Bulletin of the Seismological Society of America, Vol. 102, No. 6, pp. 2574–2584, December 2012, doi: 10.1785/0120110327

Ⓔ

Bayesian Forecast Evaluation and Ensemble Earthquake Forecasting

by Warner Marzocchi, J. Douglas Zechar,* and Thomas H. Jordan

Abstract The assessment of earthquake forecast models for practical purposes requires more than simply checking model consistency in a statistical framework. One also needs to understand how to construct the best model for specific forecasting applications. We describe a Bayesian approach to evaluating earthquake forecasting models, and we consider related procedures for constructing ensemble forecasts. We show how evaluations based on Bayes factors, which measure the relative skill among forecasts, can be complementary to common goodness-of-fit tests used to measure the absolute consistency of forecasts with data. To construct ensemble forecasts, we consider averages across a forecast set, weighted by either posterior probabilities or inverse log-likelihoods derived during prospective earthquake forecasting experiments. We account for model correlations by conditioning weights using the Garthwaite–Mubwandarikwa capped eigenvalue scheme. We apply these methods to the Regional Earthquake Likelihood Models (RELM) five-year earthquake forecast experiment in California, and we discuss how this approach can be generalized to other ensemble forecasting applications. Specific applications of seismological importance include experiments being conducted within the Collaboratory for the Study of Earthquake Predictability (CSEP) and ensemble methods for operational earthquake forecasting.

Online Material: Tables of likelihoods for each testing phase and code to analyze the RELM experiment.

Introduction

Although deterministic earthquake prediction remains an elusive goal, probabilistic earthquake forecast models have begun to quantify the temporal variation of seismic hazard and risk (Vere-Jones, 1995; Field, 2007; Schorlemmer and Gerstenberger, 2007; van Stiphout *et al.*, 2010). One recent advance has been the establishment of the Collaboratory for the Study of Earthquake Predictability (CSEP) as an international program for the prospective evaluation and comparison of forecasting models (Jordan, 2006; Gerstenberger and Rhoades, 2010; Marzocchi *et al.*, 2010; Zechar, Schorlemmer, *et al.*, 2010; Nanjo *et al.*, 2011). Another advance is the development of operational earthquake forecasting (OEF) as a set of procedures for the public dissemination of authoritative information about time-varying seismic hazard (Jordan *et al.*, 2009, 2011; Jordan and Jones, 2010). OEF requires the evaluation of earthquake forecast models for reliability and skill, and CSEP has been specifically proposed as the infrastructure for conducting these evaluations (Jordan *et al.*, 2011). Reliability is an absolute measure of performance describing the statistical agreement between the forecast probabilities of target events and the observed frequencies of those events (e.g., the mean observation conditional on a particular forecast). Skill measures the statistical performance of one model relative to another. To be useful for OEF purposes, a model must demonstrate some degree of reliability and skill. Here we consider several methodological issues related to the CSEP evaluation of OEF models.

Current CSEP experiments involve prospective blind tests of model forecasts against observations. The models forecast the distribution of future seismicity of a specified type (target earthquakes) in a fixed region during a predetermined testing period; the observations are the locations, magnitudes, and number of target earthquakes that occur during an experiment. Conceptually, each comparison of model with observation is formulated as a hypothesis test in which the null hypothesis is that the observed seismicity (a sample) is drawn from the forecast (the reference distribution). CSEP testing centers now involve many of these hypothesis tests for each experiment (see, e.g., Schorlemmer *et al.*, 2007; Zechar, Gerstenberger, and Rhoades, 2010), and the results of each test are summarized and reported as quantile scores that are equivalent to p-values. These tests follow a general procedure: one uses the forecast to simulate many catalogs and computes a parameter of interest (typically, some likelihood) for each simulated catalog; one then compares the

*Also at Department of Earth Sciences, University of Southern California, 3651 Trousdale Parkway, Los Angeles, California 90089.

same parameter for the observed catalog with the distribution of parameter values from simulations. When a p-value is small (typically, less than 0.05), the null hypothesis is rejected, indicating that the forecast is inconsistent with the observation. For example, the S(pace)-test (Zechar, Gerstenberger, and Rhoades, 2010) is based on the spatial likelihood of the observation given the forecast. If the observed spatial likelihood is smaller than more than 95% of the simulated likelihoods, one infers that the spatial component of the forecast is rejected. The typical outcome of a CSEP experiment is a report card of sorts for each forecast, comprising decisions to reject (or not reject) the null hypothesis for different components of the forecast (e.g., spatial distribution, magnitude distribution, or overall rate). Unfortunately, such an analysis does not make clear how a given forecast or set of forecasts might be practically used, say for estimating seismic hazard. Presumably, one would select the best model from a set of candidate models; because most CSEP tests are based on likelihood, the best model would be that which maximizes the likelihood of the observed target earthquakes.

Despite its widespread application throughout science, classical binary hypothesis testing has some intrinsic shortcomings (Rozeboom, 1960). One main class of drawbacks is related to the typical output of a statistical test: the decision to reject or not reject a hypothesis does not match the continuous nature of probabilistic forecasting. If one chooses a critical p-value of 0.05, a small difference, 0.04 instead of 0.06, leads to opposite conclusions that are hardly justifiable in terms of the physics of the forecasting hypothesis/model. More generally, the structure of classical hypothesis testing pushes researchers to do something more than evaluate models/hypotheses: the binary output of rejection/non-rejection implies that some decision has been made, for example, to use or not to use a candidate model. If we have N independent models under test and use a critical p-value of 0.05, the probability to reject at least one model by chance is $(1 - 0.95^N)$. As N becomes larger, it becomes increasingly likely to exclude models that are probably reliable.

Another class of shortcomings was particularly well-described by Box and Draper (1987): "…all models are wrong; the practical question is how wrong do they have to be not to be useful." With many data, a goodness-of-fit test is likely to reject all models, that is, all models will appear to be unreliable. For example, a perfectly fair coin does not exist, and so the hypothesis that heads and tails are equally likely to appear upon flipping a coin is not exactly true. On the other hand, a model may be useful even though it is not the true model: in reality, a probability of 0.5 is a good choice for betting on the outcome of a coin toss. The significance level of a model's predictive performance and the utility of this model are not the same thing; one may find that two models are statistically different (i.e., the hypothesis of equality is rejected), but the probability gain of the better model may be negligible for practical purposes.

Here we explore methods for enhancing the CSEP evaluation of earthquake forecast models and for facilitating their use in OEF. In particular, we proffer a set of Bayesian procedures to evaluate earthquake forecasts, update these evaluations during sequential testing phases, and combine different (though possibly correlated) models into ensemble forecasts in a way that takes past predictive performance into account.

Bayesian Forecast Evaluation

A typical point null hypothesis in classical statistical testing is expressed in the form $\vartheta = \vartheta_0$, where ϑ is a continuous parameter. For most parameterized models, this hypothesis will have a zero probability of being true, given that ϑ is continuous. A Bayesian test provides a posterior distribution of ϑ, and this posterior distribution is more informative than a decision of whether or not to reject the null hypothesis.

Similarly, a Bayesian approach to evaluating forecasting models replaces decisions to reject with more informative posterior distributions; such an approach is particularly useful when comparing the performances of multiple models. Suppose we observe a vector of data Ω and want to compare two competing models/hypotheses, H_1 and H_2, that comprise a complete set of mutually exclusive models/hypotheses; that is, their probabilities sum to 1, but only one of them is true. The posterior probability of each model H_i, that is, the probability that H_i is the data-generating model, is

$$P(H_i|\Omega) = \frac{P(H_i)P(\Omega|H_i)}{P(\Omega)} = \frac{P(H_i)P(\Omega|H_i)}{P(H_1)P(\Omega|H_1) + P(H_2)P(\Omega|H_2)}, \quad (1)$$

where $P(\Omega|H_i)$ is the likelihood and $P(\Omega)$ is the probability of the observations Ω. Because $P(\Omega)$ is constant for all models and can be difficult to know, equation (1) is sometimes expressed as the proportionality

$$P(H_i|\Omega) \propto P(H_i)P(\Omega|H_i). \quad (2)$$

Of course, in most real cases, no set of models can be considered complete and composed of mutually exclusive models. Moreover, if every model is wrong, it seems that $P(H_i|\Omega)$ must be zero. These objections to the Bayesian view can be addressed by interpreting $P(H_i|\Omega)$ as the posterior probability that the model H_i is the best among a set of candidate models; that is, H_i is the model that will have the maximum likelihood score using a set of independent data in a long-run experiment. The set can then be considered complete because a best model always exists (Hoeting *et al.*, 1999). $P(H_i)$ and $P(H_i|\Omega)$ are the prior and posterior probabilities, respectively, that H_i is the best model.

The Bayesian approach emphasizes model comparison. The posterior odds of H_1 and H_2 can be expressed as

$$\frac{P(H_2|\Omega)}{P(H_1|\Omega)} = \frac{P(H_2)P(\Omega|H_2)}{P(H_1)P(\Omega|H_1)} = \frac{P(H_2)}{P(H_1)} \cdot \frac{P(\Omega|H_2)}{P(\Omega|H_1)}. \tag{3}$$

The ratio

$$B_{21} = \frac{P(\Omega|H_2)}{P(\Omega|H_1)}, \tag{4}$$

is the Bayes factor (Kass and Raftery, 1995). Equation (3) suggests that the posterior odds are the product of the prior odds and the Bayes factor. In the case where two models have the same prior odds and zero degrees of freedom, the posterior odds are exactly equal to the likelihood ratio. But the Bayes factor is more general than the likelihood ratio; for example, the Bayes factor can accommodate prior uncertainties in the model parameters.

Jeffreys (1961) interpreted the Bayes factor as evidence provided by the data in favor of one model (or hypothesis). Kass and Raftery (1995) updated Jeffreys' view and offered the guide shown in Table 1. The interpretation shown in Table 1 is general, and it can be tuned for specific contexts such as criminal trials, where a very strong evidence of guilty would need a Bayes factor higher than 1000 (Evett, 1991). We note that these evidence classes scale with the logarithm of the Bayes factor.

Application to CSEP Experiments

To describe CSEP earthquake forecast experiments, we adopt the notation of Schorlemmer *et al.* (2007). The earthquake forecast is expressed as the expected number of earthquakes in multidimensional bins, with each bin representing an interval in space, time, and magnitude. (This definition can be generalized to include other dimensions such as depth and focal mechanism angles.) The forecast made by the jth model for the ith bin is denoted by λ_i^j. The forecast of the jth model for all bins is represented by a vector Λ^j with n elements, where n is the number of bins. The n-vector Ω tracks the number of earthquakes per bin, ω_i, observed during the forecast time window τ. In this view, the H^j model produces the forecast Λ^j for the time interval τ.

The CSEP suite of tests is mostly based on the concept of likelihood, that is

$$\exp[L(\Omega|\Lambda^j)] = \prod_{i=1}^{n} \Pr(\omega_i|\lambda_i^j), \tag{5}$$

Table 1
Interpretation of Jeffreys' View by Kass and Raftery (1995)

Bayes factor $\mathbf{B_{21}}$	Evidence against Model 1
1 to 3	Hardly worth mentioning
3 to 20	Positive
20 to 150	Strong
>150	Very strong

where $L(\Omega|\Lambda^j)$ is the logarithm of the joint likelihood. This quantity is the basis of the N-, L-, and R-tests described by Schorlemmer *et al.* (2007) and the M- and S-tests described by Zechar, Gerstenberger, and Rhoades (2010). The summation of the log-likelihood per bin (or, equivalently, the product shown in equation 5) implies that the number of earthquakes that occurred in one bin (ω_i) depends only on the forecast for that specific bin (λ_i^j) and not on observations in adjacent bins (ω_k where $k \neq i$). This assumption is debatable, because earthquakes interact and cluster spatially, but it may hold reasonably well in some practical applications, in particular when ω_i is small. Thus far, CSEP forecasts also assume that the expected distribution of the number of events is Poisson. In this case, equation (5) becomes

$$\begin{aligned} L(\Omega|\Lambda^j) &= \sum_{i=1}^{n} \log[\Pr(\omega_i|\lambda_i^j)] \\ &= \sum_{i=1}^{n} (-\lambda_i^j + \omega_i \ln \lambda_i^j - \ln \omega_i!). \end{aligned} \tag{6}$$

The Poisson hypothesis in the context of CSEP experiments has been discussed by Werner and Sornette (2008) and Lombardi and Marzocchi (2010). Specifically, Lombardi and Marzocchi (2010) showed that the Poisson model is not appropriate for a class of time-dependent forecast models (e.g., epidemic-type aftershock sequence models) that represent earthquake clustering, for instance, during aftershock sequences (Woessner *et al.*, 2011). The Poisson assumption may be a better approximation when the expected rates are low, but this issue has not been comprehensively investigated. For the sake of the illustrations in this article, we adopt the Poisson hypothesis, although our approach and the existing CSEP tests can be generalized to any situation in which the likelihood can be computed.

We note that the Poisson assumption has a greater impact on classical tests where a model might be rejected only because this assumption does not hold, while in the Bayesian view the same model might be penalized but would not be excluded.

In the Bayesian view of the CSEP testing phase, equation (1) becomes

$$P(\Lambda^j|\Omega) = \frac{P(\Lambda^j)\exp[L(\Omega|\Lambda^j)]}{P(\Omega)}, \tag{7}$$

where $P(\Lambda^j)$ is the prior probability of the model Λ^j, $P(\Omega)$ is the probability of observing the dataset Ω, and $P(\Lambda^j|\Omega)$ is the posterior probability of the model Λ^j. In the CSEP testing phase, the prior probability may be equal across all models for the first round of testing (if dealing with uncorrelated forecasts; a different strategy for correlated forecasts is discussed in Accounting for Forecast Correlation). In a second round of testing, the prior probability may be set to the posterior obtained from the first round. One can update the prior and posterior probabilities after any period of time, even at irregular intervals such as after every earthquake.

Because the probability $P(\Omega)$ is constant across all models, it is not necessary to know its exact value when comparing probabilities. Usually, it is very difficult to write down a closed form of this distribution (e.g., Gelman *et al.*, 1995); therefore, we choose a normalizing $P(\Omega)$ such that the sum of all posteriors for each testing round is unity, consistent with our best model interpretation of the posterior probabilities.

In the Bayesian approach for comparing models, the likelihood ratio of the classical frequentist approach is replaced by the posterior odds (equation 3) and by the Bayes factor (equation 4). The posterior odds are equal to the ratio of the probability of two competing models to be the best one in a set of models. In our specific case, the Bayes factor has the same value of the likelihood ratio, but its interpretation makes it easier to understand how much better one model is relative to others, and it overcomes intrinsic difficulties of interpreting the likelihood ratio in a classical statistical testing framework. Notably, in the classical view the significance level of the likelihood ratio can be obtained only for nested models, that is, where one model is a conceptual extension of another (see Goldstein, 2011, section 2.12). To overcome this limitation, Schorlemmer *et al.* (2007) proposed a method called the R-test to estimate the significance level of the likelihood ratio for two non-nested models, assuming that one of the two models is true. Such an assumption is unrealistic if no model can be true. Moreover, Rhoades *et al.* (2011) noted that it is unclear if the R-test compares the performances of two models or if it is instead a goodness-of-fit test.

Ensemble Forecasting

The CSEP experiments allow researchers to establish a ranking of models according to out-of-sample predictive performance (e.g., based on the likelihood of each model). Nonetheless, it is not clear how to use these results in OEF (Jordan *et al.*, 2011), where the best model has to be used (Marzocchi and Zechar, 2011). One possibility is to simply adopt the model that has performed best so far and disregard all others, but there is no guarantee that this model will be the best in the future (e.g., Oreskes *et al.*, 1994); in practice, we never know which of the candidate models will be the best in a long testing phase. We also note that the best candidate model may capture one important part of the earthquake generation process well, while others might suitably represent secondary, or at least more subtle, features. Model merging is a rational procedure to include these different aspects of the earthquake generation process in a single model (Vere-Jones, 1995). In general, model averaging is a proper way to account for uncertainty among candidate models; one familiar example is the logic tree approach to probabilistic seismic hazard assessment (e.g., Budnitz *et al.*, 1997). While this type of ensemble forecasting is fairly common in meteorology, climate studies, and hydrology, it has rarely been applied in earthquake forecasting (see, however, Rhoades and Gerstenberger, 2009; Marzocchi *et al.*, 2012).

The most natural procedure to combine different models is through a weighted average. The weight of each model should account for two main issues: (1) the correlation between the forecasts of the different models and (2) past forecast performance.

The assignment of model weights involves an unavoidable subjectivity. But we argue that an approach based on forecast performances of each model is inherently superior to simply assigning all models the same weight. There are at least three desirable features for such a weighting scheme (Garthwaite and Mubwandarikwa, 2010):

1. Models that are highly correlated with other models should be given smaller weights than those that have low correlation with other models (dilution property).
2. If a new model that is identical to an existing model is added to the set of candidate models, the weight of the duplicated model should be split and shared with the new model, and all other weights should remain as they were (strong dilution property).
3. When a new model is added to the set of candidate models, none of the weights of the others should increase (monotonicity property).

To our knowledge, no formal weighting scheme fully satisfies all these criteria. Here, we adopt a scheme that comes close to satisfying these points: the capped eigenvalue (CE) method of Garthwaite and Mubwandarikwa (2010).

Accounting for Forecast Correlation

We use the term "correlation weight" to highlight the fact that these weights are estimated before collecting any data and simply account for the correlation between model forecasts.

The correlation matrix of the model forecasts is defined as

$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1J} \\ c_{21} & c_{22} & \dots & c_{2J} \\ \dots & \dots & \dots & \dots \\ c_{J1} & c_{J2} & \dots & c_{JJ} \end{bmatrix}, \tag{8}$$

where J is the number of models and each element of the matrix is Pearson's correlation:

$$c_{lk} = \frac{\sum_{i=1}^{n}(\lambda_i^l - \bar{\lambda}^l)(\lambda_i^k - \bar{\lambda}^k)}{\sqrt{\sum_{i=1}^{n}(\lambda_i^l - \bar{\lambda}^l)^2 \sum_{i=1}^{n}(\lambda_i^k - \bar{\lambda}^k)^2}}, \tag{9}$$

where $\bar{\lambda}^l$ and $\bar{\lambda}^k$ are the average of the model forecasts Λ^l and Λ^k, respectively.

To account for forecast correlation, we employed the CE weighting scheme suggested by Garthwaite and Mubwandarikwa (2010). The scheme is based on the spectral decomposition of the correlation matrix,

$$\mathbf{C} = \mathbf{QAQ}^{\mathrm{T}}, \tag{10}$$

where $\mathbf{A}$ is the diagonal eigenvalue matrix, the columns of the matrix $\mathbf{Q}$ contain the normalized eigenvectors, and $\mathbf{Q}^{\mathrm{T}}$ is the transpose of $\mathbf{Q}$. Because the maximum value for each element of $\mathbf{C}$ is 1, each eigenvalue larger than 1 indicates that information is duplicated in some way. The CE method consists of transforming the matrix $\mathbf{A}$ to $\mathbf{A}^*$ by capping all eigenvalues at unity; if any eigenvalue in $\mathbf{A}$ is larger than 1, the corresponding element in $\mathbf{A}^*$ is set to 1, and all other elements remain the same. From $\mathbf{A}^*$, we calculate a new correlation matrix

$$\mathbf{C}^* = \mathbf{Q}\mathbf{A}^*\mathbf{Q}^{\mathrm{T}}. \tag{11}$$

The correlation weight δ^{corr} of each model is then set to

$$\delta_j^{\mathrm{corr}} = \frac{\mathrm{C}_{jj}^*}{\sum_{k=1}^{J} \mathrm{C}_{kk}^*}, \tag{12}$$

for $j = 1, \ldots, J$.

Garthwaite and Mubwandarikwa (2010) show that the CE method satisfies the dilution and monotonicity properties, but not the strong dilution property. This means that, if an identical model is added, the weights of all models may be affected. This is undesirable, but such changes are usually small, and as far as we know, no other method fully satisfies all three properties.

A Tutorial Example of Correlation-Corrected Weighting

Table 2 shows three forecasts produced by simple models. The forecast rates of Models 1 and 3 are randomly sampled from a Gaussian distribution with average 10 and standard deviation of 1; the forecast rates of Model 2 are randomly sampled from a Gaussian distribution with average 5 and standard deviation of 1. We imposed a strong correlation between Models 1 and 2 by rank ordering the forecasts in the same way. The correlation matrix is

$$\mathbf{C} = \begin{bmatrix} 1 & 0.95 & -0.54 \\ 0.95 & 1 & -0.33 \\ -0.54 & -0.33 & 1 \end{bmatrix}. \tag{13}$$

The high value of $\mathbf{C}_{12} = \mathbf{C}_{21}$ is explained by the imposed correlation between the two models. The eigenvalues of this correlation matrix are 2.25, 0.72, and 0.03. The new correlation matrix derived from the CEs (1, 0.72, 0.03) is

$$\mathbf{C}^* = \begin{bmatrix} 0.47 & 0.45 & -0.17 \\ 0.45 & 0.53 & 0.01 \\ -0.17 & 0.01 & 0.75 \end{bmatrix}, \tag{14}$$

and the final correlation weights for each model are 0.27 for Model 1, 0.30 for Model 2, and 0.43 for Model 3.

This example shows the basic features of the procedure. First, Models 1 and 2 have about three-quarters of the correlation weight of the independent Model 3. This is easily understood because Models 1 and 2 provide similar information, while Model 3 is different. Second, the correlation weights and the correlation matrix are insensitive to the number of earthquakes forecast by each model (Models 1 and 2 are highly correlated but have different averages). Third, if we consider only two models, this approach will always assign the same correlation weight (0.5) to each model.

Skill-Weighting of the Forecasts

After we have collected data from the testing phase of a forecast experiment, we update the weight of the jth model:

$$W_j = \frac{\delta_j^{\mathrm{corr}} S_j}{\sum_{k=1}^{J} [\delta_k^{\mathrm{corr}} S_k]}, \tag{15}$$

where S_j is a measure of the jth model's forecasting performance during the testing phase. One popular choice for S_j is the posterior probability of the jth model given by equation (7), which is called Bayesian Model Averaging (BMA; Hoeting *et al.*, 1999). In our CSEP example, this weighting is proportional to the cumulative likelihood $L_j = L(\Omega|\Lambda^j)$:

$$S_j^{\mathrm{BMA}} = \exp(L_j). \tag{16}$$

The log-likelihoods are negative numbers; if we order them by decreasing value, so that $|L_j| \leq |L_k|$ if $j < k$, then L_1 is the log-likelihood of the best model, and the BMA weights involve the Bayes factors relative to this best model:

$$W_j^{\mathrm{BMA}} \propto \delta_j^{\mathrm{corr}} B_{j1} = \delta_j^{\mathrm{corr}} \exp(\Delta L_j), \tag{17}$$

where $\Delta L_k = L_k - L_1$. Owing to the exponential form in (17), BMA is a strong weighting scheme, and the ensemble average can be dominated by the best model.

A second choice is a logarithmic scoring rule that sets S_j equal to the inverse of the cumulative log-likelihood magnitude:

$$S_j^{\mathrm{SMA}} = \frac{1}{|L_j|}. \tag{18}$$

Table 2
Synthetic Model Forecasts

	Forecast Rates									
Model 1	11.84	7.74	10.86	10.32	8.69	9.57	10.34	13.58	12.77	8.65
Model 2	6.42	4.80	6.41	5.71	4.94	5.67	5.73	8.03	6.49	4.88
Model 3	8.79	10.72	11.63	10.49	11.03	10.73	9.70	10.29	9.21	10.89

According to this scheme, which we call Score Model Averaging (SMA; Good, 1952), the weights can be written as

$$W_j^{\mathrm{SMA}} \propto \delta_j^{\mathrm{corr}} (1 + |\Delta L_j / L_1|)^{-1}. \tag{19}$$

When the ratio in (19) is small compared with unity, SMA implies weak weighting; that is, the ensemble average will weight all models almost equally. In general, SMA weighting can be interpreted as a scheme that accounts for both reliability, scored by L_1, and skill, scored by ΔL_j. If the reliability is low (L_1 is very negative) and the skill is not very large, the relative skill of the models is unimportant, and the SMA weighting will be weak. This is sensible: because all models are unreliable, none deserves a high weight. On the other hand, if the reliability is high ($|L_1| \le 1$) and ΔL_j is large ($|\Delta L_j| >> 1$), the weights will approach $|\Delta L_j|^{-1} = -1/\ln B_{j1}$, and the more skillful models will dominate.

A more general version of (18), denoted generalized SMA (gSMA), subtracts a constant value, $|L_0| < |L_1|$, from cumulative log-likelihoods:

$$S_j^{\mathrm{gSMA}} = \frac{1}{|L_j - L_0|} = \frac{1}{|L_1 - L_0| + |\Delta L_j|}, \tag{20}$$

The gSMA weighting can be interpreted as a scheme that can tune the role of reliability, scored by $L_1 - L_0$, with respect to skill, scored by ΔL_j. As L_0 tends to zero, the gSMA weighting approaches SMA; if L_0 tends to L_1, it approaches $|\Delta L_j|^{-1}$, and the importance of reliability becomes negligible. The reliance of this scheme on the logarithm of the Bayes factor is consistent with the evidence classes listed in the Bayesian forecast evaluation section, which are also logarithmic.

For each of these ensemble-building methods, the weight assigned to each model is a mixture of forecast correlation and past predictive skill. These weights can be updated as an experiment proceeds, and they represent a natural, rational way to assemble the best model according to the data collected so far (Vere-Jones, 1995).

Illustration of Bayesian Comparison and the Ensemble Model with RELM Forecasts

To illustrate our approach with actual forecasts and data, we considered the five-year Regional Earthquake Likelihood Models (RELM) experiment in California (Field, 2007; Schorlemmer *et al.*, 2010). (Ⓔ The codes and data used for this illustration are available in the electronic supplement to this paper.)

Of the dozen models aimed at forecasting the distribution of $M\,4.95+$ mainshocks during the RELM experiment (1 January 2006 to 31 December 2010, inclusive), only four provided a forecast that covered the entire California testing region (Schorlemmer and Gerstenberger, 2007). These were the models developed by Ebel *et al.* (2007), Helmstetter *et al.* (2007), Holliday *et al.* (2007), and Wiemer and Schorlemmer (2007). We also included two forecasts based on the TripleS model of Zechar and Jordan (2010) that were not part of the RELM experiment. The TripleS forecasts were generated in a manner that guaranteed they would be highly correlated, specifically to illustrate the CE decorrelation technique described in the Accounting for Forecast Correlation section (Ⓔ see code available in the supplement). For simplicity, we hereafter refer to all forecasts by the last name of the lead author, that is, Ebel, Helmstetter, Holliday, Wiemer, and Zechar.1 and Zechar.2 for the two TripleS forecasts. Each forecast specified the number of expected target earthquakes in bins of latitude–longitude–magnitude, and one could mask the forecast in each bin, indicating that the bin should be ignored in the final evaluation. Although the Holliday forecast masked more than 90% of the bins (Schorlemmer *et al.*, 2010), for the purposes of illustration we disregard that masking and consider all bins. We emphasize that the results we show here are therefore not indicative of the true performance of the Holliday model (or any other model) in the RELM experiment.

During the 5 years of the RELM experiment, 20 mainshock target earthquakes occurred. Zechar *et al.* (2012) considered the outcome of this experiment in terms of the likelihood-based metrics that conformed to the RELM standard tests and concluded that Helmstetter was superior: it passed all the tests and fared best in the information gain of pairwise comparisons suggested by Harte and Vere-Jones (2005) and Rhoades *et al.* (2011). In Table 3 we report the Bayes factors for each pairwise comparison of the candidate models; those values marked in bold indicate very strong evidence against the column model (Model 1). This simple

Table 3
Bayes Factors

Model 2	Model 1					
–	Ebel	Helmstetter	Holliday	Wiemer	Zechar.1	Zechar.2
Ebel	–	6×10^{-46}	2×10^{-41}	9×10^{-39}	8×10^{-41}	4×10^{-40}
Helmstetter	**2×10^{45}***	–	**4×10^{4}**	**2×10^{7}**	**10^{5}**	**7×10^{5}**
Holliday	**4×10^{40}**	2×10^{-5}	–	**400**	3	20
Wiemer	**10^{38}**	7×10^{-8}	3×10^{-3}	–	9×10^{-3}	0.05
Zechar.1	**10^{40}**	8×10^{-6}	0.3	100	–	6
Zechar.2	**2×10^{39}**	10^{-6}	0.06	20	0.2	–

*Values marked in bold indicate very strong evidence against Model 1.

Bayes-factor comparison indicates that Helmstetter is superior to the other models.

To compare the effect of having one model that is superior to all others with a more balanced population of the ensemble, we consider all following analyses twice: once with all forecasts and once without Helmstetter.

We computed the correlation weights (the initial priors) following the method described in the Accounting for Forecast Correlation section. In Tables 4 and 5 we report the correlation matrix **C** and the transformed correlation matrix **C***, respectively, for the six forecasts. For Ebel, Helmstetter, Holliday, Wiemer, Zechar.1, and Zechar.2, the resulting correlation weights are $\delta_i^{\text{corr}} = 18.6\%$, 17.8%, 18.9%, 20.4%, 11.8%, and 12.3%, demonstrating that the highly correlated Zechar.1 and Zechar.2 have been downweighted. In Table 6, we present **C*** without Helmstetter (the corresponding **C** is just Table 4 without the Helmstetter entries). For the reduced set Ebel, Holliday, Wiemer, Zechar.1, and Zechar.2, the resulting correlation weights are $\delta_i^{\text{corr}} = 21.2\%$, 21.7%, 29.3%, 13.7%, and 14.1%.

Table 4
Forecast Correlation Matrix C

	Ebel	Helmstetter	Holliday	Wiemer	Zechar.1	Zechar.2
Ebel	1.00	0.34	0.43	0.25	0.58	0.57
Helmstetter	–	1.00	0.34	0.68	0.46	0.43
Holliday	–	–	1.00	0.20	0.57	0.56
Wiemer	–	–	–	1.00	0.30	0.28
Zechar.1	–	–	–	–	1.00	0.99
Zechar.2	–	–	–	–	–	1.00

Table 5
Transformed Forecast Correlation Matrix C

	Ebel	Helmstetter	Holliday	Wiemer	Zechar.1	Zechar.2
Ebel	**0.64**	0.01	0.07	0.00	0.11	0.11
Helmstetter	–	**0.61**	0.02	0.34	0.04	0.02
Holliday	–	–	**0.65**	−0.04	0.11	0.11
Wiemer	–	–	–	**0.70**	−0.02	−0.03
Zechar.1	–	–	–	–	**0.41**	0.40
Zechar.2	–	–	–	–	–	**0.42**

Values marked in bold indicate non-normalized correlation weights.

Table 6
Transformed Forecast Correlation Matrix C

	Ebel	Holliday	Wiemer	Zechar.1	Zechar.2
Ebel	**0.63**	0.07	0.04	0.11	0.11
Holliday	–	**0.65**	−0.01	0.12	0.11
Wiemer	–	–	**0.87**	0.03	0.01
Zechar.1	–	–	–	**0.41**	0.40
Zechar.2	–	–	–	–	**0.42**

Values marked in bold indicate non-normalized correlation weights.

With 20 observed target earthquakes, there are 21 testing phases: each target earthquake signals the end of the preceding testing phase and the beginning of a new phase; the end of the experiment signals the end of the final testing phase, which by definition has no target earthquakes. For each testing phase, we computed for each forecast the posterior probability, which is normalized such that the sum of the posteriors is unity. Because the considered forecasts are time-invariant, we treated each testing phase as a forecast experiment with one observation: one target earthquake for each testing phase save the final one, which has no target earthquakes. We scaled the forecast rates to the duration of the testing phase (the time since the previous target earthquake, or, for the first testing phase, the time since the beginning of the experiment) and computed the likelihood according to equation (6). We report likelihoods for each forecast for each testing phase in Ⓔ Tables S1 and S2 (available in the supplement).

The normalized posteriors can be interpreted as the probability of each model to be the best among the models considered, and we show these posteriors as the experiment progresses in Figure 1. The posterior probabilities shown in Figure 1a support the indication from the Bayes factor comparison that Helmstetter is superior: after only a few testing phases, Helmstetter's posterior approaches unity and remains at this level for the rest of the experiment. In Figure 1b, without Helmstetter, the relative ranking of forecasts changes throughout the experiment. In the early stages, Wiemer has the highest posterior probability, but Zechar.1 has the highest posterior throughout most of the middle three years, until Holliday prevails near the end of the experiment. In other words, there is no obviously best model.

We considered the same set of forecasts and observations in constructing ensemble forecasts following the BMA, SMA, and gSMA procedures described in the Skill-Weighting of the Forecasts section. For gSMA, we arbitrarily set $L_1 - L_0 = 1$, in order to illustrate the weighting implied if we reduce the importance of the models' reliability. We anticipate that more objective strategies to select L_0 that maximize the past forecasting performances may be established; a full investigation of the role of L_0 in terms of forecasting capabilities should be conducted in future analyses.

In the first testing phase, the correlation weights are used in a weighted average of all forecasts to construct an ensemble model (i.e., $S_j = 1$ for all models), and therefore there is no difference between BMA, SMA, and gSMA. In the second testing phase, the weights take into account the results from the first testing phase; in the third testing phase, the weights depend on the first two testing phases, and so on until the end of the experiment.

In Figure 2, we show the composition of the BMA, gSMA, and SMA ensemble models for each testing phase. After only a few testing phases, the BMA forecast is almost identical to Helmstetter, because the posterior for Helmstetter is nearly unity (Fig. 2a). The logarithmic scaling of the gSMA forecast favors Helmstetter, but not nearly as strongly as the linear scaling of BMA (Fig. 2b). The SMA

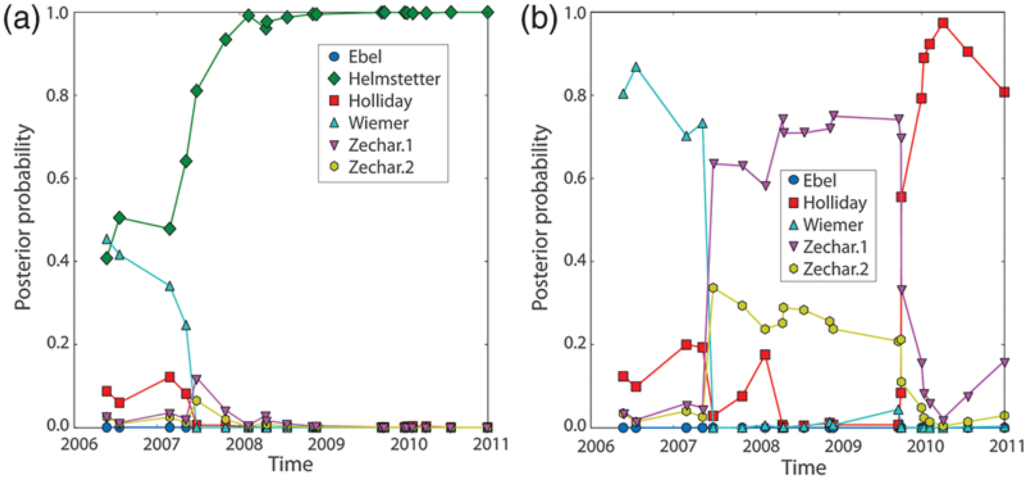

Figure 1. (a) The posterior probability for each forecast throughout the experiment; posteriors were updated whenever a new target earthquake occurred. After mid-2008, posteriors for Ebel, Holliday, Wiemer, and Zechar.2 are not visible because they fall directly below Zechar.1. (b) Same as (a), but without Helmstetter.

forecast is a nearly even mixture of all models (Fig. 2c), as expected from equation (19). We show the results of the same experiment without Helmstetter in Figure 2d–f. The weights shown in Figure 2d directly reflect the evolution of posteriors from Figure 1b.

In the standard approach currently applied in CSEP testing, one would choose the best model, the one that obtained the highest likelihood to date, at the end of the testing phase for OEF or some other application. We can compare the performance of this best model with the BMA, SMA, and gSMA forecasts. Figure 3 shows the log-likelihood (open symbols) for each testing phase and the cumulative log-likelihood (filled symbols). We begin with the second testing period, because it is the first time that the BMA, SMA, and gSMA forecasts are based on observations.

In Figure 3a,b we see that the ensemble forecasts perform about the same as, or slightly better than, the single best model. This is unsurprising for the BMA forecast, because it is nearly identical to Helmstetter (the best model throughout the experiment). The cumulative log-likelihood of Helmstetter alone

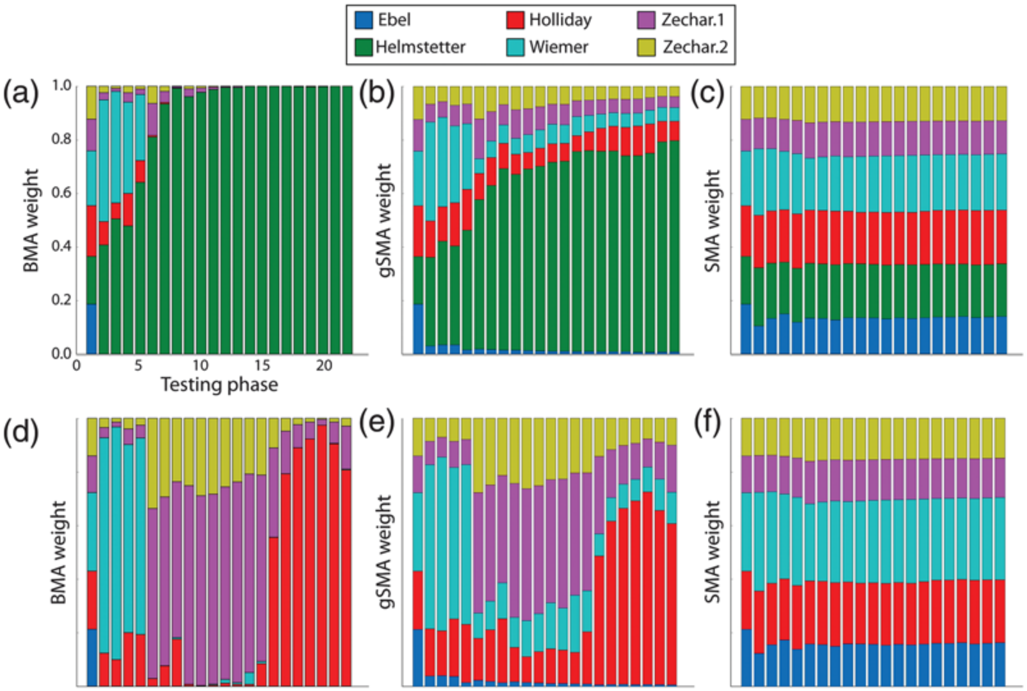

Figure 2. (a) Contribution of each forecast to the BMA forecast during each testing phase. After the first testing phase, Ebel's contribution cannot be seen because it is so close to zero; after the fifth testing phase, the ensemble is essentially identical to Helmstetter. The final column is the composition of the BMA forecast for future testing phases. (b) Same as (a) but for gSMA forecast with $L_0/L_1 = 1$. (c) Same as (a) but for SMA forecast. (d–f) Same as (a–c), but without Helmstetter.

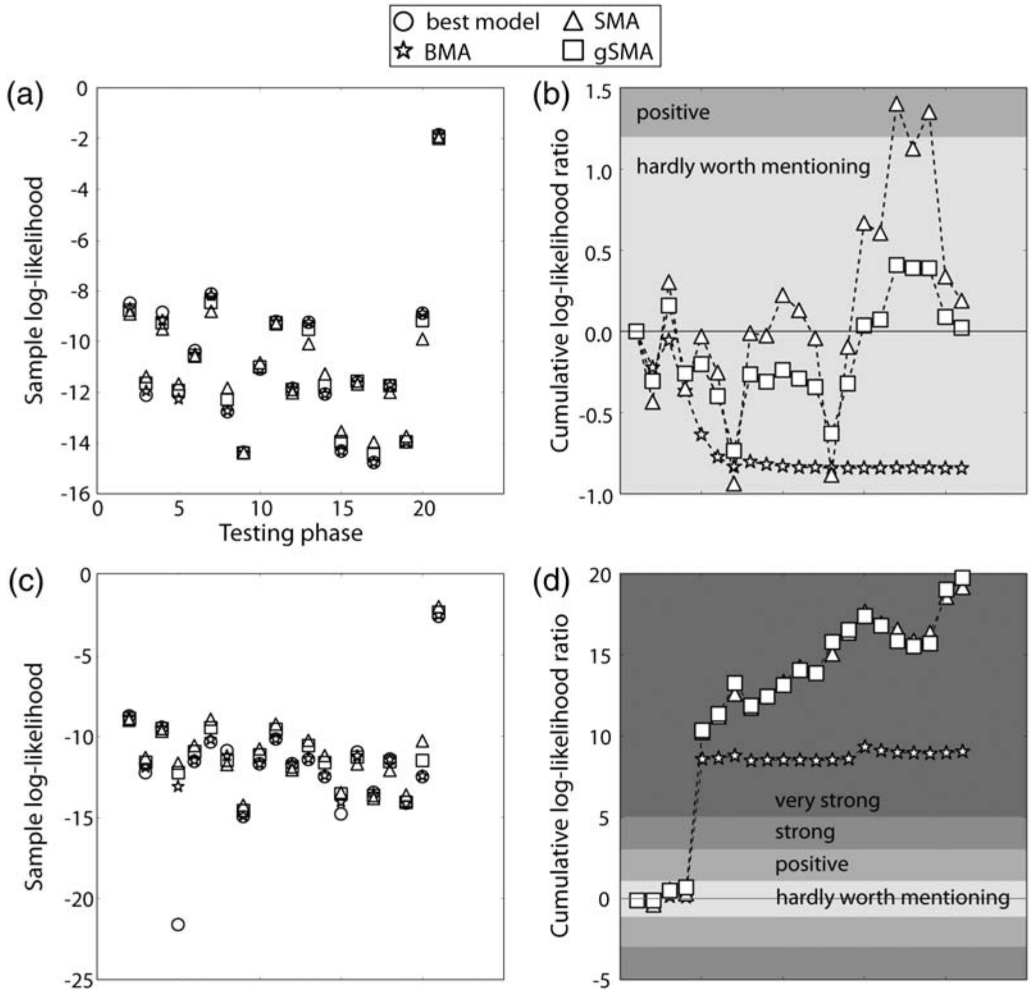

Figure 3. (a) Log-likelihood at each testing phase (after the first) for the forecast that performed best through the previous round (circles), the BMA forecast (stars), the SMA forecast (triangles), and the gSMA forecast (squares). (b) Cumulative log-likelihood ratio of each ensemble forecast to the best model. Symbols falling below the horizontal line indicate that the best model is better than the ensemble, and those above the line favor the ensemble. Shaded regions indicate the strength of the difference between the models following the Bayes factor classifications mentioned in the text. (c–d) Same as (a–b), but without Helmstetter.

(−217.8) is slightly better than the BMA forecast (−218.6), though nearly identical to that of the SMA forecast (−217.8) and not quite as good as the gSMA forecast (−217.6). (We note that these cumulative log-likelihoods do not include the results from the first round of testing because, by definition, we cannot identify one best model before any earthquakes have occurred.) That BMA is not better than Helmstetter is a simple demonstration that ensemble models are not always superior to the best candidate model evaluated *a posteriori*, especially when the performance of one model overwhelms the performances of the others. For example, in the extreme case that the data-generating model is included in the set of candidate models, the BMA ensemble would be the average of the data-generating model and noise.

On the other hand, Figure 3c,d (without Helmstetter) demonstrates that using an ensemble forecast can indeed be better than using the best model from previous rounds of testing. With this reduced set of forecasts, BMA gives a cumulative log-likelihood of −228.5, whereas using the best model from preceding rounds gives −237.6; gSMA (−222.3) and SMA (−218.4) are considerably better. In this example, the SMA and gSMA ensemble forecasts are superior to the BMA ensemble forecast and the best model. This result is particularly noteworthy because BMA is widely used.

Beyond their intrinsic appeal of being based on out-of-sample predictive performance, we conclude that BMA, SMA, and gSMA are all superior to a standard average of the models. Consider a simple example: if we add 10 models that perform terribly (log-likelihood close to minus infinity), BMA, SMA, and gSMA would assign them zero weight, while a simple average would consider these models just as reliable as much better models.

Final Remarks

The CSEP experiment is a unique attempt to quantitatively evaluate the performance of earthquake forecast models. Here, we provide new tools to measure performance and construct ensemble models that may eventually be useful for operational purposes. In particular, we suggest a Bayesian approach to the testing phase of forecast experiments, which has some advantageous features compared with the currently implemented approach (Schorlemmer *et al.*, 2007). The fundamental difference is that no models are rejected, but each one is instead assigned a relative performance probability based on the observations. This means that models that currently produce forecasts that appear by chance to be wrong will not be abandoned for future experiments.

Earthquake forecast evaluation should involve checks of absolute consistency (e.g., checking if the number of earthquakes predicted by the model is consistent with the observation) and relative comparison. The Bayesian approach emphasizes the latter and does not directly evaluate goodness of fit. For example, our approach would not give an obvious indication of a problem if all considered models were terrible. Therefore, we note that the approaches suggested here are not meant to replace, but rather to complement, the existing CSEP consistency tests (Zechar, Gerstenberger, and Rhoades, 2010). Moreover, this strategy facilitates the continuous updating of a model's posterior probability depending on new observations. In practice, we can retrieve an updated assessment about model performance according to results of the previous rounds of tests and the data observed in the current testing phase.

In addition to differences in the testing phase, and perhaps more importantly, we describe some straightforward and quantitatively justified methods to generate ensemble earthquake model forecasts. We put forward three different weighting schemes to generate an ensemble model: (1) the first model is based on the widely used BMA, (2) the second model is based on the inverse of the cumulative likelihood score (SMA), and (3) the third model allows a tuning between reliability and skill (gSMA). The results show that the forecasting performances of the ensemble models built using SMA and gSMA are superior to the performances of the best model. On the contrary, the widely used BMA performs well when one model is superior, but the BMA ensembles do not perform better than those from SMA or gSMA.

We note that the procedures outlined in this article could be applied separately to each forecast dimension (space, magnitude, overall rate, and other dimensions), so one might construct an ensemble forecast that emphasizes the best elements of each forecast. Such multidimensional ensembles might provide better forecasts than any single ensemble combination, an important consideration for operational forecasting. Moreover, this approach can be generalized and modified to incorporate other elements of Bayesian data analysis and forecast optimization. For instance, instead of emphasizing the posterior probabilities or the log-likelihood of each participating model, one might generate a multitude of weighted ensemble forecasts and compute the posterior of each ensemble (e.g., Monteith *et al.*, 2011). Echoing the view of Vere-Jones (1995), we believe that future research efforts should emphasize effective model combination more than model selection. This is likely to be of paramount importance for OEF and other practical applications.

Data and Resources

In the electronic supplement to this article, we provide all codes and data that were used in this study.

Acknowledgments

JDZ was supported in part by NSF Grant EAR-0944202 and USGS Grant G11AP20038. This work was partially funded by the EU FP7 REAKT project.

References

Box, G. E. P., and N. R. Draper (1987). *Empirical Model-Building and Response Surface*, John Wiley & Sons, Inc., New York, New York.

Budnitz, R. J., G. Apostolakis, D. M. Boore, L. S. Cluff, K. J. Coppersmith, C. A. Cornell, and P. A. Morris (1997). Senior Seismic Hazard Analysis Committee; Recommendations for Probabilistic Seismic Hazard Analysis: Guidance on Uncertainty and Use of Experts, U.S. Nuclear Regulatory Commission, U.S. Dept. of Energy, Electric Power Research Institute, NUREG/CR-6372, UCRL-ID-122160, Vol. 1–2.

Ebel, J. E., D. W. Chambers, A. L. Kafka, and J. A. Baglivo (2007). Non-Poissonian earthquake clustering and the hidden Markov model as bases for earthquake forecasting in California, *Seismol. Res. Lett.* **78,** no. 1, 57–65.

Evett, I. W. (1991). Implementing Bayesian methods in forensic science, Paper Presented at *The Fourth Valencia International Meeting on Bayesian Statistics*, Valencia, Spain, 15–20 April 1991.

Field, E. H. (2007). Overview of the working group for the development of Regional Earthquake Likelihood Models (RELM), *Seismol. Res. Lett.* **78,** no. 1, 7–16.

Garthwaite, P. H., and E. Mubwandarikwa (2010). Selection of weights for weighted model averaging, *Aust. New Zeal. J. Stat.* **52,** no. 4, 363–382.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1995). *Bayesian Data Analysis*, CRC Press, Boca Raton, Florida, 526 pp.

Gerstenberger, M. C., and D. A. Rhoades (2010). New Zealand earthquake forecast testing centre, *Pure Appl. Geophys.* **167,** 877–892.

Goldstein, H. (2011). *Multilevel Statistical Models*, John Wiley & Sons. Ltd., West Sussex, United Kingdom.

Good, I. J. (1952). Rational decisions, *J. Roy. Stat. Soc.* **14,** 107–114.

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

Helmstetter, A., Y. Y. Kagan, and D. D. Jackson (2007). High-resolution time-independent grid-based forecast for $M \geq 5$ earthquakes in California, *Seismol. Res. Lett.* **78,** no. 1, 78–86.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). Bayesian model averaging: A tutorial, *Stat. Sci.* **14,** no. 4, 382–417.

Holliday, J. R., C. C. Chen, K. F. Tiampo, J. B. Rundle, D. L. Turcotte, and A. Donnellan (2007). A RELM earthquake forecast based on pattern informatics, *Seismol. Res. Lett.* **78,** no. 1, 87–93.

Jeffreys, H. (1961). *Theory of Probability*, Third Ed., Oxford University Press, Oxford, United Kingdom, 459 pp.

Jordan, T. H. (2006). Earthquake predictability, brick by brick, *Seismol. Res. Lett.* **77,** no. 1, 3–6.

Jordan, T. H., and L. M. Jones (2010). Operational earthquake forecasting: Some thoughts on why and how, *Seismol. Res. Lett.* **81,** 571–574, doi: 10.1785/gssrl.81.4.571.

Jordan, T. H., Y.-T. Chen, P. Gasparini, R. Madariaga, I. Main, W. Marzocchi, G. Papadopoulos, G. Sobolev, K. Yamaoka, and J. Zschau (2009). *Operational Earthquake Forecasting: State of Knowledge and Guidelines for Utilization*, Report for the Council of Ministers, Italy.

Jordan, T. H., Y.-T. Chen, P. Gasparini, R. Madariaga, I. Main, W. Marzocchi, G. Papadopoulos, G. Sobolev, K. Yamaoka, and J. Zschau (2011). Operational earthquake forecasting: State of knowledge and guidelines for implementation, *Ann. Geophys.* **54,** 315–391, doi: 10.4401/ag-5350.

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

Lombardi, A. M., and W. Marzocchi (2010). The assumption of Poisson seismic-rate variability in CSEP/RELM experiments, *Bull. Seismol. Soc. Am.* **100,** no. 5A, 2293–2300.

Marzocchi, W., and J. D. Zechar (2011). Earthquake forecasting and earthquake prediction: Different approaches for obtaining the best model, *Seismol. Res. Lett.* **82,** no. 3, 442–448.

Marzocchi, W., A. Amato, A. Akinci, C. Chiarabba, A. M. Lombardi, D. Pantosti, and E. Boschi (2012). A ten-year earthquake occurrence model for Italy, *Bull. Seismol. Soc. Am.* **102,** no. 3, 1195–1213.

Marzocchi, W., D. Schorlemmer, and S. Wiemer (2010). Preface to the special volume "An earthquake forecast experiment in Italy", *Ann. Geophys.* **53,** no. 3, doi: 10.4401/ag-4851.

Monteith, K., J. L. Carroll, K. Seppi, and T. Martinez (2011). Turning Bayesian model averaging into Bayesian model combination, in *Proc. of International Joint Conference on Neural Networks*, San Jose, California, 31 July–5 August 2011, 2657–2663.

Nanjo, K. Z., H. Tsuruoka, N. Hirata, and T. H. Jordan (2011). Overview of the first earthquake forecast testing experiment in Japan, *Earth Planets Space* **63,** 159–169, doi: 10.5047/eps.2010.10.003.

Oreskes, N., K. Shraderfrechette, and K. Belitz (1994). Verification, validation, and confirmation of numerical-models in the earth-sciences, *Science* **263,** no. 5147, 641–646.

Rhoades, D. A., and M. C. Gerstenberger (2009). Mixture models for improved short-term earthquake forecasting, *Bull. Seismol. Soc. Am.* **99,** no. 2A, 636–646.

Rhoades, D. A., D. Schorlemmer, M. C. Gerstenberger, A. Christophersen, J. D. Zechar, and M. Imoto (2011). Efficient testing of earthquake forecasting models, *Acta Geophys.* **59,** no. 4, 728–747.

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test, *Psychol. Bull.* **57,** no. 5, 416–428.

Schorlemmer, D., and M. C. Gerstenberger (2007). RELM testing center, *Seismol. Res. Lett.* **78,** no. 1, 30–36.

Schorlemmer, D., M. C. Gerstenberger, S. Wiemer, D. D. Jackson, and D. A. Rhoades (2007). Earthquake likelihood model testing, *Seismol. Res. Lett.* **78,** no. 1, 17–29.

Schorlemmer, D., J. Zechar, M. Werner, E. Field, D. Jackson, and T. Jordan (2010). First results of the regional earthquake likelihood models experiment, *Pure Appl. Geophys.* **167,** no. 8–9, 859–876.

van Stiphout, T., S. Wiemer, and W. Marzocchi (2010). Are short-term evacuations warranted? Case of the 2009 L'Aquila earthquake, *Geophys. Res. Lett.* **37,** L06306, doi: 10.1029/2009GL042352.

Vere-Jones, D. (1995). Forecasting earthquakes and earthquake risk, *Int. J. Forecast.* **11,** no. 4, 503–538.

Werner, M. J., and D. Sornette (2008). Magnitude uncertainties impact seismic rate estimates, forecasts, and predictability experiments, *J. Geophys. Res. Solid Earth* **113,** no. B8, B08302.

Wiemer, S., and D. Schorlemmer (2007). Alm: An asperity-based likelihood model for California, *Seismol. Res. Lett.* **78,** no. 1, 134–140.

Woessner, J., S. Hainzl, W. Marzocchi, M. J. Werner, A. M. Lombardi, F. Catalli, B. Enescu, M. Cocco, M. C. Gerstenberger, and S. Wiemer (2011). A retrospective comparative forecast test on the 1992 Landers sequence, *J. Geophys. Res. Solid Earth* **116**, B05305.

Zechar, J. D., and T. H. Jordan (2010). Simple smoothed seismicity earthquake forecasts for Italy, *Ann. Geophys.* **53,** no. 3, 99–105.

Zechar, J. D., M. C. Gerstenberger, and D. A. Rhoades (2010). Likelihood-based tests for evaluating space-rate-magnitude earthquake forecasts, *Bull. Seismol. Soc. Am.* **100,** no. 3, 1184–1195.

Zechar, J. D., D. Schorlemmer, M. Liukis, J. Yu, F. Euchner, P. Maechling, and T. Jordan (2010). The collaboratory for the study of earthquake predictability perspective on computational earthquake science, *Concurrency Comput. Pract. Ex.* **22,** no. 12, 1836–1847.

Zechar, J. D., D. Schorlemmer, M. J. Werner, M. C. Gerstenberger, D. A. Rhoades, and T. H. Jordan (2012). Regional earthquake likelihood models I: First-order results, *Bull. Seismol. Soc. Am.* (in press).

Istituto Nazionale di Geofisica e Vulcanologia
Via di Vigna Murata 605
00143 Rome, Italy
warner.marzocchi@ingv.it
(W.M.)

Swiss Seismological Service
ETH Zurich
Sonneggstrasse 5
8092 Zurich, Switzerland
jeremy.zechar@sed.ethz.ch
(J.D.Z.)

Department of Earth Sciences
University of Southern California
3651 Trousdale Parkway
Los Angeles, California 90089
tjordan@usc.edu
(T.H.J.)

Manuscript received 21 November 2011