Earthquake forecast enrichment scores

Christine Smyth, Masumi Yamada, Jim Mori

Disaster Prevention Research Institute, Kyoto University, Japan

Abstract

The Collaboratory for the Study of Earthquake Predictability (CSEP) is a global project aimed at testing earthquake forecast models in a fair environment. Various metrics are currently used to evaluate the submitted forecasts. However, the CSEP still lacks easily understandable metrics with which to rank the universal performance of the forecast models. In this research, we modify a well-known and respected metric from another statistical field, bioinformatics, to make it suitable for evaluating earthquake forecasts, such as those submitted to the CSEP initiative. The metric, originally called a *gene-set enrichment score*, is based on a Kolmogorov-Smirnov statistic. Our modified metric assesses if, over a certain time period, the forecast values at locations where earthquakes have occurred are significantly increased compared to the values for all locations where earthquakes did not occur. Permutation testing allows for a significance value to be placed upon the score. Unlike the metrics currently employed by the CSEP, the score places no assumption on the distribution of earthquake occurrence nor requires an arbitrary reference forecast. In this research, we apply the modified metric to simulated data and real forecast data to show it is a powerful and robust technique, capable of ranking competing earthquake forecasts.

Introduction

The Collabotory for the Study of Earthquake Predictability (CSEP) is an initiative to test earthquake forecast models in a fair environment.[1] The CSEP is a rapidly expanding and dynamic experiment. The numbers of forecast models and forecast regions have increased remarkably since the start of the experiment (a full list of forecast regions is available at http://www.cseptesting.org/home). The initial forecast evaluation metrics have been revised and supplemented with new metrics. The revised evaluation metrics, in combination with the more recent metrics, allow us to gain a deeper understanding of the forecast models' capabilities.

The metrics originally employed by the CSEP to test the abilities of the submitted models include the likelihood based L-, N-, and R- tests.[2] These tests have been supplemented recently with the M- and S- tests.[3] The N-, M- and S- tests investigate the consistency of the rate, magnitude and spatial elements of the forecast with the observations. The L-test gives a broad evaluation of the forecast by combining the rate, magnitude and space elements of the forecast.[3] Finally, the R-test compares the performances of two forecasts to each other.[2] Although these tests are an essential first step towards verifying a forecast's consistency with the observations and can potentially be used to rank forecasts, they require an assumption of a Poisson (or other) distribution[2] and they do not show if a forecast is *good*.

The likelihood metrics have been complemented with Receiver Operator Characteristic (ROC) curves, Molchan diagrams[4] and the Area Skill Score (ASS).[5,6] The ROC curves are a graphical technique that compare to a random prediction. However, the random prediction baseline is not a realistic reference model and, therefore, the ROC is not advocated as a powerful technique to evaluate earthquake forecasts.[5] The Molchan diagram is closely related to the ROC curves; however, the Molchan diagram can incorporate a non-random reference model. The ASS summarizes the Molchan diagram by considering the area above the Molchan trajectory. Although these techniques are generally interpretable, they all require the specification of a reference model.

More recently, Rhoades *et al.*[7] have proposed two fast and easily interpretable alternatives to the original R-test. The first is based on the classical Student's t-test, and the second is based on the non-parametric alternative to the Student's t-test. With these proposed metrics, earthquake forecasts are compared to each other in a round robin fashion. Obviously, they cannot be used to investigate the performance of a single model; a reference forecast of some description is also required.

Clements *et al.*[8] demonstrate the applicability of modern space-time point process evaluation techniques on CSEP forecasts. The authors show how deviance residuals can be used to compare two forecasts on a bin-by-bin basis; and how Pearson residuals can highlight the differences between expected and observed values in each bin. The authors also describe how the weighted L-test can evaluate if the clustering (or lack of) within the forecast is observed within the observations. However, both the Pearson residuals and weighted L-test cannot provide information about the model in areas without earthquakes, so the authors also introduce the concept of super-thinning, which, in combination with the weighted-L test, can be used to highlight any area where the model is fitting badly.

All of these currently used or proposed evaluation metrics play an important role towards evaluating the performances of forecast models. However, there are some limitations of these models, namely the specification of an arbitrary probability distribution for each bin's observation (currently the Poisson distribution is used) and the necessity of a reference model. Given the rapid increase in the number of submitted forecast models and regions, new evaluation metrics that both clearly indicate *good* models and overcome some of the limitations of the current metrics are required. Here we describe an easily interpretable and computationally efficient metric that can be used to evaluate and rank forecasts in the spatial domain without a reference distribution and with a minimum number of assumptions. The competing models must forecast and be evaluated on the same area over the same time period. To create a metric that can be used to rank the submitted forecasts in the spatial domain with a minimal number of assumptions, we propose a well-known and respected metric commonly used in Bioinformatics. The metric, originally called a *gene set enrichment score* (GSES)[9] is described in the Enrichment Score Metric section. After describing the algorithmic details of the metric we illustrate the method on both real and simulated forecasts.

Our future goal is to submit this metric to

Correspondence: Christine Smyth, Division of Earthquake Hazards, Disaster Prevention Research Institute, Kyoto University, Gokasho, Uji, Japan 611-0011.
Tel: +81.774.38.4208 - Fax: +81.774.38.4215.
E-mail: christine@eqh.dpri.kyoto-u.ac.jp

Key words: forecast, evaluation, metric.

Acknowledgments: the authors would like to thank Associate Professor Ken'ichiro Yamashina of Tokyo University who provided data for preliminary testing. The author would also like to thank Assistant Professor Timothy Hancock of Kyoto University for pointing out the similarities between genes and earthquakes.

Conflicts of interests: this study was designed and undertaken by the authors listed. Final responsibility for the content of the manuscript lies with the first and corresponding author. There are no potential conflicts of interests.

Received for publication: 24 August 2011.
Revision received: 10 December 2011.
Accepted for publication: 21 December 2011.

Research in Geophysics 2012; 2:e2
doi:10.4081/rg.2012.e2

the various global CSEP initiatives, so that it can be used to evaluate those submitted forecast models. Therefore, the main aim of this manuscript is to explain the proposed metric in terms of its usage with CSEP-like forecasts. We direct the reader to the references for detailed descriptions of CSEP forecasts.[2,10,11] A CSEP forecast must specify the expected number of earthquakes for a set of latitude-longitude-magnitude bins. A forecast for a single latitude-longitude bin comprises the numbers of earthquakes that are expected to occur over a specified time period for a specified set of magnitude ranges within that bin. Here, for simplicity, we do not consider the forecast for each magnitude range within each geographical bin. We simply sum over all magnitude ranges for the bin and use this single number as the bin's prediction. (See Discussion and Conclusions).

The enrichment score metric

The original GSES assesses the degree to which a predefined set of genes cluster toward the top (or bottom) of a much larger list of genes that have been ranked in some fashion.[9] For example, the genes may be ranked according to their average difference in expression for two conditions (such as cancer and non-cancer) evaluated over a set of samples. Then, if the set of genes occurs towards the top (or bottom) of the list, it can be assumed that this gene set is somehow related to the class distinction between the two conditions.[9]

To modify the GSES to make it suitable for evaluating forecasts, we rank the forecast values over the experiment space. We rank the forecast values in decreasing order. We use the set of bins where earthquakes occur as our *gene set*. Then, by using the same technique as described in the original GSES article, we can assess if the set of bins where earthquakes occurred are toward the top (or bottom) of our ranked list.[9]

To calculate the degree to which our earthquake set occurs at the top of the ranked forecast list, we calculate a running-sum statistic. The running-sum statistic starts at zero. At each bin in the list, we increase the statistic if an earthquake has occurred in that bin, otherwise we decrease it. Our final evaluation of the forecast is given by the maximum deviation from zero of this statistic. We give the exact algorithmic details below (Algorithm 1).

Algorithm 1. The enrichment score algorithm.

Step 1. Rank the forecast values of the *N* forecast bins in decreasing order to form a list, *L*. If two or more forecast values are identical, randomly rank the identical values.

Step 2. Set the value of the variables P_{hit} and P_{miss} to zero. Move through the ranked list, *L*, and calculate the value of P_{hit} and P_{miss} at each bin in the list. At bin *i*:

$$P_{hit}(S,i) = \sum_{\substack{j \in S \\ j \le i}} \frac{f_j^{\,p}}{N_R};\; P_{miss}(S,i) = \sum_{\substack{j \notin S \\ j \le i}} \frac{1}{(N - N_H)} \quad (0.1)$$

where S is the set of indices indicating the positions of bins in the ranked list where earthquakes are observed; f_j is the forecast value of the j^{th} bin in the list; $N_R = \sum_{j \in S} f_j^{\,p}$;

and N_H is the number of bins where earthquakes are observed.

Step 3. Calculate the enrichment score as the maximum deviation from zero of the running sum statistic given by $P_{hit} - P_{miss}$.

We refer to our metric as an earthquake forecast enrichment score (EFES). Our EFES will assess if, over a certain time period, the forecast values at locations where earthquakes have occurred are significantly toward the top of the list (*enriched*) compared to the forecast values of bins where earthquakes did not occur. For a set of observed earthquakes that are randomly distributed throughout the *L*, the EFES will be small[9] and close to zero. The EFES is limited to lie between -1 and 1. An EFES of 1 shows that the earthquakes all occurred in bins with the highest forecast values; an EFES of -1 shows that the earthquakes all occurred in bins with the lowest forecast values.

If we set p=0, the EFES is identical to the standard Kolmogorov-Smirnov (K-S) statistic and we are testing if the distribution of forecast values for bins where earthquakes occur is equal to the distribution of forecast values for bins where earthquakes do not occur.[9] In the original GSES paper, if p=0, high GSES are obtained for sets found in the middle of the list.[9] Similarly, if we set p=0, we can potentially obtain a high EFES if our earthquakes occur close together at some point in the list, because we would be weighting each bin's forecast value equally. So as to avoid obtaining a high EFES for a set of earthquakes found in the middle of the list, which would reflect a poor forecast, we employ the weighting factor p=1, recommended by Subramanian *et al.*[9]

The EFES statistic is similar to the recent work of Rhoades *et al.*[7] who also illustrate a K-S like statistic to test if the distribution of forecast values where earthquakes are observed is equal to the distribution of forecast values over all bins. However, here we set p=1 to consider the forecast values of bins where earthquakes occur and our metric is then based on an extension of the K-S statistic (which would test if the distribution of forecast values where earthquakes are observed is equal to the distribution of forecast values where earthquakes are not observed).

Permutation testing allows for a significance value to be placed upon the score.[9] We permute the list of bins where earthquakes are observed and recalculate the enrichment score. Repeating this process we obtain a distribution of enrichment scores with which to compare our score. Therefore, the metric can be used to assess the validity of a single forecast, even if there is no available reference model. We would like to point out here that we are testing how well the forecast would perform against randomly distributed seismicity. It is well known that seismicity is not distributed randomly and another method of permutation may be more appropriate. For the examples used here, we use this simple approach.

The reader can see that the EFES metric considers only the number of earthquakes predicted in each cell and does not require an arbitrary probability distribution to be specified for the observation in each bin. One disadvantage of this approach is that if a model gave a 1 in 3 probability of zero, one, or two earthquakes in a particular bin and three earthquakes occurred, the EFES would not distinguish this forecast as incorrect, whereas a likelihood test should. However, we believe the EFES provides more flexibility by not forcing an arbitrary probability distribution.

An illustration of the metric on both real and simulated data

EFES and real forecast data

We first illustrate the EFES on the forecast model we developed and submitted to the CSEP Japan initiative.[12] Figure 1 (left) shows the forecast for the model for the year 2010. Each value is the predicted rate of M5 or greater earthquakes for each bin (0.1°×0.1°). The magnitude of the earthquake is determined by the Japan Meteorological Agency (JMA). We plot with a log scale and do not plot bins with less than 0.001 expected rate of earthquakes (although we include them in our calculations). To create this forecast we use earthquakes only to the end of 2009. The plot directly below shows where the M5 or greater earthquakes occurred during 2010. We use the same data as that used by the CSEP Japan initiative.[13] More recent data cannot be considered because of the delay in publication of the catalog.[13]

We calculate the enrichment score for the observed set of earthquakes. The corresponding enrichment score profile for this set is shown in Figure 1. The top portion of the enrichment score profile shows the running sum statistic calculated from (0.1). The y-axis shows the enrichment score. The largest deviation from zero is also marked with a green vertical line. The shape of the EFES profile will change based on the forecast and set of observed earthquakes, as shown in the later figures. Below the running sum statistic we can see the bins where earthquakes occurred. These lines are drawn three-bins-wide so that they can be seen in the plot. These bins have been sorted in decreasing order of their forecast value. Forty-eight earthquakes were observed, occurring in 43 distinct bins. Five earthquakes were observed in one bin, two earthquakes were observed in another, and all the remaining bins had only one earthquake. The EFES is 0.82. Therefore, the earthquakes tend to cluster in the higher forecast bins.

One hundred permutations were run and a score this high was only observed by chance three times. Therefore, the EFES is significant at the 5% level. We prepared a histogram of the permutations and marked our observed EFES as a green vertical line in the lower part of the enrichment score profile. By using the permutation test, this forecast can now be investigated without requiring an arbitrary reference distribution.

EFES and simulated data

The EFES was also applied to simulated data to show it is a powerful and robust technique. Forecasts were simulated for both the All Japan area, specified by the Japan CSEP group, and the California area, specified by the United States CSEP initiative.[10,13] The Japanese area has 20,062 bins; the California area is smaller with 7682 bins. The simulated scenarios are shown in Figure 2.

The first scenario is a forecast where all bins are assigned a high forecast value. Here we assigned all bins a value between 0.8 and 1. We sampled 1% of the bins, and assigned these bins an earthquake. The observed earthquakes are shown as green crosses. This simulated forecast for the Japan region is shown in Figure 2A along with the obtained ES profile. The ES profile is drawn in an identical manner to Figure 1(C). The top portion of the enrichment score profile shows the running sum statistic calculated from (0.1). Below the running sum statistic, we can see the bins where earthquakes were simulated to occur. We draw these lines three-bins-wide so that they can be seen in the plot. These bins are sorted by their forecast value, in decreasing order. We assign earthquakes to bins randomly, so we can see here that the earthquakes are distributed randomly down the list. The histogram shows the result of the permutation testing. One hundred permutations were used. The green vertical line marks the original EFES obtained directly above. Unsurprisingly, the original score was not significant.

The second scenario is almost identical to the first. However, here we assign all bins a forecast value between 0 and 0.4. The range of the forecast values is widened and the average forecast value is lower than the first scenario. The observed vector is again a random drawn from the total set of bins and we select 1% of bins to have earthquakes. As mentioned earlier, the shape of the EFES will depend on the forecast and the set of observed earthquakes, as so will the distribution of permuted EFES values. The distribution has moved to the right. However, whilst the ES is higher, it remains not significant.

The third scenario is slightly more realistic. Here we first selected 1% of the total bins and assigned these bins an earthquake. Then we simulated the forecast such that bins with an observed earthquake were assigned a value between 0.2 and 1 and bins without observed earthquakes were assigned a value between 0 and 0.8. Therefore, on average, the bins with an observed earthquake have a higher forecast value than those without. This represents the so-called *messy perfect forecast*. The ES reaches its maximum value early and the bins with earthquakes are gathered toward the top of our ranked list of earthquakes. The permutation test shows the ES is significant.

The fourth scenario is similar to the third. However, we forced a higher agreement between the forecast and the observations. Bins with an observed earthquake were assigned a value between 0.6 and 1 and bins without an earthquake were assigned a value of between 0 and 0.4. The ES is larger than that obtained with the third scenario and this result is significant.

The first four scenarios are highly unlikely to represent real submitted forecasts but they serve to illustrate some important points. If you submit a forecast similar to the first scenario, the observations will of course fall in bins with high forecast values. However, because all bins have a high-predicted value, it is a nonsensical forecast and the EFES will not be significant. If it is possible to submit the perfect forecast (or the perfect forecast in amongst a little noise), the EFES will be significant (and higher for the more accurate forecast) as witnessed with the third and fourth scenarios.

We have also simulated some scenarios that are more likely to represent submitted forecasts, where there is spatial clustering in the forecasts (Figure 2E and F) and spatial clustering in both the forecast and the observations (Figure 2G and H).

In the fifth and sixth scenarios, we first created the observations vector, by randomly sampling 1% of the total number of bins. We randomly assigned each bin a forecast between 0 and 1. Then, we divided this forecast value by the distance of the bin to the closest observation, in accordance with a predefined smoothing value. In the fifth scenario, we used a moderate amount of smoothing, and in the sixth scenario we used a harsh degree of smoothing. Therefore, in Figure 2E we can see moderate clusters within the forecast, and in Figure 2F we can see tight clusters within the forecast. The enrichment score profile for these scenarios shows that the tighter the cluster, the better the ES. In other words, the better the distinction between forecasts of bins with earthquakes and those bins without earthquakes, the better the ES. However, even with moderate clusters in our data we can still obtain significant ESs.

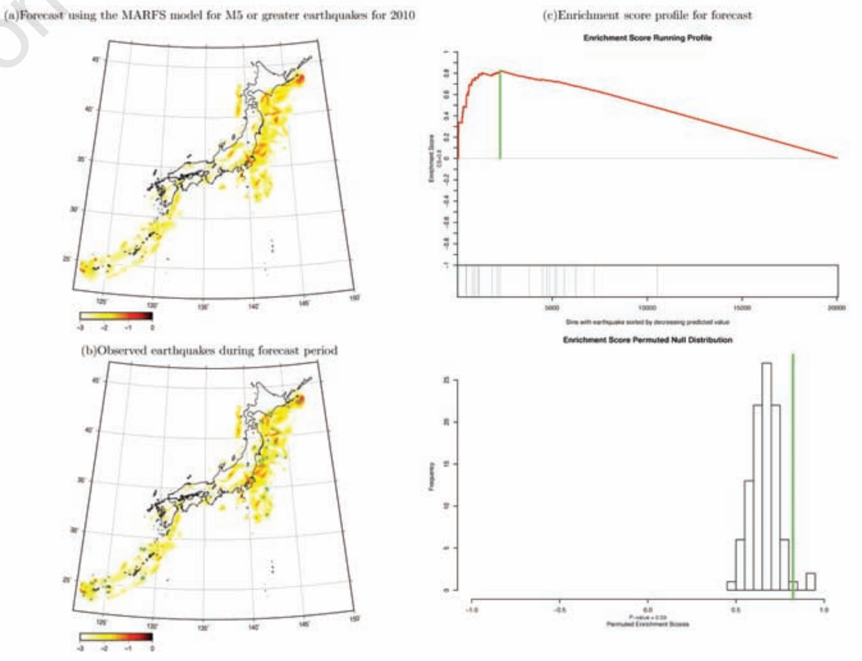

Figure 1. (A) Forecast using the MARFS model (12) for M5 or greater earthquakes for 2010. (B) The locations of observed earthquakes during the forecast period. (C) Enrichment score profile for the forecast.

The final two scenarios are identical to the preceding two. However, we also forced spatial clustering within the observations. This is a likely situation in the CSEP Japan experiment, where the catalog is not declustered before comparing it to the submitted forecasts. We see the same results as those portrayed in Figure 2E and F: the tighter the clusters around the observed clusters, the better the EFES.

The results we present in each of the figures above are obtained from a single simulation. It may be argued that we obtained fortuitous simulations that highlighted the capabilities of the EFES metric. We, therefore, repeated each of the simulations 100 times and counted the number of times the EFES is significant. The results for the Japan simulations are shown in the middle column of Table 1, and for identical scenarios for the California area in the middle column of Table 2. A comparison between the results for the California and Japan scenarios shows that the number of bins has not affected the results.

The results for 0.5% of bins randomly assigned an earthquake (left column) and 5% of bins randomly assigned an earthquake (right column) are shown in Tables 1 and 2. We can see that the results are stable. The first two scenarios are only significant on an average 5% of the time; a result which is consistent with chance. Also, the other scenarios are significant almost 100% of the time; a result that gives us confidence in the stability of the method. The results do not depend on the number of bins or the proportion of bins with earthquakes.

Discussion and Conclusions

The initial simulation runs show promising performances. We averaged the run time of the simulation analysis to calculate the computation time required to calculate an EFES and its significance with 1000 permutations. On average, it takes a little over one minute. We used a Windows desktop machine with Intel Core i5 CPU (3.20 gigahertz) and 4 gigabytes of RAM.

It is possible to compare two forecasts in the spatial domain simply by comparing their EFESs, assuming they forecast and are evaluated on the same period and area. The forecast with the higher EFES is preferable. It is possible to ascertain if the difference between the EFESs of two forecasts is significant by permutation testing. To perform this permutation test we would not permute the earthquakes at all. Rather, we would first calculate the difference between the two scores. Then, for each bin, we would randomly reassign the real forecast value of model one, to either of model one or model two. We then obtain a permuted vector for model one (and model two) which contains some forecast values from the original model one and some forecast values from the original model two. We then calculate the EFES for each permuted model and the difference between the scores. We repeat this process to get a distribution of the differences and we look to see how likely we are to get our original difference by chance.

It is possible to ascertain if the difference between the EFESs of two forecasts is significant using the permutation technique described above. However, there have been a large number of models submitted to the CSEP initiative. The probability of finding a significant result due to chance increases as we increase the number of hypotheses that we test simultaneously. Therefore, to test all possible pairwise comparisons would require an appropriate adjustment such as Bonferonni correction of the statistical significance level to deal with the issue of multiple comparisons. Bonferonni correction corrects the significant level for each individual test by the number of tests so that the probability of observing at least one significant result by chance is unaf-

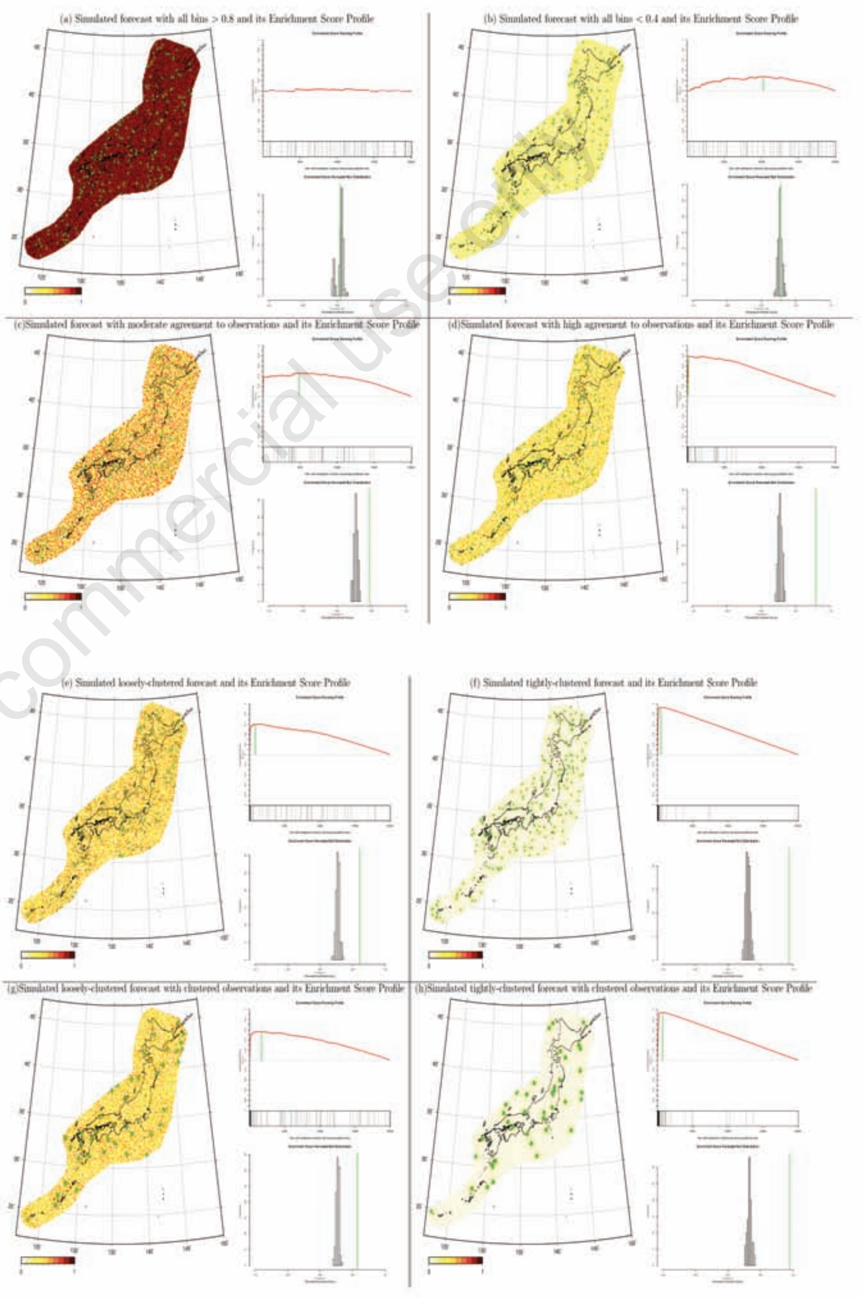

Figure 2. (A)-(H) Simulated forecasts and their respective enrichment score profiles for each of the eight scenarios described in the text.

Table 1. The number of significant simulations for each scenario applied to the Japan CSEP area (20,062 bins) with varying numbers of observed earthquakes.

	% bins with earthquake observed		
	0.5%	1%	5%
Forecast with all bins >0.8	5	8	6
Forecast with all bins <0.4	6	4	2
Moderate agreement between forecast and observations	100	100	100
High agreement between forecast and observations	100	100	100
Loosely clustered forecast with agreement to observations	100	100	100
Tightly clustered forecast with agreement to observations	100	100	100
Loosely clustered forecast and clustered observations	100	100	100
Tightly clustered forecast and clustered observations	100	100	100

Table 2. The number of significant simulations for each scenario applied to the California CSEP area (7,682 bins) with varying numbers of observed earthquakes.

	% bins with earthquake observed		
	0.5%	1%	5%
Forecast with all bins >0.8	9	7	4
Forecast with all bins <0.4	4	3	7
Moderate agreement between forecast and observations	90	99	100
High agreement between forecast and observations	100	100	100
Loosely clustered forecast with agreement to observations	100	100	100
Tightly clustered forecast with agreement to observations	100	100	100
Loosely clustered forecast and clustered observations	100	100	100
Tightly clustered forecast and clustered observations	100	100	100

fected by the number of tests performed.

The situation which came to light as a result of the analysis with real data, when more than one earthquake was observed in each bin, merits discussion. We consider our gene set as our set of bins where earthquakes occur. This means we just consider if the earthquakes occur in *red* (high probability bins) without considering how many earthquakes occurred in each bin. In fact, the formulae detailed above do not include any reference to the number of earthquakes that occurred in a bin. In this regard, the EFES is more similar to the ROC-like tests in CSEP than to the likelihood tests (which do in fact consider the actual number of observed earthquakes in a bin). In fact, the ROC also involves ranking the forecast values. The ROC curves then plot the true positive rate versus the false positive rate as an artificial alarm rate is increased.

In Nanjo *et al.,*[14] the authors suggest a technique for modifying the ROC calculation such that it is suitable for comparing forecasts with multiple earthquakes observed in a single bin. The authors consider each earthquake separately and then sum the contingency table values. However, it is impossible to do the same with the presented technique. If we were to consider each earthquake as a separate set, it is difficult to combine them in a sensible manner so that the absolute value of the resultant score is between zero and one, and is representative of the original forecast.

However, there is a simple solution if we consider bins where earthquakes occurred and calculate the root mean square error (RMSE) between the predicted and observed numbers of these bins. So, assume we have a simple experiment where six earthquakes occurred (five earthquakes in one bin, and one earthquake in another bin) and two forecasts, which are identical except for the forecasts for the bins where the earthquakes occurred. Let forecast A predict (0.9,0.1); forecast B predict (0.1,0.9) and the observed number of earthquakes be (5,1). In this scenario, both forecast A and forecast B would receive the same EFES. However, the RMSE of A is less than B, so we would assume that A is the better forecast.

In a similar sense, it should be obvious to the reader, that if we simply double the predictions for each cell and calculate the EFES, the EFES will remain the same, because the ranking of the forecast bins will be identical. More simply, the EFES does not consider the number of earthquakes of the entire forecast. Therefore, it is necessary to employ the EFES with a test that measures the consistency of the forecast rate of all bins (not only considering bins where earthquakes occurred as described in the previous paragraph) and the corresponding observed number of earthquakes. The N-test[2,3] already incorporated into the CSEP suite of tests is ideal for this purpose.

We also do not consider the forecast for each magnitude bin within each geographical cell, although the technique to calculate the score would not change if this information was included. We simply sum over all magnitude ranges for the cell and use this single number as the cell's prediction. Therefore, for the type of implementation we describe here, it is also necessary to use a test that measures the difference between the forecast magnitude distribution and the observed magnitude distribution. We recommend the M-test for this purpose.[3] In short, the EFES cannot be used to rank forecasts on its own, but must be combined with other tests.

The EFES is an easily understandable technique to assess if earthquakes occur in bins with higher forecast values than in bins with lower forecast values. The EFES is most similar to the current S-test employed in the CSEP suite of evaluation metrics; however, it does not require the specification of a distribution (usually the Poisson) for the probability of the observation in each bin. We are currently working in collaboration with the Japanese CSEP environment to prepare the metric ready for testing of real forecast models. We hope to publish the results of the EFES metric (compared to the currently used evaluation metrics) in the future. It is important to constantly reassess the evaluation metrics in use at the CSEP to ensure the most accurate evaluation of the submitted forecasts.

References

1. Zechar JD, Schorlemmer D, Liukis M, et al. The Collaboratory for the Study of Earthquake Predictability perspective on computational earthquake science. Concurr Comp - Pract E 2010;22:1836-47.
2. Schorlemmer D, Gerstenberger MC, Wiemer S, et al. Earthquake likelihood model testing. Seismol Res Lett 2007;78:17-29.
3. Zechar JD, Gerstenberger MC, Rhoades DA. Likelihood-Based Tests for Evaluating Space-Rate-Magnitude Earthquake Forecasts. Bull Seismol Soc Am 2010;100:1184-95.
4. Molchan GM. Structure of Optimal Strategies in Earthquake Prediction. Tectonophysics 1991;193:267-76.
5. Zechar JD. Evaluating earthquake predictions and earthquake forecasts: a guide for students and new researchers. Community Online Resource for Statistical Seismicity Analysis, 2010.
6. Zechar JD, Jordan TH. The Area Skill Score Statistic for Evaluating Earthquake Predictability Experiments. Pure Appl

Geophys 2010;167:893-906.
7. Rhoades DA, Schorlemmer D, Gerstenberger MC, et al. Efficient testing of earthquake forecasting models. Acta Geophys 2011;59:728-47.
8. Clements RA, Schoenberg FP, Schorlemmer D. Residual analysis methods for space-time point processes with applications to earthquake forecast models in California. Ann Appl Stat 2011. In press.
9. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545-50.
10. Schorlemmer D, Gerstenberger MC. RELM testing center. Seismol Res Lett 2007;78:30-6.
11. Schorlemmer D, Zechar JD, Werner MJ, et al. First Results of the Regional Earthquake Likelihood Models Experiment. Pure Appl Geophys 2010;167: 859-76.
12. Smyth C, Mori J. Statistical models for temporal variations of seismicity parameters to forecast seismicity rates in Japan. Earth Planets Space 2011;63:231-8.
13. Nanjo KZ, Tsuruoka H, Hirata N, Jordan TH. Overview of the first earthquake forecast testing experiment in Japan. Earth Planets Space 2011;63:159-69.
14. Nanjo KZ, Holliday JR, Chen CC, et al. Application of a modified pattern informatics method to forecasting the locations of future large earthquakes in the central Japan. Tectonophysics 2006;424:351-66.