RELM Testing Center

D. Schorlemmer
ETH Zürich, Switzerland[1]

M. C. Gerstenberger
U.S. Geological Survey, Pasadena[2]

INTRODUCTION

We describe the organizational setup and the set of rules developed to test earthquake likelihood models of the Regional Earthquake Likelihood Models (RELM) initiative. Truly prospective testing for a multitude of forecasts in an unbiased and reproducible way requires a set of rules to be obeyed by each model as well as a cyber-infrastructure: the testing center. These rules encompass free parameters, *e.g.,* the testing area and grid, forecast periods, magnitude ranges, earthquake catalogs provided to the models and used for testing, and declustering. A combination of these free parameters defines a class of models. Within RELM we distinguish between five-year models, one-year models, and one-day models, which issue their forecast for the respective periods. While five-year models provide their forecasts as numbers, one-year and one-day models must be installed in the testing center together with their source code to enable yearly or daily forecast generation. Only with installed models can results be reproduced at any time.

The declared RELM goal of testing a multitude of forecast models in a prospective (forward-looking) sense is, to our knowledge, unique in seismology. It is the first time that a group of modelers agree to submit their models to a common, community-agreed test against future seismicity. We believe that this is an important milestone for forecasting- and prediction-related research in seismology because it offers an opportunity to overcome past shortcomings and deadlocks. The primary goal of the common testing is not to find the ultimate winner but to advance our current physical understanding and/or the statistical description of the way earthquakes occur. Ultimately this will lead in small steps (or, possibly, giant leaps) to better forecast models.

Prospective testing of models in seismology is rare but a few laudable examples exist (Kagan and Jackson 1995, Evison and Rhoades 1999, Rong *et al.* 2003). However, the testing of individual models by their proponents has three major shortcomings: (1) The credibility of a test is not guaranteed, because modelers act as testers. (2) Testing procedures are not necessarily validated or commonly accepted. (3) A comparison of models and testing results is not always possible, even if models are designed for the same region, because they use different time scales, magnitude ranges, or testing schemes. Progress in forecasting-related research has in our opinion been severely hampered by these factors, and it is for these reasons that we are setting up the framework of the RELM efforts as a first community-accepted set of tests and the testing center.

Schorlemmer *et al.* (2007, this issue) describe the theory of the RELM group (http://www.relm.org) testing procedure for grid-based probabilistic earthquake forecasts. Although the mathematical side of it is fairly straightforward, it is only a generic description and its implementation depends on the goals one wants to achieve with the test. A number of free parameters in the testing have to be specified. They include the classes of models, testing area and grid, declustering, etc. The so called "rules of the game" (*i.e.*, the rules that each model has to obey in order to be an accepted model) must be prescribed to allow for truly prospective, reproducible, and comparable testing. The free parameters need to be carefully chosen to preserve as much of the characteristics of the tested models as possible and to maximize the information content in the results. Location-specific definitions, such as catalog quality, testing area, etc. must be made.

In this paper we present the framework used for testing earthquake likelihood models developed within RELM. This framework is focused on the RELM goals: to create California-wide forecasts and test them against each other to improve the understanding of the underlying physics of earthquake generation and hazard calculation in general. The underlying principles of testing and the testing center, however, are general and could be applied to any region.

One of the important steps in defining any test of multiple models is the definition of model classes so that they group models of the same forecast type. A model might address the next hour, day, month, year, or decade, and these differences not only require different approaches in the testing center but also are of different value for society and are likely to be influenced by different fundamental physical processes. Comparisons of models across different classes are not possible: A model forecasting the next day cannot be fairly tested against a model that forecasts the next five years. The daily model has the advantage of knowing the most recent developments, but it has less value

1. Now at: University of Southern California, Department of Earth Sciences.
2. Now at: GNS Science, New Zealand.

for long-term hazard assessment because it does not allow for longer-term planning (one cannot change building codes on a daily basis). To have enough models to test within each class, it is mandatory that modelers agree on few model classes; within RELM, we have agreed upon three classes: long-term (five-year and one-year models) and short-term (one-day models).

All necessary rules must be implemented in a transparent way to make it possible to validate the testing procedures. To make the most robust and reproducible test possible, it is necessary to coordinate all tests of the models, centralize the evaluation of results, document each step of testing and the results, and remove the responsibility of individual modelers to provide each daily forecast.

TESTING CENTER

Numerous tests of models have already been performed by various scientists (*e.g.,* Molchan 1990, Molchan and Kagan 1992, Kagan and Jackson 1994, Jackson 1996, Molchan 1997, Vere-Jones 1998, Evison and Rhoades 1999, Console 2001, Rong 2002, Jones and Jones 2003, Helmstetter and Sornette 2003, Marzocchi *et al.* 2003, Daley and Vere-Jones 2004, Schorlemmer *et al.* 2004, Gerstenberger *et al.* 2004, Zechar and Jordan 2006). Researchers in atmospheric science also performed tests of tornado forecasts (see Jolliffe and Stephenson 2003 and references therein), which with some modifications can be used as blueprints for earthquake forecast tests. Usually, only one model was tested against a null hypothesis that had been chosen by the modeler. So far, no comprehensive comparison of different models has been performed; a framework for doing so has not been available to scientists. In the testing center, a large number of models will be tested together to determine their consistency with the observed data as well as to measure their relative performance against observed earthquakes. Additionally, these tests are designed to be fully prospective: they will be conducted for a total period of five years.

With the testing center, we gain several advantages:

Transparency. The testing center will catalog all data provided by modelers (*e.g.*, GPS data), by the testing center (*e.g.*, earthquake catalogs), or produced by the models (*e.g.*, forecasts). Based on the class of models, each forecast will either be provided as a set of numbers or as the computer code to produce the forecasts. All computer codes will be managed in a versioning system, *e.g.*, CVS (http://www.nongnu.org/cvs/) or Subversion (http://subversion.tigris.org). These systems store program codes and any changes to them. This setup enables us to document the code and track any potential changes to it (such as fixing a catastrophic bug). All procedures and program codes will be published on the testing section of the RELM Web site.

Controlled environment. All seismicity-based models will be provided with identical earthquake catalog data from the Advanced National Seismic System (ANSS) (http://www.anss.org) as an independent source to make the results comparable and to minimize data bias. The prospective tests will be performed with a delay of approximately three weeks to ensure the catalog data is as accurate as possible. The testing center is designed to minimize the possibility that any modeler can consciously, or unconsciously, bias his forecast to the earthquake data; all model procedures are installed in the testing center and are controlled by the testing center. Once in the testing center, the modelers have no access to their models. Additionally, these circumstances allow for truly prospective tests in a retrospective manner. Any data used for forecast generation must be supplied by the modelers before the starting time of the forecast. Such data will get a time-stamp so that the data may be reused in additional testing runs. The storage of these additional data sets is an important part of the documentation; within the testing center, it should be possible to regenerate a forecast, of any model, for any period, as long as the period is past the submission deadline of the models.

Comparability. All models within a class will not only be tested against observed data for consistency but also against each other. The testing procedure allows not only for comparing the overall results between models but also for additional, more detailed analyses, *e.g.*, spatial performance comparison for a limited area or for specific magnitude bands.

Reproducibility. A very important feature of the testing center is the ability to rerun tests at later times with either alternative options (*e.g.*, different magnitude ranges, different areas, etc.) or even with alternative tests. Also, with this flexibility, bugs in the testing procedure do not invalidate previous test results but simply require a rerun with improved code.

TESTING CLASSES

To compare models with different characteristics within one test, the models need to comply to one set of rules, or a testing class. The testing classes defined for RELM reflect both scientific and public interest and needs, including applications of forecasts. The classes correspond to long-term (five-year and one-year) and short-term (one-day) models. Long-term models are quasistationary and assume that earthquake rates are relatively stable over about a year or longer. They may actually be time-dependent models; however, the forecasts are stable over the forecasted period of one or five years. Short-term models are inherently time-dependent and assume that rates vary from day to day because of stress changes and other variations possibly resulting from past earthquakes.

Long-term models are relevant for public policy, construction planning, and setting insurance rates and priorities for remediation, all of which require seismic hazard estimates valid for years or more. Some long-term models are fundamentally time-dependent. For example, some renewal models assert that large-earthquake probability decreases substantially after a large event and only gradually recovers over a period of decades or centuries.

Short-term models are needed for emergency response and public information announcements. These models incorporate probabilities that depend on time and distance from previous earthquakes, usually exploiting power-law time dependence evident in aftershock sequences. It is difficult to apply these

TABLE 1
Classes of Models. The magnitude range's upper limit of 9 means that the last bin covers the magnitude range 8.95 ≤ M < 10. Any other magnitude bins are of the size of ΔM = 0.1

Class	Forecast Period	Aftershocks	Magnitude Range	Submission
I	5 years	Not included	$M \in [5; 9]$	Numbers
II	5 years	Included	$M \in [5; 9]$	Numbers
III	1 day	Included	$M \in [4; 9]$	Code
IV	1 year	Not included	$M \in [5; 9]$	Code
V	1 year	Included	$M \in [5; 9]$	Code

models in any fixed time interval because of the implicit scale invariance. Because earthquake rates change so fast after an earthquake, only an automatic algorithm for updating rates is adequate to implement the full time-dependence of these models.

Long-term Models

Long-term models issue their forecasts for a lowest magnitude of $\mathbf{M} = 5$ (expressed as a bin covering the magnitude range $4.95 \le \mathbf{M} < 5.05$). The maximum forecast magnitude is $\mathbf{M} = 9.0$ and the bin covers the magnitude range $8.95 \le \mathbf{M} < 10$ to allow for probabilities of events with very large magnitudes. Because most long-term models do not include aftershocks in their forecasts, one class allows for the removal of aftershocks from the test catalog; details of this procedure are provided in a later section. Modelers producing five-year models will not be required to submit their computer code; however, the forecasts are required to be submitted before the testing period begins.

A second class of long-term models (see table 1) will include aftershocks for the consistency test. Otherwise, the testing parameters of this class are the same as for the aforementioned class, and the tests will be carried out the same way.

A special case of long-term models are one-year models. They forecast quasistationary seismicity for one-year periods (see table 1) but have to be considered time-dependent as they change their forecast each year and do not issue their forecast for the full five-year period of testing. Because of this inherent time-dependence, these models also need to be installed in the testing center. As with the five-year models, including aftershocks is optional for the one-year models.

Short-term Models

Short-term models account for changes in seismic activity over time. Usually they adjust their forecast based on recent seismicity. Within the RELM framework, short-term models shall issue their forecast daily for a one-day period (see table 1). The magnitude range is extended down to $\mathbf{M} = 4$ for this class of models. The daily forecasts require that the models' codes are installed in the testing center and can be run independently of the modelers. Other than the different magnitude range, we will apply the same tests as we do to the long-term models.

Classes Setup (Numbers versus Code)

Any model that issues a forecast during the testing period needs to be installed as an executable computer code in the testing center. This allows us to reproduce any result or to recompute a test using a different testing routine. This setup enables the testing center to not only repeat previous computations but also to introduce a time delay in forecast generation to allow the models to use catalog data with reduced errors for their forecast generation. If the time-dependent models are not installed in the testing center, the modelers would have to provide their forecast on a daily basis. This is impractical in many respects.

TIME LINES (REVISED DATA)

All forecasts must be tested against revised seismicity data. There is no value in performing tests using flawed data to examine the performance of the models. Thus, tests need to be performed with a certain time delay. As shown in figure 1, a time-dependent model has to generate its forecast at the time t_0, using the catalog of the period before t_0. The model forecasts the seismicity of the period from t_0 to t_1. Neither the revised catalog of the learning period is available at t_0 nor the revised catalog for testing at t_1. Thus, any model has to wait until t_C for the revised catalog of the learning period to be able to generate its forecast (waiting period 1). Simultaneously, testing can only be performed at time t_T, when the revised catalog for the period t_0 to t_1 becomes available (waiting period 2). To simplify the procedure in the testing center, forecast generation and testing should be performed at t_T.

These time delays require that any model that generates its forecasts after the beginning of the testing period has to be installed at the testing center for automatic processing without modeler interaction. This allows the operators of the testing center to generate the model's forecasts at any time without introducing a possible bias, because the period for which the forecast is generated has already passed. Only with an installed model's code can a truly prospective testing be guaranteed, although the tests are performed with a delay.

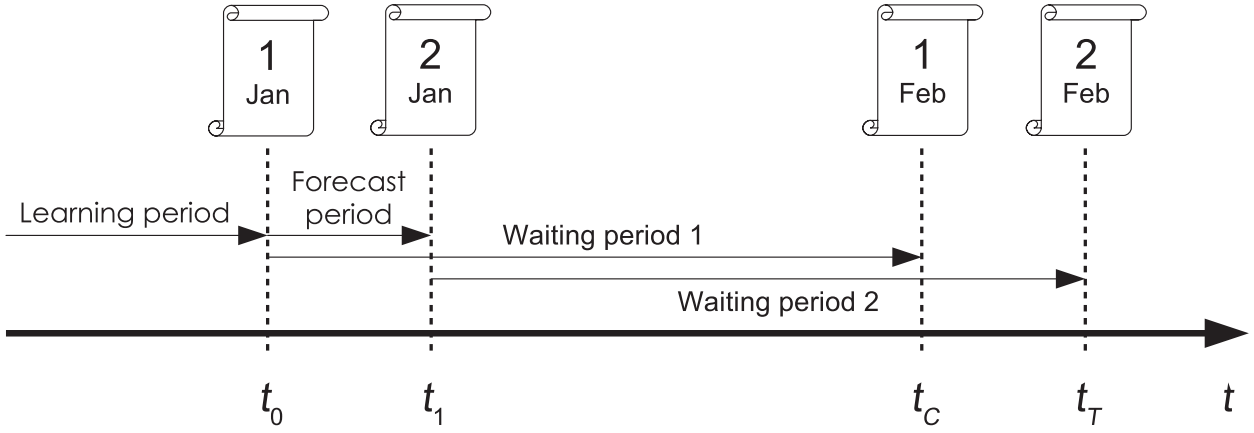
With the model's code installed at the testing center, we can perform truly prospective tests at any time for any period after the submission deadline. This ensures that we can recompute all tests in case of erroneous testing routines, catalog flaws, and changes or extension of the testing algorithm.

CALIFORNIA SPECIFICS

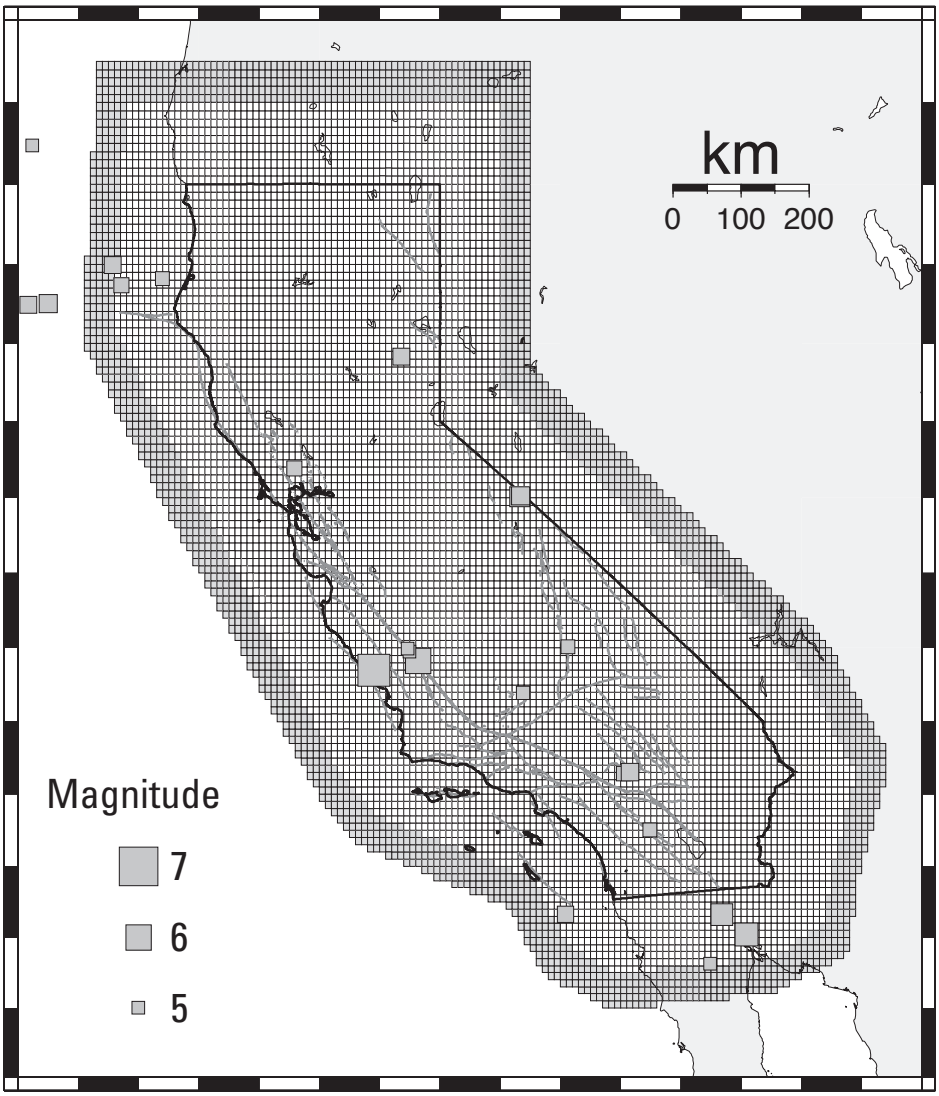
Testing Area and Collection Area

The testing area is designed to cover all of California and a boundary zone of about one degree around California (figure 2). The boundary zone is of importance for hazard calculations; events in this zone can potentially affect the hazard within California.

Any seismicity-based time-dependent model needs earthquake data for generating its forecast. This data is provided by the testing center and will cover the collection area, which is the testing area extended additionally by about 0.5 degrees. This



▲ **Figure 1.** Time line of forecasting and testing. Any model generates its forecast at t_0 using all data from the learning period. The forecast is valid for the period from t_0 to t_1. The revised catalog for forecast generation is available after the waiting period 1 at t_C. Testing can be performed after the waiting period 2 at t_T.



▲ **Figure 2.** Testing and collection area. The white squares indicate spatial cells of the testing area. The cells extending the testing area to the collection area are drawn in gray. Main faults are indicated with gray lines. The squares mark earthquakes of magnitude **M** ≥ 5 of the ANSS catalog in the period 2000–2005.

TABLE 2
Coordinates of the Polygon Encompassing the Testing Area

Latitude	Longitude
43.0	–125.2
43.0	–119.0
39.4	–119.0
35.7	–114.0
34.3	–113.1
32.9	–113.5
32.2	–113.6
31.7	–114.5
31.5	–117.1
31.9	–117.9
32.8	–118.4
33.7	–121.0
34.2	–121.6
37.7	–123.8
40.2	–125.4
40.5	–125.4

extension has been chosen because seismicity outside the testing area may influence seismicity inside the testing area.

The completeness of the ANSS catalog is $M_c \leq 3.7$ over the entire area. The polygons of the testing and collection areas are listed in tables 2 and 3, respectively. During the first hours of a large aftershock sequence, we can expect the completeness magnitude to be larger than 3.7; therefore we will only test events within three magnitude units of the main-shock magnitude for the first 12 hours following the main shock.

The Earthquake Catalog

We will use the ANSS catalog for testing. Information about errors in magnitude and location will be included in the testing as described in Schorlemmer *et al.* (2007, this issue). The catalog will be rebinned to $\Delta M = 0.1$. No focal mechanisms will be tested during the initial stage of the testing center. Potentially, we will add focal mechanism bins at a later time.

Declustering

In general, long-term models do not account for aftershock sequences. This creates a problem in testing because distinguishing between independent and dependent (aftershock) events is not an exact science. However, many attempts have been made to separate dependent events from independent, *e.g.*, by Gardner and Knopoff (1974) and Reasenberg (1985). Within the seismological community, the algorithm by Reasenberg 1985 is commonly used; however, there is no agreement on the parameter settings required by this method. Due to the fact that we do not know how to identify independent events, any parameter combination is seemingly as arbitrary as any other. To avoid using only one arbitrary parameter combination, we create more than 10,000 "declustered" catalogs by performing a Monte Carlo search over a limited parameter range. The

TABLE 3
Coordinates of the Polygon Encompassing the Collection Area

Latitude	Longitude
43.5	–125.7
43.5	–118.5
39.7	–118.5
36.1	–113.6
34.6	–112.6
34.3	–112.6
32.7	–113.1
31.8	–113.2
31.2	–114.5
31.0	–117.1
31.1	–117.4
31.5	–118.3
32.4	–118.8
33.3	–121.3
34.0	–122.0
37.5	–124.3
40.0	–125.9
40.5	–125.9
43.0	–125.7

TABLE 4
Parameters of the Declustering Algorithm by Reasenberg (1985) and their Respective Ranges Used for Generating Independence Probabilities for Each Event

Parameter	Range
r_{Fact}	5–20
x_k	0–1
τ_{min}	0.5–2.5
τ_{max}	3–15
P	0.9–0.99

boundaries for the parameters have been chosen based on simulation runs where we compared results obtained using the default parameters with results based on other parameter combinations. We limit the parameter range so that the declustered catalogs remain reasonable; all parameter ranges are listed in table 4.

Before declustering, we limit the catalog to the collection area and remove events below $\mathbf{M} = 3$ because the algorithm requires homogeneous completeness. The choice of $\mathbf{M} = 3$ as completeness magnitude is justified by the fact that potential aftershocks of this size can be considered fully detected or at least detected to a level where the algorithm can identify them as aftershocks. Given the principles in the declustering algorithm by Reasenberg (1985), aftershock sequences are "glued" together by smaller events. Thus, there is a tradeoff between the

ability to concat the aftershock sequence and the area covered because of spatial completeness variations. Using $M_c = 3.0$ as completeness magnitude allows for declustering of the entire collection area and enables sufficient "gluing" of aftershock sequences. From these 10,000 declustered catalogs, we compile one catalog in which each event is assigned an independence probability based on the number of catalogs that contained this particular event. This approach accounts for the inherent uncertainties in the parameter estimation, and the compiled catalog reflects these uncertainties with the independence probabilities.

CONCLUSIONS

With the recent advances and multitude of models becoming available in earthquake forecasting and prediction, it has become clear that a standard is necessary for forecast testing. Through the design of this testing procedure we have learned that a test must be statistically robust, not violate the spirit of the forecast, and be practical; these tasks are not easily combined together. For example, a 30-year hazard forecast may not necessarily be best served by a testing period of five years, a time period during which representative earthquakes might not occur. Unfortunately a longer testing period would be longer than the life of most models tested, not to mention the interest of many scientists. Additionally, a deterministic prediction does not necessarily fit into a testing scheme we have presented; however, with the probabilistic seismic hazard assessment (PSHA) and related goals of RELM, the testing as described provides a fair test for the largest number of models.

When reviewing the literature of earthquake forecast/prediction testing it is not always easy to interpret the results, nor is it easy to understand how particular results are relevant to other such predictions and tests. We have aimed to design a procedure and center that is a step toward easing these difficulties. The advantages of such a center are many:

- Establish an expandable framework for testing earthquake forecasts and predictions.
- Community-discussed and generally accepted testing procedure.
- Easily reproducible test results (with archived codes, forecasts and data).
- Test results that can be compared across multiple models.
- Real-time results available to the community via the Web.
- Ability to perform prospective tests retrospectively on models implemented within the test center; such models will be certified to be unbiased by new data.
- Extensively documented models, testing procedures and data.
- Freely available and usable testing code.
- Transparency and ability for outside evaluation of the testing procedure and code.

The testing center as it stands is only a first step. The framework established easily lends itself to the addition of needed components such as non-grid-based tests of fault-based models and tests of deterministic predictions (Jolliffe and Stephenson 2003, Zechar and Jordan 2006). With the testing center now established, we expect these and other additions to be added in the future and encourage other modelers to become involved.

ACKNOWLEDGMENTS

We want to thank E. Field for his longstanding effort to make RELM a successful project and for his support for the testing center. We also thank D. D. Jackson, L. Jones, and Y. Y. Kagan for their support and numerous debates. We profited from discussions with M. Bebbington, K. Felzer, A. Helmstetter, T. H. Jordan, P. Maechling, D. A. Rhoades, J. Rundle, D. Sornette, D. Vere-Jones, S. Wiemer, and J. Woessner. This paper is contribution 1471 of the Geophysical Institute, ETH Zürich. This research was supported by National Science Foundation Grant EAR-0409890 and by the Southern California Earthquake Center. SCEC is funded by NSF Cooperative Agreement EAR-0106924 and USGS Cooperative Agreement 02HQAG0008. The SCEC contribution number for this paper is 1036.

REFERENCES

Console, R. (2001). Testing earthquake forecast hypotheses. *Tectonophysics* **338**, 261–268.

Daley, D. J., and D. Vere-Jones (2004). Scoring probability forecasts for point processes: The entropy score and information gain. *Journal of Applied Probability* **41A** (Special issue SI), 297–312.

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

Gardner, J. K., and L. Knopoff (1974). Is the sequence of earthquakes in southern California, with aftershocks removed, Poissonian? *Bulletin of the Seismological Society of America* **64**(5), 1,363–1,367.

Gerstenberger, M. C., S. Wiemer, and L. Jones (2004). *Real-time forecasts of tomorrow's earthquakes in California: a new mapping tool.* USGS Open-File Report 2004-1390.

Helmstetter, A. and D. Sornette (2003). Predictability in the epidemic-type aftershock sequence model of interacting triggered seismicity. *Journal of Geophysical Research* **108**(B10), B2483, doi:10.1029/2003JB002485.

Jackson, D. D. (1996). Hypothesis testing and earthquake prediction. *Proceedings of the National Acadamy of Sciences of the United States of America* **93**, 3,772–3,775.

Jolliffe, I. T., and D. B. Stephenson, eds. (2003). *Forecast Verification*. Chichester, England and Hoboken, NJ: John Wiley & Sons Ltd.

Jones, R. H., and A. L. Jones (2003). Testing skill in earthquake predictions. *Seismological Research Letters* **74**, 753–760.

Kagan, Y. Y., and D. D. Jackson (1994). Long-term probabilistic forecasting of earthquakes. *Journal of Geophysical Research* **99**(B7), 13,685–13,700.

Kagan, Y. Y., and D. D. Jackson (1995). New seismic gap hypothesis: Five years after. *Journal of Geophysical Research* **100**(B3), 3,943–3,959.

Marzocchi, W., L. Sandri, and E. Boschi (2003). On the validation of earthquake-forecasting models: The case of pattern recognition algorithms. *Bulletin of the Seismological Society of America* **93**(5), 1,994–2,004.

Molchan, G. M. (1990). Strategies in strong earthquake prediction. *Physics of the Earth and Planetary Interiors* **61**, 84–98.

Molchan, G. M. (1997). Earthquake prediction as a decision-making problem. *Pure and Applied Geophysics* **149**, 233–248.

Molchan, G. M., and Y. Y. Kagan (1992). Earthquake prediction and its optimization. *Journal of Geophysical Research* **97**, 4,823–4,838.

Reasenberg, P. (1985). Second-order moment of central California seismicity, 1969–1982. *Journal of Geophysical Research* **90**(B7), 5,479–5,495.

Rong, Y. (2002). Evaluation of earthquake potential in China. Ph.D. diss., Univ. of California, Los Angeles.

Rong, Y., D. D. Jackson, and Y. Y. Kagan (2003). Seismic gaps and earthquakes. *Journal of Geophysical Research* **108**(B10), 2,471, doi:10.1029/2002JB002334.

Schorlemmer, D., S. Wiemer, M. Wyss, and D. D. Jackson (2004). Earthquake statistics at Parkfield: 2. Probabilistic forecasting and testing. *Journal of Geophysical Research* **109**(B12), B12308, doi: 10.1029/2004JB003235.

Schorlemmer, D., M. Gerstenberger, S. Wiemer, and D. D. Jackson (2007). Earthquake likelihood model testing. *Seismological Research Letters* **78**, 17–29.

Vere-Jones, D. (1998). Probability and information gain for earthquake forecasting. *Computational Seismology* **30**, 248–263.

Zechar, J. and T. H. Jordan (2006). Testing alarm-based earthquake prediction strategies (abstract). *Seismological Research Letters* 77, 258–259.

University of Southern California
Department of Earth Sciences
3651 Trousdale Parkway
Los Angeles, California 90089-0740 USA
danijel@usc.edu
(D.S.)

GNS Science
1 Fairway Drive
Avalon, P. O. Box 30-368
Lower Hutt, New Zealand
m.gerstenberger@gns.cri.nz
(M.C.G.)