Geophys. J. Int. (2014) **196,** 422–431
Advance Access publication 2013 October 11

doi: 10.1093/gji/ggt369

Assessing annual global *M*6+ seismicity forecasts

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

[1] *Istituto Nazionale di Geofisica e Vulcanologia, Via di Vigna Murata,* 605*, I-*00143 *Roma, Italy. E-mail:* matteo.taroni@ingv.it
[2] *Swiss Seismological Service, ETH Zurich, Sonneggstrasse* 5*, NO H* 3*, CH-*8092 *Zurich, Switzerland*
[3] *Department of Earth Sciences, University of Southern California,* 3651 *Trousdale Pkwy, Los Angeles, CA* 90089*, USA*

Accepted 2013 September 10. Received 2013 September 10; in original form 2013 June 11

SUMMARY

We consider a seismicity forecast experiment conducted during the last 4 yr. At the beginning of each year, three models make a 1-yr forecast of the distribution of large earthquakes everywhere on the 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; Earthquake interaction, forecasting, and prediction; Statistical seismology.

1 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 nine target earthquakes have occurred since the experiment began on 2009 August 1, 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 three models participating as of 2009 January 1 (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 catalogues and to determine the testing centre 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 timescales (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 paper, 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 yr 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.

2 MODELS

In this paper we consider three classes of models: those that have been running in the US CSEP testing centre, 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 centre; at the beginning of each calendar year, these codes generate a 1-yr 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 × 1° latitude cells in which the number of expected earthquakes with moment magnitude $M_w \geq 5.95$ and depth $\leq$30 km in the next year is forecast. As in other CSEP experiments, every modeller has agreed that the Poisson distribution should be used to represent the uncertainty in the annual rate in each cell.

2.1 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 centre until 2010, so we can only present results for KJSS in 2010, 2011 and 2012. In the Supporting Information, 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 2-D 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 catalogue 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 non-zero 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.

2.2 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). However, Marzocchi *et al*. (2012) suggested that rather than selecting the best model from those in a testing centre, 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),

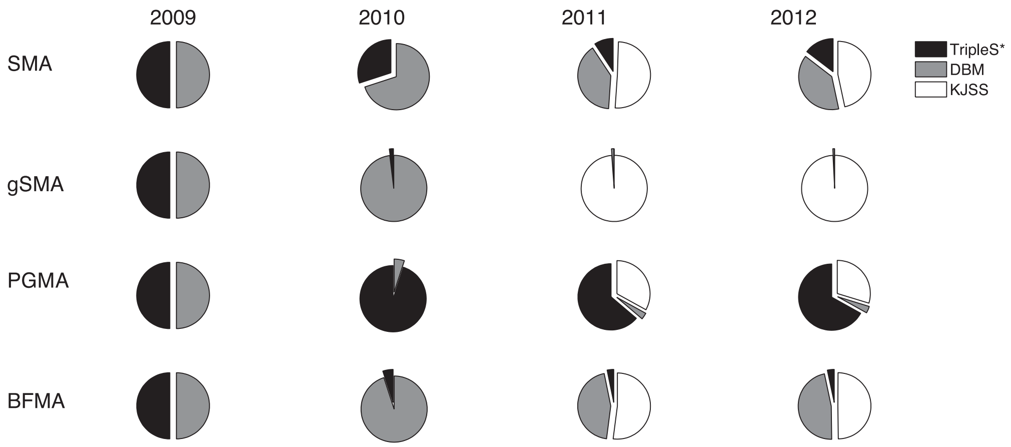

Figure 1. Composition of ensemble models for each annual experiment. In 2009, KJSS was not available and there was no prior performance result, so every ensemble was simply 50 per cent TripleS, 50 per cent DBM.

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:

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

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

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 2010, 2012); see Section 4.3 for details. PGMA assigns a weight to the i-th model according to this formula:

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

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 per cent.

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 (TBF) which in a purely prospective experiment is the total likelihood ratio (Kass & Raftery 1995). One calculates the TBF of a model by summing the log-Bayes factor of the model with respect to all the others. The corresponding weight is

$$\omega_i^{\mathrm{BFMA}} = 1 + \beta \cdot \mathrm{TBF}_i, \tag{4}$$

where TBF_i is the TBF of the i-th model, $\beta = \frac{0.90}{|\mathrm{TBF}_{\min}|}$ and $\mathrm{TBF}_{\min}$ 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 four ensemble models are identical and give 50 per cent weight to the DBM forecast and 50 per cent 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 three models contribute to build the four ensemble models; in this case we have three 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 Fig. 1.

2.3 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 2 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.

3 DATA

Experiment participants agreed to use the Global Centroid Moment Tensor (GCMT) catalogue (Dziewonski *et al.* 1981; Ekström *et al.* 2005) for model development and assessment. Compared with other global earthquake catalogues, the GCMT catalogue 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 M_{w} from the total moment M_0 reported in the catalogue. Kagan (2003) suggested that the catalogue is complete for earthquakes occurring at depths no greater than 70 km at M_{w} 5.3. For the prototype global experiment, participants agreed that target earthquakes would be all those with magnitude not smaller than M_{w} 5.95 and depths not greater than 30 km—we did not decluster the catalogue. In the first 4 yr of this experiment, the number of target events was 92, 103, 108 and 91, respectively.

4 ASSESSMENT TECHNIQUES

The statistical tests used in CSEP testing centres 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 Regional Earthquake Likelihood Models (RELM) assessment, Schorlemmer *et al*. (2007) proposed that a model that failed a consistency test should be 'rejected'. However, 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 2012 November 26 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 centre 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 ($M_{\mathrm{w}} \geq 5.95$), 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). However, 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 (IG) 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 behaviour rather than mean behaviour. The rationale for using the Sign-test is that for asymmetric distributions the median is a more appropriate indicator of the centre 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. Moreover, 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.

4.1 Consistency tests

4.1.1 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

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

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

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 per cent significance.

4.1.2 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:

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

where $\{L_{\mathrm{S}}\}$ is the set of simulated likelihoods, L is the likelihood of the model with respect to the observed catalogue 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 per cent significance.

4.1.3 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 similarly summarized by a quantile scores, ζ:

$$\zeta = \frac{\#\{S_x \mid S_x \leq S, S_x \in S_{\mathrm{S}}\}}{\#\{S_{\mathrm{S}}\}}, \tag{8}$$

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

4.2 Comparison tests

The *T*-test, *W*-test and Sign-test are based on the 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 IG for the *i*-th event is defined as:

$$\mathrm{IG}_i(\mathrm{A}, \mathrm{B}) = \ln(\lambda_{\mathrm{A}_i}) - \ln(\lambda_{\mathrm{B}_i}) - \frac{\Lambda_{\mathrm{A}} - \Lambda_{\mathrm{B}}}{N}, \tag{9}$$

where λ_{A_i} is the rate of model A in the bin where the *i*-th earthquake occurs, λ_{B_i} 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.

4.2.1 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 IGs 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.* (2011): when the sample IGs are not normally distributed, we use the *W*-test instead of the *T*-test.

4.2.2 W-test

The Wilcoxon signed-rank test (e.g. Siegel 1956) is a non-parametric alternative to Student's paired two-sample *t*-test. The null hypothesis of the *W*-test is that the median IG 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.

4.2.3 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 IGs are symmetric (Gibbons & Chakraborti 2003). Hence, the Sign-test is more widely applicable, but it is less powerful than the other tests when the IG distribution is symmetric (Gibbons & Chakraborti 2003).

4.2.4 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:

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

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 $\log(\mathrm{BF}) > 0$ 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.

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

Log(BF)	Evidence against M1
0–1.1	Hardly worth mentioning
1.1–3	Positive
3–5	Strong
>5	Very strong

4.2.5 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

$$R^1_j = -1 + n\frac{p^1_j}{\sum_{i=1}^{n} p^i_j}, \tag{11}$$

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.

5 RESULTS

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

No CSEP model passes all three consistency tests (*N*-, *L*- and *S*-test) in the 4 yr considered (2009, 2010, 2011 2012). TripleS models spatial clustering and fails nine of 12 consistency tests, while the very simple UNIF model fails only five 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 IGs to see which tests apply (Fig. 3). We begin with the Lilliefors test to check the assumption that IGs are normally distributed. If this assumption is not violated, we apply the *T*-test, and otherwise we

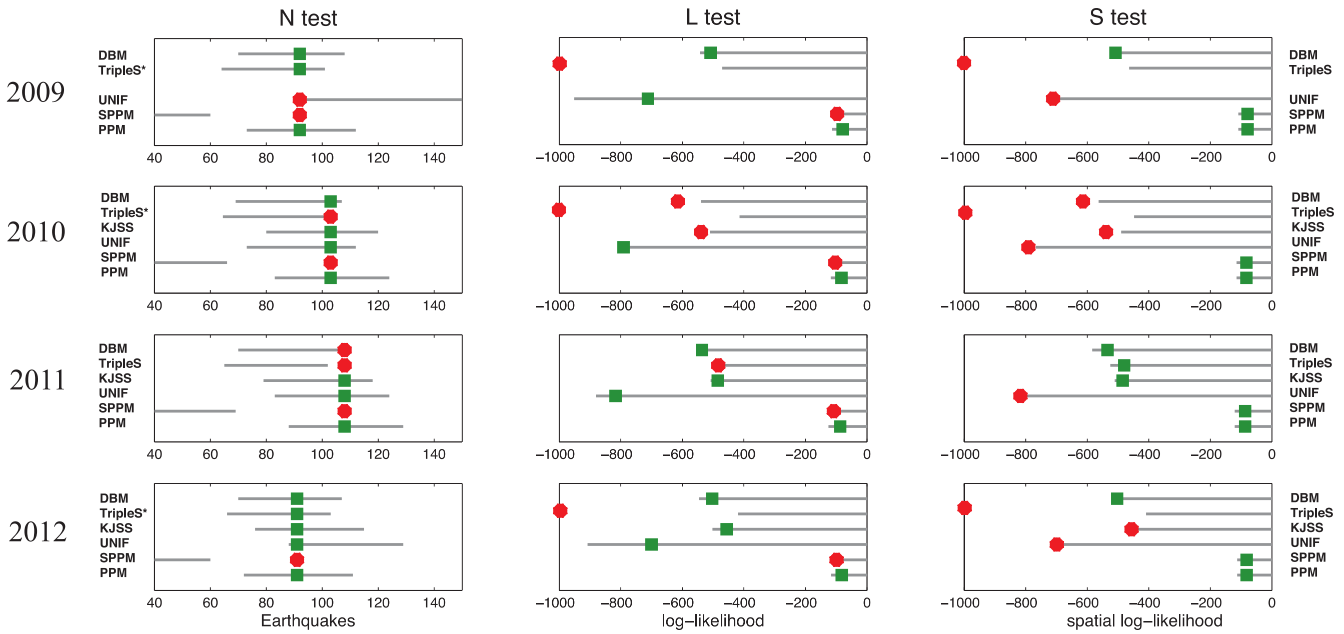

Figure 2. Results of CSEP consistency tests for CSEP and reference models. For the *N*-test we consider the total number of earthquakes forecast by each model (the total rate), for *L*-test and *S*-test we consider the log-likelihood and the space log-likelihood of each model, respectively. The grey line shows the non-rejection region, at 5 per cent 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.

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 IGs 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 IGs 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.

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.

	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 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.

	DBM versus TripleS*		DBM versus KJSS		TripleS* versus 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
2010–2012	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

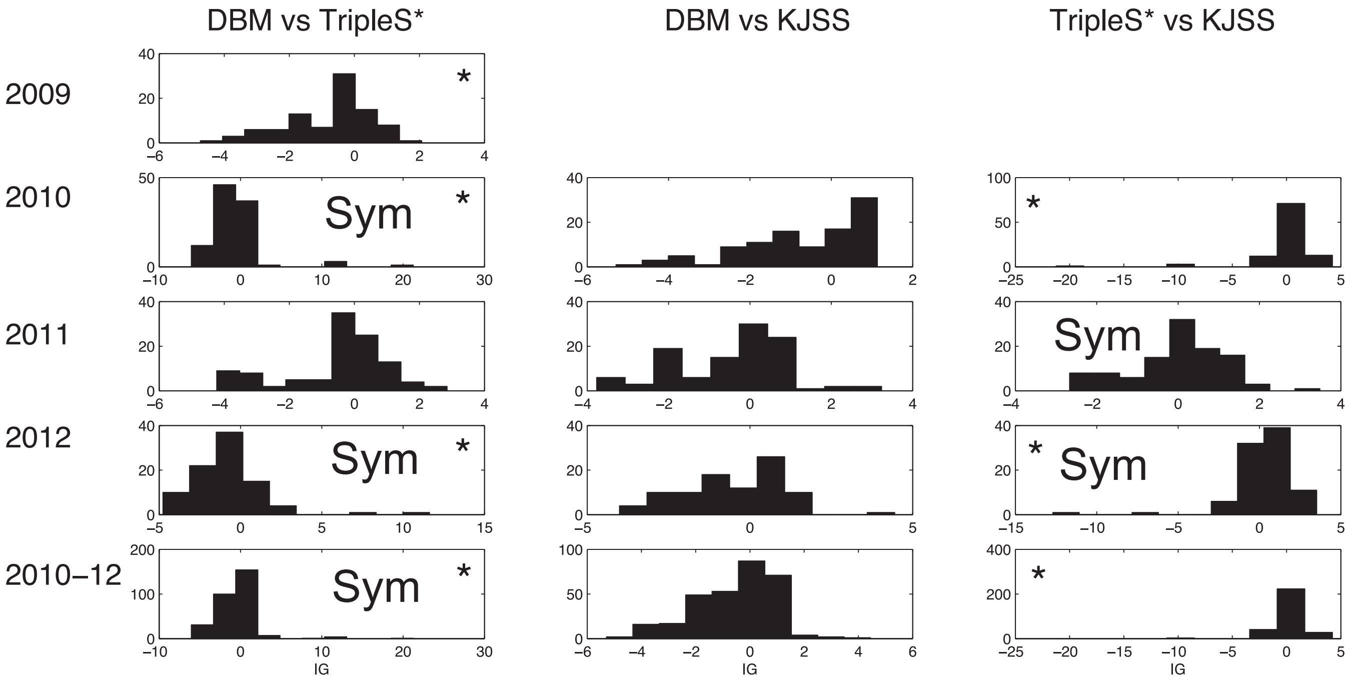

Figure 3. Information gain histograms for each pair of models (DBM versus TripleS*, DBM versus KJSS, TripleS* versus KJSS) for 2009, 2010, 2011, 2012 and all years considered jointly. The null hypothesis of a normal distribution was rejected at <0.05 significance with a Lilliefors test for all model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

Table 4. Rank of the CSEP models for 2009, 2010, 2011 and 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 do not need to use the TripleS* (with the rate correction) instead the TripleS.

	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
2010–2012 (302)	1. DBM, TripleS*, KJSS	1. KJSS 2. DBM 3. TripleS	1. TripleS 2. KJSS 3. DBM

The most striking feature of Table 4 is that the model ranking is not the same for each comparison test. However, 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 4 yr 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 IG 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 IG results are ambiguous; DBM is better than TripleS according to the *W*-test,

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.

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

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 IG 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 4 yr 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—that is, 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. However, 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.

6 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.

6.1 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. However, 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). However, 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 M_w 5.95.

6.2 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 are 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 reasons 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.

6.3 Looking for the 'best' model

Marzocchi & 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.

7 CONCLUSIONS: SUGGESTIONS FOR THE NEXT GENERATION OF GLOBAL SEISMICITY FORECAST EXPERIMENTS

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. However, 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 paper 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):

(1) The GCMT catalogue should be used, and forecasts could target events as small as M_w 5.5;

(2) Unlike the forecasts in this study, each cell should specify a magnitude distribution, rather than lumping earthquakes of different sizes together;

(3) Any assessment that can reveal model features to be improved, or differences between models, should be used;

(4) Catalogue 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 (Gutenberg & Richter 1954), and

(i) a global Gutenberg–Richter b-value or
(ii) b-value per spatial cell;

(3) Probability in every space/magnitude/number voxel (any nonspecified 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 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 epicentres (or epicentroids), because these are commonly estimated by network agencies and it is relatively straightforward to assign a point to a grid cell. However, 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.

ACKNOWLEDGEMENT

This research was partially carried out in the framework of REAKT Project (Strategies and tools for Real-Time EArthquake RisK ReducTion) founded by the European Community via the Seventh Framework Program for Research (FP7), contract no. 282862.

REFERENCES

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*, American Association for the Advancement of Science.

Dixon, W.J. & Mood, A.M., 1946. The statistical sign test, *J. Am. 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 Paciffic, *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. Geophys.*, **36,** 51–60.

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

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

Jordan, T.H., 2006. Earthquake predictability, brick by brick, *Seism. Res. Lett.,* **77**(1), 3–6.

Jordan, T.H. *et al.*, 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. Int.*, **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*, October 15–18, Strasbourg, France, pp. 586–593.

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

Knuth, D., 1974. Structured programming with GOTO statements, *Comput. Surv.*, **6,** 261–301.

Lilliefors, H.W., 1967. On the Kolmogorov-Smirnov test for normality with mean and variance unknown, *J. Am. Stat. Assoc.*, **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, doi:10.1029/2006JB004568.

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), B08317, doi:10.1029/2007JB005472.

Marzocchi, W. & Zechar, J.D., 2011. Earthquake forecasting and earthquake prediction: different approaches for obtaining the best model, *Seism. 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, R.H., Fligner, M.A., Policello, G.E. & Wolfe, D.A., 1980. An asymptotically distribution-free test for symmetry versus asymmetry, *J. Am. Stat. Assoc.*, **75,** 168–172.

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. seism. 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 Geophys.*, **59,** 728–747.

Rong, Y., Jackson, D.D. & Kagan, Y.Y., 2003. Seismic gaps and earthquake, *J. geophys. Res.*, **108,** 2471, doi:10.1029/2002JB002334.

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., 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., 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.

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, doi:10.1029/2007JB005427.

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, August 19–24.

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.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this paper:

The figures show the log-rate of each cell and the target events for TripleS, DBM and KJSS model for all years (http://gji.oxfordjournals.org/lookup/suppl/doi:10.1093/gji/ggt369/-/DC1).

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.