions of the so-called *T* and *W* comparison tests, and we illustrate some methods to do so. As an extension of model assessment, we explore strategies to combine forecasts, and we discuss the implications for operational earthquake forecasting. Finally, we make suggestions for the next generation of global seismicity forecast experiments.

Key words:

Probabilistic forecasting; statistical testing of models; Earthquake interaction, forecasting, and prediction; statistical seismology.

Introduction

"Prediction is very difficult, especially when it is about the future." This statement, attributed to Niels Bohr (or Yogi Berra, depending on whom you ask), highlights the importance of testing a model out-of-sample: checking if a model can forecast data that were not used to build the model. In a recent article, Marzocchi & Zechar (2011) emphasized the dual importance of this type of forecasting for seismology: from a philosophical point of view, forecasting is the cornerstone of scientific knowledge (AAAS, 1989); and from a practical perspective, forecasting is crucial for forming sound risk mitigation strategies. In other words, forecast experiments allow us to understand what we really know about earthquake occurrence processes, and they also guide our efforts to provide the best model for reducing risks. For philosophical and practical ends, thorough model assessment is essential, and this is the main goal of the Collaboratory for the Study of Earthquake Predictability (CSEP), an international cooperation of researchers conducting numerical seismicity forecast experiments (Jordan 2006; Zechar et al. 2010). After many decades of *ex post facto* earthquake predictions and individual case studies exploring precursory seismicity patterns, earthquake scientists now broadly view prospective forecast experiments, such as those that CSEP conducts, as the only 'true' test of a model. This follows the pioneering work of: Y. Kagan & L. Knopoff on statistical earthquake forecasting (Kagan & Knopoff 1977, 1987); Y. Kagan & D. Jackson on the seismic gap hypothesis (Kagan & Jackson 1991, 1995; Rong et al. 2003) and smoothed seismicity forecasting (Kagan & Jackson 1994, 2000); and F. Evison & D. Rhoades in the development of the precursory swarm hypothesis (Evison & Rhoades 1993, 1997, 1999) and practical applications of forecast models (Rhoades & Evison 1989). One disadvantage of prospective seismicity forecasting is that relatively short regional experiments might result in small samples, meaning that one might have to wait to obtain 'meaningful' results or that one earthquake sequence may dominate the results of an experiment. (Let us ignore the problem of unambiguously delimiting a sequence.) For example, in the CSEP-Italy experiment (Schorlemmer et al. 2010a), only 9 target earthquakes have occurred since the experiment began on August 1, 2009, which makes it difficult to make robust inferences. To address this deficiency, one can do what D. Jackson calls 'trading space for time': consider larger regions to obtain larger samples. Again following the trail blazed by Kagan & Jackson (1994), researchers began a prototype experiment in the western Pacific in late 2008, with 3 models participating as of 1 Jan 2009 (Eberhard et al. 2012). At that time, these researchers agreed to participate in a prototype global experiment with the same three models. The current western Pacific and global experiments are prototypes in the sense that only a few researchers are participating and model development has been minimal: the models were adapted from regional CSEP experiments with few changes. Along with accumulating larger

samples, the primary motivations for the prototype global experiment were to explore the availability and reliability of global earthquake catalogs and to determine the testing center features needed for future large-scale, multi-investigator experiments.

A global seismicity forecast experiment can be thought of as a sandbox in which to test hypotheses related to seismogenesis and earthquake triggering. In addition to advancing basic understanding of large earthquake nucleation, a global experiment with broad participation has the potential to impact seismic risk reduction. For instance, such an experiment may indicate the best models to be used for operational earthquake forecasting (OEF) at different time scales (Jordan et al., 2011). Findings from a global CSEP experiment could also influence development of related projects such as the Global Earthquake Model (http://www.globalquakemodel.org).

Because a global seismicity forecast experiment has the potential to impact basic research and seismic risk mitigation, the prototype CSEP experiment is worthy of careful consideration. In this article, we present results from the assessment of the three participating models. In this context, we are also interested in exploring how CSEP model assessment works, how it could be improved, and how CSEP models could be applied in the context of OEF.

In the following section, we describe the participating models, methods for combining them, and conceptual reference models that are useful for understanding the assessment results. In sections 3 and 4, we describe the data and assessment techniques, respectively, used in this study. We present the results of the first 4 years of the prototype global experiment in section 5. In Section 6, we discuss the results and suggest strategies for improving CSEP assessment and selecting the best model for operational purposes. We conclude by making recommendations for a full-fledged global experiment.

Models

In this article we consider three classes of models: those that have been running in the US CSEP testing center, those that are formed via combinations of the first group, and hypothetical reference models used to impart understanding of the model assessments. For brevity, we call these groups the CSEP models, the ensemble models, and the reference models, respectively. Only the CSEP models formally participated in the experiment; the others were constructed after the experiment ended

The CSEP models are fully-specified stochastic models represented by codes running in the testing center; at the beginning of each calendar year, these codes generate a one-year forecast. We assess the resulting forecasts by comparing them with each other and with the observed seismicity. We apply the same assessments to the ensemble and reference models.

Every forecast consists of 64 800 1° longitude x 1° latitude cells in which the number of expected earthquakes with moment magnitude Mw≥5.95 and depth ≤30 km in the next year is forecast. As in other CSEP experiments, every modeler has agreed that the Poisson distribution should be used to represent the uncertainty in the annual rate in each cell.

CSEP models

The three CSEP models in the prototype global experiment are the same as those in the western Pacific experiment (Eberhard et al. 2012): DBM, the double branching model (Marzocchi & Lombardi 2008); KJSS, the Kagan and Jackson smoothed seismicity model (Kagan & Jackson 2000, 2010); and TripleS, a time-invariant simple smoothed seismicity model (Zechar & Jordan 2010). Although the prototype global experiment began in 2009, the KJSS model was not implemented in the testing center until 2010, so we can only present results for KJSS in 2010, 2011, and 2012. In the supplement to this article, we show map-views of each forecast and the observed target earthquakes.

The DBM attempts to model two types of temporal clustering. One represents the well-known short-term clustering that characterizes classical aftershock sequences (Ogata 1988, 1999) and the other is related to clustering that was found on a longer timescale and may be due to the post-seismic effects of earthquakes or other long-term modulation of seismic activity. The KJSS model is a probabilistic model based on smoothed seismicity with an anisotropic smoothing kernel; this type of kernel uses an orientation function that depends on the presumed fault plane of the earthquake being smoothed. The kind of focal mechanism is also taken into account. In contrast to the DBM model, KJSS uses a tapered version of the Gutenberg-Richter relation (Kagan & Jackson 2000). TripleS uses a two-dimensional Gaussian smoothing kernel with only one parameter to smooth past seismicity and construct a predictive density. One peculiarity of this model is that in its implementation all earthquakes in the catalog are used, even the ones below the completeness magnitude; this is intended to allow for a more accurate spatial description of seismicity. Another peculiarity is that, owing to an optimization in the model code (see Zechar & Jordan 2010, eq. 5, and recall Knuth's (1974) warning that "premature optimization is the root of all evil"), TripleS forecasts have many cells with zero expected earthquakes. This implies that earthquakes are impossible in these cells. Because some earthquakes happened in cells with zero expectation in 2009, 2010 and 2012, the TripleS models in these years have likelihood equal to zero. To better understand the performance of this model, we consider a modified TripleS model (TripleS*), replacing the rates in these zero cells with a rate of 10^{-300}, which is the smallest nonzero number that is representable on the computer we used for assessment. We do this so that we can explore the

forecasting capabilities of TripleS, even though, officially, this model fails all CSEP tests that are based on likelihood. We note that these zero rates did not have an effect in the CSEP-Italy experiment for which the TripleS code was developed. We suggest that future applications of TripleS adopt a minimum rate greater than zero for each cell.

Ensemble models

While CSEP experiments primarily emphasize the scientific study of seismicity in isolation—that is, divorced of its ultimate impact on society–the results of these experiments can have practical implications. For example, these experiments can inform how we select the best model (Marzocchi & Zechar 2011). But Marzocchi et al. (2012) suggested that rather than selecting the best model from those in a testing center, one can create ensemble models based on CSEP experiment outcomes; in particular, they also showed that an ensemble model can outperform the single best model. Along the same lines, Rhoades & Gerstenberger (2009) conducted similar analyses in which they mixed a long-term model and a short-term model and obtained a yet more informative model. In this study, we investigated four types of ensemble models distinguished by how they were constructed: score model averaging (SMA), generalized score model averaging (gSMA), Bayes factor model averaging (BFMA), and parimutuel gambling model averaging (PGMA). The first two models were described and illustrated by Marzocchi et al. (2012), as was Bayesian model averaging (BMA), which is based on the Bayesian posterior probability. Despite its widespread use, Marzocchi et al. (2012) found that BMA is not particularly well-suited to seismicity forecast experiments because the resulting ensemble average is often dominated by the single best model, regardless of its reliability.

These ensemble models are built by calculating a weighted average of the rates in each cell of the CSEP forecasts. SMA uses a weight that is inversely proportional to a model's log-likelihood, while gSMA uses a weight that is inversely proportional to the difference between a model's log-likelihood and a reference value (if set to 0, gSMA is the same SMA; see Marzocchi et al., 2012). Following Marzocchi et al. (2012), we choose the reference value to be the best model's log-likelihood:

$$\omega_i^{SMA} = \frac{1}{|L_i|} \tag{1}$$

$$\omega_i^{gSMA} = \frac{1}{|L_i - L_0| + 1} \tag{2}$$

where L_i is the log-likelihood of the *i*-th model and L_0 is the log-likelihood of the best performing model; the value 1 in the denominator of (2) ensures finite weight for the best model.
In contrast to SMA and gSMA, PGMA does not use model likelihood; instead it uses the results of a parimutuel gambling analysis that is inherently comparative (Zechar & Zhuang 2012); see Section 4.3 for details. PGMA assigns a weight to the *i*-th model according to this formula:

$$\omega_i^{PGMA} = 1 + \alpha \cdot V_i \tag{3}$$

where V_i is the parimutuel gambling score of the model, $\alpha = \frac{0.90}{|V_{\max}|}$, and V_{max} is the maximum loss among all models; in this case, therefore, a model's weight can be reduced at most by 90%.

The fourth ensemble model, BFMA, blends concepts from the others: it is based on likelihood but imposes a penalty similar to that used for PGMA. Each model is weighted according to its total Bayes factor which in a purely prospective experiment is the total likelihood ratio (Kass & Raftery, 1995). One calculates the total Bayes factor of a model by summing the log-Bayes factor of the model with respect to all the others. The corresponding weight is:

$$\omega_i^{BFMA} = 1 + \beta \cdot TBF_i \tag{4}$$

where TBF_i is the total Bayes factor of the i-th model, $\beta = \frac{0.90}{|TBF_{\min}|}$, and $TBF_{\min}$ is the minimum *TBF* among all models.

For PGMA and BFMA, the value of 0.90 in the penalty term is arbitrary, but it is necessary to have a number less than 1 to keep the weights positive.
For 2009, because we have no measure of past model performance, so all 4 ensemble models are identical and give 50% weight to the DBM forecast and 50% weight to TripleS. For 2010, we constructed the ensemble models using the DBM and TripleS performances in 2009 (KJSS does not have any performance for 2009). For 2011 and 2012, all 3 models contribute to build the 4 ensemble models; in this case we have 3 models that compose the ensemble so we also adjust the coefficient to account for the correlation of the forecasts (Marzocchi et al. 2012 paragraph 4.1). The composition of each ensemble model for each year is shown in Figure 1.

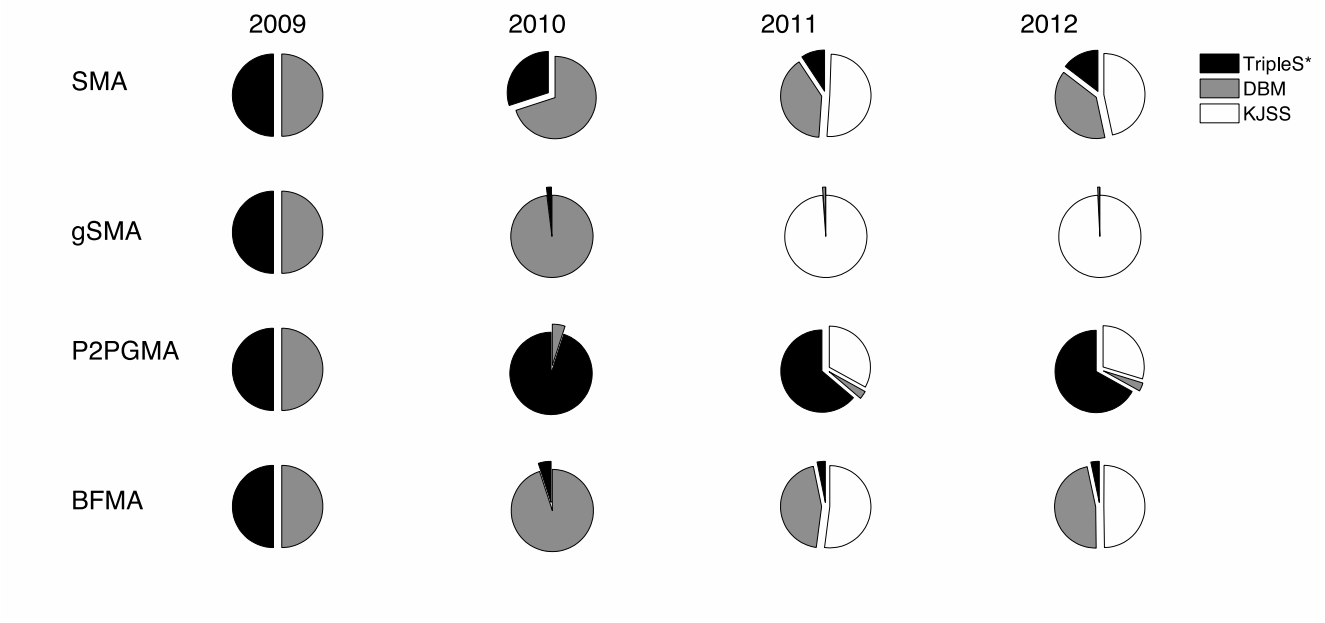


Figure 1.

Composition of ensemble models for each annual experiment. In 2009, KJSS was not available and there was no prior performance result, so every ensemble was simply 50% TripleS, 50% DBM.

Reference models

To better understand the consistency and comparison tests, we consider three simple reference models (following Werner et al. 2010): the Uniform model (UNIF), which has the same rate for each cell (scaled to the area of the cell) and a total rate that is equal to the number of target earthquakes in the previous year; the Perfect Poisson model (PPM), which in each cell has a rate equal to the number of observed target earthquakes in that cell, and the Semi-Perfect Poisson model (SPPM) which is obtained by dividing PPM by two everywhere. We note that the PPM does not have likelihood equal to 1, because in the CSEP experiment Poisson uncertainty is assumed (a likelihood of one could only be achieved with Dirac distribution in each cell); nevertheless, it has the best possible performance given the Poisson restriction.

Data

Experiment participants agreed to use the Global Centroid Moment Tensor (GCMT) catalog (Dziewonski et al. 1981; Ekström et al. 2005) for model development and assessment. Compared with other global earthquake catalogs, the GCMT catalog is homogeneous: the time, location, and size of each earthquake are estimated using the same procedure for each earthquake. In this study, we use the epicentroid and, following Eberhard et al. (2012), we calculate the moment magnitude Mw from the total moment M_0 reported in the catalog. Kagan (2003) suggested that the catalog is

complete for earthquakes occurring at depths no greater than 70 km at Mw 5.3. For the prototype global experiment, participants agreed that target earthquakes would be all those with magnitude not smaller than Mw 5.95 and depths not greater than 30 km—we did not decluster the catalog. In the first four years of this experiment, the number of target events was 92, 103, 108, and 91, respectively.

Assessment techniques

The statistical tests used in CSEP testing centers can be grouped in two categories: consistency and comparison. The purpose of consistency tests is to determine whether the observed distribution of target earthquakes is consistent with a given forecast. When discussing consistency tests in the context of RELM assessment, Schorlemmer et al. (2007) proposed that a model that failed a consistency test should be 'rejected.' But the consistency and comparison tests can yield a counterintuitive situation: a model that fails a consistency test may, in a comparison test, be deemed better than a model that passes all consistency tests. To understand this apparent paradox, consider the following: in the 2009-2010, 2010-2011, and 2011-2012 seasons, Kevin Durant led the National Basketball Association in scoring, averaging 30.1, 27.7, and 28.0 points per game, respectively. On 26 November 2012 against the lowly Charlotte Bobcats, Durant scored only 18 points, nowhere near his 2012-2013 season average of 28.1. And yet no other player on either team scored as many points. In this example, Durant would be judged the best scorer among the players, but the observation would be inconsistent with expectations.

Currently CSEP testing center use the following consistency tests: the N(umber)-test, the L(ikelihood)-test, the S(pace)-test, and the M(agnitude)-test (Zechar et al. 2010). Because the forecasts in this study only have one magnitude bin per spatial cell ($Mw \geq 5.95$), we disregard the M-test, which is used to assess the magnitude distribution of a forecast. In planning the RELM experiment, the likelihood ratio was suggested for testing the hypothesis that two models have equal forecast skill (Schorlemmer et al., 2007). But Rhoades et al. (2011) highlighted flaws with the corresponding so-called R-test and suggested applying classical tests to the rate-corrected average information gain per earthquake in the T- and W-tests. In this study, we check the assumptions of the T- and W-tests and, when the assumptions appear to be violated, especially the symmetry of the distribution, we suggest using the Sign-test (Dixon & Mood 1946) that emphasizes median behavior rather than mean behavior. The rationale for using the Sign-test is that for asymmetric distributions the median is a more appropriate indicator of the center of the distribution with respect to the mean,

it is less sensitive to outliers. We also consider two assessment methods that do not involve statistical hypothesis tests: the Bayes factor and parimutuel gambling.

Note that we do not intend to replace any of the metrics currently being used in CSEP experiments, but we do urge researchers to consider the assumptions of the tests. Moreoever, the approaches we suggest are intended to be informative in situations where the assumptions of the existing tests are violated, and they allow one to answer a wider range of questions that could be raised by different stakeholders.

Consistency tests

N-test

This test compares the total number of earthquakes forecast (N_{fore}) with the observed number (N_{obs}); the N-test result is summarized by two quantile scores, δ_1 and δ_2, that are

$$\delta_1 = 1 - F((N_{obs} - 1) \mid N_{fore}) \tag{5}$$

$$\delta_2 = F(N_{obs} \mid N_{fore}) \tag{6}$$

where $F(.)$ is the cumulative Poisson distribution. If one of these scores is below the critical threshold value, the forecast is deemed to be overpredicting or underpredicting, respectively; because the N-test is a two-sided test, using a critical value of 0.025 corresponds to 5% significance.

L-test

This test compares the likelihood of a model with a set of likelihoods simulated to be consistent with the model being assessed (Zechar et al. 2010); the L-test result is summarized by a quantile score, γ:

$$\gamma = \frac{\#\{L_x \mid L_x \leq L, L_x \in L_S\}}{\#\{L_S\}} \tag{7}$$

where $\{L_S\}$ is the set of simulated likelihoods, L is the likelihood of the model with respect to the observed catalog, and $\#\{A\}$ indicates the number of elements in a set $\{A\}$. If γ is below the critical threshold value, the forecast is deemed to be inconsistent with the space-rate distribution of the observation; because the L-test is a one-sided test (it has been noted that very high values of γ

should not be used to judge a forecast (Schorlemmer et al. 2007)), a critical value of 0.05 corresponds to 5% significance.

S-test

This test is very similar to the L-test, but it is applied to a forecast after normalizing it so the total forecast rate matches the observed number of earthquakes, thereby isolating the spatial component of the forecast. After normalizing the forecast, the comparison is the same as in the L-test and the S-test result is likewise summarized by a quantile scores, ζ:

$$\zeta = \frac{\#\{S_x \mid S_x \le S, S_x \in S_S\}}{\#\{S_S\}} \tag{8}$$

where $\{S_S\}$ is the set of simulated spatial likelihoods and S is the likelihood of the spatial forecast relative to the observed catalog. If ζ is below the critical threshold value, the spatial forecast is deemed inconsistent.

Comparison tests

The T-test, W-test, and Sign-test are based on the information gain (IG) of one model relative to another (Rhoades et al. 2011, Eberhard et al. 2012). This measure is intended to indicate which of two models is more informative and for Poisson forecasts, the information gain for the i-th event is defined as:

$$IG_i(A,B) = \ln(\lambda_{A_i}) - \ln(\lambda_{B_i}) - \frac{\Lambda_A - \Lambda_B}{N} \tag{9}$$

where λ_{Ai} is the rate of model A in the bin where the i-th earthquake occurs, λ_{Bi} is the same for model B, Λ_A is the total rate for the model A, Λ_B is the same for model B, and N is the total number of target earthquakes.

T-test

In the context of CSEP experiments, the T-test is an application of Student's paired two-sample t-test (Student 1908) to the IGs of two models. The null hypothesis is that the information gains are independent samples from a normal population with zero mean. This null hypothesis suggests that the models are equally informative. The T-test assumes that the IGs are normally distributed; we

check this assumption using the Lilliefors test (Lilliefors 1967), which is a variant of the one-sample Kolmogorov-Smirnov test. When the normality assumption does not hold, the Central Limit theorem guarantees that the *T*-test becomes increasingly accurate as $N \to \infty$. In this study we have for some cases (year 2009) fewer than 100 samples, raising some doubts on the validity of the Central Limit theorem. To be conservative, we follow the suggestion of Rhoades et al. (2010): when the sample IGs are not normally-distributed, we use the W-test instead of the T-test.

W-test

The Wilcoxon signed-rank test (e.g. Siegel 1956) is a nonparametric alternative to Student's paired two-sample t-test. The null hypothesis of the W-test is that the median information gain is zero. The W-test assumes that the IGs are symmetrically distributed; we use the Triples test (Randles et al. 1980) to check this assumption. If the Triples test indicates that the samples are not symmetric, both the W-test and the T-test are not appropriate.

Sign test

The Sign test (Dixon & Mood 1946) is another non-parametric alternative to Student's paired two-sample T-test. The null hypothesis for the Sign test is that the medians of the two models likelihood are equal. In contrast to the T- and W-test, the Sign-test does not assume that the information gains are symmetric (Gibbons and Chakraborti, 2003). Hence, the Sign-test is more widely applicable, but it is less powerful than the other tests when the information gain distribution is symmetric (Gibbons and Chakraborti, 2003).

Bayes factor (likelihood ratio)

In a prospective experiment, the likelihood ratio and the Bayes factor are equal. Nonetheless, the Bayesian interpretation of this ratio allows us to overcome the problems inherent to the likelihood ratio tests (Rhoades et al., 2011; Marzocchi et al., 2012) and to score the forecasting capabilities of each model. Specifically, if we have two models A and B, and a set of observations Ω, the posterior odds of A and B can be expressed as:

$$\frac{P(A \mid \Omega)}{P(B \mid \Omega)} = \frac{P(A)}{P(B)} \frac{P(\Omega \mid A)}{P(\Omega \mid B)} = \frac{P(A)}{P(B)} BF(A,B) \tag{10}$$

where *BF(A,B)* is the Bayes factor (Kass & Raftery, 1995) of A versus B; when two models have the same prior and zero degrees of freedom, as in CSEP experiments, the posterior odds are exactly equal to BF, that is the likelihood ratio. Generally speaking, when log(BF) > 0 model A is more

supported by the data than B. Kass & Raftery (1995) proposed a guide to interpreting the numerical values of the Bayes factor; we reproduce this in Table 1. We stress that the Bayes factor is not a test, like the T-, W- or Sign-test; it is only a metric to compare model performance. This allows us to rank the performance of the models in a straightforward way.

log(BF)	**Evidence against M1**
0 to 1.1	Hardly worth mentioning
1.1 to 3	Positive
3 to 5	Strong
> 5	Very strong

Table 1. Guide to interpreting the log-Bayes factor (after Kass & Raftery 1995).

Parimutuel gambling score

Unlike the other comparison tests discussed in this subsection, the parimutuel gambling approach simultaneously compares all models, rather than considering each pair. The parimutuel gambling analysis also yields a ranking of models determined by the total amount 'won' by each model in the following game: for each cell of the grid, each model bets one credit (one unit of probability spread across the expected number of earthquakes), and at the end of the experiment is returned an amount proportional to the total number of competing models and to the probability of occurrence provided for that cell. The score for the first model in the *j*-th cell is:

$$R_j^1 = -1 + n\frac{p_j^1}{\sum_{i=1}^{n} p_j^i}$$

(11)

where n is the total number of models and p_j is the respective probability for that cell. The parimutuel gambling score is obtained by summing these returns over all cells. Unlike the measures based on likelihood, this type of score is much less sensitive to a target earthquake occurring in a cell with very low or zero rate, because each model can lose at most one credit in any cell.

Results

The results of the consistency tests and the comparison tests are reported in Figure 2 and Tables 2 and 3.

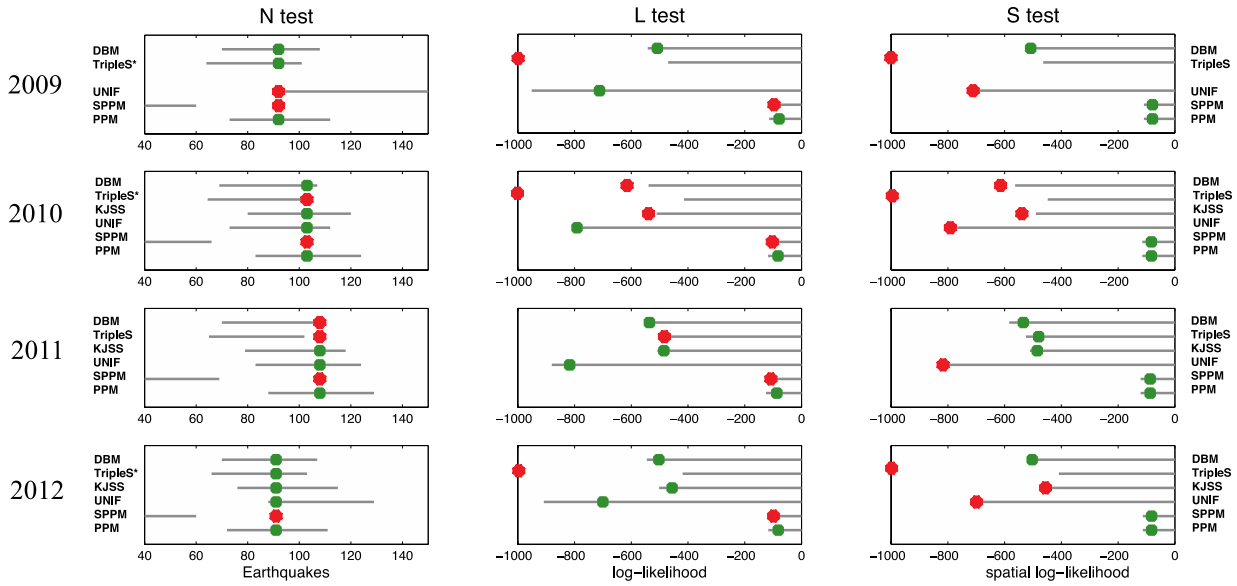


Figure 2. Results of CSEP consistency tests for CSEP and reference models. For the N test we consider the total number of earthquakes forecast by each model (the total rate), for L test and S test we consider the log-likelihood and the space log-likelihood of each model, respectively. The gray line shows the non-rejection region, at 5% significance; green squares show the values that have passed the test, while the red circles show the values that did not pass the test. A red circle on the left edge of the box indicates a very low value that falls outside the x-axis scale.

	2009	**2010**	**2011**	**2012**
DBM		L S	N	
TripleS*	L S	N L S	N L	L S
KJSS	**n/a**	L S		S
SMA		L		
gSMA		L		S
PGMA		N L S		
BFMA		L S	N	
UNIF	N S	S	S	S
PPM				
SPPM	N L	N L	N L	N L

Table 2. Results of consistency tests for CSEP, ensemble and reference models in 2009, 2010, 2011 and 2012. The letters indicate the test that has been rejected at a 0.05 significance level.

	DBM vs. TripleS*		DBM vs. KJSS		TripleS* vs. KJSS	
	Sign- or W-test	Bayes factor	Sign- or W-test	Bayes factor	Sign- or W-test	Bayes factor
2009	TripleS* (Sign) (-2.90 -0.40 0.75)	DBM 608	n/a	n/a	n/a	n/a
2010	DBM (W) (-3.45 -0.99 1.83)	DBM 1980	= (-2.68 -0.37 0.91)	KJSS 75	TripleS* (Sign) (-1.84 0.36 2.26)	KJSS 2056
2011	= (-3.30 -0.11 1.08)	TripleS 52	= (-2.33 -0.15 0.88)	KJSS 50	= (-1.66 0.07 1.49)	= 2.3
2012	DBM (W) (-3.30 -0.51 0.65)	DBM 616	= (-2.75 -0.41 1.16)	KJSS 46	KJSS (W) (-1.15 0.36 1.95)	KJSS 663
2010-12	DBM (W) (-3.36 -0.33 1.23)	DBM 2544	= (-2.63 -0.23 1.00)	KJSS 171	TripleS* (Sign) (-1.70 0.23 1.64)	KJSS 2716

Table 3. Results of the comparison tests. In each cell, we report the better model according to the W- or Sign-test, followed by the better model according to the Bayes factor analysis. We use '=' to indicate that the models are not significantly different (W-, Sign-test) or that the differences are hardly worth mentioning or positive (Bayes factor < 3). In the second line of each cell we report the 10^{th}, 50^{th} and 90^{th} percentile of IG and the absolute value of the log-Bayes factor.

No CSEP model passes all 3 consistency tests (N-, L-, and S-test) in the 4 years considered (2009, 2010, 2011, 2012). TripleS models spatial clustering and fails 9 of 12 consistency tests, while the very simple UNIF model fails only 5 and almost always passes the N-test and the L-test. Moreover, SPPM never passes the N-test or the L-test, despite obtaining a much higher likelihood (annual mean log-likelihood -109.3) than all CSEP models and UNIF (annual mean log-likelihood -755.1). (SPPM fails the N-test because it was designed to fail the N-test—recall that its forecast values are one-half of the observed values everywhere. And it fails the L-test because, as Schorlemmer et al. (2010b) noted, the L- and N-tests are dependent.) .

In considering the consistency tests for the RELM experiment, Marzocchi et al. (2012) pointed out that all models are wrong, so it must be only a matter of collecting enough data to show a statistical discrepancy and thereby fail a model with a consistency test. Here, we see that the consistency tests may be also misleading: some "good" models (e.g., SPPM and TripleS) may be dismissed more

often than "bad" models (e.g. UNIF). Hence, the consistency tests should not be used to reject any model but rather to understand where (the number of event, spatial or magnitude forecasts) a model can be improved.

For comparison tests, we explore the distribution of information gains to see which tests apply (Fig. 3). We begin with the Lilliefors test to check the assumption that information gains are normally distributed. If this assumption is not violated, we apply the T-test, and otherwise we use the Triples test to check the symmetry of the distribution. If the symmetry assumption does not hold, we apply the Sign-test. For all of the experiments in this study, we found that information gains are not normally distributed so we did not apply the T-test. The Bayes factor and the parimutuel gambling score do not include assumptions about the distribution of information gains and so we apply them in each experiment.

In Table 4 we report the ranking of the CSEP models according to the different comparison tests. For the W-, and Sign-test, the model with the first rank performs significantly better than the other models; a model with rank 2 performs significantly better than the model ranked 3 or higher. If the result of the applied test is not statistically significant, the models have the same rank. For the Bayes factor test, a model has a better rank if it scores "Positive", "Strong" or "Very strong" (see Table 1). Parimutuel gambling rankings do not include statements of statistical significance.

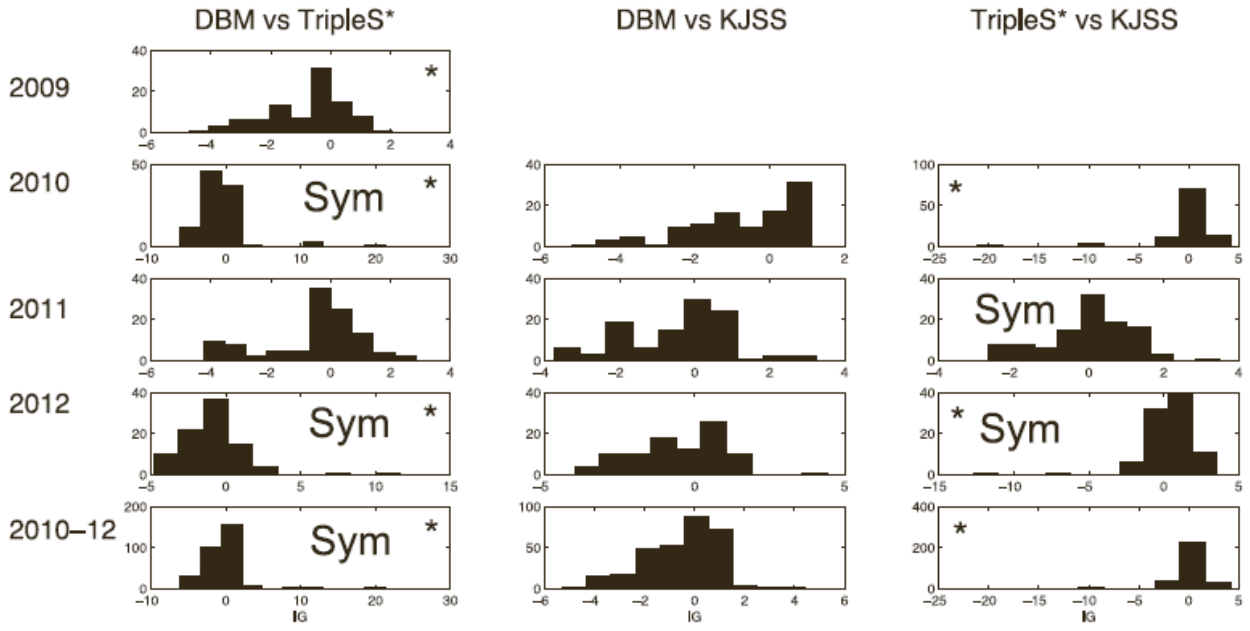


Figure 3. Information gain histograms for each pair of models (DBM vs TripleS*, DBM vs KJSS, TripleS* vs KJSS) for 2009, 2010, 2011, 2012, and all years considered jointly. The null hypothesis of a normal distribution was rejected at < 0.05 significance with a Lilliefors test for all model pairs.

Plots with the label “Sym” are of distributions for which the null hypothesis of a symmetric distribution was not rejected at < 0.05 significance with a Triples test. Plots with an asterisk have one or more outliers beyond the x-axis scale shown here.

	Sign- or W-test	**Bayes factor**	**PGS**
2009 (92)	1. TripleS* 2. DBM	1. DBM 2. TripleS*	1. TripleS 2. DBM
2010 (103)	1. DBM, TripleS*, KJSS	1. KJSS 2. DBM 3. TripleS	1. TripleS 2. KJSS 3. DBM
2011 (108)	1. DBM, TripleS, KJSS	1. TripleS, KJSS 2. DBM	1. TripleS 2. KJSS 3. DBM
2012 (91)	1. DBM, KJSS 2. TripleS*	1. KJSS 2. DBM 3. TripleS*	1. TripleS 2. KJSS 3. DBM
2010-12 (302)	1. DBM, TripleS*, KJSS	1. KJSS 2. DBM 3. TripleS	1. TripleS 2. KJSS 3. DBM

Table 4. Rank of the CSEP models for 2009, 2010, 2011, 2012 and the cumulative 2010-2012 using the Sign- or W-test, the Bayes factor, and parimutuel gambling score. Below the year we report the number of target earthquakes. For the parimutuel gambling score we don’t need to use the TripleS* (with the rate correction) instead the TripleS.

The most striking feature of Table 4 is that the model ranking is not the same for each comparison test. But this is not entirely unexpected: each comparison looks at different features of model performance. For example, the gambling score comparison is not sensitive to events that occur in cells with a very low rate, while the Bayes factor, based on joint likelihood, is. This explains why

TripleS is ranked as the top model for all four years using the gambling score test, but not according to the Bayes factor. Indeed, TripleS is the worst model in 2009, 2010 and 2012 according to the Bayes factor because one or a few target earthquakes occurred in cells where TripleS assigned very low rates. These ranking differences are closely linked to deciding what it means to be the 'best' model (Marzocchi & Zechar 2011): is it more important for a forecast not to miss any earthquake, or for a forecast to be very good for most earthquakes? Of course, one goal of building ensemble models is to try to balance these desires, and this balance is also related to the general problem of decision-making in the context of earthquake preparedness (Kantorovich & Keilis-Borok 1991).
The W- and Sign-tests emphasize the median of the information gain distribution, and therefore they are slightly less sensitive to extreme values that can arise from target earthquakes occurring in cells with very low forecast rates. This explains the difference between the Sign-test and Bayes factor rankings in 2009. In 2010 the information gain results are ambiguous; DBM is better than TripleS according to the W-test, while the Sign-test indicates that DBM and KJSS are not significantly different and TripleS is better than KJSS. In 2011 the models are not significantly different according to the information gain tests (Table 4). The same kind of ambiguity persists when the period 2010-2012 is considered.
In Table 5 we show the likelihood of CSEP models and of different ensemble models. The results indicate that the ensemble models are almost always better than the CSEP models. This is even true in 2009, when the ensemble model is simply made by averaging TripleS and DBM: the likelihood of the ensemble model is much higher than the likelihood of TripleS or DBM. In the four years considered in the prototype global experiment, the PGMA ensemble obtains the highest likelihood of any model. This can be explained by considering the performance of the individual CSEP models. The TripleS model is best—i.e., has the highest forecast rate—for many target earthquakes but is terrible for a few target earthquakes; those few target earthquakes cause the very low likelihood of TripleS in 2009, 2010 and 2012. But the PGMA ensemble does not harshly penalize TripleS for these few events and it therefore gives TripleS substantial weight; it also makes up for those shortcomings by mixing in the other CSEP models.
The gSMA model is the worst ensemble model because it tends to assign most weight to the model that has the highest cumulative likelihood, underweighting all the other models. Marzocchi et al (2012) showed that gSMA weighting scheme takes into account only the relative scoring, without considering the absolute performances of each model. In other words, two models that have the same difference in cumulative likelihoods have the same weights regardless of the absolute value of their likelihoods. On the contrary, SMA and BFMA ensembles take into account the absolute performances of each model and perform better than gSMA.

2009	2010	2011	2012
EM: -451.5	**PGMA: -513.6**	**PGMA: -469.2**	**PGMA: -419.5**
DBM: -507.9	**SMA: -536.5**	**SMA: -478.0**	**SMA: -438.3**
TripleS*: -1160	KJSS: -539.5	**BFMA: -482.0**	**BFMA: -449.6**
	BFMA: -584.8	TripleS: -483.1	**gSMA: -455.4**
	gSMA: -612.6	**gSMA: -484.8**	KJSS: -456.1
	DBM: -615.1	KJSS: -485.5	DBM: -502.7
	TripleS*: -2649	DBM: -535.5	TripleS*: -1119

Table 5. CSEP and Ensemble models ordered by their joint log-likelihood for 2009, 2010, 2011 and 2012 (best model first). For 2009 we consider only one ensemble model (EM), which is an equal mixture of DBM and TripleS. Ensemble models are shown in bold.

Discussion

The statistical analysis of the annual global forecasts since 2009 highlights several interesting aspects related to the performances of the models, the testing procedures, and the definition of the "best" model to be used for practical purposes.

Model performance

The results of this study show that our impression of model performance heavily relies on the metric used to evaluate them. For example, TripleS is intuitively a good model: of the CSEP models, it often has the highest rate where target earthquakes occurred. But any score based on the log-likelihood tends to penalize TripleS because it grossly fails in forecasting a few target earthquakes. In particular, any score based on log-likelihood deems TripleS to be worse than the most basic earthquake occurrence model where earthquakes may occur everywhere with the same probability (UNIF). Using another metric, such as the gambling score, TripleS appears to be the best model.

Among the CSEP models, only DBM is truly time-dependent and takes into account long-term variation of the seismicity (Lombardi & Marzocchi, 2007). But DBM fares no better than the time-

independent models (KJSS and TripleS) using annual forecasts. This suggests that time-dependent models based on earthquake clustering may improve forecasts only when regularly updated after an event. On the other hand, the spatial distribution of seismicity appears to be a key issue. We speculate that TripleS has a better spatial forecast for most earthquakes because it uses also the spatial distribution of events smaller than Mw 5.95.

Improving the testing procedures

The analyses carried out in this study highlight several key issues that should be taken into account to improve CSEP experiments. First, the consistency tests used in CSEP experiments cannot be used to 'reject' any model, but they are important to identify the weaknesses of a forecast model. For example, one model may appear unreliable because it grossly fails the N-test but it may score quite well in the S-test, indicating a good spatial forecasting capability. This is the case for SPPM. Second, we agree with Marzocchi et al. (2012) that model comparisons are enlightening and should be emphasized in future CSEP experiments. These comparisons are fundamental for ranking models according to their forecast performance. There is a wide range of possible tests and none is best in every situation. We should choose the comparison test based on what aspect of the model we're interested in assessing, and we should check the underlying assumptions. Another basic difference in comparison tests is about their intrinsic nature. Classical statistical tests consider the null hypothesis of equal model performance. If we have more than two models, the results of the classical statistical tests are not well-suited to establish a simple ranking; in fact in table 4 for these reason we have a lot of apparent ties. The Bayes factor and gambling score are more suitable to provide a ranking of models; this avoids the ambiguities that may arise with classical tests results.

Looking for the "best" model

Marzocchi and Zechar (2011) suggested that the term "best" model may have different meanings to different potential users. For operational purposes, Jordan et al. (2011) suggested that the model used has to be "authoritative." The results obtained here are in agreement to what has been found in Marzocchi et al. (2012) and suggest that the term authoritative may be attributed to a suitable ensemble model. In particular, Marzocchi et al. (2012) show that the forecasts of a sound ensemble model perform better than the forecast made by the best performing model at the time of the forecast during the RELM experiment; here, we show that a sound ensemble model produce forecasts that are always better than the ones of each single model. Such empirical findings seem to indicate that, when only rough models are available, a sound ensemble model is less likely to fail dramatically. As a rule of thumb, we conclude that the kind of weighting scheme is not dramatically

important and the appropriate choice may depend on the specific nature of the forecast models considered.

Conclusions

Conducting a full-fledged global seismicity forecast experiment should be a high priority for CSEP scientists and anyone interested in learning more about large earthquake occurrence, seismic hazard, and seismic risk. But a global seismicity forecast experiment should not be a mere aggrandizement of the regional CSEP experiments. The results of this paper and of other papers recently published (Eberhard et al., 2012) suggest a number of ways in which future CSEP experiments might be different. In particular, we conclude this article by describing our vision of a global seismicity forecast experiment.

Relative to the current experiment configuration, we agree with the following changes suggested by Eberhard et al. (2012):

- The GCMT catalog should be used, and forecasts could target events as small as Mw 5.5;
- Unlike the forecasts in this study, each cell should specify a magnitude distribution, rather than lumping earthquakes of different sizes together
- Any assessment that can reveal model features to be improved, or differences between models, should be used
- Catalog uncertainties should be accounted for in model development and model assessment

Eberhard et al. (2012), following the analyses of Werner & Sornette (2008) and Lombardi & Marzocchi (2010), also discussed the 'Poisson assumption' used in CSEP experiments. This is the assumption that earthquake counts follow a Poisson distribution, and making this assumption simplifies model development and model assessment: participants only have to specify one number per cell. In addition to being wrong in most cases—earthquake counts have super-Poisson variance—this assumption can conceal that each forecast specifies not only an expectation in each cell, but a complete probability mass function. In other words, CSEP forecasts are fully-specified stochastic models in the sense that one can compute the likelihood of any observation.

To preserve the ease of model development but allow greater flexibility, we suggest these alternative forecast formats for a future global seismicity forecast experiment:

1. Expected rate in every space/magnitude bin
2. Gutenberg-Richter *a*-value in every spatial cell and
    a. a global Gutenberg-Richter *b*-value, or
    b. *b*-value per spatial cell

3. Probability in every space/magnitude/number voxel (any non-specified voxels are assumed to have zero probability)

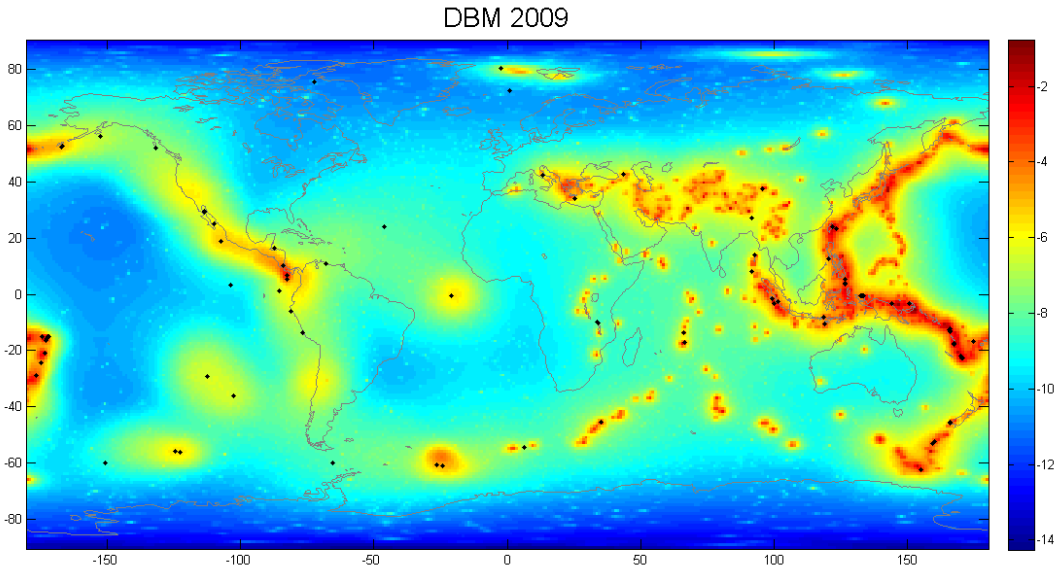
If you use one of the first two formats, you must also specify an analytical forecast uncertainty: you could choose Poisson, or you could choose negative binomial. If you choose negative binomial, which is characterized by two parameters, you could choose to have the other global parameter value automatically estimated from historical seismicity by CSEP, or you could provide this extra parameter value globally, or at the cell level, or at the bin level. Or you could choose another analytical distribution so long as it is calculable at the voxel level.

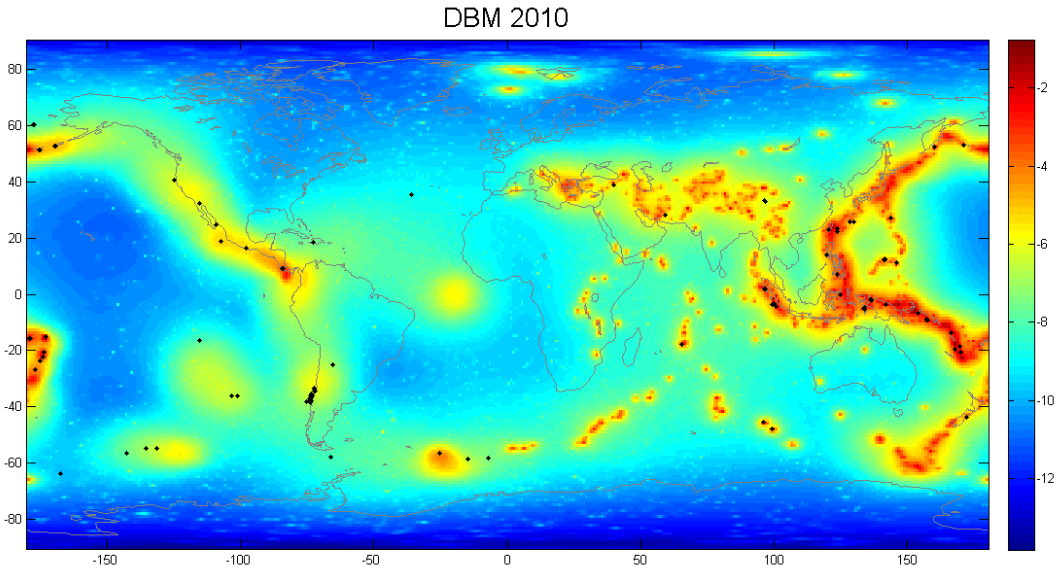
Certainly one could argue that, even with this added flexibility, the forecast format is rather restrictive. A global seismicity forecast should be inclusive, and any model or scientist that produces falsifiable predictive statements about earthquake occurrence should be considered. We should strike a balance between the limiting the number of forecast formats (so many forecasts can be readily compared) and maximizing participation, both in number of participants and distinct earthquake occurrence hypotheses.

In the context of a global experiment, it may also be worthwhile to expand the scope of what is forecast. So far, it has been the space-time-size distribution of epicenters (or epicentroids), because these are commonly estimated by network agencies and it is relatively straightforward to assign a point to a grid cell. But a point source does not adequately represent large earthquakes, and the shaking that results from earthquakes is of far greater practical importance. Models that are developed for seismicity forecasting could be extended to forecast measures of ground shaking.

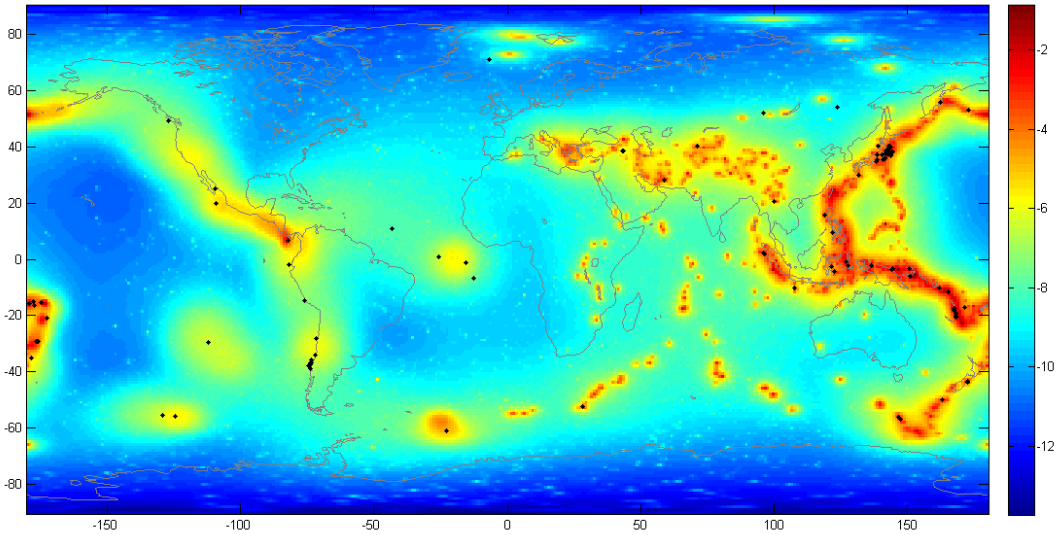
Online supplementary material:

the next figures show the log-rate of each cell and the target events for TripleS, DBM and KJSS model for all years.

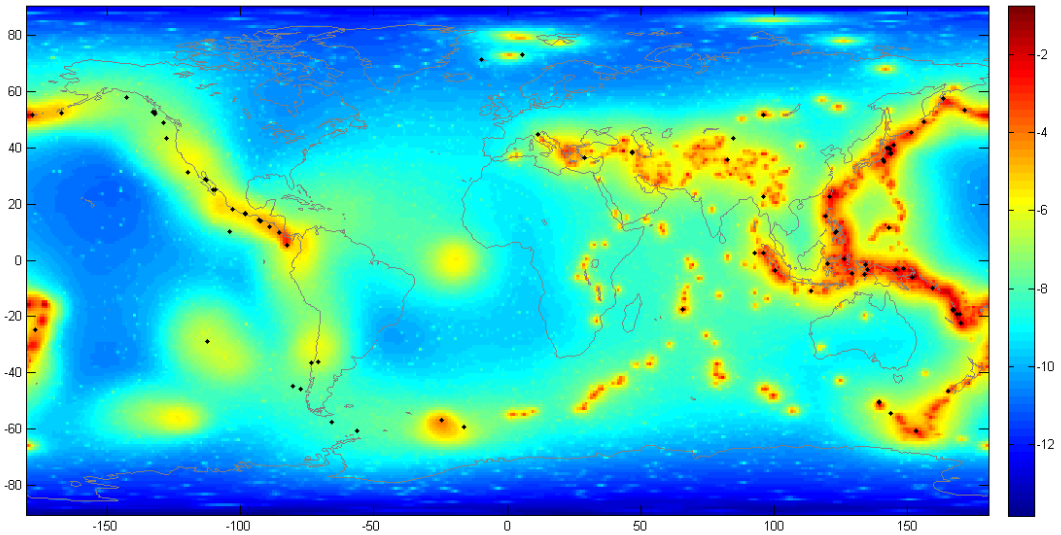




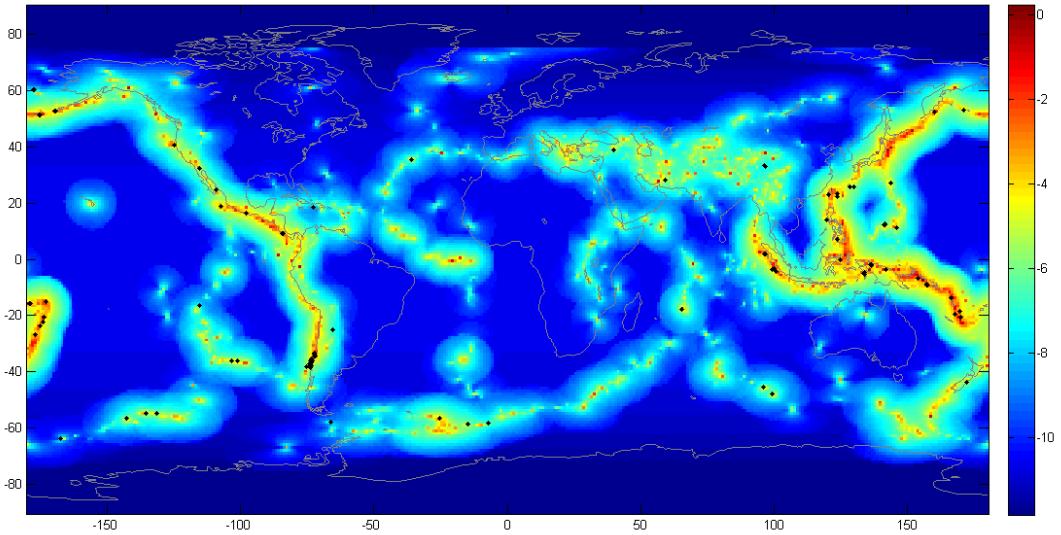
DBM 2011



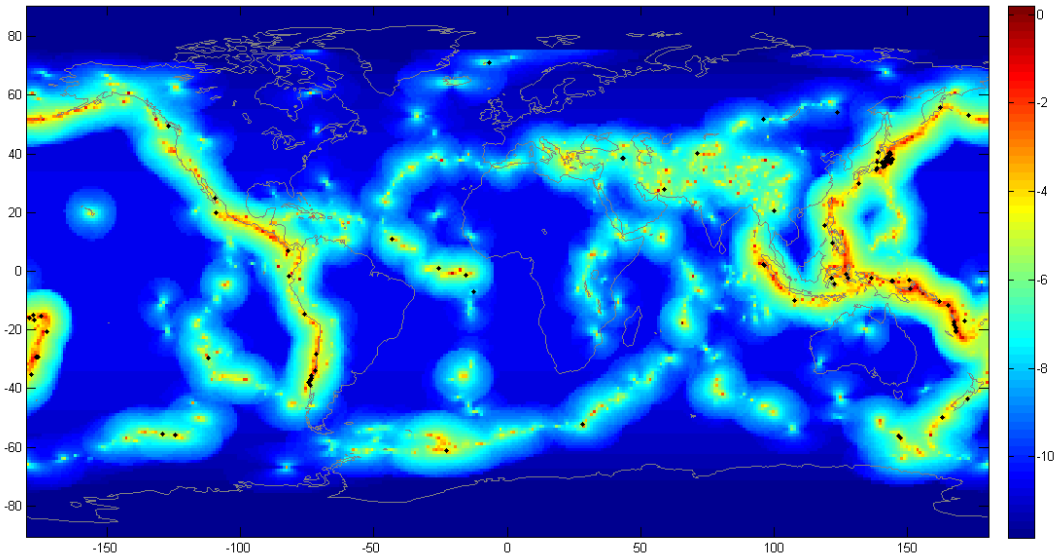
DBM 2012



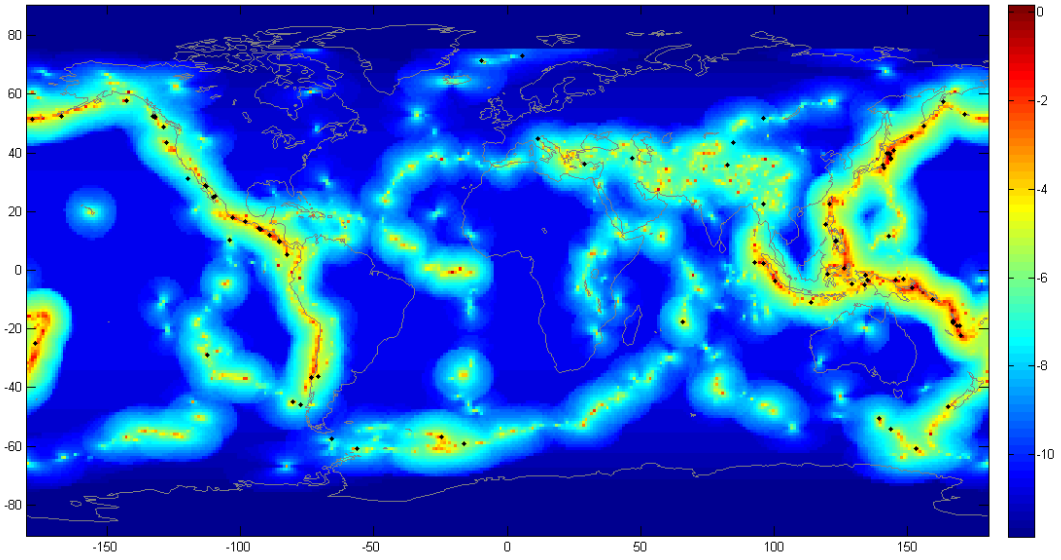
KJSS 2010



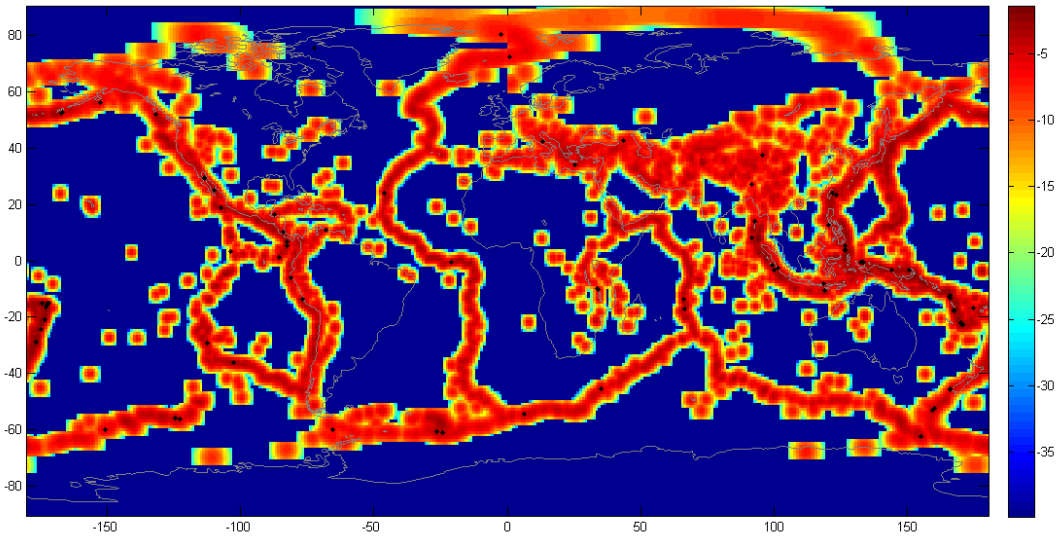
KJSS 2011



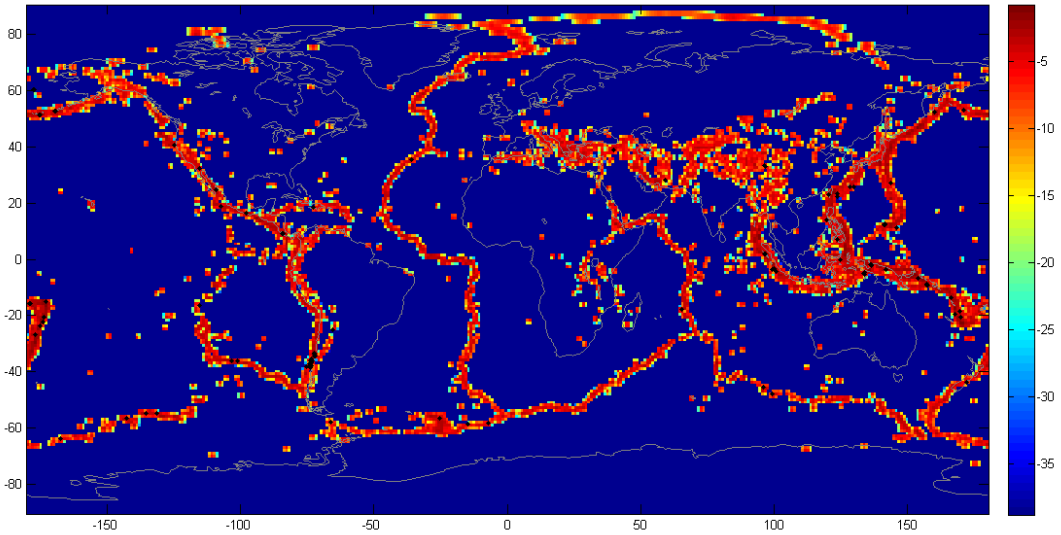
KJSS 2012



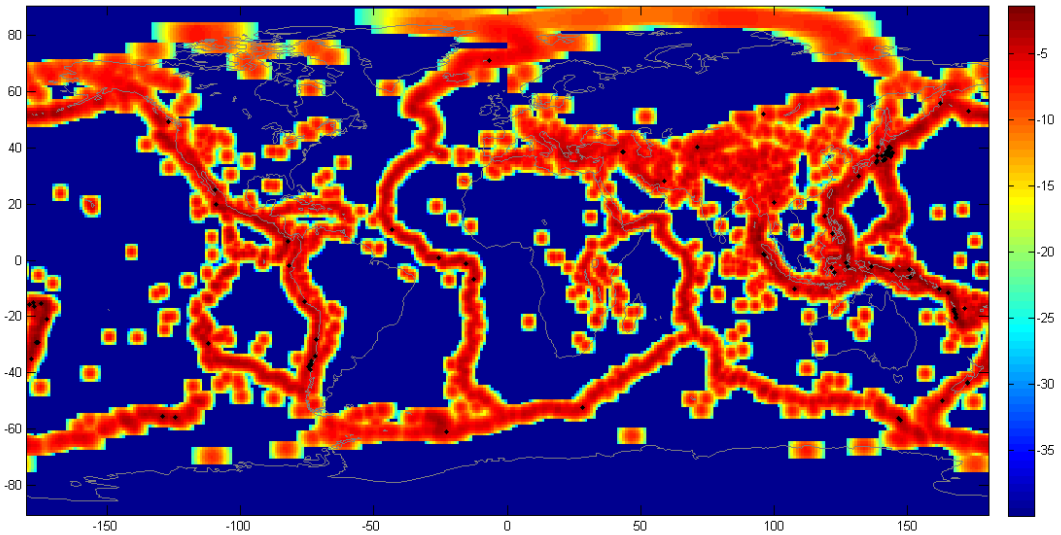
TripleS 2009

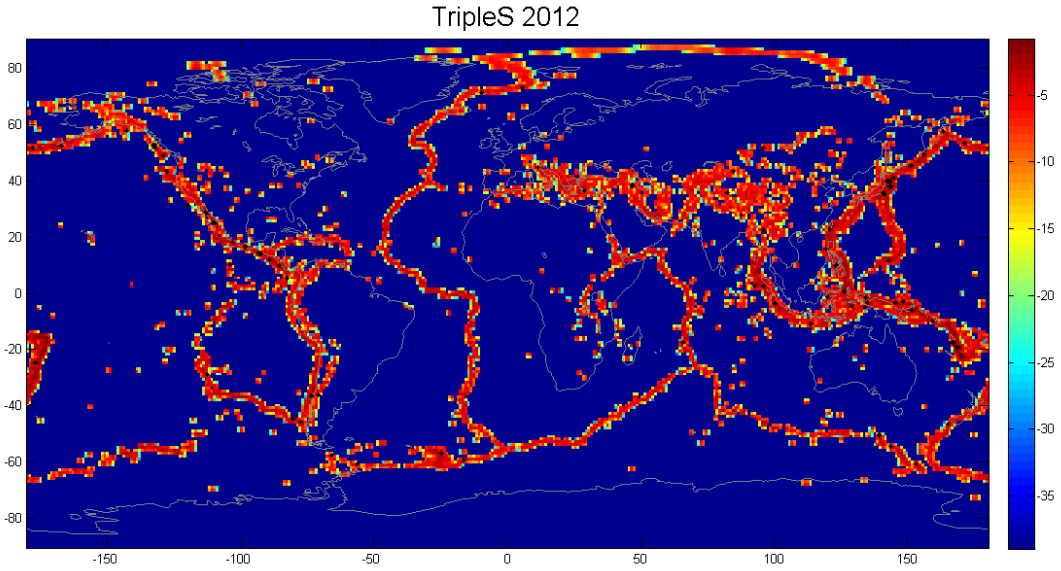


TripleS 2010



TripleS 2011





Bibliography:

American Association for the Advancement of Science (AAAS), 1989. *Science for All Americans: A Project 2061 Report on Literacy Goals in Science, Mathematics and Technology*. Washington, DC: American Association for the Advancement of Science.

Dixon, W. J., & Mood, A. M., 1946. The statistical sign test. *J. Amer. Statist. Assoc*., **41**, 557-566.

Dziewonski, A.M., Chou, T.-A. & Woodhouse, J.H., 1981. Determination of earthquake source parameters from waveform data for studies of global and regional seismicity, *J. geophys. Res.,* **86**(B4), 2825-2852.

Eberhard, D.A., Zechar, J.D. & Wiemer, S., 2012. A prospective earthquake forecast experiment in the western Pacific, *Geophys. J. Int.*, **190**(3), 1579-1592.

Ekström, G., Dziewonski, A., Maternovskaya, N. & Nettles, M., 2005. Global seismicity of 2003: centroid moment-tensor solutions for 1087 earthquakes, *Phys. Earth Planet. Inter.,* **148**(2-4), 327-351.

Evison, F.F. & Rhoades, D.A., 1993. The precursory earthquake swarm in New Zealand: hypothesis tests, *New Zeal. J. Geol. Geop.,* **36**, 51-60.

Evison, F.F. & Rhoades, D.A., 1997. The precursory earthquake swarm in New Zealand: hypothesis tests II, *New Zeal. J. Geol. Geop.,* **40**, 537-547.

Evison, F.F. & Rhoades, D.A., 1999. The precursory earthquake swarm in Japan: hypothesis test, *Earth Planets Space,* **51**, 1267-1277.

Gibbons, J.D. & Chakraborti, S., 2003. *Nonparametric statistical inference*, Marcel Dekker Inc., New York.

Gutenberg, B. & Richter, C.F., 1954. *Seismicity of the Earth and Associated Phenomena,* 2nd edn, Princeton Univ. Press, Princeton, NJ.

Harte, D. & Vere-Jones, D., 2005. The entropy score and its uses in earthquake forecasting, *Pure appl. Geophys.,* **162**(6-7), 1229-1253.

Jordan, T. H., Chen, Y.-T., Gasparini, P., Madariaga, R., Main, I., Marzocchi, W., Papadopoulos, G., Sobolev, G., Yamaoka, K., & Zschau, J., 2011. Operational Earthquake Forecasting: state of knowledge and guidelines for utilization. *Ann. Geophys*., **54**, 315-391.

Kagan, Y., 2003. Accuracy of modern global earthquake catalogs, *Phys. Earth planet. Inter.,* **135**, 173-209.

Kagan, Y. Y., & Jackson, D.D., 1991. Seismic gap hypothesis: Ten years after, *J. Geophys. Res*., **96**, 21,419-21,431.

Kagan, Y. Y., & Jackson, D.D., 1994. Long-term probabilistic forecasting of earthquakes, *J. Geophys. Res*., **99**, 13,685-13,700.

Kagan, Y. Y., & Jackson, D.D., 1995. New seismic gap hypothesis: Five years after, *J. Geophys. Res*., **100**, 3943-3959.

Kagan, Y.Y. & Jackson, D.D., 2000. Probabilistic forecasting of earthquakes, *Geophys. J. Inter.,* **143**, 438-453.

Kagan, Y.Y. & Jackson, D.D., 2010. Earthquake forecasting in diverse tectonic zones of the globe, *Pure appl. Geophys.,* **167**(6-7), 709-719.

Kagan, Y.Y. & Knopoff, L., 1977. Earthquake risk prediction as a stochastic process, *Phys. Earth Planet. Inter.,* **14**(2), 97-108.

Kagan, Y.Y. & Knopoff, L., 1987. Statistical short-term earthquake prediction, *Science,* **236**, 1563-1567.

Kantorovich, L.V., & Keilis-Borok, V.I., 1991. Earthquake prediction and decision making: social, economic, legislative and civil defense domains. In: *Proceedings of International Conference on Earthquake Prediction: State-of-the-Art*, 15-18 October, Strasbourg, France, pp. 586-593.

Kass, R. E., & Raftery A.E., 1995. Bayes factors*, J. Am. Stat. Association,* **90***,* 773-795.

Knuth, D., 1974. Structured programming with GOTO statements, *Computing Surveys*, **6**, 261-301.

Lilliefors, H. W., 1967. On the Kolmogorov-Smirnov test for normality with mean and variance unknown, *J. Am. Stat. Association*., **62**, 399-402.

Lombardi, A.M., & Marzocchi, W., 2007. Evidence of clustering and nonstationarity in the time distribution of large worldwide earthquakes. *J. Geophys. Res*., **112**, B02303.

Lombardi, A.M., & Marzocchi, W., 2010. The ETAS model for daily forecasting of Italian seismicity in the CSEP experiment. *Ann. Geophys*., **53**, 155-164.

Marzocchi, W. & Lombardi, A.M., 2008. A double branching model for earthquake occurrence, *J. Geophys. Res.,* **113**(B8), 1-12.

Marzocchi, W. & J. D. Zechar, 2011. Earthquake forecasting and earthquake prediction: Different approaches for obtaining the best model. *Seismol. Res. Lett.,* **82**, 442-448.

Marzocchi, W., Zechar, J.D. & Jordan, T.H., 2012. Bayesian Forecast Evaluation and Ensemble Earthquake Forecasting, *Bull. Seism. Soc. Am.,* **102**, 2574-2584.

Ogata, Y., 1988. Statistical models for earthquake occurrences and residual analysis for point processes, *J. Am. Stat. Assoc.,* **83**, 9-27.

Ogata, Y., 1999. Seismicity analysis through point-process modeling: a review, *Pure appl. Geophys.,* **155**, 471-507.

Randles, Ronald H., Fligner M.A., Policello, G.E., Wolfe, D.A., 1980. An Asymptotically Distribution-Free Test for Symmetry Versus Asymmetry, *J. Am. Stat. Association*, **75**, 168-72.

Rhoades, D.A. & Evison, F.F., 1989. Time-variable factors in earthquake hazard, *Tectonophysics,* **167**, 201-210.

Rhoades, D. A., & Gerstenberger, M.C., 2009. Mixture models for improved short-term earthquake forecasting, *Bull. Seismol. Soc. Am*. **99**, 636-646.

Rhoades, D.A., Schorlemmer, D., Gerstenberger, M.C., Christophersen, A., Zechar, J.D. & Imoto, M., 2011. Efficient testing of earthquake forecasting models, *Acta Geophysica,* **59**, 728-747.

Rong, Y., Jackson, D.D., & Kagan, Y.Y., 2003. Seismic gaps and earthquake, *J. Geophys. Res*., **108**, 2471.

Schorlemmer, D., Christophersen, A., Rovida, A., Mele, F., Stucchi, M., & Marzocchi, W., 2010a. Setting up an earthquake forecast experiment in Italy, *Ann. Geophys*., **53**, 1-9.

Schorlemmer, D., Gerstenberger, M.C., Wiemer, S., Jackson, D.D. & Rhoades, D.A., 2007. Earthquake likelihood model testing, *Seism. Res. Lett.,* **78**, 17-29.

Schorlemmer, D., Zechar, J.D., Werner, M.J., Field, E.H., Jackson, D.D. & Jordan, T.H., 2010b. First results of the Regional Earthquake Likelihood Models experiment, *Pure appl. Geophys.,* **167**, 859-876.

Siegel, S., 1956. *Non-Parametric Statistics for the Behavioral Sciences*, McGraw-Hill, New York.

Student, 1908. The Probable Error of a Mean, *Biometrika,* **6**, 1-24.

Werner, M. J., & Sornette, D.D., 2008. Magnitude uncertainties impact seismic rate estimates, forecasts, and predictability experiments, *J. Geophys. Res*., **113**, B08302.

Werner, M.J., Zechar, J.D., Marzocchi, W., Wiemer, S. & Nazionale, I., 2010. Retrospective evaluation of the five-year and ten-year CSEP Italy earthquake forecasts, *Ann. Geophys.,* **53**, 11-30.

Zechar, J.D. & Jordan, T.H., 2010. Simple smoothed seismicity earthquake forecasts for Italy, *Ann. Geophys.,* **53,** 99-105.

Zechar, J.D. & Zhuang, J., 2010. Risk and return: evaluating RTP earthquake predictions, *Geophys. J. Int*., **182**, 1319-1326.

Zechar, J.D., & Zhuang, J., 2012. Betting against the house and peer-to-peer gambling: A Monte Carlo view of earthquake forecasting, presented at 2012 General Assembly, ESC, Moscow, Russia, 19–24 August.

Zechar, J.D., Gerstenberger, M.C. & Rhoades, D.A., 2010. Likelihood-based tests for evaluating space-rate-magnitude earthquake forecasts, *Bull. Seism. Soc. Am.,* **100**, 1184-1195.

Some Thoughts on Declustering in Probabilistic Seismic Hazard Analysis

W. Marzocchi, M. Taroni
Istituto Nazionale di Geofisica e Vulcanologia, Via di Vigna Murata 605, 00143 Roma, Italy

Abstract

In this paper we discuss in depth one of the basic procedures that stand behind probabilistic seismic hazard analysis (PSHA), i.e., the declustering of the seismicity rates. First, we explore the technical, scientific and practical motivations which led to introducing the declustering of seismicity rates. Then, we show that for PSHA declustering is essential only to minimize a spatial distorsion of the earthquake occurrence process, but, conversely, it may lead to significant underestimation of the true seismic hazard. This underestimation precludes the possibility to test PSHA against real observations, and it may lead to underestimate the seismic risk, whenever seismic hazard maps are used for risk assessment. Finally, we propose a methodology that can be used in PSHA to avoid this potential bias.

Keywords. Seismic hazard, declustering, exceedance probability

Introduction

Probabilistic seismic hazard analysis (PSHA) is the main scientific component that stands at the basis of the definition of a suitable building code in many countries. In essence, PSHA is a forecast of how intense the ground motion will probably be at one specific site on Earth's surface during a future interval of time. PSHA is articulated by means of a complex procedure that involves many scientific issues and often a lot of subjective expert opinions. In this paper we discuss one specific scientific issue related to PSHA. PSHA practice requires that seismicity rates are 'declustered', i.e., triggered events are removed from the calculation, leaving only the largest independent earthquakes.

Cornell (1968) introduced the PSHA concepts that are still widely used today. In this framework, the seismic hazard in each arbitrary location is represented by an hazard curve that represents the exceedance rate of one specific ground motion parameter z in a reference period of time; hereafter, we set this reference period of time; hereafter, we set this reference period to one year, thus the rate is always meant to be an annual rate. In a general form, the exceedance rate reads

$$\lambda_z(Z > z) = \int_{\Xi} P[Z > z|\vec{\xi}]\ \lambda_s\left(\vec{\xi}\right)d\vec{\xi} \qquad (1)$$

where λ_z is the exceedance rate for the ground motion parameter z, λ_s is the seismicity rate above a threshold magnitude (usually PSHA considers only earthquakes that can produce a significant ground motion), and $\vec{\xi} \in \Xi$ is a vector that contains all relevant earthquake source parameters like, for instance, the location and the magnitude (other parameters like focal mechanism may be included here). $P[Z > z|\vec{\xi}]$ is the so-called ground motion prediction equation (GMPE) that gives the exceedance probability of the ground motion value z given the knowledge of source parameters.

PSHA is usually expressed in probabilistic terms, for example, indicating the specific ground motion parameter value that has 10% of probability to be exceeded in a time interval of 50 years. So, it is necessary to pass from the rate of equation 1 to probability. This transition is made assuming that earthquakes have a Poisson distribution in time, and so the number of exceedance in terms of ground motion.

The exceedance probability for an arbitrary site is then

$$P_z(Z > z) = 1 - \exp[-\lambda(Z > z)\Delta\tau] \qquad (2)$$

Here $\Delta\tau$ = 50 years. Since the beginning, almost all PSHA have considered that the seismicity rate must be Poisson, i.e., the earthquakes have to be time independent. However, it is well know that

the real seismic catalogs show significant departures from the Poisson distribution, being usually much more clustered in time and space (Kagan, 2010). So, the usual PSHA practice demands to decluster the seismicity in order to get a set of time independent earthquakes. There is a wide range of declustering techniques (VanStiphout et al., 2011; 2012).

This large variety reects the intrinsic difficulty to define what an aftershock is. In PSHA practice, the most basic declustering methods are usually applied (e.g. Gardner and Knopoff, 1974), because they produce declutered catalogs that resemble a Poisson dataset. The other more sophisticated declustering techniques produce declustered catalogs that are not usually distributed according to a Poisson process (Van Stiphout et al., 2012).

Taking into account that the declustering procedure brings to an underestimation of the seismicity rates that will be observed in the future and the apparent controversy and subjectivity in deciding what an aftershock is, we wonder if the declustering is really necessary for PSHA. Reading the literature we think that the basic rationales of declusterting seismicity rates are:

1. The Poisson assumption holds for declustered catalogs and it is necessary to move from rates to probabilities (see equations 1-2)

2. Keeping the largest events, we remove earthquakes which have a negligible contribution in terms of ground shaking, and in terms of earthquake risk as well (Lomnitz, 1966); consequently, declustering is assumed to induce a negligible effect for engineering practice.

3. Non declustered seismic catalogs give a biased view of the true spatial variability seismicity rates

In this paper we discuss all of these issues and we show that a more accurate PSHA should not consider declustered seismicity rates.

The Poisson assumption for PSHA and the distribution of the seismicity rate

Consider a practical PSHA calculation for one particular site. Here we assume that $\lambda_s\left(\vec{\xi}\right)$ of equation 1 can be written as

$$\lambda_s\left(\vec{\xi}\right) \equiv \lambda_e(m, \vec{x}) \qquad (3)$$

where m is the magnitude of the source event and $\vec{x}$ the geographical location of its epicenter. The seismicity rate can be calcolated in a grid of cells, so equation 1 can be written as

$$\lambda_z(Z > z) = \sum_{i=1}^{N}\sum_{k=1}^{M} \lambda_i a_{ik} P_{ik} \qquad (4)$$

where N is the total number of cells in which the whole region is divided, M is the total number of magnitude bins, λ_i is the cumulative seismicity rate in the specified space cell, a_{ik} is the percentage of the rate λ_i in each magnitude bin (this value depends by the frequency-magnitude distribution) and P_{ik} is the GMPE that provides the probability that each event in the i-th geographical cell and of the k-th class of magnitude will produce a ground shacking parameter Z > z in the specific site considered. Note that in equation 4 and hereafter the Einstein summation convention of repeatable indices is suppressed.

If the seismicity rate λ_i is the rate of a Poisson distribution, the GMPE operates as a random selection of events, then the rate of ground motion exceedance due to events in the k-th class of magnitude and occurring in the i-th geographical cell is a Poisson distribution with rate $\lambda_i a_{ik} P_{ik}$ (Cornell, 1968). The total exceedance rate $\lambda_z(Z > z)$ is a sum of Poisson independents random variables, so it follows a Poisson distribution.

What happens if λ_i is the rate of a non-poissonian distribution?

Earthquakes are not independent because they are clustered in space and time, so we expect that the seismicity rate and the exceedance rate of one specific site (equation 4) may depend on time. Consider a sequence of a possible dependent Bernoulli variables (0 or 1) that represents the exceedance of one specific ground shaking threshold in one specific site, X_t (t=1, … , N_T) where N_T is the number of time windows of length $\Delta\tau$ in which we subdivide the whole time interval considered T ($\Delta\tau$ can contain only 0 or 1 event), with probability $\theta_t = P(X_t = 1|X_{t-1} = x_{t-1}, \dots)$; here the index t mimics a possible time variability of the quantities in different time windows. Now let X^*_t (t=1, … , N_T) the sequence of an independent Bernoulli variables with a constant probability $P(X_t^* = 1) = \pi = \lambda_z \Delta\tau$. The Le Cam's theorem (Le Cam, 1960; Serfling, 1975) evaluates the Poisson approximation for dependent Bernoulli variables. It reads

$$\sum_{j=0}^{\infty} \left| P(S_{TOT} = j) - \Lambda_X^j \frac{e^{-\Lambda_X}}{j!} \right| < 2\sum_{t=1}^{N_T} \theta_t^2 + 2\sum_{t=1}^{N_T} \epsilon_t \qquad (5)$$

Where $S_{TOT} = \sum_{t=1}^{N_T} X_t$, $\Lambda_X = \sum_{t=1}^{N_T} \theta_t$, $\epsilon_t = E(|\theta_t - \pi|)$, and E(.) is the expectation value. The first right hand side addend accounts for the fact that each single Bernoulli variable does not come from a Poisson distribution, while the second right hand side addend accounts for the possible

correlation between variables. Equation 5 implies that a sum of non-negative random variables may be treated as a Poisson variable if X_t have a su_ciently high probability of taking 0 value (i.e., 1-θ_t), and sufficiently weak mutual dependence (Sering, 1975).

It is useful to consider the case of the Le Cam's theorem applied to PSHA practice where the exceedance probability threshold is often set to 10% (or less) for a time interval of 50 years (or more). In this case, the probability is small enough to assume $\Lambda_X \sim P(Z > z) = 0.1$, and the first right hand side addend can be approximated as

$$2\sum_{t=1}^{N_T} \theta_i^2 < 2(\Lambda_X)^2 \approx 0.02 \qquad (6)$$

The evaluation of the second right hand side addend of equation 5 requires more elaborated thoughts. Let us consider a simplified case in which an earthquake can induce another earthquake on the next time window with a probability *a*. So, the total number of earthquakes generated by one event is

$$b = \sum_{j=1}^{N_T} a^j = \left(\frac{1 - a^{N_T}}{1 - a}\right) - 1 \qquad (7)$$

The value of b represents the ratio between triggered events and the total number of events. This ratio can be obtained by dividing the number of earthquakes in a declustered catalog and the undeclustered catalog. If we take b = 0.43 (a typical figure that derives from real seismic catalogs after GK declustering), inverting equation 7 we obtain $a \approx 0.3$. This means that every earthquake has about 30% of probability to trigger another earthquake. This correlation is not mapped in a similar correlation of exceedances. The total number of exceedances is smaller than the total number of earthquakes, and it depends on the threshold of ground shaking used; higher thresholds lead to smaller numbers of exceedances. In our simplified example, the average conditional exceedance probability given an exceedance at t-1 (X_{t-1}=1) is $\theta_t = \beta \cdot a = \beta \cdot 0.3$, where β is the average proportion of ground shaking exceedances over the total number of earthquakes that may affect the site. The value of β is usually very low in PSHA practice, because we usually have many earthquakes in a 50 years period, but in one site we expect on average 0.1 number of exceedances.

Thus, $|\theta_t - \pi|$ will be approximatively zero when $X_{t-1} = 0$ (i.e., $\theta_t = \pi$), and it will be ($\beta \cdot a - \pi$) when $X_{t-1} = 1$ that has probability $\pi = \frac{\Lambda_X}{N_T}$.

$$2\sum_{t=1}^{N_T}\epsilon_t = 2\sum_{t=1}^{N_T}E(|\theta_t-\pi|) = 2\sum_{t=1}^{N_T}(\beta\cdot a-\pi)\frac{\Lambda_X}{N_T} = 2(\beta\cdot a-\pi)\Lambda_X \le 0.2\cdot a\cdot\beta \quad (8)$$

Using a = 0.3 and very cautelative values for β (for instance β = 0.1 means that we have in average 10 earthquakes and 1 exceedance), we always get the second right hand side addend of equation 5 less than 0.01, making the correlation between exceedances unnoticeable. Note that if the ground shaking threshold is very low then $\beta\rightarrow 1$ (all earthquakes will overcome the threshold) and $a > 0.3$ (a higher percentage of earthquakes that can contribute to the number of exceedances will be removed by declustering); in this case, the second right hand side addend would not be anymore negligible. In conclusion, in most of pratical PSHA applications, where exceedance probabilities smaller than 10% are required and the ground shaking thresholds are not very small, π and ϵ_t are small enough to assume that the exceedance times follows a Poisson distribution with Λ_X, independently of the distribution and time dependence of the number of earthquakes.

Influence of declustering on PSHA

Here we show that the statement keeping the largest events, we also keep the largest ground shaking for a specific site is not necessarily true. In figure 1 we report the probability that an aftershock having different degrees of magnitude less than the mainshock gives a ground motion larger than the mainshock. Here we have used the GMPE proposed by Cauzzi and Faccioli (2008) without considering the information on focal mechanism. The distance of mainsock and aftershock from the site plays also a role, so we plot this probability as a function of the ratio of the distance of the mainshock from the site over the distance of the aftershock from the site (a large number means that the aftershock is closer to the site with respect to the mainshock). From the figure we observe that for a unit magnitude difference between mainshock and aftershock and if the mainshock and aftershock occur at the same place (distance ratio equals to 1), there is about 20% probability that the aftershock produces a larger ground shaking. This probability becomes even larger if the aftershocks get closer to the site reaching 60% when the aftershock is at 1/3 of the distance from the mainshock. We think that these numbers are hardly negligible for practical calculations.

One implicit assumption of our thought is that the same GMPE holds for mainshocks and aftershocks. Some practitioners argue that aftershocks might have a different GMPE because their stress drop is smaller on average. This may make the inuence of aftershocks less important. Yet, this assumption has not yet proved with real observations, and real recent cases, like the Christchurch

earthquake and the second largest shock on May 29 of the Emilia sequence (Fry et al., 2011; Scognamiglio et al.,2012) showed that aftershocks may induce a significantly high ground motion.

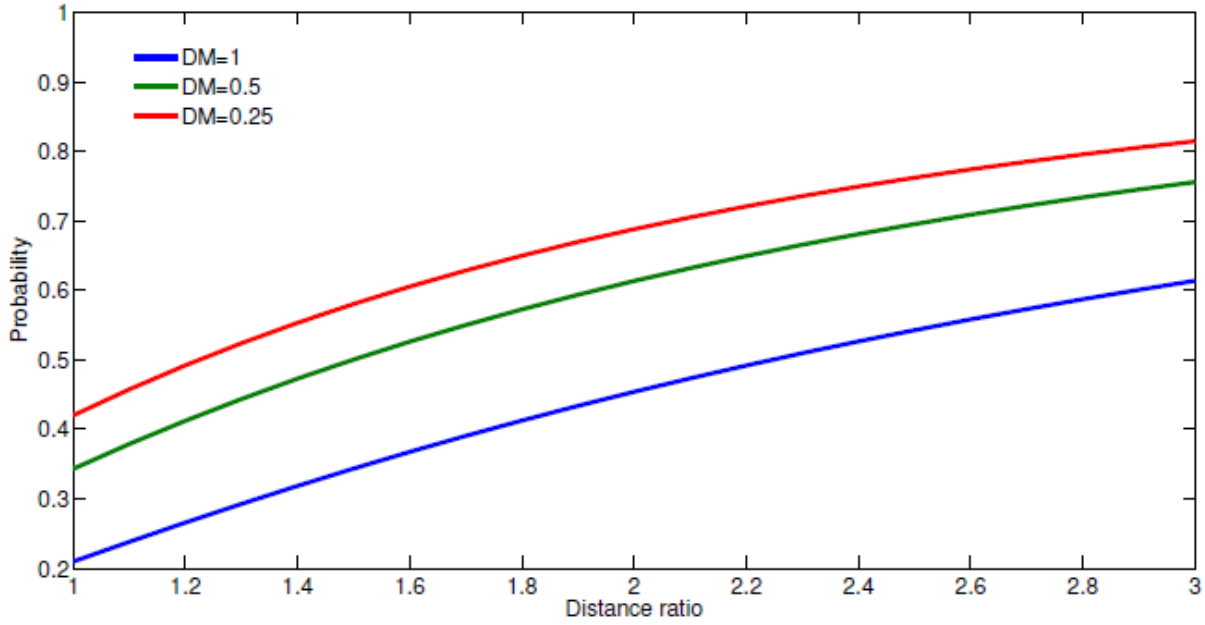


Figure 1: Probability that the peak ground acceleration (PGA) caused by an aftershock with magnitude M_0+DM is higher that the PGA caused by the mainshock with magnitude M_0, as a function of the distance ratio. The distance ratio is the ratio between the distance of the mainshock and the aftershock from the site.

Declustering or not?

In the previous sections we have showed that the declustering is not a necessary technical requirement, and that it may lead to a significant underestimation of the true hazard (Boyd, 2012; Iervolino et al., 2013). On the other hand, the seismicity rates obtained by the real seismic catalog represent a biased view of the real long-term seismicity rates. In fact, clusters of large earthquakes do not occur always in the same way (Kagan and Jackson, 2000; Faenza et al., 2003; Cinti et al., 2004); using a nondeclustered seismic catalog, the seismicity rate may be overestimated for the regions that recently experienced a strong seismic cluster with respect to areas that did not experience clusters in the time interval covered by the catalog. In other terms, the limited length of the seismic catalogs does not allow us to have a complete view of all areas where clusters may occur in the future, and of their frequency of occurrence.

As a consequence, we argue that declustering is a necessary step in order to avoid this bias, provided that the seismicity rates will be eventually corrected for declustering. Here, we show the effects of this procedure in a simplified example for the Italian territory. Here, we do not aim at replicating the seismic hazard map in Italy (Gruppo di Lavoro MPS 2004), but we intend to show the e_ects of maps with and without declustering and corrections in a realistic case. In particular, the seismic hazard map for Italy is calculated with these components (table 1 contains the parameters used in this application):

• the seismic zonation used for the actual seismic hazard map in Italy (Meletti et al., 2007);

• the CPTI11 (Rovida et al., 2011) seismic catalog; earthquakes with $M_W \geq 4.8$ in the time interval 1901-2006.

• the Gardner and Knopoff (GK) declustering technique has been used to decluster CPTI11;

• the b-value for each zone is considered the same that we calculate for the whole Italian territory by the seismic catalog and we set a maximum magnitude equal to 8 for the whole territory.

• the Cauzzi and Faccioli (2008) GMPE.

Figure 2a shows the seismic hazard map in Italy using these components. Figure 2b-c report the map corrected for declustering, and the map obtained by the catalog without the declustering, respectively. Among the components used for this application, the choice of the seismic catalog is probably the most important to discuss. CPTI11 starts only from the 1901, while the CPTI04 seismic catalog used for the national seismic hazard map covers a longer time interval (CPTI Working Group 2004); on the other hand, CPTI11 is mostly nondeclustered (some aftershocks are still missing) while the CPTI04 catalog is already a declustered seismic catalog.

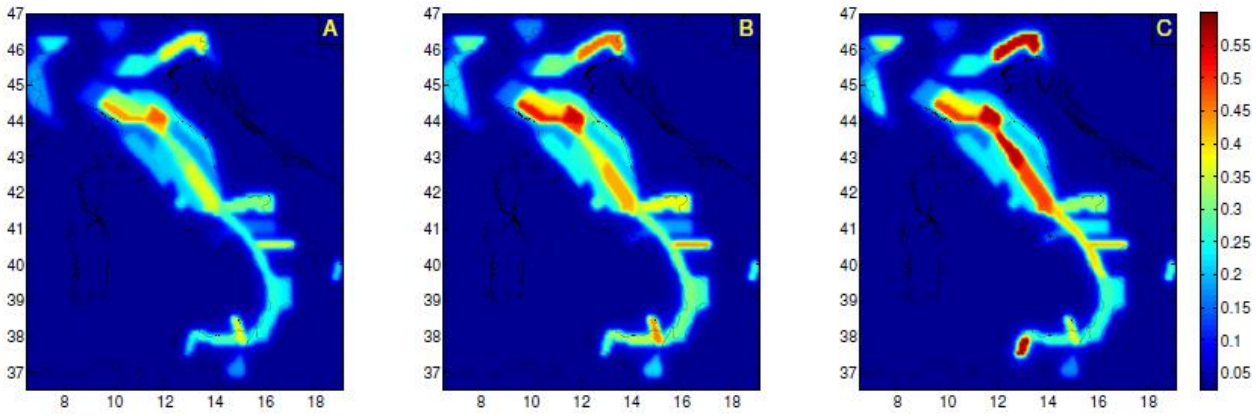


Figure 2: Seismic hazard map (10% of exceedance probability in 50 years) using (A) the declustered seismicity rates, (B) the corrected seismicity rates, and (C) the observed (undeclustered) seismicity rates.

We apply the GK declustering that the GK declustering is the most used because the residuals (the declustered catalog) resembles Poissonian distributed (VanStiphout et al., 2012). GK declustering leads to some effects worth noting: first, the overall seismicity rate is lowered (from 502 events to 357); second, the real frequency-magnitude distribution is distorted, because GK removed only the smaller earthquakes (Kagan, 2010); the spatial distribution tends to be modified because the spatial location of large aftershocks and foreshocks is not considered.

The latter is implicitly accounted for when the spatial smothing of declustered events is applied (the characteristic dimension of the spatial filter is usually larger than the aftershock/foreshock area). The other two issues are not usually systematically compensated in PSHA practice. We argue that the easiest way to correct for declustering is to multiply the declustering seismicity rate of each cell of the grid λ_i by a factor γ that is defined as

$$\gamma = \frac{\sum \lambda_i^*}{\sum \lambda_i} \qquad (9)$$

where λ_i^* is the real seismicity rate in each cell and the summation is made for all the cells of the spatial grid. In other words, the new non-declustered seismicity rates are the declustered seismicity rates multiply by a factor in a way that the total seismicity rate matches to the whole observed seismicity rate. In this way we are assuming that each cell of the grid has similar clustering properties. However, we can easily generalize equation 9 to take into account possible spatial heterogeneities of the clustering properties (e.g., Boyd 2012). For example, we can define a spatial factor $\gamma_i^* = \gamma \times \omega_i$ where ω_i is the normalized clustering capability of the i-th cell. Then, the cumulative rate can be subdivied for each magnitude bin, using the frequency-magnitude distribution (for instance, the same b-value of the G-R law) of the nondeclustered seismic catalog. This procedure differs in few basic points from the method proposed by Boyd (2012) and Iervolino et al. (2013). In essence, their methods are based on the identi_cation of independent clusters, where each cluster is described by the probability to overcome a specific ground motion threshold in one specific site. This probability is larger than the one due only to the mainshock, because it considers also the foreshocks and aftershocks. In practice, this new probability is estimated by simulating the ground motion of the clusters in space (Boyd, 2012) and time through the Omori law (Iervolino et al., 2013). Noteworthy, this procedure precludes the possibility that the ground motion threshold can be overcome more than once inside the same cluster, unavoidably leading to an underestimation of the real exceedance rate.

Our procedure has some basic features worth remarking. First, we consider the possibility to get more than one exceedance inside the same cluster, thus avoiding a systematic underestimation.

Second, we do not use simulations to correct for the missing rate due to declustering. As a matter of fact, the parameters of any aftershock/foreshock model do not have a real physical meaning, but they depend on how the catalog has been declustered. Different declustering techniques remove significantly different percentages of earthquakes, hence the parameters found in literature should be used with extreme caution. In essence, our procedure assumes that the basic constraint for the correction of the seismicity rate is to preserve the total number of earthquakes observed (equation 9), without specifying when the events occur and their spatial distribution that is already accounted for the spatial smoothing of the mainshocks. This implicitly assumes that the number of earthquakes observed in the catalog allows an accurate estimation of the average number of earthquakes in the time period of interest.

The effects of declustering can be appreciated looking at figure 3 where we show the ratio of the exceedance probabilities for the corrected and declustered seismic hazard map using a spatially uniform correction factor given by equation 9. The probabilities are calculated for ground acceleration thresholds z = 0.1; 0.2, and 0.3g. The figures show that the corrected hazard maps have probabilities that can be up to 35% higher than the correspoding probabilities for the declustered hazard map. Noteworthy, the percentage of probability increase is not constant through space, while γ is. This is due to two different issues. The first one is the saturation effect of the probability for the cells with high rate; in fact, the correction factor is multiplicative on the exceedance rates that are unbounded, while the exceedance probability is bounded to 1. So, we expect to have a smaller ratio of probabilities for the regions with higher probabilities. This is shown in figure 3 where the smaller ratios are relative to the regions of the Appennines where the hazard is higher (cf. figure 2). A second factor that inuences the rate is the decrease of the b-value due to the declustering. A decrease of the b-value implies that the declustering affects more the probability of smaller events rarther than the probability of large earthquakes. So, in regions of low hazard (e.g., the Tyrrhenian offshore region) that are mostly affected by large inland earthquakes we expect to have a smaller ratio of probabilities (see figure 3). If we consider both effects, the higher ratios are expected in areas of medium hazard. In figure 4 we show the differences between the PGA values with exceeding probability of 10% in 50 years calculated with the corrected seismicity rates and with the declustered rates. The regions with higher differences are the ones with the higher earthquake rates. Probably more interesting is the map of figure 5 where we show the percentage of the variation of the peak ground acceleration at 10% and 50 years using the corrected and declustered hazard analysis. The regions with higher percentages are the same where we observe the higher ratios of the exceedance probabilities. Interestingly, the maximum peak ground acceleration can have variations up to 35-40% of the value calculated from the declustered seismicity rates.

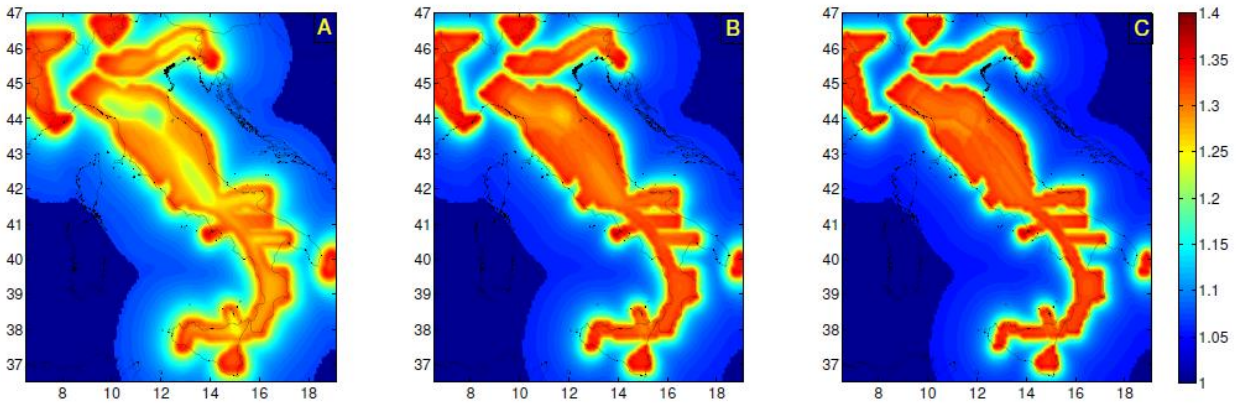


Figure 3: Ratio of the exceedance probabilities calculated using the corrected and declustered seismicity rates for a ground motion acceleration of (A) 0.1g, (B) 0.2g, and (C) 0.3g

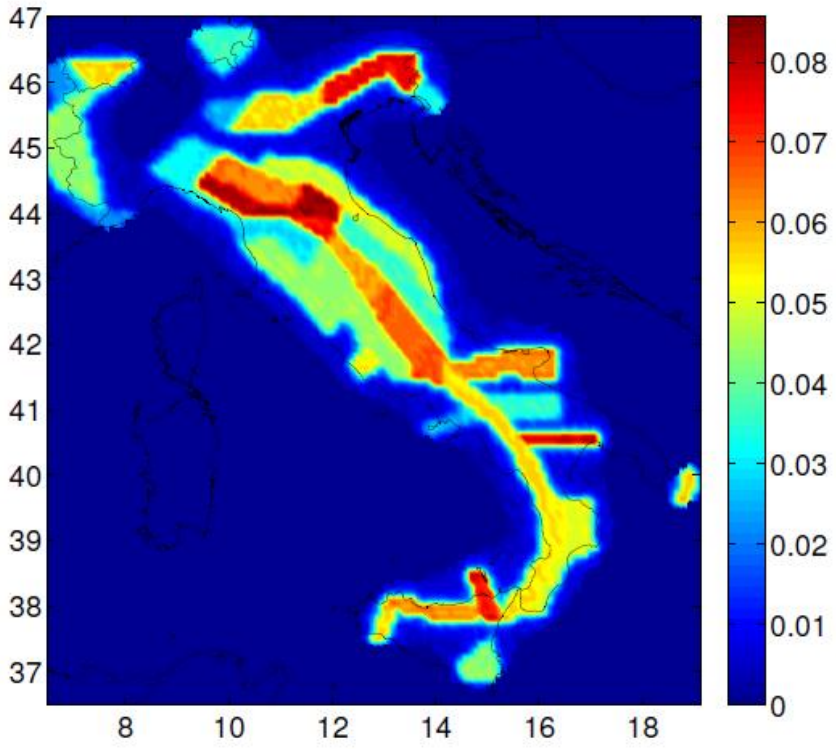


Figure 4: Differences between PGA values with exceeding probability of 10% in 50 years calculated with the corrected seismicity rates and with the declustered rates.

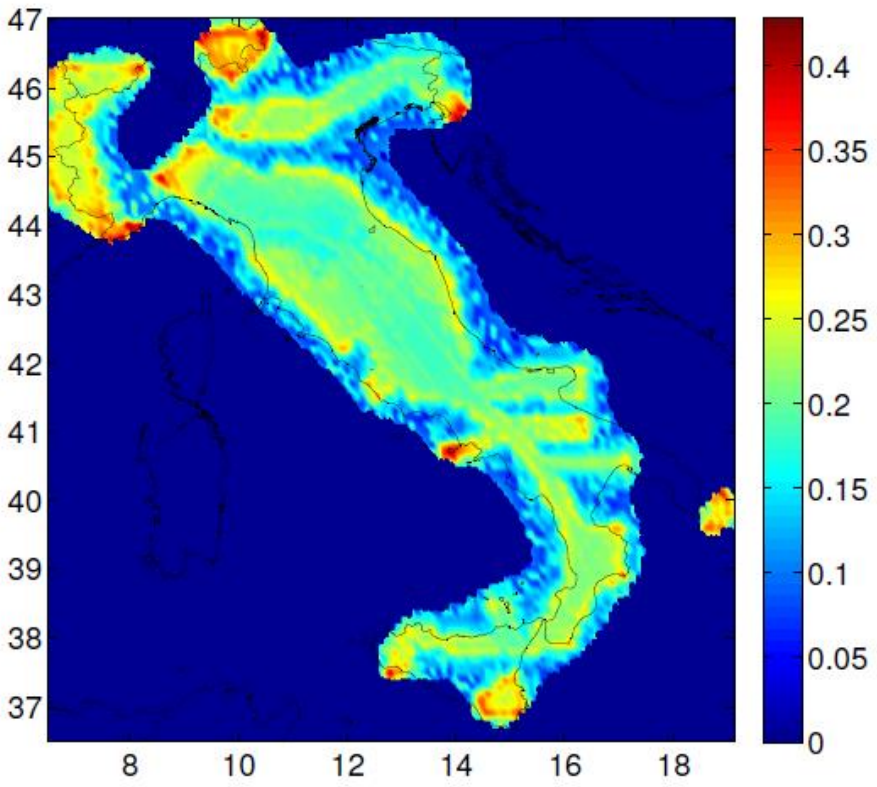


Figure 5: The same as figure 4, but normalized to the value of the declustered map.

Conclusions

We have discussed one of the basic and very common procedures of the probabilistic seismic hazard analysis, i.e., the declustering of seismicity rates. The declustering is meant to achieve two main goals: (i) to get a Poisson distribution for the (declustered) earthquakes in order to move from exceedance rates to probabilities; (ii) to remove the spatial bias of undeclustered catalogs that is caused by the fact that the few clusters recorded in the catalog are not representative of the whole distribution of possible clusters.

In this paper we have showed that the Poisson distribution for earthquake is a sufficient, but not necessary condition to have a Poisson distribution for the exceedances. In fact, for most of practical applications (where the exceedance rate is equal or smaller than 0.10 and the ground shaking thresholds are not negligible), a more realistic clustered distribution of earthquakes lead to a distribution of exceedances that is still well approximated by a Poisson distribution. This implies that declustering is necessary only to remove the spatial bias, but it is also necessary that the rates have to be corrected for compensating the seismicity rate reduction due to declustering. In fact, the seismic hazard analysis based on declustered seismicity rates leads to underestimate significantly

the exceedance probabilities, and, consequently, it may lead often to reject hazard maps in a testing phase and to underestimate the seismic risk.

One of the main risk reduction action linked to the seismic hazard map is the definition of the building code. However, we think that a discussion of the appropriateness of the actual building code definition has to be made by regulators, not by seismologists (Marzocchi, 2013). Building codes are usually defined for ground acceleration thresholds that have 10% of probability to be overcome in the next 50 years. These parameters (10% in 50 years) were not chosen because they own a specific physical meaning, but they represent a balance between safety and feasibility, because too stringent regulations may lead to prohibitive costs for the new buildings and for retrofitting. Our results show that the ground acceleration thresholds for building code adopted using hazard maps based on declustered seismicity rates have a higher exceedance probability than 10% in 50 years. In principle, regulators may either keep the existing ground acceleration values increasing the exceedance probabilities associated to them, or maintaining the exceedance probabilities to 10% in 50 years, and increasing the ground acceleration thresholds.

Data and Resources

We have used the CPTI11 database that is available online at http://emidius.mi.ingv.it/CPTI.

The last access is in March 2013.

Bibliography

Boyd, O. S. (2012). Including Foreshocks and Aftershocks in Time-Independent Probabilistic Seismic-Hazard Analyses, Bull. Seismol. Soc. Am. 102 909-917.

Cauzzi, C., and Faccioli, E. (2008). Broadband (0.05 to 20 s) prediction of displacement response spectra based on worldwide digital records, J. Seismol. 12 453-475.

Cinti, F. R., Faenza, L., Marzocchi, W., and Montone, P. (2004). Probability map of the next Mw 5.5 earthquakes in Italy, Geochem. Geophys. Geosyst. 5 Q11003, doi:10.1029/2004GC000724.

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

CPTI Working Group. (2004). Catalogo Parametrico dei Terremoti Italiani, versione 2004 (CPTI04). INGV, Bologna. Available on-line at http://emidius.mi.ingv.it/CPTI (last accessed March 2013).

Faenza, L., Marzocchi, W., and Boschi, E. (2003). A non-parametric hazard model to characterize the spatio-temporal occurrence of large earthquakes; an application to the Italian catalogue, Geophys. J. Int. 155 521-531.

Fry, B., Benites, R., and Kaiser, A. (2011). The character of accelerations in the Mw 6.2 Christchurch earthquake, Seismol. Res. Lett. 82 846-852.

Gardner, J. K., and Knopoff, L. (1974). Is the sequence of earthquakes in southern California, with aftershocks removed, Poissonian?, Bull. Seismol. Soc. Am. 64 1363-1367.

Gasperini, P., Stucchi, M., and Vannucci, G. (2004) Zonazione sismogenetica ZS9. App. 2 al Rapporto Conclusivo.

Gruppo di Lavoro, M. P. S. (2004). Redazione della mappa di pericolosit sismica prevista dallOrdinanza PCM 3274 del 20 marzo 2003. Rapporto Conclusivo per il Dipartimento della Protezione Civile, INGV, Milano-Roma,5.

Iervolino, I., Giorgio, M., and Polidoro, B. Probabilistic seismic hazard analysis for seismic sequences, Vienna Congress on Recent Advances in Earthquake Engineering and Structural Dynamics 2013, C. Adam, R. Heuer, W. Lenhardt & C. Schranz (Editors) 28-30 August 2013, Vienna, Austria Paper No. 66

Jackson, D. D., and Kagan, Y. Y. (1999). Testable earthquake forecasts for 1999, Seismol. Res. Lett. 70 393-403.

Kagan, Y. Y. (2010). Earthquake size distribution: Power-law with exponent?. Tectonophysics 490 103-114.

Le Cam, L. (1960). An approximation theorem for the Poisson binomial distribution, Pacific J. Math. 10 1181-1197.

Lomnitz, C. (1966). Statistical prediction of earthquakes, Rev. Geophys. 4 377-393.

Marzocchi, W. (2013). Seismic hazard and public safety, Trans. Am. Geophys. Union 94 240-241.

Ogata, Y. (2011). Significant improvements of the space-time ETAS model for forecasting of accurate baseline seismicity, Earth Planets Space 63 217.

Reasenberg, P. (1985). Second-order moment of central California seismicity, 1969-1982, J. Geophys. Res. 90 5479-5495.

Rovida, A., Camassi, R., Gasperini, P., and Stucchi, M. (2011). CPTI11, la versione 2011 del Catalogo Parametrico dei Terremoti Italiani. Milano, Bologna, http://emidius.mi.ingv.it/CPTI (last accessed March 2013).

Scognamiglio, L., Margheriti, L., Mele, F. M., Tinti, E., Bono, A., De Gori, P., ...and Quintiliani, M. (2012). The 2012 Pianura Padana Emiliana seimic sequence: locations, moment tensors and magnitudes, Ann. Geophys. 55.

Serfling, R. J. (1975). A general Poisson approximation theorem, Ann. Probab. 3 726-731.

Van Stiphout, T., Schorlemmer, D., and Wiemer, S. (2011). The effect of uncertainties on estimates of background seismicity rate, Bull. Seismol. Soc. Am. 101 482-494.

Van Stiphout, T., Zhuang, J., and Marsan, D. (2012). Seismicity declustering, Community Online Resource for Statistical Seismicity Analysis.

Zhuang, J., Ogata, Y., and Vere-Jones, D. (2002). Stochastic declustering of space-time earthquake occurrences, J. Am. Stat. Assoc. 97 369-380.

Accounting for Epistemic Uncertainty in PSHA: Logic Tree and Ensemble Modeling

W. Marzocchi[1], M. Taroni[1], J. Selva[2]
1- Istituto Nazionale di Geofisica e Vulcanologia, Via di Vigna Murata 605, 00143 Roma, Italy
2- Istituto Nazionale di Geofisica e Vulcanologia, Via D. Creti 12, 40128 Bologna, Italy

Abstract

The logic tree scheme is the probabilistic framework that has been widely used in the last decades to incorporate epistemic uncertainties into probabilistic seismic hazard analysis (PSHA). Notwithstanding its vast popularity, the use of the logic tree in a PSHA context has few important conceptual drawbacks. Some of these drawbacks have been reported in the past, but a careful evaluation of their impact on PSHA is still lacking. Specifically, in this work we explore in detail the most controversial issues related to the application of the logic tree on PSHA and suggest a possible alternative framework to incorporate formally epistemic uncertainty in the calculations. First, we show that the output of a logic tree is often misrepresented; for example, the use of percentiles (median included) is not consistent with the probabilistic framework of the logic tree scheme. Second, we show that the output of a logic tree model necessarily represents a biased view of the real hazard, posing significant challenges for testing the seismic hazard model. Nonetheless, we also show that in most of practical applications the logic tree is not implemented in its standard (and proper) way, but de facto as a tool to create an ensemble model. In this perspective, the term logic tree is no longer necessary and should be avoided because its use gives only a basis for criticism. Finally, we show that the ensemble modeling accounts for epistemic uncertainties in a sound probabilistic framework and it provides PSHA outputs that are in a testable format.

Introduction

Any reliable hazard assessment has to incorporate in a proper way the most relevant uncertainties, or, using the D. Rumsfeld's words, all known unknowns. Many hazard practitioners distinguish uncertainties of different nature in a convenient way, adopting the term aleatory uncertainty/variability to describe the intrinsic irreducible variability of the natural process, and epistemic uncertainty to characterize all reducible uncertainties due to our limited knowledge about the process.
Notwithstanding the popularity, many authors take the view that this subdivision does not have a theoretical significance because, as far as our knowledge of the system increases, all uncertainties become epistemic. Recently, Marzocchi and Jordan (2014) suggest that a clear and univocal distinction between aleatory and epistemic uncertainty can be made only in the framework of a well defined experimental concept. In general, the experimental concept defines collections of data, observed and not yet observed, that are judged to be exchangeable (Draper et al., 1993); for instance, when we decide the set of data that will be used to test the model, we are implicitly defining an experimental concept. In this view, the distinction between aleatory and epistemic uncertainty is not only of practical convenience, but it is of primary importance to make any probabilistic assessment testable, and consequently scientific (see Marzocchi and Jordan, 2014 for more details).

In probabilistic seismic hazard analysis (PSHA), a typical experimental concept consists of the collection of the number of exceedances (the number of times in which a specific ground motion parameter overcomes a preselected threshold) in one specific site or in one region that are then considered an exchangeable dataset. In this experimental concept, epistemic uncertainty is represented by the lack of knowledge of what is the right model to describe this set of data, both in terms of functional mathematical formulation and/or of its parameters, and the probability is interpreted as the frequency of this exchangeable dataset.

The inclusion of the epistemic uncertainty is tackled using the logic tree structure as originally suggested by Kulkarni et al. (1984). In essence, all epistemic uncertainties related to PSHA should be represented by different branches in a graphical logic tree structure. Despite the use of the logic tree scheme has become de rigoeur (Bommer and Scherbaum, 2008), it is well known that there are

potential pitfalls that should be taken seriously into account, and there is still discussion on the real meaning of the logic tree output (cf. Abrahamson and Bommer, 2005; McGuire et al., 2005). In this paper, we explore in detail how the logic tree concept is applied into PSHA practice, pointing out potential problems and possible solutions. In particular, we introduce a procedure based on ensemble modeling that has a formal probabilistic interpretation, and it avoids the pitfalls intrinsic to the use of the logic tree in PSHA practice.

The probabilistic structure of the logic tree

The logic tree is one of the several flavors of graphical probability trees (issue tree, event tree, fault tree, etc) that aim to dissect a specific problem into its basic components. The structure and the kind of the tree are defined according to its practical use. Yet, we can identify one common feature for all the tree structures: the branches emerging from each node must represent a mutually exclusive and collectively exhaustive (MECE) set of events. The most important consequence is that one path of the tree must represent the truth. The need to have a MECE set of events is imposed by how probabilities are combined in the probability tree. For example, consider a logic tree where we are interested in the probability of one specific event E and the terminal branches of the tree mimic different ways (different models) that can be used to get such a probability. The final assessment is given by the law of total probability that reads

$$\Pr(E) = \sum_{i=1}^{N} \Pr(E \cap H_i) = \sum_{i=1}^{N} \Pr(E \mid H_i) \Pr(H_i) \qquad (1)$$

where Pr(E) is the probability of the event of interest, Pr(E|Hi) is the conditional probability of the event E given the model Hi, and Pr(Hi) is the probability that the model Hi is the true model. As long as equation 1 is used in the probability tree, the models Hi must represent a set MECE.
The functioning of the logic tree can be grasped through a simple example (figure 1) where there are two boys (Tim and Tom) having two and three coins each. The Tim's coin are biased, i.e., Pr(head) is 0.4 and 0.3. The Tom's coins are biased as well, having Pr(head) equal to 0.7, 0.7, and 0.8. If we do not know who will toss the next coin (Tim and Tom have the same probability) and the coin that will be used (each coin has the same probability to be thrown), the tree has five terminal branches with different weights, i.e., each one of the Tim's branches has Pr(Hi) =0.25, while the Tom's branches have weight Pr(Hi) $= 0.1\overline{6}$. In this case, using equation 1, we get Pr(E)=0.54. This value has a frequentist interpretation. If we run a simulation in which, for each run, we select randomly the boy who will toss the coin and the coin to be tossed, the expected long-term frequency of head is 0.54. This example does not pretend to be exhaustive of the functioning

of any possible probability tree, but it underlies the basic features that we are going to discuss in the next sections of the paper.

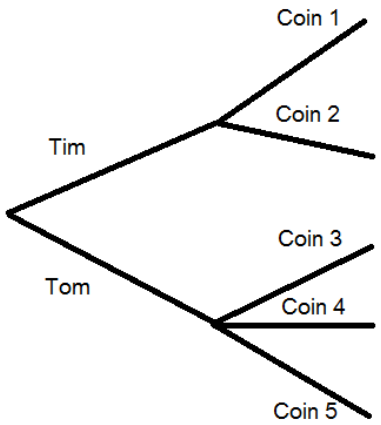


Figure 1. Logic tree of the Tim & Tom example

The logic tree in PSHA

In PSHA, the different branches of the logic tree are meant to describe the epistemic uncertainties related to the different components (nodes) of the hazard model. Since no practitioner believes that one of the path of the logic tree represents the true hazard, the MECE assumption is pragmatically resumed replacing the term true with the one that should be used (Scherbaum and Kuhen, 2011). However, we note that since the beginning (Kulkarni et al., 1984) the logic tree has been sometimes used in PSHA in a peculiar way. For instance, it has been used to define percentiles of the expected ground shaking probability, reporting the average value (given by equation 1) and a confidence interval defined through the percentiles. In some cases, it has been argued that the hazard should be represented by a percentile instead of the average (Abrahamson and Bommer, 2005). The use of percentiles is questioned by other practitioners (McGuire et al., 2005; Musson, 2005) who assert that a logic tree can provide only one number; using their words, the mean hazard is the hazard that is obtained over epistemic uncertainties. In this paper we argue that both views can lead to misunderstandings and conceptual problems.

The essence of these conceptual problems is captured by comparing the PSHA practice with a case in which the logic tree can be applied in a proper way. In particular, let us to come back to the example of figure 1 where equation 1 applies and it is certainly meaningful. In that case, the average probability represents the long-term frequency of the event conditioned to the fact that each run will follow a different path of the tree. This is fundamentally different from the PSHA case, where we expect that the branch that should be used is always the same after each earthquake. It is

like to toss always the same coin by the same boy. In this case we would be sure that the average of 0.54 is wrong. In brief, we can be sure that the average hazard will be certainly wrong because it will never coincide with the branch that should be used.

Often the logic tree is used to provide an interval of values instead of a single value. The interval is usually bounded by two percentiles, like, for example, the 10-th and 90-th percentiles. In practice, this interval is meant to describe a range that has a 0.8 probability to include the branch that should be used. In our view, this peculiar use of the logic tree makes sense in practice, but it is hardly justifiable in terms of the logic tree structure described in the previous section. For instance, what is the meaning of a confidence interval that can be obtained by the logic tree reported in figure 1? In that case, the average has a specific meaning, but an interval defined by the percentiles of the different branches would not have sense from a probabilistic point of view.

In the following section we discuss in detail a more proper probabilistic framework to interpret such an interval.

Logic tree or ensemble modeling?

Reading the literature on the logic tree applications to PSHA we feel that practitioners use the logic tree to sample, rather than to fully describe the epistemic uncertainty. This is not a trivial semantic difference. This means that practitioners use the terminal branches of the logic tree like a sample of a parent (unknown) distribution of the expected exceedance frequency. Noteworthy, the definition of the parent distribution for a set of values can be considered as equivalent to the definition of an ensemble model (Marzocchi et al., 2012)

In figure 2, we show a simple example of seismic hazard analysis made with a logic tree and the interpretation of the same hazard model based on the ensemble modeling. Figure 2a shows the logic tree that is composed by two different seismicity rates, two b-values, two maximum magnitude, and two ground motion prediction equations (GMPEs) (Cauzzi & Faccioli, 2008, Bindi et al., 2011) for a 16-branches tree. Panel 2b reports the discrete distribution of the exceedance probability for all branches and the parent distribution that can be obtained by the values of the sixteen branches. At first, we note that the ensemble model has a continuous distribution, so it does not consider the sixteen values as belonging to a MECE dataset. Indeed, the most probable values are not the ones of the branches. Second, it considers as possible (even unlikely) values that are larger (smaller) than the highest (lowest) of the sixteen values. In this case, we have modeled the parent distribution using a Beta distribution that is particularly suitable for describing random variables bounded between 0 and 1 and having a unimodal distribution. This assumption can be checked quantitatively through classical goodness-of-fit tests. In case, other more appropriate distributions can be used.

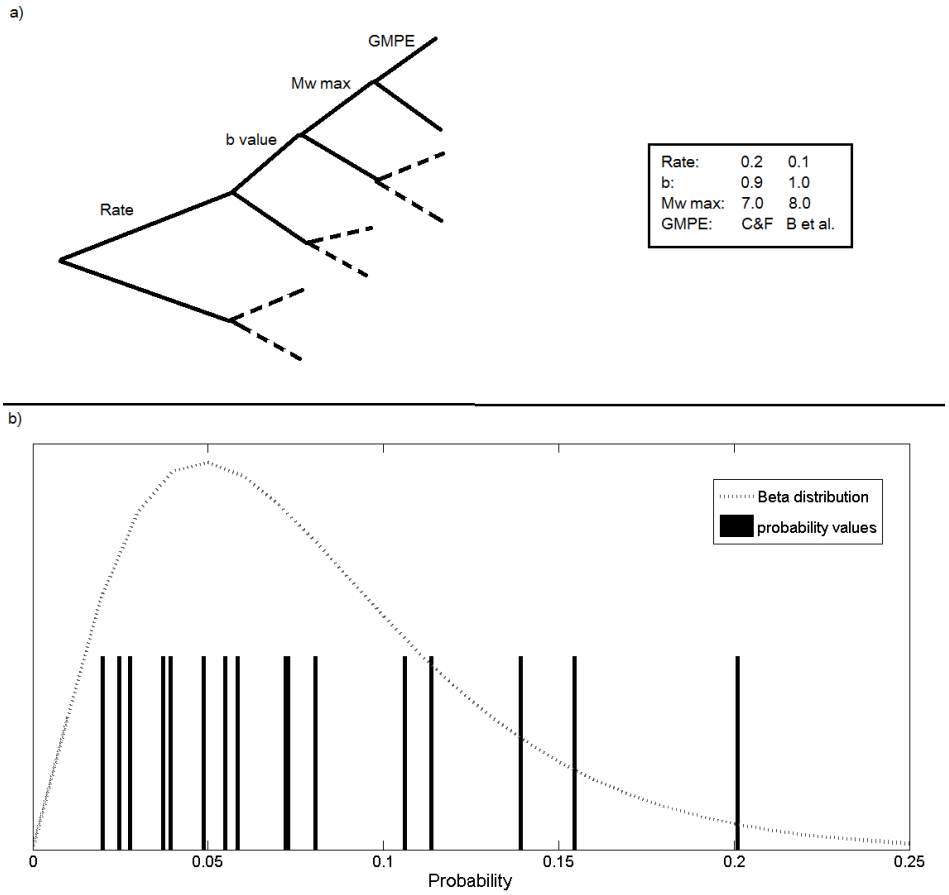


Figure 2. Panel a) logic tree of the seismic hazard example; panel b) the 16 values of the final branches of the logic tree and the Beta distribution made by these values.

The exceedance frequencies of the different branches usually incorporate a significant amount of expert opinions, for instance, when the weights of the branches are subjectively assigned by experts. The use of subjective expert opinion raised several critics to the scientific meaning of PSHA. However, we think that the use of different degrees of subjective expert opinion in hazard modeling does not pose any specific problem. In fact, Marzocchi and Jordan (2014) show that all critics on the use of subjective expert opinion in scientific modeling are based on the wrong dichotomy subjective/unscientific (non testable). This wrong dichotomy is well depicted by the experiment set up by Sir Francis Galton early in the 20th century. During a tour of the English countryside, he recorded 787 guesses of the weight of an ox made by a farming crowd and found that the average was correct to a single pound (Galton, 1907). The collection of subjective measurements he tabulated passes the appropriate Student-t test (retrospectively, of course, since this test wasn't invented until 1908). It is worth noting a difference between expert farmers and PSHA experts. In PSHA, the experts' opinion depicts the epistemic uncertainty at a particular epoch, which may be

larger or smaller depending on how the experts are able to sample the appropriate information space. In Galton's experiment, a farmer looks individually at the ox, reaching his estimate independently of his colleagues. As more farmers participate, they add new data; i.e., the epistemic uncertainty is reduced by averaging out their guesses, because each experts' opinion counts like a real measure. At a particular epoch, adding more PSHA experts will better determine, but usually not reduce, the epistemic uncertainties, because they do not make independent observations, but work from a common body of knowledge.

We underline that the ensemble modeling has several remarkable advantages and avoids the pitfalls of the logic tree. First, the structure of the logic tree makes complex the inclusion of epistemic uncertainty coming from different types of modeling. For example, some GMPEs require an estimation of the focal mechanism that is not provided by all available earthquake occurrence models. In practice, if we want to use such GMPEs, we are forced to exclude a priori some reliable earthquake occurrence model just because it does not provide a forecast of the focal mechanism. Second, the logic tree structure may lead to some inconsistencies among the branches. For example, when a logic tree structure is particularly complex, one of the alternatives in a specific node may be not consistent with alternatives of previous nodes. This happens when the nodes are not completely independent. For example, in figure 2, it may be argued that the seismicity rate and the b-value are not independent, so, one particular b-value may be not consistent with one of the seismicity rate adopted (Musson, 2005). Ensemble modeling does not have these limitations. Different measures of the exceedance probability (frequency) can be obtained in different ways, using a logic tree or using completely independent approaches, leaving more flexibility to incorporate all epistemic uncertainty. Noteworthy, in the ensemble modeling the correlation among models can be included and properly managed. To this purpose, Marzocchi et al. (2012) proposed a method to assign the weights of each model taking into account also the correlation among the forecasts.
Third, as said before, the logic tree requires a MECE set of models, while the ensemble modeling does not. Besides to be more appropriate from a conceptual point of view, relaxing the MECE assumption has also the important advantage to make easier the interpretation of the weight of each model that basically represents the 'confidence' of the experts about the output of that specific model, once the correlation of models has been taken into account.
Four, the use of ensemble modeling legitimates the use of percentiles in a meaningful way. For example, the interval bounded by the 10-th and 90-th percentiles represents the range of values having a 0.8 probability to include the real value (Marzocchi and Jordan, 2014).

Ensemble modeling in PSHA

To build an ensemble model, we may simply collect the output of independent models and use them to build a parent distribution for the seismic hazard. This is what has been done in Figure 2 where the sixteen branch values have been used to define the ensemble Beta distribution. In principle, we may also merge two different ensemble models, one for the earthquake occurrence, and one for the ground motion prediction equations. Mathematically, this may be accomplished replacing the single seismic and exceedance rates in the classical seismic hazard equation with statistical distributions. In this case, the exceedance rate of the ground motion parameter z at one specific site is calculated through

$$[\lambda(Z > z)] = \int_{\Xi} [P(Z > z \mid \dot{\xi})][\lambda(\dot{\xi})] d\dot{\xi} \qquad (2)$$

where $\dot{\xi} \in \Xi$ is a vector that contains all relevant earthquake source parameters like, for instance, the location and the magnitude; $\lambda(\cdot)$ is the rate of occurrence; $P(Z > z \mid \dot{\xi})$ is the GMPE that gives the exceedance probability of the ground motion value z given the knowledge of source parameters. In equation 2 the square brackets stand for a distribution. For example, the rate of occurrence is not anymore given by a single number, but by an ensemble distribution that represents the epistemic uncertainty over that quantity. The distribution on the left hand side can be obtained simulating random values from the two distributions on the right hand side.

Conclusions

In this work we have discussed the methods to incorporate epistemic uncertainties into PSHA. We have argued that the logic tree, which is by far the most used technique, has few important drawbacks that make it inappropriate to describe the epistemic uncertainty in PSHA and that have drawn harsh critics in the past to the use of PSHA.

We have introduced the basic principles of a method, based on ensemble modeling, that overcomes the pitfalls of the logic tree and should be preferred to describe the epistemic uncertainties in PSHA. Fundamentally, the ensemble modeling is based on a more appropriate probabilistic scheme that allows us to overcome all major obstacles to the use of the logic tree. Moreover, the ensemble modeling is more flexible than the logic tree to incorporate different types of modeling. As final consideration, we have noted that the logic tree is usually implemented in PSHA practice in an unusual way that makes it similar to an ensemble modeling.

Bibliography:

Abrahamson, Norman A., and Julian J. Bommer. "Probability and uncertainty in seismic hazard analysis." Earthquake Spectra 21.2 (2005): 603-607.

Bindi, D., et al. "Ground motion prediction equations derived from the Italian strong motion database." Bulletin of Earthquake Engineering 9.6 (2011): 1899-1920.

Bommer, Julian J., and Frank Scherbaum. "The use and misuse of logic trees in probabilistic seismic hazard analysis." Earthquake Spectra 24.4 (2008): 997-1009.

Cauzzi, Carlo, and Ezio Faccioli. "Broadband (0.05 to 20 s) prediction of displacement response spectra based on worldwide digital records."Journal of Seismology 12.4 (2008): 453-475.

Draper D, Hodges J, Mallows C, Pregibon D (1993). Exchangeability and data analysis (with discussion). Journal of the Royal Statistical Society Series A, 156, 9.

Galton, Francis. "Vox populi (the wisdom of crowds)." Nature 75 (1907).

Kulkarni, R. B., R. R. Youngs, and K. J. Coppersmith. "Assessment of confidence intervals for results of seismic hazard analysis." Proceedings of the Eighth World Conference on Earthquake Engineering. Vol. 1. 1984.

Marzocchi W., Jordan T.H. (2014). Testing for Ontological Errors in Probabilistic Forecasting Models of Natural Systems. Submitted to Proc. Nat. Acad. Sci.

Marzocchi, Warner, J. Douglas Zechar, and Thomas H. Jordan. "Bayesian forecast evaluation and ensemble earthquake forecasting." Bulletin of the Seismological Society of America 102.6 (2012): 2574-2584.

McGuire, Robin K., C. Allin Cornell, and Gabriel R. Toro. "The case for using mean seismic hazard." Earthquake Spectra 21.3 (2005): 879-886.

Musson, R. M. W. "Against fractiles." Earthquake Spectra 21.3 (2005): 887-891.

Scherbaum, Frank, and Nicolas M. Kuehn. "Logic tree branch weights and probabilities: Summing up to one is not enough." Earthquake Spectra 27.4 (2011): 1237-1251.