Schorlemmer, D., Werner, M., Marzocchi, W., Jordan, T. H., Ogata, Y., Jackson, D. D., Mak, S., Rhoades, D. A., Gerstenberger, M. C., Hirata, N., Liukis, M., Maechling, P., Strader, A., Taroni, M., Wiemer, S., Zechar, J. D., & Zhuang, J. (2018). The Collaboratory for the Study of Earthquake Predictability: Achievements and Priorities. *Seismological Research Letters*, *89*(4), 1305-1313. https://doi.org/10.1785/0220180053

Seismological Research Letters

The Collaboratory for the Study of Earthquake Predictability: Achievements and Priorities

--Manuscript Draft--

Manuscript Number:	SRL-D-18-00053R2
Full Title:	The Collaboratory for the Study of Earthquake Predictability: Achievements and Priorities
Article Type:	Focus Section - CSEP: New Results and Future Directions
Corresponding Author:	Danijel Schorlemmer GFZ German Research Centre for Geosciences Potsdam, GERMANY
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	GFZ German Research Centre for Geosciences
Corresponding Author's Secondary Institution:	
First Author:	Danijel Schorlemmer
First Author Secondary Information:	
Order of Authors:	Danijel Schorlemmer
	Maximilian J. Werner
	Warner Marzocchi
	Thomas H. Jordan
	Yosihiko Ogata
	David D. Jackson
	Sum Mak
	David A. Rhoades
	Matthew C. Gerstenberger
	Naoshi Hirata
	Maria Liukis
	Philip J. Maechling
	Anne Strader
	Matteo Taroni
	Stefan Wiemer
	Jeremy D. Zechar
	Jiancang Zhuang
Order of Authors Secondary Information:	
Manuscript Region of Origin:	GERMANY
Suggested Reviewers:	Edward Field field@usgs.gov
	Seth Stein seth@earth.northwestern.edu
	Marco Pagani marco.pagani@globalquakemodel.org

Opposed Reviewers:	

The Collaboratory for the Study of Earthquake Predictability: Achievements and Priorities

Danijel Schorlemmer*[1], Maximilian J. Werner[2], Warner Marzocchi[3], Thomas H. Jordan[4], Yosihiko Ogata[5], David D. Jackson[6], Sum Mak[1], David A. Rhoades[7], Matthew C. Gerstenberger[7], Naoshi Hirata[8], Maria Liukis[9], Philip J. Maechling[4], Anne Strader[1], Matteo Taroni[3], Stefan Wiemer[10], Jeremy D. Zechar[11], and Jiancang Zhuang[5]

[1]GFZ German Research Centre for Geosciences, Potsdam, Germany
[2]University of Bristol, Bristol, UK
[3]Istituto Nazionale di Geofisica e Vulcanologia, Rome, Italy
[4]University of Southern California, Los Angeles, USA
[5]Institute for Statistical Mathematics, Tachikawa, Japan
[6]University of California Los Angeles, Los Angeles, USA
[7]GNS Science, Lower Hutt, New Zealand
[8]Earthquake Research Institute, University of Tokyo, Tokyo, Japan
[9]Jet Propulsion Laboratory, Pasadena, USA
[10]Swiss Seismological Service, ETH Zurich, Zurich, Switzerland
[11]Axis, Zurich, Switzerland

---

*Corresponding author

Abstract

The Collaboratory for the Study of Earthquake Predictability (CSEP) is a global cyberinfrastructure for prospective evaluations of earthquake forecast models and prediction algorithms. CSEP's goals are to improve our understanding of earthquake predictability, advance forecasting model development, test key scientific hypotheses and their predictive power, and to improve seismic hazard assessments. Since its inception in California in 2007, the global CSEP collaboration has been conducting forecast experiments in a variety of tectonic settings and at the global scale, and now operates four testing centers on four continents to automatically and objectively evaluate models against prospective data. These experiments have provided a multitude of results that are informing operational earthquake forecasting systems and seismic hazard models, and they have provided new, and sometimes surprising, insights into the predictability of earthquakes and spurned model improvements. CSEP has also conducted pilot studies to evaluate ground-motion and hazard models. Here, we report on selected achievements from a decade of CSEP, and we present our priorities for future activities.

Introduction

Earthquake forecasts and ground-motion models are the key ingredients to one of the most important products of seismological research: seismic hazard assessments. To better capture and assess the epistemic uncertainties of earthquake forecast models, the Southern California Earthquake Center (SCEC) and the United States Geological Survey (USGS) started the Regional Earthquake Likelihood Models (RELM) project. In the early 2000s, RELM initiated the development and rigorous prospective testing of a suite of such models for California [*Field*, 2007, and articles in the same special issue]. Each participating model's forecast was submitted to the testing group before 1 January 2006, the starting time of the 5-year prospective testing period. This concept of rigorous and prospective testing quickly gained support, and SCEC started the Collaboratory for the Study of Earthquake Predictability (CSEP) with funding provided by the W. M. Keck Foundation [*Jordan*, 2006]. Its first achievement was the development of the testing center software system [*Schorlemmer and Gerstenberger*, 2007; *Zechar et al.*, 2010b] for the RELM experiment [*Field*, 2007; *Schorlemmer et al.*, 2007; *Zechar et al.*, 2013; *Strader et al.*, 2017]. Over the following years, CSEP has expanded to four international testing centers that collectively test over four hundred models and model versions in a variety of tectonic settings and on a global scale. Besides California, testing centers are located in New Zealand [*Gerstenberger and Rhoades*, 2010], Japan [*Tsuruoka et al.*, 2012] and Europe [*Marzocchi et al.*, 2010], while a Chinese testing center is under development [*Mignan et al.*, 2013], see Figure 1. In 2011, the Global Earthquake Model (GEM) Foundation provided funds to develop procedures and metrics for evaluating intensity-prediction equations (IPEs), ground-motion prediction equations (GMPEs), and hazard models at a new testing center at the German Research Centre for Geosciences (GFZ) with the goal to integrate these in the CSEP framework. The centers have produced a plethora of results. Here, we present a selection of highlights and broader achievements from a decade of CSEP. We also outline CSEP's priorities for the future.

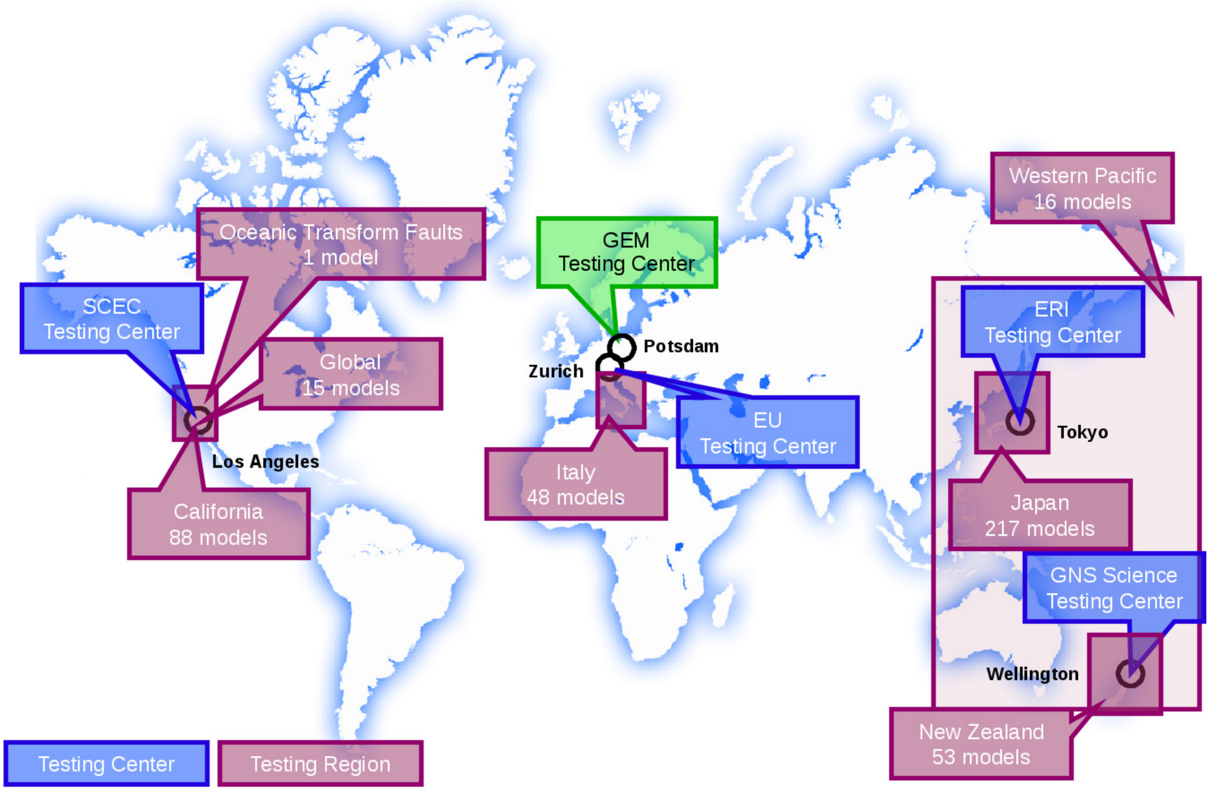

Figure 1: Map showing the locations of CSEP testing centers and testing regions. The SCEC testing center in Los Angeles is operating the testing regions of California, western Pacific, oceanic transform faults (in the Pacific) and the global experiment. The EU testing center in Zurich operates the testing region of Italy, the New Zealand testing center in Wellington the New Zealand experiment, and the Japan testing center the three testings regions in Japan. The GEM testing center in Potsdam develops ground-motion and hazard-related testing procedures and implemented case studies but, unlike the other centers, does not run earthquake forecast experiments.

The Philosophy behind CSEP

The fundamental idea of CSEP is simple in principle but complex in practice: forecasting models should be tested against future observations to assess their performance, thereby ensuring an unbiased test of the forecasting power of a model. The more common retrospective tests (testing a model's forecast against past data or parts of past data not used in the forecast) or pseudo-prospective tests (dividing past data into a learning dataset and an observational dataset so that time-dependent causality is preserved) bear the problem that features of the observations used for testing might have been known to the modeler and included in the model consciously or unconsciously.

The CSEP concept of prospective testing requires scientists to express their hypotheses and models quantitatively for testing against pre-agreed datasets, and to comply with agreed test procedures and metrics. For each experiment, the test area, its subdivision into spatial cells and magnitude bins, the type of forecast (usually number of earthquakes expected during a pre-defined period), the input data, the observations, and the metrics are defined through a community process: modelers have to fully specify (with zero degrees of freedom) their forecast according to standards. Observations come from authoritative sources, agreed upon in advance, and are used without any further or a posteriori interpretation by the modelers or testers, ensuring full independence from the testing process. The standardization also allows for comparative testing as all models participating in one experiment produce compatible forecasts, covering the same region, magnitude range, and testing period. Models producing time-varying forecasts are compiled and installed from source codes registered in the testing center to allow for automated and repeated forecast generation.

The CSEP approach showcases a wide range of plausible forecasts and their comparison. Previously, comparisons were often difficult because of the preferences of individual researchers for specific regions, testing periods, magnitude scales, or datasets. CSEP thereby elicits otherwise implicit assumptions and requires that abstract ideas are made concrete and testable, and reduces various cognitive inference biases (e.g. confirmation or hindsight bias). The history of earthquake prediction is riddled with controversies, disputes and biased inferences and although vigorous scientific debate continues, peer review is not sufficient to settle many of these disputes. CSEP has set an international standard for transparent, reproducible, and prospective experiments against the reproducibility crisis in science and created an infrastructure for more objective debates.

A Decade of CSEP: An Overview of Achievements

New Insights Into Earthquakes and Their Predictability

The longest-running experiment in CSEP covers the 5-year RELM forecasts for California. This experiment has been continued with unchanged forecasts after the initial 5-year period (1 January 2006–1 January 2011). It provided evidence that the locations of past shocks, particularly the many small (M2+) ones recorded by dense networks, can contain more predictive skill of moderate to strong earthquakes over a 5- to 10-year period than many other forecast approaches, including geological (fault-based), geodetic, and tectonic models [*Schorlemmer et al.*, 2010c; *Zechar et al.*, 2013; *Strader et al.*, 2017]. One of the participating forecasts, the Uniform California Earthquake Rupture Forecast version 2 (UCERF2), is particularly important because it provided government agency hazard estimates that set

Californian building codes and insurance rates, and underlies catastrophe models [*Field et al.*, 2009]. UCERF2 was consistent with observed moderate-to-strong seismicity during 2007–2016 and had greater forecast skill than most other RELM forecasts [*Strader et al.*, 2017]. Evaluation of the new UCERF3 [*Field et al.*, 2014, 2015, 2017] will be a major future CSEP activity.

Models based on geodetic strain-rate data have shown promise. The RELM forecast by *Shen et al.* [2007] for southern California was about as informative as UCERF2 in forecasting M5+ shocks. The strongest evidence, however, is based on two years of testing global forecasts: the GEAR1 model [*Bird et al.*, 2015], a hybrid model of the global strain rate map and smoothed seismicity, outperformed both of its individual components (Strader et al., this issue). Retrospective test results from New Zealand also support the predictive skill of strain-rate data converted to seismicity rates [*Rhoades et al.*, 2017].

CSEP is testing statistical clustering models in California, Italy, New Zealand, Japan, and globally. Multiple versions of the Epidemic Type Aftershock Sequence (ETAS) models demonstrated reliable forecasts of the 2011 M9 Tohoku earthquake sequence [*Nanjo et al.*, 2012; *Ogata et al.*, 2013]. Importantly, measured probability gains during major aftershock sequences are consistent with theoretical gains of two to three orders of magnitude over time-independent models [*Taroni et al.*, this issue; *Cattania et al.*, this issue; *Woessner et al.*, 2011; *Rhoades et al.*, this issue]. CSEP also identified the most skillful version of the Every Earthquake a Precursor According to Scale (EEPAS) model, that is based on the precursory scale increase phenomenon, during the period 2009–2012 in California [*Schneider et al.*, 2014] and 2009–2017 in New Zealand [*Rhoades et al.*, this issue].

Physics-based models, i. e. models that use physical concepts like rate-and-state [*Di-*

eterich, 1994] behavior or Coulomb-stress changes [*King et al.*, 1994] for forecasting rather than being based purely on statistics, have drawn a lot of attention in the past decade. The performance of the first generation of such models of aftershock sequences was poor in a retrospective evaluation during the 1992 M7.3 Landers earthquake sequence [*Woessner et al.*, 2011]. The authors concluded that Coulomb/rate-state models [e. g., *Stein*, 1999] were substantially less informative than several ETAS and STEP [*Gerstenberger et al.*, 2005] models. Subsequent model development, however, has led to dramatic improvements: the second generation of Coulomb-based models suggests much improved skill and reliability in a retrospective test of the 2010–2012 Canterbury, New Zealand, earthquake sequence [*Cattania et al.*, this issue]. These results are encouraging for the prospects of physics-based forecasting.

One of the main CSEP priorities for the future is to test also ground-motion and seismic hazard models. A pilot study has explored the feasibility to carry out CSEP-type experiments in these domains. The analysis on IPEs in Italy showed that the global model by *Allen et al.* [2012] performed well for Italian earthquakes, comparable to the best local model. Among the local models, some newer models based on more data did surprisingly not perform better than older ones based on the same functional form [*Mak et al.*, 2015]. This is contrary to the belief that using more and newer data per se will necessarily lead to better models, underlining the need for future independent and prospective testing experiments. A similar observation was made in the GMPE pilot study in Japan, where the newest NGA-West2 global model [*Campbell and Bozorgnia*, 2014] outperformed pre-NGA local models on which the Japanese hazard model is based [*Mak et al.*, in press], supporting again the notion of testing rather then assuming that models created specifically for local conditions are per se better.

The final element in the chain are hazard models. While site-specific hazard is often the focus of many applications, *Mak et al.* [2014] showed that the statistical power of testing a site-specific hazard model is in general very low and thus only a regional hazard model can be meaningfully tested. Testing the last four US National Seismic Hazard Maps [*Petersen et al.*, 2014] in a prospective sense, *Mak and Schorlemmer* [2016] showed in their pilot study that in the central and eastern US the model is consistent with observed peak-ground-acceleration (PGA) and spectral accelerations (SA) at 1s, while in California the model is consistent with the observation for PGA but overpredicts the hazards for SA at 1s. However, given the long forecasting horizon of the hazard models, long-term testing is needed to increase the power of these results.

New Insights Into Model Evaluation Methods

CSEP developed a suite of new, community-endorsed tests and metrics that probe forecasts from different perspectives, and identify strengths and weaknesses by highlighting discrepancies between forecast and data [*Schorlemmer et al.*, 2007; *Zechar et al.*, 2010a; *Werner et al.*, 2011]. Some initially promising tests have been replaced by others [e. g., *Rhoades et al.*, 2011]. CSEP stimulated innovation in performance metrics, e.g. those based on point process residuals [*Clements et al.*, 2011; *Gordon et al.*, 2015], gambling and betting frameworks [*Zhuang*, 2011; *Zechar and Zhuang*, 2010, 2014], and an extension of Molchan error diagrams [*Zechar and Jordan*, 2008]. Strengthening the evaluation methods further remains a CSEP priority [e. g., *Werner and Sornette*, 2008; *Lombardi and Marzocchi*, 2010; *Molchan et al.*, 2017].

CSEP stimulated new ensemble modeling techniques, which aim to combine multiple forecasts optimally to exploit complementary strengths. Techniques include Bayesian model averaging and other additive models [*Marzocchi et al.*, 2012; *Taroni et al.*, 2013], as well as multiplicative models [*Rhoades et al.*, 2014; *Bird et al.*, 2015]. Ensemble models can also express epistemic uncertainty arising from data incompleteness, parameter uncertainty, and model uncertainty. For example, *Omi et al.* [2015] concluded that Bayesian ensemble forecasts were more reliable than forecasts that did not consider epistemic uncertainty.

In the hazard domain, a new metric for GMPE testing has been proposed, based on the widely used LLH score [*Scherbaum et al.*, 2009]. It is applicable to model GMPEs with complicated correlation structure [*Mak et al.*, 2017]. *Mak and Schorlemmer* [2016] also applied a formal test of the number of exceedances to hazard forecasts, paving the way to future hazard-testing expriments within the CSEP framework.

Future CSEP Activities

CSEP activities during the next decade will be guided by three main objectives: expanding the data space, expanding the model space, and testing key hypotheses and questions.

(1) Expanding the data space. The main limitation in the testing of earthquake forecasts is the lack of data. CSEP will extend spatial coverage by encouraging forecast testing in other regions with good earthquake catalogs (e. g., seismic belts of Asia and South America), as well as globally. It will extend temporal coverage by expanding its retrospective testing capabilities to take advantage of well-recorded aftershock sequences and other datasets, including information on large, infrequent earthquakes from pre-instrumental historical and

paleoseismic observations.

Another limitation is the data quality, i. e. the errors in the estimates of occurrence times, epicenter locations, and magnitudes, but also missing small events in earlier periods of aftershock sequences and in places of low earthquake detectability. CSEP analyzed data quality in test regions [*Schorlemmer et al.*, 2010b,a, 2018] but is still in need of models to assess the difference between catalogs and actual seismicity. Such models can quantify uncertainties in model evaluations.

Finally, CSEP will address the important question of the minimum duration of an experiment to derive conclusions about model performances with sufficient power. While some models can be rendered wrong with an earthquake considered impossible by the model, positive statements about model performances, in particular of long-term models, can technically only be made after the forecasting period has passed completely. Such an approach is not feasible for e. g. 50-year models and a shorter but sufficient period needs to be determined for meaningful and practical tests. This question touches on the practical limits of testability of models and will involve the developments of alternative approaches like component-based testing of models or model reformulations to match observables that can be obtained.

(2) Expanding the model space, focusing on new types of forecasts. Earthquake forecasting is a rapidly growing scientific endeavor, motivated by the needs of long-term PSHA and shorter-term operational earthquake forecasting (OEF). CSEP will promote this research by striving to test the most advanced and innovative earthquake forecasts.

3D models CSEP has thus far evaluated epicentral forecasts of shallow earthquakes, rather than hypocenter distributions. However, 3D forecasts are needed to assess hypotheses and seismic hazard in structurally complex tectonic settings, such as subduction zones.

The 3D Kanto experiment provides a blueprint for such activities. It covers the densely-populated metropolitan area of Tokyo down to depths of 100km, where three tectonic plates meet. Interactions among the inter-plate and intraplate earthquakes are not well captured in 2D, and preliminary results show an advantage of 3D models [*Tsuruoka*, 2017].

Ensemble forecasting Recent studies on hybrid/ensemble models of several different types (additive, multiplicative, maximum, and using different weighing schemes) concluded that these models can sometimes outperform individual models based on a single idea or data source [*Rhoades and Gerstenberger*, 2009; *Rhoades and Stirling*, 2012; *Marzocchi et al.*, 2012; *Taroni et al.*, 2013; *Rhoades*, 2013; *Steacy et al.*, 2014; *Rhoades et al.*, 2014, 2015, 2016, 2017], and are never much worse than the best individual model, which is not known a priori. CSEP will support methods to test combinations of two or more existing models or to assimilate new gridded covariates into existing models. Likewise, component-based combinations (e. g. taking the smoothing kernel of one model and the spatial magnitude distribution of another model) can be explored, either through ensemble techniques or on the model source-code level to improve capturing of model uncertainties.

Fault-based models Models that explicitly incorporate known faults are thought to provide better long-term forecasts than models lacking such information [*Field et al.*, 2009]. Fault-based models rely on fault geometry to forecast large fault ruptures. The *association problem*, matching of a future observed rupture with a specific hypothetical rupture, is currently unsolved because finite ruptures are not consistently reported by a community-agreed independent source. Thus to compare future earthquakes against fault-based models like UCERF3 [*Field et al.*, 2014], CSEP will need to develop new methods.

Event-based models CSEP models forecast earthquake rates in each space-time-magnitude bin independently of the earthquakes in all other bins assuming a Poisson distribution. It has been recognized early that earthquake occurrence is clustered and does not follow a Poisson distribution [*Schorlemmer et al.*, 2007]. Clustering implies that earthquakes are not independent of previous events. In Japan, 1-year forecasts became meaningless after the 2011 Tohoku earthquake because its triggered events dominated the seismicity. This dependency can be accounted for by models and experiments that allow forecast updates after each event, in contrast to regular time intervals.

Physics-based models A major CSEP objective is to improve forecasting accuracy by harnessing the explanatory power of rupture physics. The Canterbury experiment [*Cattania et al.*, this issue] also highlighted the difficulties of prospectively testing stress-transfer models that must be updated with slip models during a seismic sequence. Further experiments using well-recorded aftershock sequences are planned. On a different scale, simulators like RSQSim [*Dieterich and Richards-Dinger*, 2010; *Richards-Dinger and Dieterich*, 2012] are employing rupture physics and are capable of simulating very long (more than a million years) earthquake catalogs that are, in principal, suitable for producing time-dependent forecasts on all relevant time scales. This will require the inclusion of off-fault seismicity and, more important, schemes for initializing the fault-system simulations with stress states consistent with the observed earthquake history, which is a difficult, unsolved problem. Testing such forecasts will also require a solution to the association problem.

Complete probabilistic models A proper model validation requires a full description of all uncertainties [*Marzocchi and Jordan*, 2014, 2017]. CSEP will overcome these limitations by considering a more complete description of a model's forecast, allowing it to specify not only the expected number in each bin but also the distribution of the number of target earthquakes in each bin and the correlations between bins to account for epistemic uncertainties. A wider range of test statistics, describing various features of the earthquake process, will also be possible in this framework.

Ground-motion and hazard models Testing ground-motion models will need to extend the association problem with more rupture-specific parameters provided by an authoritative source. Similar to the complete probabilistic models, testing hazard models needs to take into account spatial (and temporal, for time-dependent hazard models) correlations of models. These correlations will be included in the test, especially for hypothesis tests with well-defined mathematical meaning. The first step will be a test of the Japanese national seismic hazard model.

Precursor models Some studies concluded that geodetic and electromagnetic anomalies can be exploited for earthquake forecasting, even though the information gain is low [*Zhuang et al.*, 2005]. Tailored, prospective experiments are necessary for an assessment of forecast improvements through possible precursory models.

External forecasts Thus far, CSEP has been evaluating internal forecasts, namely those generated by model software compiled and installed within its testing centers. This ensures reproducibility and transparency within a controlled environment, and means that the model under evaluation is not a moving target. However, CSEP also aims to support the evaluation of select External Forecasts and Predictions (EFPs), such as operational forecasts issued (elsewhere) by government agencies or predictions from precursor models that cannot be installed within CSEP. External forecasts and predictions seldom fit the requirements of CSEP forecasts. Solution are to 'collapse' CSEP forecasts to the same format of the external forecasts or to tailor an experiment to the forecast. This will require automated transfer protocols for verified and unambiguous forecasts and predictions, along with versioning of underlying models to document model changes. CSEP's internal models can serve as benchmarks. However, the problem of possible biases of non-documented forecasts remain.

(3) Testing key hypotheses and questions. Formal testing provides a valuable tool for probing, improving, and possibly discarding fundamental assumptions about earthquake behaviour. Many scientific questions could be refined by carefully formulated forecast models, especially if a tailored experiment is specified simultaneously.

- *Are big earthquakes fundamentally different from smaller ones in their clustering, scaling behavior or long-term behavior?* Scaling relations between rupture dimensions and moment often suggest a break at a certain magnitude, presumably related to seismogenic depth. How can these observations be exploited to improve predictive skill? Regional and global tests against a null hypothesis could help answer these questions.

- *What is the magnitude distribution of earthquakes on a single master fault?* A Gutenberg-Richter distribution, or something else? Do on-fault and off-fault earthquakes have the same size limits? Effective tests would require a good definition of 'on-fault' over a region and suffcient time to supply large on-fault events.

- *Elastic rebound?* Do large mainshocks reduce the probability of other ones nearby (rebound model), or do they increase the probability preferentially (traditional ETAS model)?

- *Are moderate earthquakes more likely to trigger big ones if they are near 'ripe' major faults?* If so, how much more likely? Can we identify 'sleeping giants', or places where prior probability is high? As above, large regions and suffcient time would be required.

- *Do b-values (or other features of relative magnitude distribution) as a possible proxy to stress have predictive power?* Do they help forecast locations and focal mechanisms of future events? Tailored experiments on b-value anomalies could provide an analysis of the change in forecasting power when including b-values.

- *Is the location of small earthquakes the best predictor of the location of coming bigger ones?* Or do rate-state Coulomb models add significant new information? This question has been pursued in aftershock studies, with improved results [*Cattania et al.*, this issue]. In Japan, inland background seismicity rates of the HIST-ETAS model [e. g., *Ogata*, 2011, 2017] correlate well with future and historical (599-1884) large earthquakes. Challenges include approximating the initial stress conditions, and accurately modeling the stresses. Because each event changes the conditions, forecasts must adapt automatically without human interaction.

- *Can foreshocks be discriminated?* One way to solve this question is by combining an existing space-time forecast model with a magnitude-frequency model of a foreshocks forecast [*Ogata and Katsura*, 2014; *Nomura and Ogata*, this issue] for comparison with an independent Gutenberg-Richter magnitude sequence. Another way would be in a tailored test to assign each event a foreshock probability and compare it with future activity.

Conclusions

CSEP is building a community of earthquake forecasting researchers, who share data, models, ideas, and evaluation approaches. CSEP has set an international standard for conducting forecast experiments and evaluating the predictive power of models and hypotheses. Through insistence on prospective testing, quantitative metrics, independent authoritative data streams, transparency, and reproducibility, CSEP has reduced subjective biases from evaluations of earthquake forecast models and prediction algorithms. This has inspired other communities to follow suit, including induced seismicity [e. g., *Király-Proag et al.*, 2016] and earthquake early warning [*Böse et al.*, 2014].

CSEP has also explored the current limits of predictability and of testing forecasts or their components. Meaningful evaluations of hypotheses about the long-term behavior of large earthquakes may take decades or centuries in regional fault systems, necessitating global models for testing hypotheses such as characteristic earthquakes, segmentation, and quasi-periodic recurrences. Such hypotheses inform important seismic hazard models in California, Italy, Japan, and Europe; however, the dearth of large earthquakes in individual regions is a major limitation of evaluations. For the same reason, models of expected maximum magnitude on a fault (segment) are not readily testable [*Holschneider et al.*, 2011, 2014].

Despite these fundamental problems, CSEP's model evaluations have influenced and improved seismic source models for hazard estimates. In California, the performance of the *Helmstetter et al.* [2007] RELM model led to the inclusion of adaptive smoothing of the locations of small quakes in UCERF3 [*Field et al.*, 2014], while the demonstrated skill of the ETAS model class underpins the UCERF3-ETAS model [*Field et al.*, 2017]. In New Zealand, short-term and medium-term models under CSEP evaluation were used to provide operational forecasts and hazard estimates during and after the 2010–2012 Canterbury and 2016 Kaikoura sequences [*Gerstenberger et al.*, 2014, 2016; *Rhoades et al.*, 2016]. In Japan, real-time aftershock forecasts at the National Research Institute for Earth Science and Disaster Resilience in Japan provide information for the government [*Omi et al.*, 2016]. Finally, the Italian OEF system for the Civil Protection Agency employs an ensemble of CSEP-tested models [*Marzocchi et al.*, 2014; *Iervolino et al.*, 2015]. These examples suggest that CSEP evaluations are leading to safer and better informed societies through dynamic earthquake probabilities, and a better decision-making basis for building codes and retrofitting priorities.

Data and Resources

No data were used in this paper.

Acknowledgments

The authors would like to thank Peter Bird, Zhigang Peng, and an anonymous reviewer for their helpful comments to improve the paper. We also want to thank the wider CSEP community for their participation and all their work. Finally, we thank the open source community for providing many tools used in this work.

CSEP was established under a grant from the W. M. Keck Foundation and has been supported by the Southern California Earthquake Center under NSF Cooperative Agreement EAR-1033462 and USGS Cooperative Agreement G12AC20038. SCEC contribution number 8036. This work was supported by the New Zealand Strategic Science Investment Fund, by the Global Earthquake Model Foundation and the King Abdullah University of Science and Technology (KAUST) research Grant URF/1/2160-01-01.

References

Allen, T. I., D. J. Wald, and C. B. Worden (2012), Intensity attenuation for active crustal regions, *J. Seismol.*, *16*, 409–433, doi:10.1007/s10950-012-9278-7.

Bird, P., D. D. Jackson, Y. Y. Kagan, C. Kreemer, and R. S. Stein (2015), GEAR1: A global earthquake activity rate model constructed from geodetic strain rates and smoothed seismicity, *Bull. Seismol. Soc. Am.*, *105*(5), 2538–2554.

Böse, M., et al. (2014), CISN ShakeAlert: An earthquake early warning demonstration system for California, in *Early Warning for Geological Disasters*, edited by F. Wenzel and J. Zschau, Advanced Technologies in Earth Sciences, pp. 49–69, Springer, Berlin, Heidelberg.

Campbell, K. W., and Y. Bozorgnia (2014), NGA-West2 ground motion model for the average horizontal components of PGA, PGV, and 5% damped linear acceleration response spectra, *Earthquake Spectra*, *30*(30), 1087–1115, doi:10.1193/062913EQS175M.

Cattania, C., et al. (this issue), Evaluation of coulomb-based seismicity forecasting models during the 2010-2012 Canterbury, New Zealand, earthquake sequence, *Seismol. Res. Lett.*

Clements, R. A., F. P. Schoenberg, and D. Schorlemmer (2011), Residual analysis methods for space-time point processes with applications to earthquake forecast models in California, *Annals of Applied Statistics*, *5*(4), 2549–2571, doi:10.1214/11-AOAS487.

Dieterich, J. (1994), A constitutive law for rate of earthquake production and its application to earthquake clustering, *J. Geophys. Res.*, *99*(B2), 2601–2618.

Dieterich, J. H., and K. B. Richards-Dinger (2010), Earthquake recurrence in simulated fault systems, *Pure Appl. Geophys.*, *167*(8-9), 1087–1104, doi:10.1007/s00024-010-0094-0.

Field, E. H. (2007), Overview of the working group for the development of Regional Earthquake Likelihood Models (RELM), *Seismol. Res. Lett.*, *78*(1), 7–16.

Field, E. H., K. R. Milner, J. L. Hardebeck, M. T. Page, N. van der Elst, T. H. Jordan, A. J. Michael, B. E. Shaw, and M. J. Werner (2017), A spatiotemporal clustering model for the third Uniform California Earthquake Rupture Forecast (UCERF3-ETAS):

Toward an operational earthquake forecast, *Bull. Seismol. Soc. Am.*, *107*(3), 1049–1081, doi:10.1785/0120160173.

Field, E. H., et al. (2009), Uniform California Earthquake Rupture Forecast, version 2 (UCERF 2), *Bull. Seismol. Soc. Am.*, *99*(4), 2053–2107, doi:10.1785/0120080049.

Field, E. H., et al. (2014), Uniform California Earthquake Rupture Forecast, version 3 (UCERF3)—the time-independent model, *Bull. Seismol. Soc. Am.*, *104*(3), 1122–1180.

Field, E. H., et al. (2015), Long-term time-dependent probabilities for the third Uniform California Earthquake Rupture Forecast (UCERF3), *Bull. Seismol. Soc. Am.*, *105*(2A), 511–543, doi:10.1785/0120140093.

Gerstenberger, M. C., and D. A. Rhoades (2010), New Zealand Earthquake Forecast Testing Centre, *Pure and Applied Geophysics*, *167*(8-9), 877–892, doi:10.1007/s00024-010-0082-4.

Gerstenberger, M. C., S. Wiemer, L. M. Jones, and P. A. Reasenberg (2005), Real-time forecasts of tomorrow's earthquakes in California, *Nature*, *435*(7040), 328–331, doi:10.1038/03622.

Gerstenberger, M. C., G. McVerry, D. A. Rhoades, and M. Stirling (2014), Seismic hazard modeling for the recovery of Christchurch, *Earthquake Spectra*, *30*(1), 17–29.

Gerstenberger, M. C., D. A. Rhoades, and G. H. McVerry (2016), A hybrid time-dependent probabilistic seismic-hazard model for Canterbury, New Zealand, *Seismol. Res. Lett.*, *87*(6), 1311–1318.

Gordon, J. S., R. A. Clements, F. P. Schoenberg, and D. Schorlemmer (2015), Voronoi residuals and other residual analyses applied to CSEP earthquake forecasts, *Spatial Statistics*, *14*, 133–150, doi:10.1016/j.spasta.2015.06.001.

Helmstetter, A., Y. Y. Kagan, and D. D. Jackson (2007), High-resolution time-independent grid-based forecast for $M \geq 5$ earthquakes in California, *Seismol. Res. Lett.*, *78*(1), 78–86, doi:10.1785/gssrl.78.1.78.

Holschneider, M., G. Zoeller, and S. Hainzl (2011), Estimation of the maximum possible magnitude in the framework of a doubly truncated Gutenberg-Richter model, *Bull. Seismol. Soc. Am.*, *101*(4), 1649–1659, doi:10.1785/0120100289.

Holschneider, M., G. Zoeller, R. Clements, and D. Schorlemmer (2014), Can we test for the maximum possible earthquake magnitude?, *J. Geophys. Res.*, *119*(3), 2019–2028, doi:10.1002/2013JB010319.

Iervolino, I., E. Chioccarelli, M. Giorgio, W. Marzocchi, G. Zuccaro, M. Dolce, and G. Manfredi (2015), Operational (short-term) earthquake loss forecasting in Italy, *Bull. Seismol. Soc. Am.*, *105*(4), 2286–2298.

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

King, G. C. P., R. S. Stein, and J. Lin (1994), Static stress changes and the triggering of earthquakes, *Bull. Seismol. Soc. Am.*, *84*, 935–953.

Király-Proag, E., J. D. Zechar, V. Gischig, S. Wiemer, D. Karvounis, and J. Doetsch (2016), Validating induced seismicity forecast models—induced seismicity test bench, *J. Geophys. Res.*, *121*(8), 6009–6029.

Lombardi, A. M., and W. Marzocchi (2010), The assumption of Poisson seismic-rate variability in CSEP/RELM experiments, *Bull. Seismol. Soc. Am.*, *100*(5A), 2293–2300.

Mak, S., and D. Schorlemmer (2016), A comparison between the forecast by the United States national seismic hazard maps with recent ground motion records, *Bull. Seismol. Soc. Am.*, *106*(4), 1817–1831, doi:10.1785/0120150323.

Mak, S., R. A. Clements, and D. Schorlemmer (2014), The statistical power of testing probabilistic seismic-hazard assessments, *Seismol. Res. Lett.*, *85*(4), 781–783, doi:10.1785/0220140012.

Mak, S., R. A. Clements, and D. Schorlemmer (2015), Validating intensity prediction equations for Italy by observations, *Bull. Seismol. Soc. Am.*, *105*(6), 2942–2954, doi:10.1785/0120150070.

Mak, S., R. A. Clements, and D. Schorlemmer (2017), Empirical evaluation of hierarchical ground-motion models: Score uncertainty and model weighting, *Bull. Seismol. Soc. Am.*, *107*(2), 949–965, doi:10.1785/0120160232.

Mak, S., F. Cotton, M. Gerstenberger, and D. Schorlemmer (in press), An evaluation of the applicability of NGA-West2 ground-motion models for Japan and New Zealand, *Bull. Seismol. Soc. Am.*

Marzocchi, W., and T. H. Jordan (2014), Testing for ontological errors in probabilistic forecasting models of natural systems, *Proceedings of the National Academy of Sciences of the United States of America*, *111*(33), 11,973–11,978, doi:10.1073/pnas.1410183111.

Marzocchi, W., and T. H. Jordan (2017), A unified probabilistic framework for seismic hazard analysis, *Bull. Seismol. Soc. Am.*, *107*(6), 2738–2744, doi:10.1785/0120170008.

Marzocchi, W., D. Schorlemmer, and S. Wiemer (2010), Preface, *Annals of geophysics*, *53*(3).

Marzocchi, W., J. D. Zechar, and T. H. Jordan (2012), Bayesian forecast evaluation and ensemble earthquake forecasting, *Bull. Seismol. Soc. Am.*, *102*(6), 2574–2584.

Marzocchi, W., A. M. Lombardi, and E. Casarotti (2014), The establishment of an operational earthquake forecasting system in Italy, *Seismol. Res. Lett.*, *85*(5), 961–969.

Mignan, A., C. Jiang, J. D. Zechar, S. Wiemer, Z. Wu, and Z. Huang (2013), Completeness of the Mainland China earthquake catalog and implications for the setup of the China Earthquake Forecast Testing Center, *Bull. Seismol. Soc. Am.*, *103*(2A), 845–859.

Molchan, G., L. Romashkova, and A. Peresan (2017), On some methods for assessing earthquake predictions, *Geophys. J. Int.*, *210*(3), 1474–1480.

Nanjo, K. Z., et al. (2012), Predictability study on the aftershock sequence following the 2011 off the Pacific coast of Tohoku, Japan, earthquake: first results, *Geophys. J. Int.*, doi:10.1111/j.1365-246X.2012.05626.x.

Nomura, S., and Y. Ogata (this issue), Foreshock discrimination and short-term mainshock prediction based on magnitude differences and spatiotemporal distances, *Seismol. Res. Lett.*

Ogata, Y. (2011), Significant improvements of the space-time etas model for forecasting of accurate baseline seismicity, *Earth, Planets and Space*, *63*(3), 217–229, doi: 10.5047/eps.2010.09.001.

Ogata, Y. (2017), On spontaneous seismicity rate in Japan inland, in *Report of the Coordinating Committee for Earthquake Prediction*, vol. 97, pp. 9–12, The Coordinating Committee for Earthquake Prediction.

Ogata, Y., and K. Katsura (2014), Comparing foreshock characteristics and foreshock forecasting in observed and simulated earthquake catalogs, *J. Geophys. Res.*, *119*(11), 8457–8477, doi:10.1002/2014JB011250.

Ogata, Y., K. Katsura, G. Falcone, K. Z. Nanjo, and J. Zhuang (2013), Comprehensive and topical evaluations of earthquake forecasts in terms of number, time, space, and magnitude, *Bull. Seismol. Soc. Am.*, *103*(3), 1692–1708, doi:10.1785/0120120063.

Omi, T., Y. Ogata, Y. Hirata, and K. Aihara (2015), Intermediate-term forecasting of aftershocks from an early aftershock sequence: Bayesian and ensemble forecasting approaches, *J. Geophys. Res.*, *120*(4), 2561–2578, doi:10.1002/2014JB011456.

Omi, T., Y. Ogata, K. Shiomi, B. Enescu, K. Sawazaki, and K. Aihara (2016), Automatic aftershock forecasting: A test using real-time seismicity data in Japan, *Bull. Seismol. Soc. Am.*, *106*(6), 2450–2458, doi:10.1785/0120160100.

Petersen, M. D., et al. (2014), Documentation for the 2014 update of the United States national seismic hazard maps, *Open-File Report 20141091*, U. S. Geological Survey.

Rhoades, D. A. (2013), Mixture models for improved earthquake forecasting with short-to-medium time horizons, *Bull. Seismol. Soc. Am.*, *103*(4), 2203–2215, doi: 10.1785/0120120233.

Rhoades, D. A., and M. C. Gerstenberger (2009), Mixture models for improved short-term earthquake forecasting, *Bull. Seismol. Soc. Am.*, *99*(2A), 636–646, doi: 10.1785/0120080063.

Rhoades, D. A., and M. W. Stirling (2012), An earthquake likelihood model based on proximity to mapped faults and cataloged earthquakes, *Bull. Seismol. Soc. Am.*, *102*(4), 1583–1599, doi:10.1785/0120110326.

Rhoades, D. A., D. Schorlemmer, M. C. Gerstenberger, A. Christophersen, J. D. Zechar, and M. Imoto (2011), Efficient testing of earthquake forecasting models, *Acta Geophysica*, *59*(4), 728–747, doi:10.2478/s11600-011-0013-5.

Rhoades, D. A., M. C. Gerstenberger, A. Christophersen, J. D. Zechar, D. Schorlemmer, M. J. Werner, and T. H. Jordan (2014), Regional Earthquake Likelihood Models II: Information gains of multiplicative hybrids, *Bull. Seismol. Soc. Am.*, *104*(6), 3072–3083, doi:10.1785/012014003.

Rhoades, D. A., A. Christophersen, and M. C. Gerstenberger (2015), Multiplicative earthquake likelihood models based on fault and earthquake data, *Bull. Seismol. Soc. Am.*, *105*(6), 2955–2968.

Rhoades, D. A., M. Liukis, A. Christophersen, and M. C. Gerstenberger (2016), Retrospective tests of hybrid operational earthquake forecasting models for Canterbury, *Geophys. J. Int.*, *204*(1), 440–456.

Rhoades, D. A., A. Christophersen, and M. C. Gerstenberger (2017), Multiplicative earthquake likelihood models incorporating strain rates, *Geophys. J. Int.*, *208*(3), 1764–1774.

Rhoades, D. A., A. Christophersen, M. C. Gerstenberger, M. Liukis, F. Silva, M. Marzocchi, M. J. Werner, and T. H. Jordan (this issue), Highlights from the first ten years of the new zealand earthquake forecast testing center, *Seismol. Res. Lett.*

Richards-Dinger, K., and J. H. Dieterich (2012), RSQSim earthquake simulator, *Seismol. Res. Lett.*, *83*(6), 983–990, doi:10.1785/0220120105.

Scherbaum, F., E. Delavaud, and C. Riggelsen (2009), Model selection in seismic hazard analysis: An information-theoretic perspective, *Bull. Seismol. Soc. Am.*, *99*(6), 3234–3247, doi:10.1785/0120080347.

Schneider, M., R. A. Clements, and D. Schorlemmer (2014), Likelihood- and residual-based evaluation of medium-term earthquake forecast models for California, *Geophys. J. Int.*, *198*, 1307–1318, doi:10.1093/gji/ggu178.

Schorlemmer, D., and M. Gerstenberger (2007), RELM Testing Center, *Seismol. Res. Lett.*, *78*(1), 30–36.

Schorlemmer, D., M. Gerstenberger, S. Wiemer, D. D. Jackson, and D. A. Rhoades (2007), Earthquake likelihood model testing, *Seismol. Res. Lett.*, *78*(1), 17–29.

Schorlemmer, D., A. Christophersen, A. Rovida, F. Mele, M. Stucchi, and W. Marzocchi (2010a), Setting up an earthquake forecast experiment in Italy, *Annals of Geophysics*, *53*(3), doi:10.4401/ag-4844.

Schorlemmer, D., F. Mele, and W. Marzocchi (2010b), A completeness analysis of the national seismic network of Italy, *J. Geophys. Res.*, *115*, B04308, doi:10.1029/2008JB006097.

Schorlemmer, D., J. D. Zechar, M. J. Werner, E. H. Field, D. D. Jackson, T. H. Jordan, and the RELM Working Group (2010c), First results of the Regional Earthquake Likelihood Models experiment, *Pure and Applied Geophysics*, doi:10.1007/s00024-010-0081-5.

Schorlemmer, D., N. Hirata, Y. Ishigaki, K. Doi, K. Z. Nanjo, H. Tsuruoka, T. Beutin, and F. Euchner (2018), Earthquake detection probabilities in Japan, *Bull. Seismol. Soc. Am.*, doi:10.1785/0120170110.

Shen, Z., D. D. Jackson, and Y. Y. Kagan (2007), Implications of geodetic strain rate for future earthquakes, with a five-year forecast of M5 earthquakes in southern California, *Seismol. Res. Letts.*, *78*(1), 116–120, doi:10.1785/gssrl.78.1.116.

Steacy, S., M. C. Gerstenberger, C. A. Williams, D. A. Rhoades, and A. Christophersen (2014), A new hybrid coulomb/statistical model for forecasting aftershock rates, *Geophys. J. Int.*, *196*(2), 918–923, doi:10.1093/gji/ggt404.

Stein, R. S. (1999), The role of stress transfer in earthquake occurrence, *Nature*, *402*(6762), 605–609, doi:10.1038/45144.

Strader, A., M. Schneider, and D. Schorlemmer (2017), Prospective and retrospective evaluation of five-year earthquake forecast models for California, *Geophys. J. Int.*, *211*(1), 239–251, doi:10.1093/gji/ggx268.

Taroni, M., J. D. Zechar, and W. Marzocchi (2013), Assessing annual global M6+ seismicity forecasts, *Geophys. J. Int.*, *196*(1), 422–431, doi:10.1093/gji/ggt369.

Taroni, M., W. Marzocchi, D. Schorlemmer, M. J. Werner, S. Wiemer, J. D. Zechar, L. Heiniger, and F. Euchner (this issue), Prospective csep evaluation of 1-day, 3-month, and 5-year earthquake forecasts for italy, *Seismol. Res. Lett.*

Tsuruoka, H. (2017), Earthquake predictability experiment based on CSEP project - trial of forecast experiments in Japan -, in *Report of the Coordinating Committee for Earthquake Prediction*, vol. 98, pp. 460–464, The Coordinating Committee for Earthquake Prediction.

Tsuruoka, H., N. Hirata, D. Schorlemmer, F. Euchner, K. Z. Nanjo, and T. H. Jordan (2012), CSEP testing center and the first results of the earthquake forecast testing experiment in Japan, *Earth Planets Space*, *64*, 661–671, doi:10.5047/eps.2012.06.007.

Werner, M. J., and D. Sornette (2008), Magnitude uncertainties impact seismic rate estimates, forecasts and predictability experiments, *J. Geophys. Res.*, *113*(B8), B08302, doi:10.1029/2007JB005427.

Werner, M. J., A. Helmstetter, D. D. Jackson, and Y. Y. Kagan (2011), High-resolution long-term and short-term earthquake forecasts for California, *Bull. Seismol. Soc. Am.*, *101*(4), 1630–1648.

Woessner, J., et al. (2011), A retrospective comparative forecast test on the 1992 Landers sequence, *J. Geophys. Res.*, *116*, B05305, doi:10.1029/2010JB007846.

Zechar, J. D., and T. Jordan (2008), Testing alarm-based earthquake predictions, *Geophys. J. Int.*, *172*(2), 715–724, doi:10.1111/j.1365-246X.2007.03676.x.

Zechar, J. D., and J. Zhuang (2010), Risk and return: evaluating reverse tracing of precursors earthquake predictions, *Geophys. J. Int.*, *182*(3), 1319–1326.

Zechar, J. D., and J. Zhuang (2014), A parimutuel gambling perspective to compare probabilistic seismicity forecasts, *Geophys. J. Int.*, *199*(1), 60–68.

Zechar, J. D., M. C. Gerstenberger, and D. A. Rhoades (2010a), Likelihood-based tests for evaluating space-rate-magnitude earthquake forecasts, *Bull. Seismol. Soc. Am.*, *100*(3), 1184–1195, doi:10.1785/0120090192.

Zechar, J. D., D. Schorlemmer, M. Liukis, J. Yu, F. Euchner, P. J. Maechling, and T. H. Jordan (2010b), The Collaboratory for the Study of Earthquake Predictability perspective on computational earthquake science, *Concurrency and Computation: Practice and Experience*, *22*(12), 1836–1847, doi:10.1002/cpe.1519.

Zechar, J. D., D. Schorlemmer, M. J. Werner, M. C. Gerstenberger, D. A. Rhoades, and T. Jordan (2013), Regional Earthquake Likelihood Models I: First-order results, *Bull. Seismol. Soc. Am.*, *103*(2A), 787–798, doi:10.1785/0120120186.

Zhuang, J. (2011), Gambling scores for earthquake predictions and forecasts, *Geophys. J. Int.*, *181*(1), 382–390.

Zhuang, J., D. Vere-Jones, H. Guan, Y. Ogata, and L. Ma (2005), Preliminary analysis of observations on the ultra-low frequency electric field in a region around Beijing, *Pure Appl. Geophys.*, *162*, 1367–1396, doi:10.1007/s00024-004-2674-3.