Pure Appl. Geophys. 167 (2010), 859–876

DOI 10.1007/s00024-010-0081-5

Pure and Applied Geophysics

First Results of the Regional Earthquake Likelihood Models Experiment

Danijel Schorlemmer,[1] J. Douglas Zechar,[1,2] Maximilian J. Werner,[3] Edward H. Field,[4] David D. Jackson,[5] Thomas H. Jordan,[1] and The RELM Working Group

Abstract—The ability to successfully predict the future behavior of a system is a strong indication that the system is well understood. Certainly many details of the earthquake system remain obscure, but several hypotheses related to earthquake occurrence and seismic hazard have been proffered, and predicting earthquake behavior is a worthy goal and demanded by society. Along these lines, one of the primary objectives of the Regional Earthquake Likelihood Models (RELM) working group was to formalize earthquake occurrence hypotheses in the form of prospective earthquake rate forecasts in California. RELM members, working in small research groups, developed more than a dozen 5-year forecasts; they also outlined a performance evaluation method and provided a conceptual description of a Testing Center in which to perform predictability experiments. Subsequently, researchers working within the Collaboratory for the Study of Earthquake Predictability (CSEP) have begun implementing Testing Centers in different locations worldwide, and the RELM predictability experiment—a truly prospective earthquake prediction effort—is underway within the U.S. branch of CSEP. The experiment, designed to compare time-invariant 5-year earthquake rate forecasts, is now approximately halfway to its completion. In this paper, we describe the models under evaluation and present, for the first time, preliminary results of this unique experiment. While these results are preliminary—the forecasts were meant for an application of 5 years—we find interesting results: most of the models are consistent with the observation and one model forecasts the distribution of earthquakes best. We discuss the observed sample of target earthquakes in the context of historical seismicity within the testing region, highlight potential pitfalls of the current tests, and suggest plans for future revisions to experiments such as this one.

Key words: Statistical seismology, earthquake predictability, earthquake statistics, earthquake forecasting and testing, seismic hazard.

1. Introduction

The Regional Earthquake Likelihood Model (RELM) working group formed in 2000 and was supported by the Southern California Earthquake Center (SCEC) and the United States Geological Survey (USGS). The group's main purpose was to improve seismic hazard assessment and to increase understanding of earthquake generation processes. Seismic hazard analysis requires two fundamental components: an earthquake forecast that describes the probabilities of earthquake occurrence in a spatio-temporal volume; and a ground-motion model that transforms each forecasted event into a site-specific estimate of ground-shaking. RELM participants focused on the former component and developed several earthquake forecast models (Bird and Liu, 2007; Console *et al.*, 2007; Ebel *et al.*, 2007; Gerstenberger *et al.*, 2007; Helmstetter *et al.*, 2007; Holliday *et al.*, 2007; Kagan *et al.*, 2007; Petersen *et al.*, 2007; Rhoades, 2007; Shen *et al.*, 2007; Ward, 2007; Wiemer and Schorlemmer, 2007). These models span a broad range of input data and methods: most are based on past seismicity, however some incorporate geodetic data and/or geological insights. See Field (2007) and the special volume of Seismological Research Letters for more details on the RELM project.

In addition to developing forecast models, RELM also explored comparative testing strategies and established a plan for conducting these tests.

The members of the RELM Working Group are listed in the Acknowledgments section.

[1] Department of Earth Sciences, Southern California Earthquake Center, University of Southern California, 3651 Trousdale Parkway, Los Angeles, CA 90089-0740, USA. E-mail: ds@usc.edu
[2] Lamont-Doherty Earth Observatory, Columbia University, P.O. Box 1000, Palisades, NY 10964, USA.
[3] Swiss Seismological Service, ETH Zurich, Sonneggstrasse 5, 8092 Zurich, Switzerland.
[4] United States Geological Survey, 525 S. Wilson Avenue, Pasadena, CA 91106, USA.
[5] Department of Earth and Space Sciences, University of California Los Angeles, Los Angeles, CA 90095, USA.

The group developed a suite of likelihood tests (Schorlemmer *et al.*, 2007) to be implemented within a Testing Center, a facility in which earthquake forecast models are installed as software codes and in which all necessary tests are conducted in an automated and fully prospective fashion (Schorlemmer and Gerstenberger, 2007). By the end of the 5-year project, 19 earthquake forecasts were submitted for prospective testing in the period of 1 January 2006, 00:00–1 January 2011, 00:00. These forecasts were not installed as software codes in the Testing Center because the RELM group decided to use simple forecast tables; nevertheless, the processing is fully automated and does not require human interaction. All other models in the Testing Center, including the RELM 1-day models, are installed as codes.

Following the conclusion of the RELM project, the Collaboratory for the Study of Earthquake Predictability (CSEP) was formed as a venue to expand upon the RELM experiment and to establish and maintain a Testing Center (Jordan, 2006). CSEP is built upon a global partnership to promote rigorous earthquake predictability experiments in various tectonic environments. In addition to establishing new testing regions, CSEP is developing new testing methods, introducing new kinds of earthquake forecast models, and improving upon the testing rules suggested by the RELM working group. The U.S. branch of CSEP inherited all RELM earthquake forecasts, as well as the task of testing them according to the rules outlined by Schorlemmer *et al.* (2007) in a Testing Center designed according to Schorlemmer and Gerstenberger (2007).

All models developed by RELM participants forecast earthquakes in a testing area that covers the state of California and all regions within about one degree of its borders. This test region was chosen to include any earthquake that might cause shaking within the state of California (Schorlemmer and Gerstenberger, 2007). The RELM working group proposed two major classes of forecasts: 1 day and 5 years (Schorlemmer and Gerstenberger, 2007). In contrast to daily or yearly periodicity in weather, earthquakes do not follow obvious seasonal or cyclical patterns that could be used to scientifically justify the chosen durations. Rather, the classes are end-user-oriented: The 5-year class is relevant for seismic hazard calculations, while the 1-day class allows a closer look at aftershock hazard forecasts and potential short-term precursor detection. Daily forecasts can make use of all seismicity up to and including the previous day to adapt to new earthquakes and to re-calibrate the model, whereas the 5-year forecasts are fixed at the beginning of the experiment and never updated. Because of this fundamental difference in the setup, models were either submitted for the 1-day class or the 5-year class. Forecasts submitted to the 5-year class were taken to be time-invariant. We briefly describe the models below; a detailed summary of the models is given by Field (2007) while the full descriptions of each model can be found in the individual articles in the special volume of Seismological Research Letters (see Table 1).

One of the main goals of RELM was to test models comparatively; to compare models, a significant standardization of the forecasts was necessary. Therefore, all testing rules, the testing period, the testing area, and the earthquake catalog and its processing were defined by Schorlemmer and Gerstenberger (2007) and agreed upon by the members of the RELM working group. This standardization also required that all RELM models provide grid-based forecasts: earthquake rates specified in latitude/longitude/magnitude bins, and characterized by Poisson uncertainty. Models that declare alarms or forecast fault ruptures were not considered, as no testing method was developed or specified for these kinds of forecasts.

In this paper we describe the different model classes and present the results from the first 2.5 years of testing the time-invariant 5-year RELM forecasts. Because the forecasts were specified as being time-invariant, all forecast rates were halved for the results presented here. We emphasize, however, that these results are preliminary because the forecasts were specified as 5-year forecasts. As more earthquakes occur, the results will likely change. Nevertheless, the results indicate which models are consistent with the observations to date and which models have so far performed best in comparative testing.

Table 1

RELM models being evaluated within the Testing Center

Model	Testing class	Forecasted number of earthquakes	Fraction of area covered by forecast (%)	Reference
EBEL-ET-AL.MAINSHOCK	*5-year mainshock*	8.6703 (8.6705)	47.37	EBEL *et al.* (2007)
EBEL-ET-AL.MAINSHOCK.CORRECTED	*5-year mainshock*	9.2431 (9.2433)	51.74	EBEL *et al.* (2007)
HELMSTETTER-ET-AL.MAINSHOCK	*5-year mainshock*	10.5760	100.00	HELMSTETTER *et al.* (2007)
HOLLIDAY-ET-AL.PI	*5-year mainshock*	14.4205 (15.0164)	8.29	HOLLIDAY *et al.* (2007)
KAGAN-ET-AL.MAINSHOCK	*5-year mainshock*	5.9998 (5.9998)	44.39	KAGAN *et al.* (2007)
SHEN-ET-AL.MAINSHOCK	*5-year mainshock*	5.2369 (5.2369)	44.39	SHEN *et al.* (2007)
WARD.COMBO81	*5-year mainshock*	9.4812 (16.0582)	26.72	WARD (2007)
WARD.GEODETIC81	*5-year mainshock*	12.1498 (27.9849)	26.72	WARD (2007)
WARD.GEODETIC85	*5-year mainshock*	6.9972 (16.1169)	26.72	WARD (2007)
WARD.GEOLOGIC81	*5-year mainshock*	8.3332 (9.0760)	26.72	WARD (2007)
WARD.SEISMIC81	*5-year mainshock*	7.9605 (11.1136)	26.72	WARD (2007)
WARD.SIMULATION	*5-year mainshock*	3.7261 (4.1027)	26.72	WARD (2007)
WIEMER-SCHORLEMMER.ALM	*5-year mainshock*	11.8693	100.00	WIEMER and SCHORLEMMER (2007)
BIRD-LIU.NEOKINEMA	*5-year mainshock+aftershock*	27.9514	100.00	BIRD and LIU (2007)
EBEL-ET-AL.AFTERSHOCK	*5-year mainshock+aftershock*	36.4017 (36.4026)	47.37	EBEL *et al.* (2007)
EBEL-ET-AL.AFTERSHOCK.CORRECTED	*5-year mainshock+aftershock*	37.5664 (37.5674)	51.74	EBEL *et al.* (2007)
HELMSTETTER-ET-AL.AFTERSHOCK	*5-year mainshock+aftershock*	17.7012	100.00	HELMSTETTER *et al.* (2007)
KAGAN-ET-AL.AFTERSHOCK	*5-year mainshock+aftershock*	7.9910 (7.9910)	44.39	KAGAN *et al.* (2007)
SHEN-ET-AL.AFTERSHOCK	*5-year mainshock+aftershock*	7.3236 (7.3236)	44.39	SHEN *et al.* (2007)

All models were submitted before 1 January 2006, except for the EBEL-ET-AL.MAINSHOCK.CORRECTED model and the EBEL-ET-AL.AFTERSHOCK.CORRECTED model, which were submitted 12 November 2006. The forecasted number of earthquakes reported here is the number forecasted in all unmasked cells, followed parenthetically by the number forecasted in all cells (see Masking subsection in the text). The fraction of the area covered by forecast is the portion of the study region for which the model makes an unmasked forecast

2. *Models*

2.1. *5-Year Models*

The forecasts submitted to the 5-year class represent a broad spectrum of models, each of which is built on its own set of scientific hypotheses pertaining to the occurrence of earthquakes. Most of the models use past seismicity as the primary data set for model calibration and parameter value estimation, and they then extrapolate historical seismicity rates into the future. However, some models make use of geological, geodetic, and/or tectonic data.

Large earthquakes are followed by dozens to hundreds of earthquakes in their immediate wake. If a very large event were to occur in California tomorrow, its triggered earthquakes would likely dominate the statistics of the entire 5-year period. Because mainshocks and dependent aftershocks cannot be identified by some physical measurement, a compromise was made to accommodate models which forecast independent mainshocks only. Two forecast subclasses were created: one for forecasts of mainshocks only (*mainshock* models) and one for forecasts of all earthquakes (*mainshock+aftershock* models). SCHORLEMMER and GERSTENBERGER (2007) and SCHORLEMMER *et al.* (2007) provide details on the declustering procedure that is used at the testing center to create catalogs of mainshocks against which the *mainshock* models are tested. Both classes forecast rates of earthquakes with magnitude greater than or equal to 4.95 with a binning of 0.1 magnitude units (resulting in magnitude bins of [4.95, 5.05), [5.05, 5.15), etc., with a final bin starting at magnitude 8.95 with no upper limit) and a spatial binning of $0.1° \times 0.1°$ with the cell boundaries aligned to the full degrees. The observed magnitude is taken to be the magnitude reported in the Advanced National Seismic System (ANSS) catalog, disregarding the magnitude scale.

2.2. *Mainshock Models*

Twelve *mainshock* models were submitted to RELM; these were formally registered and published

on the RELM website (http://relm.cseptesting.org, see also Table 1 and Figs. 1 and 2). Of these, many were generated by smoothing past seismicity under different assumptions. The models EBEL-ET-AL.MAINSHOCK and EBEL-ET-AL.MAINSHOCK.CORRECTED (see below for the explanation of the double entry), developed by EBEL *et al*. (2007), average the 5-year rate of $M \geq 5$ earthquakes in 3° by 3° cells from a declustered catalog from 1932 until 2004 and use a Gutenberg-Richter distribution for computing rates per magnitude. The model KAGAN-ET-AL.MAINSHOCK (KAGAN *et al*., 2007) smooths past earthquakes using a longer catalog dating back to 1800 and it accounts for the spatial extent of large earthquake ruptures. Rates are calculated using a tapered Gutenberg-Richter distribution with corner magnitude 8. HELMSTETTER *et al*. (2007) extend this approach to their HELMSTETTER-ET-AL.MAINSHOCK model by including past $M \geq 2$ events since 1984 in the smoothing, by optimizing the smoothing, and by accounting for the spatial variability of the completeness magnitude. The model WARD.SEISMIC81 (WARD, 2007) is also based on smoothing past earthquakes, in this case going back to 1850.

WIEMER and SCHORLEMMER (2007) estimated the *a* and *b* values of the Gutenberg-Richter distribution in each latitude/longitude cell to test the hypothesis that spatial variations in these values designate stationary

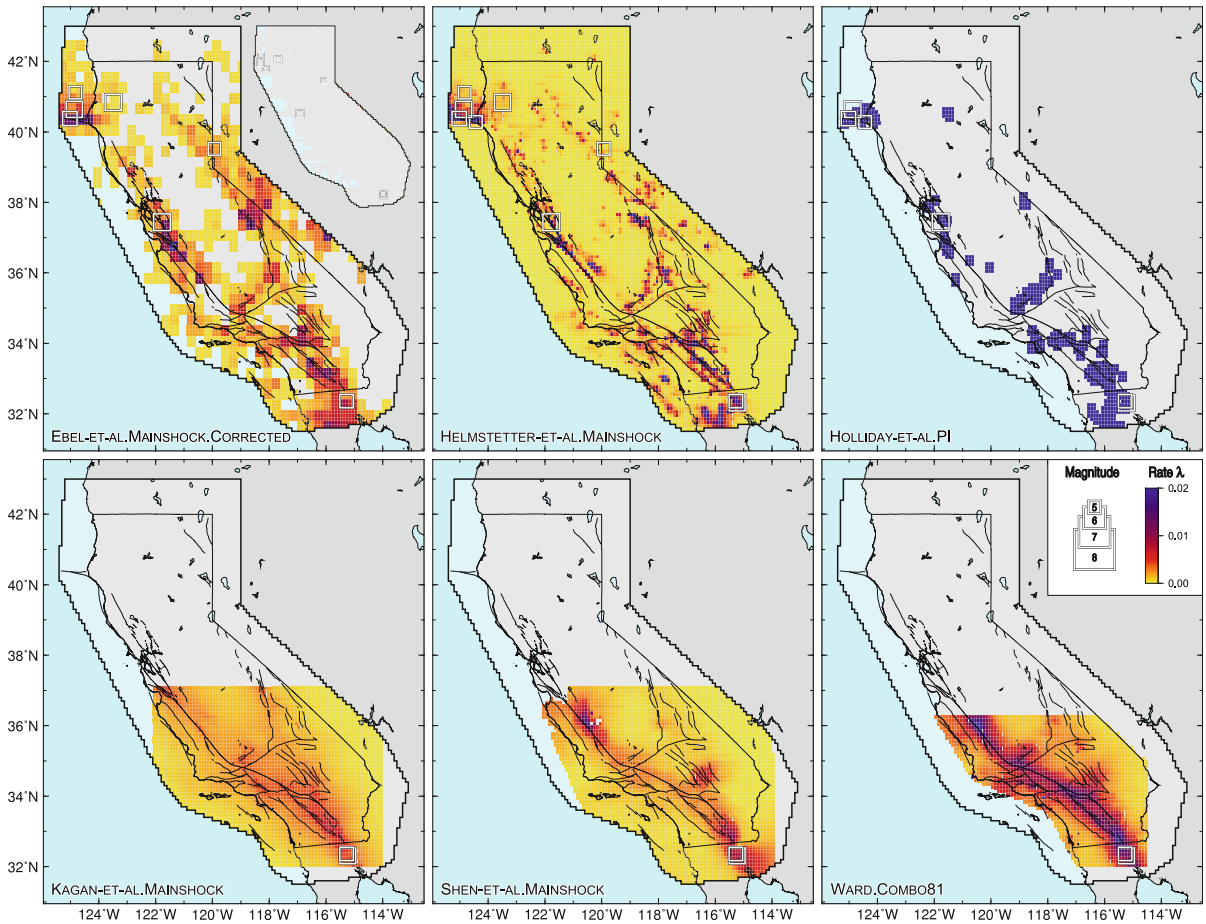

Figure 1

Forecast maps of *5-year mainshock* models. Colors indicate the forecast rate of all events with $M \geq 4.95$ (unmasked areas only), reducing the latitude/longitude/magnitude forecasts to latitude/longitude forecasts by summing over the magnitude bins. The observed target earthquakes are shown as *white squares*; only those earthquakes occurring in unmasked cells are shown for each model. Models from left to right: (*first row*) EBEL-ET-AL.MAINSHOCK.CORRECTED with EBEL-ET-AL.MAINSHOCK as inset, HELMSTETTER-ET-AL.MAINSHOCK, and HOLLIDAY-ET-AL.PI. (*second row*) KAGAN-ET-AL.MAINSHOCK, SHEN-ET-AL.MAINSHOCK, and WARD.COMBO81

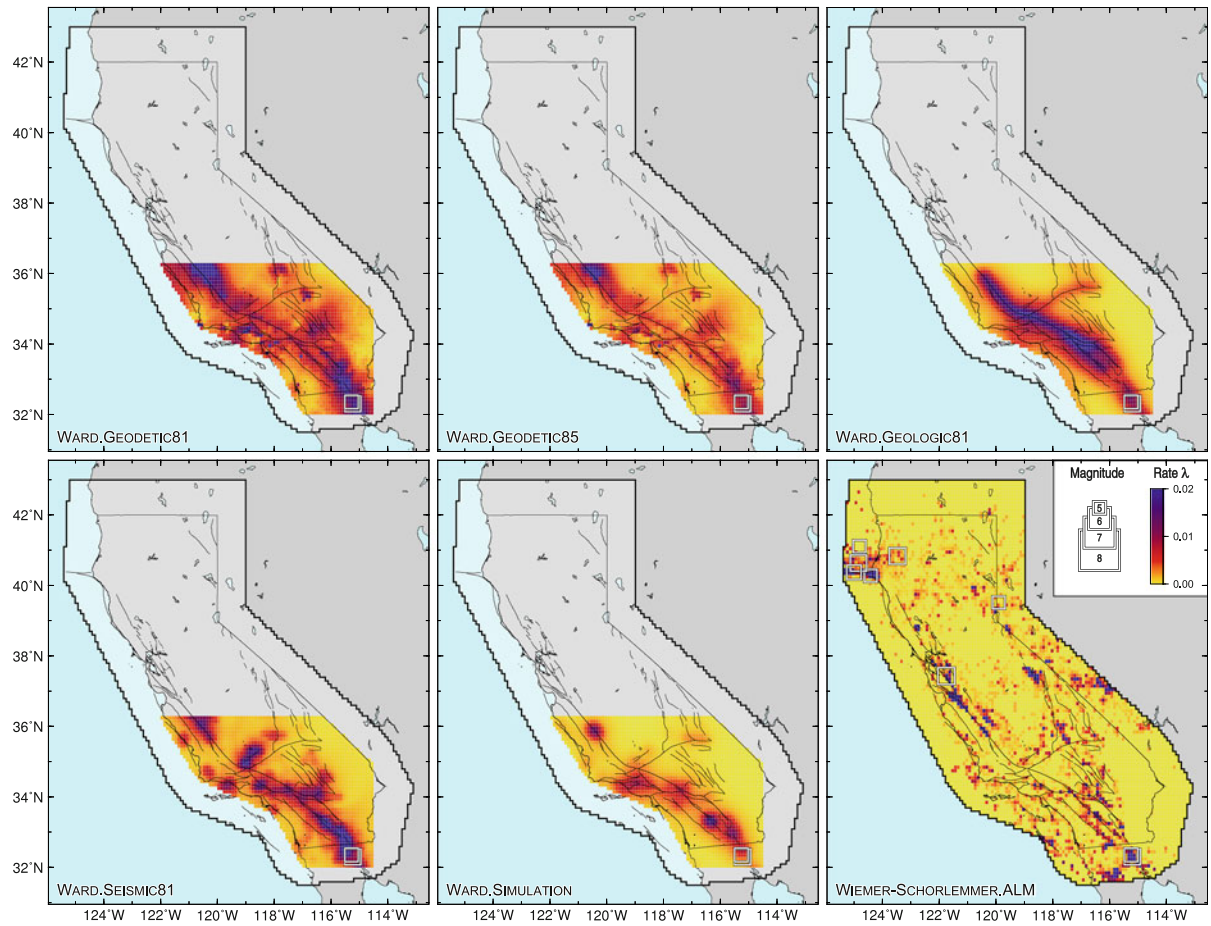

Figure 2

Forecast maps of *5-year mainshock* models. Colors indicate the forecast rate of all events with $M \geq 4.95$ (unmasked areas only), reducing the latitude/longitude/magnitude forecast to latitude/longitude forecasts by summing over the magnitude bins. The observed target earthquakes are shown as *white squares*; only those earthquakes occurring in unmasked cells are shown for each model. Models from left to right: (*first row*) WARD.GEODETIC81, WARD.GEODETIC85, and WARD.GEOLOGIC81. (*second row*) WARD.SEISMIC81, WARD.SIMULATION, and WIEMER-SCHORLEMMER.ALM

asperities that govern the relative frequency of large and small earthquakes (the WIEMER-SCHORLEMMER.ALM model). The model HOLLIDAY-ET-AL.PI, submitted by HOLLIDAY *et al.* (2007), is based on the assumption that regions of strongly fluctuating seismicity will be the regions of future large earthquakes.

Some models include data other than past earthquake observations. Three models are based solely on geodetic data. In one, SHEN-ET-AL.MAINSHOCK, SHEN *et al.* (2007) assumed that the earthquake rate is proportional to the horizontal maximum shear strain rate. The magnitude rates are obtained from a spatially-invariant tapered Gutenberg-Richter distribution with corner magnitude 8.02. A second model, WARD.GEODETIC81 by WARD (2007), uses a larger data set and a different technique to map strain rates to seismicity rates. The sole difference between this and the third model, WARD.GEODETIC85 by WARD (2007), is the maximum magnitude in the truncated Gutenberg-Richter distribution (8.1 and 8.5, respectively).

WARD (2007) also provided a mainshock model based solely on geological data (WARD.GEOLOGIC81). The model is constructed by mapping fault slip rates into a smoothed geological moment rate density and then into seismicity rate, again assuming a spatially invariant truncated Gutenberg-Richter distribution. The model WARD.SIMULATION is based on simulations of velocity-weakening friction on a fixed fault

network representing California. The model WARD. COMBO81 presents the average of the seismic, geodetic, and geological models by WARD (2007).

2.3. Mainshock+Aftershock Models

Six *mainshock+aftershock* models were submitted to RELM (see Table 1 and Fig. 3). Of these, all but one are modifications of corresponding *mainshock* forecasts: EBEL *et al*. (2007), KAGAN *et al*. (2007), HELMSTETTER *et al*. (2007) and SHEN *et al*. (2007) calibrated their *mainshock+aftershock* forecast to a complete catalog while their *mainshock* forecasts were calibrated based on a declustered catalog of past seismicity. The model BIRD-LIU. NEOKINEMA by BIRD and LIU (2007) is based on a local kinematic model of surface velocities derived from geodetic, tectonic, geological, and stress-direction data. The velocities are mapped into seismic moment rate and then into long-term seismicity rate.

2.4. Corrected Forecast Groups

Two additional 5-year model classes were introduced to account for corrected versions of the models by EBEL *et al*. (2007). In their initial submission, the

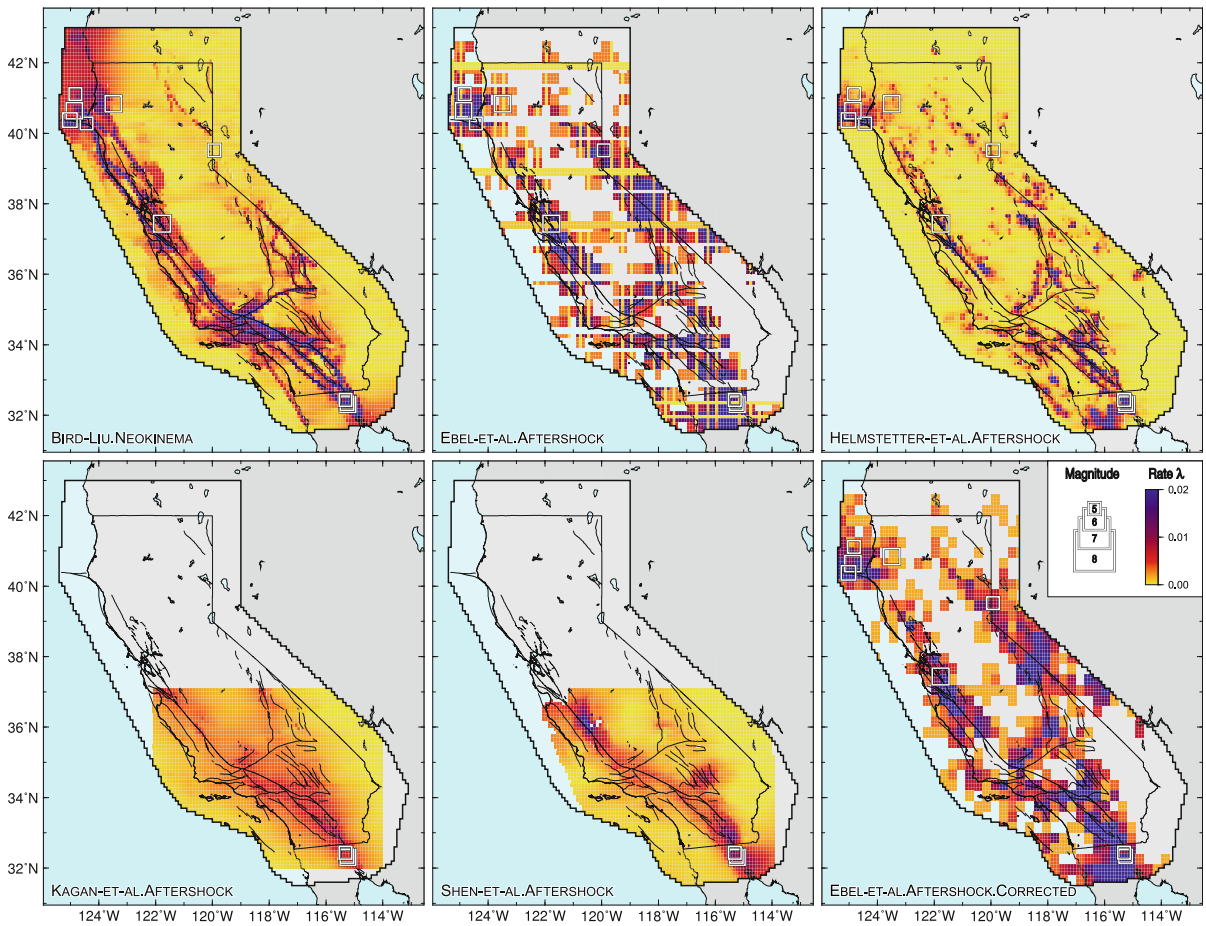

Figure 3

Forecast maps of all *5-year mainshock+aftershock* models. Colors indicate the forecast rate of all events with $M \geq 4.95$ (unmasked areas only), reducing the latitude/longitude/magnitude forecasts to latitude/longitude forecasts by summing over the magnitude bins. The observed target earthquakes are shown as *white squares*; only those earthquakes occurring in unmasked cells are shown for each model.. Models from left to right: (*first row*) BIRD-LIU.NEOKINEMA, EBEL-ET-AL.AFTERSHOCK, and HELMSTETTER-ET-AL.AFTERSHOCK. (*second row*) KAGAN-ET-AL. AFTERSHOCK, SHEN-ET-AL.AFTERSHOCK, and EBEL-ET-AL.AFTERSHOCK.CORRECTED. The EBEL-ET-AL.AFTERSHOCK.CORRECTED model was submitted on 12 November 2006 and is therefore tested against a smaller set of earthquakes

forecasts were erroneous at some locations; they were replaced by a corrected version on 12 November 2006. Because of the logic of truly prospective testing, the *mainshock* class and the *mainshock+aftershock* class were expanded into two groups each. The first group includes all initial RELM submissions and compares them to observations from 1 January 2006 forward, while the second group (denoted by a "corrected" suffix) covers all initial submissions and the corrected version of the model by Ebel *et al*. (2007). Because the corrected versions were submitted later, testing for this group started at the submission date of the corrected versions.

For any further model addition or correction, a new group will be introduced. Such a group would consist of all existing models and the new submissions, and the starting date for testing would be the submission date of the new contributions.

3. Testing Center

The Testing Center is a multi-computer system running the CSEP Testing Center software. It is divided into four main components: the development system, the integration system, the operational system, and the web presentation system (Zechar *et al*., 2009). The development system is used for software development of the Testing Center software and for model development and installation. After Testing Center software and respective models successfully run on the development system, their functionality is tested on the integration system. Each day this system checks out all necessary software codes and performs unit and acceptance tests for all software programs. This step is introduced to mimic the operational system and to detect possible problems before codes are transferred to the operational system. The operational system has the same setup as the integration system, however the codes are only updated every three months according to the release schedule of new versions of the Testing Center software. On the operational system, all tests are performed according to different scheduling depending on the model groups. All results are copied to the web presentation system from which they can be retrieved.

The design of the Testing Center followed the four main goals as outlined by Schorlemmer and Gerstenberger (2007):

Transparency. All computer codes are managed in a version control repository and are freely available. Thus, all changes to the codes are documented and a web-based collaboration system allows everyone to monitor the software development. The Testing Center codes are published under the open-source General Public License, and the majority of the models which were submitted as codes are open-source codes and can be used by other researchers. The RELM 5-year models were submitted as simple forecast files which are also freely available on the RELM website (http://relm.cseptesting.org). The Testing Center also catalogs all data files used for generating and testing forecasts. Any of these files is freely available.

Controlled Environment. The Testing Center ensures truly prospective tests of all submitted models with the same data. Any model submission gets time-stamped and will only be tested for periods after the submission date. Such an environment is needed for continuous testing of short-term models like the RELM *1-day* model class. Because modelers cannot modify their models after submission, no conscious or unconscious bias of a modeler is introduced into the forecasts.

Comparability. One of the major purposes of the Testing Center is the comparative testing of models. Models are tested for consistency with the observation and against each other (given the observation) to assess their comparative performance.

Reproducibility. Full reproducibility of any result is perhaps the most important feature of the Testing Center. Each data set used for computing a test is stored in the system. Thus, any forecast and any input data set can be reproduced and the tests can be recomputed at any time. Each test computation also stores the system configuration for full reproducibility.

3.1. Tests for Evaluating the Earthquake Forecasts

Schorlemmer *et al*. (2007) proposed a suite of statistical tests to evaluate probabilistic earthquake

forecasts. Similar tests were discussed by JACKSON (1996) and used by KAGAN and JACKSON (1994, 1995) for the evaluation of long-term forecasts of large earthquakes. In the language of statistical hypothesis testing, the tests fall into the class of significance tests: Assuming a null hypothesis (a given forecast model), the distribution of an observable test statistic is simulated; if the observed test statistic (e.g., the number of earthquakes) falls into the upper or lower tail of the distribution, the null hypothesis is rejected. The predictive distributions are constructed from model-dependent Monte Carlo simulations and hence are not assumed to be asymptotically normal. DALEY and VERE-JONES (2004) and HARTE and VERE-JONES (2005) explored performance evaluations based on the entropy score and the information gain.

Three tests are used to evaluate the RELM forecasts: the first two—the L(ikelihood)-Test and the N(umber)-Test—measure the consistency of the forecasts with the observations, while the third—the likelihood R(atio)-Test—measures the relative performance of one model against another. Each of these tests compares forecast rates with observed rates, and although they make slightly different measurements, these tests are not independent metrics.

For the RELM models, the forecast in each bin is the expected Poisson earthquake rate (the mean seismicity rate), which is usually a very small floating point number (e.g., 10^{-4}). To evaluate the likelihood of the model forecast given an observation (which is an integer, usually 0 or 1), the discrete Poisson distribution with mean equal to the forecast is used. For simplicity, the forecasts are stated in terms such that all observations in bins are independent, allowing probabilities to factorize.

3.2. The Number- or N-Test

The N(umber)-Test measures the consistency of the total forecasted rate with the total number of observed earthquakes, summed over all bins. The results of the N-Test indicate whether a forecast has predicted too many earthquakes, too few earthquakes, or a number of earthquakes that is considered to be consistent with the observed number. For example, consider a model which predicted $\lambda = 28.4$ earthquakes in the total space-time-magnitude testing region, and assume that, like the RELM models we consider, the forecast is characterized by Poisson uncertainty. If $\omega = 30$ events were observed during the experiment, the model obtains a quantile score of $\delta = \mathrm{Poi}(\omega = 30|\lambda = 28.4) = 0.66$ (here Poi stands for the Poisson cumulative distribution function). A model may be rejected if δ is very small (e.g., less than 0.025) or very large (e.g., greater than 0.975), which would indicate that the observed number of earthquakes falls into the far upper or far lower end of the forecast distribution, respectively. This indicates that the number of observed earthquakes is unlikely given the model forecast and, hence, the forecast is inconsistent with the observation. The N-Test disregards the spatial and magnitude distributions of the forecast and the observations, emphasizing each forecast's rate model.

3.3. The Likelihood- or L-Test

The L(ikelihood)-Test measures the consistency of a forecast with the observed rate and distribution of earthquakes. In each latitude-longitude-magnitude bin, the log-likelihood of an observation, given the forecast, is computed (again assuming the Poisson distribution). The log-likelihoods are then summed over all bins. To understand whether this sum—the observed log-likelihood—is consistent with what would be expected if the model were correct, many synthetic catalogs consistent with the model forecast are simulated, and their log-likelihoods calculated. This process produces a distribution of log-likelihoods, assuming that the model of interest is the "true" model. The statistic γ measures the proportion of simulated log-likelihoods less than the observed log-likelihood. If γ is low (e.g., less than 0.05), then the observed log-likelihood is much smaller than what would be expected given the model's veracity. The observation may therefore be considered inconsistent with the model. If γ is very high, the observed likelihood is considerably higher than expected, given the model forecast's veracity. In this case, however, it may be that a model predicted the distribution of earthquakes well but smoothed its forecast too much, and therefore high γ values are not considered grounds for model rejection. For example, consider the case when earthquakes occur only in a

model's most highly-ranked bins—those bins with the highest forecast rates. If the model is smooth, simulations consistent with the model would produce more diffuse seismicity than that observed, yielding simulated catalogs with events in bins with lower forecast rates, and thus a very high γ. Considering this effect, the L-Test is one-sided.

3.4. The Likelihood-Ratio- or R-Test

The likelihood R(atio)-Test consists of a pairwise-comparison between forecasts (e.g., forecasts i and j). The observed log-likelihood is calculated for each model forecast, and the difference—the observed likelihood ratio—indicates which model better fits the observations. To understand whether this difference is significant, a null hypothesis that model i is correct is adopted and synthetic catalogs consistent with this model are produced. The likelihood ratio is calculated for each simulated catalog. If the fraction α^{ij} of simulated likelihood ratios less than the observed likelihood ratio is very small (e.g., less than 0.05), the observed likelihood ratio is deemed significantly small enough to reject model i. So that no single forecast is given an advantage, this procedure is applied symmetrically. That is, synthetic catalogs are also simulated assuming model j to be true, and these simulations are used to estimate α^{ji}. Comparing each model with all other models results in a table of α values.

3.5. Masking

Several models are based on data that are not available throughout the entire testing area, and some researchers felt their model was not applicable everywhere in the testing area. For a forecast to cover fully the testing area, a model needs an additional "background" model to fill the gaps. RELM requested that all submitted models cover the entire testing area, although modelers were permitted to mask the area in which they were unable to create their forecast according to their scientific ideas. Thus, the area of the genuine forecast can be identified during testing, although it is also possible to evaluate a model over the entire testing area if a background model is chosen. Currently, only the unmasked areas are tested in the Testing Center; that is, a forecast is only evaluated over bins which are unmasked. For the R-Test, only bins which are unmasked in both forecasts are considered.

3.6. Uncertainties in Observations

The earthquake catalog data used to test forecasts contain measurement uncertainties. To account for these uncertainties in the tests, Schorlemmer *et al.* (2007) proposed generating "modified" catalogs. Each event's location and magnitude is modified using an error distribution suggested by the catalog compilers. Additionally, in the case of mainshock catalogs, declustering according to Reasenberg (1985) is applied using parameters that are sampled as described by Schorlemmer and Gerstenberger (2007). For each observed catalog, 1000 modified catalogs are generated, and these modified catalogs help to estimate the uncertainty of the test results resulting from the uncertainties of earthquake data.

4. Results

In this section we report preliminary summary results for the first half of the ongoing 5-year RELM experiment in California. Detailed results are available at http://us.cseptesting.org, where they are archived and regularly updated. We remind the reader that these results are preliminary, as they are based on only the first half of the 5-year experiment in progress.

4.1. Observed Earthquakes

Twelve earthquakes with magnitude greater than or equal to 4.95 were reported in the ANSS catalog in the RELM testing region during the first half of the ongoing 5-year experiment. Table 2 lists the properties of these target events. Among the details in Table 2 is the estimated independence probability for each earthquake, computed by a Monte-Carlo application (Schorlemmer and Gerstenberger, 2007) of the Reasenberg (1985) declustering algorithm. For example, the first target earthquake has an independence probability, P_I, of 21%, indicating that the

Table 2

Observed target earthquakes of magnitude $M_{ANSS} \geq 4.95$ in the testing area

No.	Origin Time (UTC)	Latitude	Longitude	M_{ANSS}	P_I	Mainshock
1	24 May 2006, 4:20	32.31	−115.23	5.37	0.21	Yes
2	19 Jul. 2006, 11:41	40.28	−124.43	5.00	1.00	Yes
3	26 Feb. 2007, 12:19	40.64	−124.87	5.40	1.00	Yes
4	9 May 2007, 7:50	40.37	−125.02	5.20	1.00	Yes
5	25 Jun. 2007, 2:32	41.12	−124.82	5.00	1.00	Yes
6	31 Oct. 2007, 3:04	37.43	−121.77	5.45	1.00	Yes
7	9 Feb. 2008, 7:12	32.36	−115.28	5.10	0.04	Yes
8	11 Feb. 2008, 18:29	32.33	−115.26	5.10	0.96	No
9	12 Feb. 2008, 4:32	32.45	−115.32	4.97	0.02	No
10	19 Feb. 2008, 22:41	32.43	−115.31	5.01	0.26	No
11	26 Apr. 2008, 06:40	39.52	−119.93	5.00	1.00	Yes
12	30 Apr. 2008, 3:03	40.84	−123.50	5.40	1.00	Yes

P_I denotes the independence probability as derived from Monte Carlo declustering simulations. The final column indicates whether the event is considered a mainshock by the REASENBERG (1985) declustering method with standard California parameters and is used to evaluate forecasts in the *5-year mainshock* group

declustering algorithm identified this earthquake as belonging to a cluster in 79% of the declustering iterations, each using a different, Monte Carlo-sampled set of algorithm parameters from a range of plausible values. The independence probabilities were used during evaluation of the *mainshock* and *mainshock.corrected* forecast group models; as mentioned in the previous section, the tests estimate the effect of observation uncertainties by generating modified catalogs, and the independence probability determines in what percentage of the modified catalogs a given earthquake appears.

For the *5-year mainshock* forecast class, only a subset of the events in Table 2 are considered. This subset is determined by applying the REASENBERG (1985) declustering algorithm to the original observed catalog, using standard California parameters. Those events that are not declustered are considered mainshocks and are used to evaluate the *5-year mainshock* forecasts.

An investigation of historical seismicity rates in the RELM testing region indicates that the observed sample of 12 earthquakes (with nine of them mainshocks) in a 2.5-year period is relatively small, but not significantly so. We analyzed the rate of all $M \geq 4.95$ earthquakes from 1 January 1932 to 30 June 2004 using the ANSS catalog. To compare with the experimental observation, we divided this time period into 29 non-overlapping periods of 2 years and 6 months duration; the rates in each period are shown in Fig. 4a. On average, 15.45 earthquakes (with 10.59 of them being mainshocks) were observed during each 2.5-year period, with a sample standard deviation of 9.99. As suggested by JACKSON and KAGAN (1999) (see also (VERE-JONES, 1970; KAGAN, 1973)), we found that the number of earthquakes in each period is better fit by a negative binomial distribution than a Poisson distribution—that is, the best-fit negative binomial distribution obtains a lower Akaike Information Criterion (AIC) value (AKAIKE, 1974) (206.4) than the best-fit Poisson distribution (278.2). The best-fitting negative binomial distribution also provides a marginally better fit to the mainshock rate distribution: the negative binomial model obtains an AIC value of 167.3, whereas the Poisson model obtains an AIC of 168.5. The seismicity rate data and the best fits are shown in Fig. 4b. We find the best-fit negative binomial distribution is described by parameter values $(\tau, \nu) = (2.83, 0.15)$; under this model, the probability to obtain fewer than 12 earthquakes is 41.01%. Accordingly, under the best-fit model for mainshock rates, the probability to obtain fewer than nine mainshocks is 32.91%. Despite our finding that the negative binomial distribution better fits historical rates of seismicity, RELM forecasts were formulated as having Poisson uncertainty, and therefore the tests applied to the models are based on Poisson statistics.

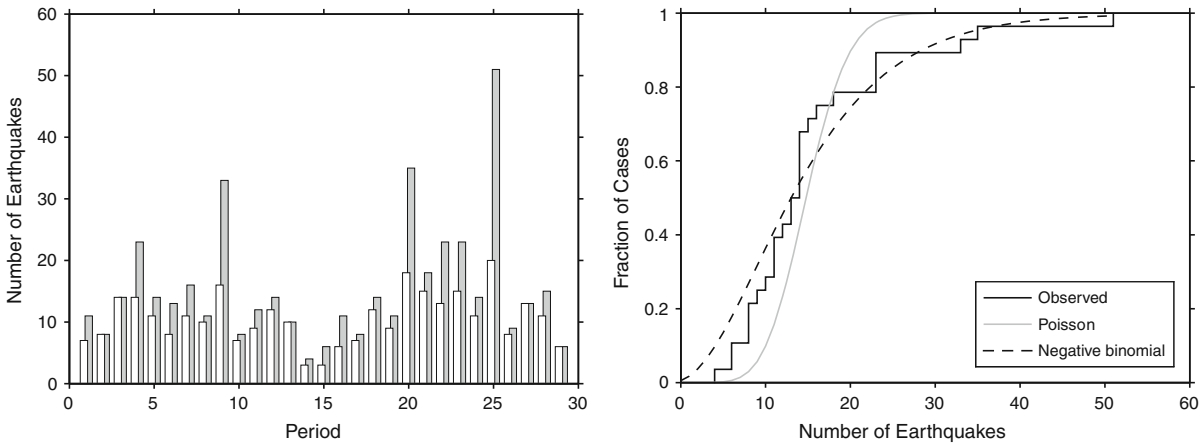

Figure 4
Earthquake rates in California from 1 January 1932 to 30 June 2004. (*left*) *Bar graph* showing the number of earthquakes in 29 non-overlapping periods of 2 years and 6 months duration. *White and gray bars* indicate the number of earthquakes in the declustered catalog, thus mainshocks only, and complete catalog, respectively. (*right*) Cumulative distribution function of the earthquakes rates in the complete catalog from the left frame. The *solid black line* indicates the observation, the *solid gray line* indicates the Poissonian distribution of rate $\lambda = 15.45$, the *dashed black line* indicates the best-fit negative binomial distribution

4.2. Mainshock Models

The summary results for the *mainshock* forecast class are given in Tables 3, 4, and 5. Table 3 lists the quantile scores for the L- and N-Tests. The RELM working group decided a priori to use a significance value of 5%; in the case of the two-sided N-Test, this corresponds to critical values of 2.5% and 97.5%; bold values in the tables indicate that the corresponding forecast is inconsistent with the observed target earthquake catalog. Recall that the γ quantile score, associated with the L-Test, describes how well a forecast matches the observed distribution of earthquakes. A very low γ score is means for rejecting a model, while a very high γ score is suspect, but not grounds for rejection. On the other hand, an extremely low or extremely high δ quantile score—characterizing the overall rate of earthquakes but not including any spatial information—yields rejection. From Table 3 we see that the observations during the first half of the RELM experiment are inconsistent—at the a priori significance level—with the HOLLIDAY-ET-AL.PI, WARD.COMBO81, WARD.GEODETIC81, WARD.GEOLOGIC81, and WARD.SEISMIC81 forecasts. All of these models have overpredicted in the first half of the experiment as indicated by their small δ values. (See also Fig. 5 for a visual comparison of predicted and observed number of earthquakes per model.)

Table 3

L-Test and N-Test results for the mainshock forecast class

Model	γ	δ
EBEL-ET-AL.MAINSHOCK	0.149	0.503
HELMSTETTER-ET-AL.MAINSHOCK	0.723	0.391
HOLLIDAY-ET-AL.PI	0.992	[**0.011**]
KAGAN-ET-AL.MAINSHOCK	0.974	0.063
SHEN-ET-AL.MAINSHOCK	0.969	0.107
WARD.COMBO81	0.998	[**0.004**]
WARD.GEODETIC81	1.000	[**0.000**]
WARD.GEODETIC85	0.987	0.030
WARD.GEOLOGIC81	0.998	[**0.011**]
WARD.SEISMIC81	0.993	[**0.014**]
WARD.SIMULATION	0.725	0.282
WIEMER-SCHORLEMMER.ALM	0.637	0.256

The statistics γ and δ measure the proportion of simulated likelihoods/numbers less than the observed likelihood/number. Bold values indicate that the observed target earthquake catalog is inconsistent with the corresponding forecast

Table 4 shows the contribution of each earthquake to the resulting likelihoods per model and highlights for each earthquake the model with the highest forecast rate in the respective bin—in other words, which model best forecast the earthquake. The WIEMER-SCHORLEMMER.ALM model provides the highest forecast rate for four earthquakes, and the

Table 4

Result details for the mainshock forecast class

Model		Earthquake								
		1	2	3	4	5	6	7	11	12
		$M5.37$	$M5.00$	$M5.40$	$M5.20$	$M5.00$	$M5.45$	$M5.10$	$M5.00$	$M5.40$
Ebel-et-al.Mainshock	λ	$9.55 \cdot 10^{-8}$	$3.56 \cdot 10^{-3}$	$3.39 \cdot 10^{-4}$	n/a	$1.15 \cdot 10^{-4}$	$7.64 \cdot 10^{-7}$	$9.55 \cdot 10^{-8}$	$5.74 \cdot 10^{-4}$	$9.55 \cdot 10^{-8}$
	L	-16.16	-5.64	-7.99		-9.07	-14.08	-16.16	-7.46	-16.16
Helmstetter-et-al.Mainshock	λ	$4.59 \cdot 10^{-3}$	$6.45 \cdot 10^{-3}$	$2.92 \cdot 10^{-4}$	$4.14 \cdot 10^{-3}$	$2.06 \cdot 10^{-4}$	$9.86 \cdot 10^{-4}$	$8.50 \cdot 10^{-3}$	$8.20 \cdot 10^{-5}$	$1.44 \cdot 10^{-4}$
	L	-5.39	-5.05	-8.14	-5.49	-8.49	-6.92	-4.78	-9.41	-8.85
Holliday-et-al.PI	λ	$1.85 \cdot 10^{-3}$	$4.66 \cdot 10^{-3}$	$1.85 \cdot 10^{-3}$	$2.94 \cdot 10^{-3}$	n/a	$1.47 \cdot 10^{-3}$	$3.70 \cdot 10^{-3}$	n/a	n/a
	L	-6.29	-5.37	-6.29	-5.83		-6.52	-5.60		
Kagan-et-al.Mainshock	λ	$3.57 \cdot 10^{-4}$	n/a	n/a	n/a	n/a	n/a	$7.12 \cdot 10^{-4}$	n/a	n/a
	L	-7.94						-7.25		
Shen-et-al.Mainshock	λ	$7.21 \cdot 10^{-4}$	n/a	n/a	n/a	n/a	n/a	$1.44 \cdot 10^{-3}$	n/a	n/a
	L	-7.24						-6.54		
Ward.Combo81	λ	$1.12 \cdot 10^{-3}$	n/a	n/a	n/a	n/a	n/a	$2.08 \cdot 10^{-3}$	n/a	n/a
	L	-6.80						-6.18		
Ward.Geodetic81	λ	$1.33 \cdot 10^{-3}$	n/a	n/a	n/a	n/a	n/a	$2.48 \cdot 10^{-3}$	n/a	n/a
	L	-6.62						-6.00		
Ward.Geodetic85	λ	$7.67 \cdot 10^{-4}$	n/a	n/a	n/a	n/a	n/a	$1.43 \cdot 10^{-3}$	n/a	n/a
	L	-7.17						-6.55		
Ward.Geologic81	λ	$9.76 \cdot 10^{-4}$	n/a	n/a	n/a	n/a	n/a	$1.82 \cdot 10^{-3}$	n/a	n/a
	L	-6.93						-6.31		
Ward.Seismic81	λ	$1.04 \cdot 10^{-3}$	n/a	n/a	n/a	n/a	n/a	$1.94 \cdot 10^{-3}$	n/a	n/a
	L	-6.87						-6.25		
Ward.Simulation	λ	$1.87 \cdot 10^{-4}$	n/a	n/a	n/a	n/a	n/a	$1.63 \cdot 10^{-5}$	n/a	n/a
	L	-8.59						-11.02		
Wiemer-Schorlemmer.ALM	λ	$5.47 \cdot 10^{-3}$	$5.17 \cdot 10^{-3}$	$3.45 \cdot 10^{-4}$	$2.48 \cdot 10^{-3}$	$1.47 \cdot 10^{-8}$	$1.64 \cdot 10^{-3}$	$1.01 \cdot 10^{-2}$	$2.54 \cdot 10^{-5}$	$4.32 \cdot 10^{-4}$
	L	-5.21	-5.27	-7.97	-6.00	-18.03	-6.41	-4.61	-10.58	-7.75

Contributions of each target earthquake to the log-likelihoods, L, and the forecast rate, λ, of each model for the corresponding bins are shown. For each earthquake, the model with the highest and lowest forecast for the respective bin is highlighted in light gray and dark gray, respectively. Some models do not provide a forecast for the entire space-magnitude testing area and some earthquakes fall into these masked bins, indicated by n/a. Earthquake numbers correspond to those listed in Table 2

Table 5

R-Test results for the mainshock forecast class

	Model	1	2	3	4	5	6	7
1	Ebel-et-al.Mainshock	–	**[0.000]**	**[0.000]**	**[0.000]**	**[0.000]**	**[0.000]**	**[0.000]**
2	Helmstetter-et-al.Mainshock	0.943	–	0.412	0.189	0.703	0.544	0.480
3	Kagan-et-al.Mainshock	0.965	**[0.000]**	–	**[0.010]**	0.326	0.369	**[0.000]**
4	Shen-et-al.Mainshock	0.944	**[0.007]**	0.783	–	0.964	0.586	**[0.000]**
5	Ward.Geodetic85	0.916	**[0.000]**	0.110	**[0.001]**	–	0.156	**[0.000]**
6	Ward.Simulation	0.939	**[0.000]**	**[0.001]**	**[0.001]**	**[0.002]**	–	**[0.000]**
7	Wiemer-Schorlemmer.ALM	0.547	**[0.000]**	0.130	0.123	0.799	0.614	–

All models which are consistent with the observation in the L- and N-Tests are compared and their corresponding α-values are shown. If printed in bold, the row model (labeled to the left) should be rejected in favor of the column model (labeled at the top). The results show that all models can be rejected in favor of the Helmstetter-et-al.Mainshock model

Helmstetter-et-al.Mainshock model has the highest forecast rate for three earthquakes. The Ebel-et-al. Mainshock and Holliday-et-al.PI models provide the highest forecast rate for one earthquake each.

The R-Test results for the *mainshock* forecast class are shown in Table 5 and provide a comparative evaluation of the forecasts. This table lists the α quantile scores for each pairwise comparison; for simplicity, we exclude the pairwise comparisons that would include the models shown to be inconsistent by the L- and/or N-Tests. Scores indicating that the corresponding model can be rejected are shown in bold. In this case, such a score indicates that the row model (labeled to the left) should be rejected in favor of the column model (labeled at the top). For example, the α value in the first row and second column indicates that the Ebel-et-al.Mainshock forecast should be rejected in favor of the

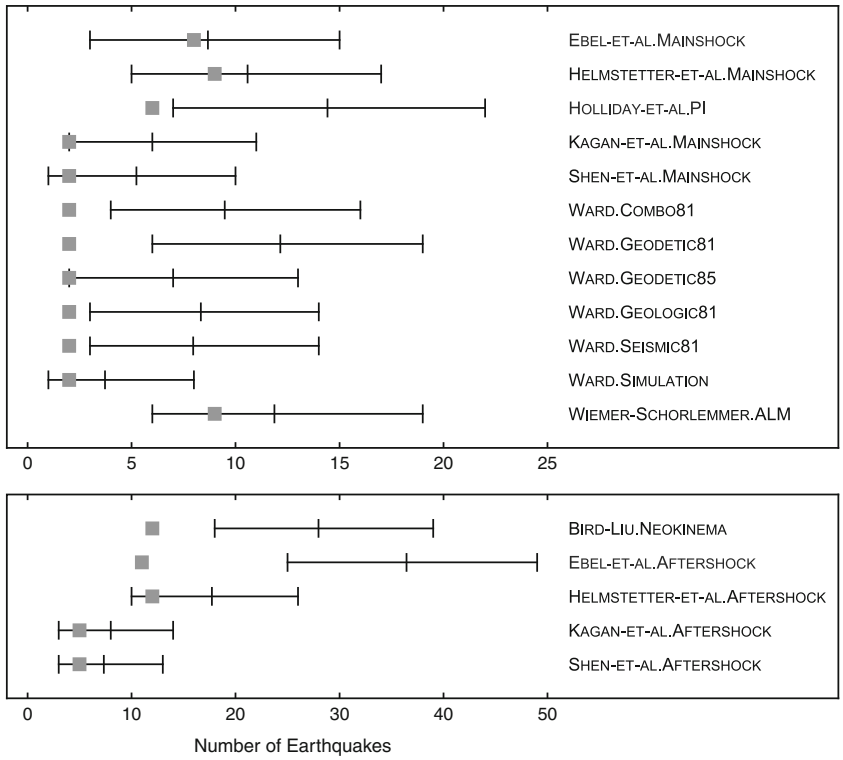

Figure 5
Visual comparison of predicted and observed number of earthquakes per model in the *mainshock* and *mainshock+aftershock* forecast classes. For each model, the *bar* indicates the range of observed earthquake rates that would be consistent with the model, given a Poissonian distribution. The *gray squares* indicate observations per model considering the coverage of the model. If the *gray square* overlaps with the *bar*, the model is consistent with the observation

HELMSTETTER-ET-AL.MAINSHOCK forecast. From this table, we find that only the HELMSTETTER-ET-AL.MAINSHOCK forecast is not rejected (because all other rows contain at least one bold value). Moreover, all models are rejected in favor of the HELMSTETTER-ET-AL.MAINSHOCK forecast (all scores in the second column are bold).

4.3. Mainshock Corrected

As mentioned in the Models section, the *mainshock.corrected* forecast group contains all the same forecasts as the *mainshock* forecast class with one exception: the EBEL-ET-AL.MAINSHOCK.CORRECTED forecast is added and implicitly replaces the EBEL-ET-AL.MAINSHOCK forecast. For consistency, the experiment for this forecast group began on 12 November 2006, so it contains only earthquakes 3–11 from Table 2. The summary results for this forecast group are shown in Tables 6 and 7. In this forecast group, the L- and N-Test results indicate that the observed earthquake distribution is consistent with all forecast models except the WARD.COMBO81 and WARD.GEODETIC81 models, which overpredicted the number of events (Table 6). The R-Test results are similar to the results for the *mainshock* forecast class and indicate that only the HELMSTETTER-ET-AL.MAINSHOCK forecast is not rejected in any pairwise comparison (Table 7).

4.4. Mainshock+Aftershock Models

The summary results for the *mainshock+aftershock* forecast class are shown in Tables 8, 9, and 10. N-Test results show that the BIRD-LIU.NEOKINEMA model and the EBEL-ET-AL.AFTERSHOCK model have each predicted too many earthquakes in the experiment to date (see also Fig. 5). The R-Test results

Table 6

L-Test and N-Test results for the mainshock.corrected forecast group

Model	γ	δ
Ebel-et-al.Mainshock	0.085	0.661
Ebel-et-al.Mainshock.Corrected	0.769	0.300
Helmstetter-et-al.Mainshock	0.434	0.613
Holliday-et-al.PI	0.984	0.042
Kagan-et-al.Mainshock	0.968	0.098
Shen-et-al.Mainshock	0.969	0.145
Ward.Combo81	0.998	[**0.015**]
Ward.Geodetic81	0.997	[**0.003**]
Ward.Geodetic85	0.984	0.058
Ward.Geologic81	0.992	0.028
Ward.Seismic81	0.990	0.034
Ward.Simulation	0.708	0.301
Wiemer-Schorlemmer.ALM	0.335	0.488

The statistics γ and δ measure the proportion of simulated likelihoods/numbers less than the observed likelihood/number. Bold values indicate that the observed target earthquake catalog is inconsistent with the corresponding forecast

show that only the Helmstetter-et-al.Aftershock forecast is not rejected in any pairwise comparison.

4.5. Mainshock+Aftershock Corrected

As with the *mainshock* and *mainshock.corrected* forecast groups, the *mainshock+aftershock.corrected* forecast group was added to the *mainshock+aftershock* forecast class. The Ebel-et-al.Aftershock.Corrected forecast is added and implicitly replaces the Ebel-et-al.Aftershock forecast. For consistency, the experiment for this forecast group began on 12 November 2006. The summary results for this forecast group are shown in Tables 11 and 12.

Table 8

L-Test and N-Test results for the mainshock+aftershock forecast class

Model	γ	δ
Bird-Liu.Neokinema	1.000	[**0.001**]
Ebel-et-al.Aftershock	1.000	[**0.000**]
Helmstetter-et-al.Aftershock	0.949	0.104
Kagan-et-al.Aftershock	0.895	0.193
Shen-et-al.Aftershock	0.896	0.262

The statistics γ and δ measure the proportion of simulated likelihoods/numbers less than the observed likelihood/number. Bold values indicate that the observed target earthquake catalog is inconsistent with the corresponding forecast

As in the *mainshock+aftershock* forecast group, the N-Test results show that the Ebel-et-al.Aftershock model has predicted too many earthquakes in the experiment to date, as has the Ebel-et-al.Aftershock.Corrected model. The R-Test results show that only the Helmstetter-et-al.Aftershock forecast is not rejected in any pairwise comparison.

5. Discussion

The science of earthquake predictability is an active field with many unsolved problems, including the question of best practices for formulating and

Table 7

R-Test results for the mainshock.corrected forecast group

	Model	1	2	3	4	5	6	7	8	9	10	11
1	Ebel-et-al.Mainshock	–	[**0.000**]	[**0.000**]	[**0.000**]	[**0.000**]	[**0.000**]	[**0.000**]	[**0.000**]	[**0.000**]	[**0.000**]	[**0.000**]
2	Ebel-et-al.Mainshock.corrected	0.840	–	[**0.003**]	0.406	0.089	0.034	0.278	0.270	0.385	0.445	0.085
3	Helmstetter-et-al.Mainshock	0.926	0.351	–	0.509	0.339	0.185	0.573	0.536	0.681	0.579	0.630
4	Holliday-et-al.PI	0.489	[**0.004**]	[**0.001**]	–	[**0.003**]	[**0.003**]	[**0.003**]	[**0.006**]	[**0.003**]	0.035	[**0.000**]
5	Kagan-et-al.Mainshock	0.886	0.333	[**0.012**]	0.527	–	0.045	0.453	0.409	0.477	0.478	[**0.007**]
6	Shen-et-al.Mainshock	0.869	0.440	[**0.025**]	0.529	0.676	–	0.974	0.576	0.711	0.654	[**0.010**]
7	Ward.Geodetic85	0.788	0.135	[**0.002**]	0.631	0.123	[**0.004**]	–	0.225	0.283	0.245	[**0.001**]
8	Ward.Geologic81	0.701	0.087	[**0.002**]	0.636	0.050	[**0.013**]	0.086	–	0.125	0.190	[**0.004**]
9	Ward.Seismic81	0.722	0.104	[**0.005**]	0.732	0.080	[**0.022**]	0.165	0.210	–	0.247	[**0.002**]
10	Ward.Simulation	0.761	[**0.001**]	[**0.000**]	[**0.010**]	[**0.004**]	[**0.001**]	[**0.009**]	[**0.009**]	[**0.005**]	–	[**0.000**]
11	Wiemer-Schorlemmer.ALM	0.473	[**0.000**]	[**0.000**]	0.286	0.134	0.138	0.600	0.539	0.679	0.651	–

All models are compared and their corresponding α values are shown. If printed in bold, the row model (labeled to the left) should be rejected in favor of the column model (labeled at the top). The results show that all models can be rejected in favor of model Helmstetter-et-al.Mainshock

Table 9

Result details for the mainshock+aftershock forecast class

Model		Earthquake 1 M5.37	2 M5.00	3 M5.40	4 M5.20	5 M5.00	6 M5.45	7/8 M5.10/5.10	9/10 M4.97/5.01	11 M5.00	12 M5.40
Bird-Liu.Neokinema	λ	$2.08 \cdot 10^{-3}$	$3.57 \cdot 10^{-3}$	$7.66 \cdot 10^{-4}$	$5.71 \cdot 10^{-3}$	$1.82 \cdot 10^{-3}$	$7.86 \cdot 10^{-4}$	$4.01 \cdot 10^{-3}$	$2.32 \cdot 10^{-3}$	$8.76 \cdot 10^{-5}$	$2.23 \cdot 10^{-4}$
	L	-6.18	-5.64	-7.18	-5.17	-6.31	-7.15	-11.74	-12.83	-9.34	-8.41
Ebel-et-al.Aftershock	λ	$3.43 \cdot 10^{-6}$	$1.70 \cdot 10^{-2}$	$1.45 \cdot 10^{-3}$	n/a	$5.48 \cdot 10^{-4}$	$2.74 \cdot 10^{-5}$	$3.43 \cdot 10^{-6}$	$2.74 \cdot 10^{-3}$	$2.74 \cdot 10^{-3}$	$3.43 \cdot 10^{-6}$
	L	-12.58	-4.09	-6.54		-7.51	-10.50	-25.86	-12.50	-5.90	-12.58
Helmstetter-et-al.Aftershock	λ	$7.71 \cdot 10^{-3}$	$1.14 \cdot 10^{-2}$	$4.90 \cdot 10^{-4}$	$7.13 \cdot 10^{-3}$	$3.63 \cdot 10^{-4}$	$1.63 \cdot 10^{-3}$	$1.48 \cdot 10^{-2}$	$4.03 \cdot 10^{-3}$	$1.45 \cdot 10^{-4}$	$2.41 \cdot 10^{-4}$
	L	-4.87	-4.49	-7.62	-4.95	-7.92	-6.42	-9.13	-11.73	-8.84	-8.33
Kagan-et-al.Aftershock	λ	$4.76 \cdot 10^{-4}$	n/a	n/a	n/a	n/a	n/a	$9.50 \cdot 10^{-4}$	$1.45 \cdot 10^{-3}$	n/a	n/a
	L	-7.65						-14.61	-13.77		
Shen-et-al.Aftershock	λ	$1.01 \cdot 10^{-3}$	n/a	n/a	n/a	n/a	n/a	$2.02 \cdot 10^{-3}$	$2.61 \cdot 10^{-3}$	n/a	n/a
	L	-6.90						-13.11	-12.59		

Contributions of each target earthquake to the log-likelihoods, L, and the forecast rates, λ, of each model for the respective bins. For each earthquake, the model with the highest and lowest forecast for the respective bin is highlighted in light gray and dark gray, respectively. Earthquakes 7 and 8 as well as 9 and 10 occurred in the same bin and are therefore combined in this table. Some models do not provide a forecast for the entire space-magnitude testing area and some earthquakes fall into these masked bins, indicated by n/a. Earthquake numbers correspond to those listed in Table 2

Table 10

R-Test results for the mainshock+aftershock forecast class

	Model	1	2	3
1	Helmstetter-et-al.Aftershock	–	0.372	0.091
2	Kagan-et-al.Aftershock	[**0.000**]	–	[**0.000**]
3	Shen-et-al.Aftershock	[**0.001**]	0.902	–

All models which are consistent with the observation in the L- and N-Tests are compared and their corresponding α values are shown. If printed in bold, the row model (labeled to the left) should be rejected in favor of the column model (labeled at the top). The results show that all models can be rejected in favor of model Helmstetter-et-al.Aftershock

Table 11

L-Test and N-Test results for the mainshock+aftershock.corrected forecast class

Model	γ	δ
Bird-Liu.Neokinema	0.984	0.027
Ebel-et-al.Aftershock	0.994	[**0.000**]
Ebel-et-al.Aftershock.Corrected	1.000	[**0.000**]
Helmstetter-et-al.Aftershock	0.692	0.394
Kagan-et-al.Aftershock	0.783	0.402
Shen-et-al.Aftershock	0.706	0.479

The statistics γ and δ measure the proportion of simulated likelihoods/numbers less than the observed likelihood/number. Bold values indicate that the observed target earthquake catalog is inconsistent with the corresponding forecast

Table 12

R-Test results for the mainshock+aftershock.corrected forecast group

	Model	1	2	3	4
1	Bird-liu.Neokinema	–	[**0.000**]	0.034	[**0.002**]
2	Helmstetter-et-al.Aftershock	0.067	–	0.433	0.159
3	Kagan-et-al.Aftershock	0.083	[**0.001**]	–	[**0.004**]
4	Shen-et-al.Aftershock	0.377	[**0.005**]	0.928	–

All models which are consistent with the observation in the L- and N-Tests are compared and their corresponding α values are shown. If printed in bold, the row model (labeled to the left) should be rejected in favor of the column model (labeled at the top). The results show that all models can be rejected in favor of model Helmstetter-et-al.Aftershock

evaluating earthquake forecasts. The RELM effort, as one of the first large-scale, prospective, and cooperative predictability experiments, can provide lessons along these lines. RELM experiment participants decided to specify their forecasts as the expected rate of earthquakes in latitude/longitude/magnitude bins, and they decided that the forecasts should be interpreted as having Poisson uncertainty. As we showed in the Observed Earthquakes subsection (and as shown by Jackson and Kagan, 1999), seismicity rates are better fit by a negative binomial distribution than a Poisson distribution; therefore it may be worthwhile for future forecasts to specify an additional parameter per bin (or per forecast) that allows for negative binomial uncertainty. Preferably, a forecast should specify a discrete probability distribution in each bin. This approach would not require the agreement of all participants on one particular distribution to be used for testing and it would also allow for propagating uncertainties of input data into the forecast (Werner and Sornette, 2008). The tests and forecast format that RELM decided to use are relatively simple yet

powerful. Nevertheless, they are not without flaws; for example the assumption that observations in each space-time-magnitude bin are independent may sometimes be violated, particularly in the wake of a large earthquake.

Some of these issues will be addressed by considering alternative forecast formats, e.g., by allowing models to specify the likelihood distribution to be used. Moreover, CSEP is incorporating modifications to the current tests and other tests, e.g., alarm-based tests that do not require a specific rate or uncertainty model (MOLCHAN, 1990; MOLCHAN and KAGAN, 1992; KAGAN, 2007; MOLCHAN and KEILIS-BOROK, 2008; ZECHAR and JORDAN, 2008).

The stability of RELM test results—including those presented here—is not easy to understand comprehensively. We made efforts to address stability of the L-Test by exploring a hypothetical predictability experiment. For a given forecast, we determined the bin with the lowest forecast rate, and we generated a modified catalog by adding to the observed catalog one additional event placed in this bin. This additional event represents the most unexpected occurrence according to the model, and we were curious to see if this one event could cause a forecast to be rejected if it otherwise was not rejected. We applied the L-Test to each forecast and the corresponding modified catalog and compared the resulting γ statistic with the observed γ reported in the tables throughout the Results section. We find that there is no simple relationship: some forecasts were rejected while others were not, and rejection depended on the peakedness of a forecast. For example, if a forecast has a very high ratio between its highest and lowest forecast values (i.e., it is very peaked), the most unexpected event has a much stronger effect on the L-Test result than otherwise. In other words, stability of test results is model-dependent, and this issue should be considered carefully in future experiments.

Another aspect of result stability is the duration of the experiment. Five years will most likely not be long enough for a comprehensive and final test result, as it can be questioned how representative the seismicity of these particular five years is. One effect of this problem can be seen in the results of the *mainshock* and *mainshock.corrected* forecast groups. While in the former group five models are rejected based on N-Test results, only two are rejected in the latter group. The exclusion of about 11 months from testing changes the L-Test considerably. However, the results of the R-Test suggest in both cases that the HELMSTETTER-ET-AL.MAINSHOCK cannot be rejected by any other model.

The fact that some forecasts masked a significant portion of the entire testing area led to the problem that eight of the twelve *mainshock* forecasts were tested against only two earthquakes. Four of these eight were rejected due to overpredicting the number of events. Although only two earthquakes occurred in the unmasked area, this low number indicates that the models are not consistent with the observation as the models expected far more events.

Although the RELM project ended in 2005, efforts to develop testing methods, implement these methods into Testing Center software systems, and expand the scope of experiments to other seismically active regions are ongoing, as is the experiment considered in this study. CSEP, the successor of RELM, took over the entire operation and development and is becoming a global reference project for earthquake predictability research.

Standardization can be considered one of the most important achievements of the RELM project and the Testing Center. The substantial consensus of RELM participants on the tests, rules, and processes is more than just a nucleus for other efforts. The Testing Center software is currently deployed to facilities in New Zealand, Europe, and Japan, and the rules set in California are adopted throughout all new Testing Centers. The next major step will become the unification of all efforts into a global testing program which was made possible only through the successful standardization.

Acknowledgments

This research was supported by the Southern California Earthquake Center (SCEC). SCEC is funded by NSF Cooperative Agreement EAR-0106924 and USGS Cooperative Agreement 02HQAG0008. Tests were performed within the W. M. Keck Collaboratory for the Study of Earthquake Predictability (CSEP)

Testing Center at SCEC, which was made possible by the generous financial support of the W. M. Keck Foundation. We thank the following for their contributions to RELM working group model development: J. A. Baglivo, P. Bird, K. W. Campbell, T. Cao, F. Catalli, D. W. Chambers, C. Chen, R. Console, A. Donnellan, J. E. Ebel, G. Falcone, A. D. Frankel, M. C. Gerstenberger, A. Helmstetter, J. R. Holliday, L. M. Jones, A. L. Kafka, Y. Y. Kagan, Z. Liu, M. Murru, M. D. Petersen, D. A. Rhoades, Y. Rong, J. B. Rundle, Z. Shen, K. F. Tiampo, D. L. Turcotte, S. N. Ward, and S. Wiemer. For stimulating discussion and various contributions to the RELM working group, we thank the following: M. Bebbington, D. Bowman, W. Ellsworth, K. Felzer, S. Hough, D. Sornette, R. Stein, M. Stirling, D. Vere-Jones, and J. Woessner. We thank F. Euchner, N. Gupta, V. Gupta, P. J. Maechling, and J. Yu for computational assistance. We also thank the open source community for the Linux operating system and the many programs used to create the Testing Center. We especially thank M. Liukis for her enthusiastic computational support and Testing Center development. Maps were created using the Generic Mapping Tools (Wessel and Smith, 1998). We thank the editor D. A. Rhoades, J. E. Ebel, and an anonymous reviewer for thoughtful reviews that enhanced the paper. M. J. Werner was supported by the EXTREMES project of the Competence Center Environment and Sustainability of ETH. The SCEC contribution number for this paper is 1230.

References

Akaike, H. (1974) *A New Look at the Statistical Model Identification*, IEEE Trans. Automatic Control *19*, 716–723.

Bird, P. and Liu, Z. (2007), *Seismic hazard inferred from tectonics: California*, Seismol. Res. Lett. *78*, 37–48, doi:10.1785/gssrl.78.1.37.

Console, R., Murru, M., Catalli, F., and Falcone, G. (2007), *Real time forecasts through an earthquake clustering model constrained by the rate-and-state constitutive law: Comparison with a purely stochastic ETAS model*, Seismol. Res. Lett. *78*, 49–56, doi:10.1785/gssrl.78.1.49.

Daley, D. J. and Vere-Jones, D. (2004), *Scoring probability forecasts for point processes: The entropy score and information gain*, J. Appl. Probab. *41A*, 297–312.

Ebel, J. E., Chambers, D. W., Kafka, A. L., and Baglivo, J. A. (2007), *Non-Poissonian earthquake clustering and the hidden Markov model as bases for earthquake forecasting in California*, Seismol. Res. Lett. *78*, 57–65, doi:10.1785/gssrl.78.1.57.

Field, E. H. (2007), *Overview of the working group for the development of regional earthquake likelihood models (RELM)*, Seismol. Res. Lett. *78*, 7–16, doi:10.1785/gssrl.78.1.7.

Gerstenberger, M. C., Jones, L. M., and Wiemer, S. (2007), *Short-term aftershock probabilities: Case studies in California*, Seismol. Res. Lett. *78*, 66–77, doi:10.1785/gssrl.78.1.66.

Harte, D. and Vere-Jones, D. (2005), *The entropy score and its uses in earthquake forecasting*, Pure Appl. Geophys. *162*, 1229–1253, doi:10.1007/s00024-004-2667-2.

Helmstetter, A., Kagan, Y. Y., and Jackson, D. D. (2007), *High-resolution time-independent grid-based forecast for $M \geq 5$ earthquakes in California*, Seismol. Res. Lett. *78*, 78–86, doi:10.1785/gssrl.78.1.78.

Holliday, J. R., Chen, C., Tiampo, K. F., Rundle, J. B., Turcotte, D. L., and Donnellan, A. (2007), *A RELM Earthquake forecast based on Pattern Informatics*, Seismol. Res. Lett. *78*, 87–93, doi:10.1785/gssrl.78.1.87.

Jackson, D. D. (1996), *Hypothesis testing and earthquake prediction*, Proc. Natl. Acad. Sci. USA *93*, 3772–3775.

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

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

Kagan, Y. Y. (1973), *Statistical methods in the study of seismic processes*, Bull. Int. Stat. Inst. *45*, 437–453.

Kagan, Y. Y. (2007), *On earthquake predictability measurement: information score and error diagram*, Pure Appl. Geophys. *164*, 1947–1962, doi:10.1007/s00024-007-0260-1.

Kagan, Y. Y. and Jackson, D. D. (1994), *Long-term probabilistic forecasting of earthquakes*, J. Geophys. Res. *99*, 13685–13700.

Kagan, Y. Y. and Jackson, D. D. (1995), *New seismic gap hypothesis: Five years after*, J. Geophys. Res. *100*, 3943–3959.

Kagan, Y. Y., Jackson, D. D., and Rong, Y. (2007), *A testable five-year forecast of moderate and large earthquakes in southern California based on smoothed seismicity*, Seismol. Res. Lett. *78*, 94–98, doi:10.1785/gssrl.78.1.94.

Molchan, G. M. (1990), *Strategies in strong earthquake prediction*, Phys. Earth Planet. Inter. *61*, 84–98, doi:10.1016/0031-9201(90)90097-H.

Molchan, G. M. and Kagan, Y. Y. (1992), *Earthquake prediction and its optimization*, J. Geophys. Res. *97*, 4823–4838.

Molchan, G. M. and Keilis-Borok, V. (2008), *Earthquake prediction: probabilistic aspect*, Geophys. J. Int. *173*, 1012–1017, doi:10.1111/j.1365-246X.2008.03785.x.

Petersen, M. D., Cao, T., Campbell, K. W., and Frankel, A. D. (2007), *Time-independent and time-dependent seismic hazard assessment for the State of California: Uniform California Earthquake Rupture Forecast Model 1.0*, Seismol. Res. Lett. *78*, 99–109, doi:10.1785/gssrl.78.1.99.

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

Rhoades, D. A. (2007), *Application of the EEPAS model to forecasting earthquakes of moderate magnitude in southern California*, Seismol. Res. Lett. *78*, 110–115, doi:10.1785/gssrl.78.1.110.

Schorlemmer, D. and Gerstenberger, M. C. (2007), *RELM Testing Center*, Seismol. Res. Lett. *78*, 30–36, doi:10.1785/gssrl.78.1.30.

Schorlemmer, D., Gerstenberger, M. C., Wiemer, S., Jackson, D. D., and Rhoades, D. A. (2007), *Earthquake Likelihood Model Testing*, Seismol. Res. Lett. *78*, 17–29, doi:10.1785/gssrl.78.1.17.

Shen, Z., Jackson, D. D., and Kagan, Y. Y. (2007), *Implications of geodetic strain rate for future earthquakes, with a five-year forecast of M 5 earthquakes in southern California*, Seismol. Res. Lett. *78*, 116–120, doi:10.1785/gssrl.78.1.116.

Vere-Jones, D. (1970), *Stochastic models for earthquake occurrence*, J. Roy. Stat. Soc. Series B (Methodological) *32*, 1–62.

Ward, S. N. (2007), *Methods for evaluating earthquake potential and likelihood in and around California*, Seismol. Res. Lett. *78*, 121–133, doi:10.1785/gssrl.78.1.121.

Werner, M. J., and Sornette D. (2008), *Magnitude uncertainties impact seismic rate estimates, forecasts, and predictability experiments*, J. Geophys. Res. *113*, B08302, doi:10.1029/2007JB005427.

Wessel, P. and Smith, W. (1998), *New, improved version of Generic Mapping Tools released*, EOS Trans. AGU *79*, 579.

Wiemer, S. and Schorlemmer, D. (2007), *ALM: An Asperity-based Likelihood Model for California*, Seismol. Res. Lett. *78*, 134–140, doi:10.1785/gssrl.78.1.134.

Zechar, J. D. and Jordan, T. (2008), *Testing alarm-based earthquake predictions*, Geophys. J. Int. *172*, 715–724, doi:10.1111/j.1365-246X.2007.03676.x.

Zechar, J. D., Schorlemmer, D., Liukis, M., Yu, J., Euchner, F., Maechling, P. J., and Jordan, T. H. (2009), *The Collaboratory for the Study of Earthquake Predictability perspective on computational earthquake science*, Concurrency and Computation: Practice and Experience, doi:10.1002/cpe.1519.

(Received August 14, 2008, revised February 21, 2009, accepted March 10, 2009, Published online May 11, 2010)