Final Report

2014 CSEP/USGS/GEM Workshop: Next Steps for Testing Operational Earthquake Forecasts and Seismic Hazard Models

December 5, 2014

Organizers: Maximilian Werner (Bristol University), Danijel Schorlemmer (GFZ Potsdam), Thomas Jordan (USC), Andy Michael (USGS) and Morgan Page (USGS)

Date: September 6, 2014
Location: Hilton Palm Springs Resort, Palm Springs, California, USA
Attendees: by invitation only.
Website: http://www.scec.org/workshops/2014/csep/index.html
(including a full list of attendees and links to presentations)

OVERVIEW: The Collaboratory for the Study of Earthquake Predictability (CSEP), operated by the Southern California Earthquake Center (SCEC), provides a research cyber-infrastructure for independent and prospective testing of earthquake forecasts. As such, CSEP is well situated to evaluate operational forecasting models of earthquake potential and ground motions by the USGS, GEM and other international governmental and non-governmental organizations. The ongoing development and implementation of operational models, however, entail new requirements for CSEP's infrastructure, methods and experiment design.

The purposes of this workshop were: (i) to assess the evolving needs of agencies for CSEP-based testing of OEF and seismic hazard models, (ii) to disseminate and review recent CSEP and GEM Testing & Evaluation (T&E) results, (iii) to assess the adequacy of CSEP's current methods and infrastructure in light of evolving needs, and (iv) to gather community input on the next steps for testing OEF and seismic hazard models.

The workshop brought together CSEP personnel, agency representatives, and scientists interested in the scientific and operational aspects of earthquake and ground-motion forecasting and testing. This one-day workshop included sessions on the following topics:

1. Status and requirements for OEF and seismic hazard models
2. Current CSEP capabilities and review of earthquake forecasts under testing
3. Status and requirements for the short-term UCERF3 and GEM Global Earthquake Activity Rate (GEAR) models
4. Results from short-term and global CSEP experiments
5. Results from the retrospective Canterbury, New Zealand, experiment
6. Operationalization: real-time forecasting and data considerations
7. Status and requirements for seismic hazard models (GMPEs, IPEs, NSHMP)
8. Review of seismic hazard models under CSEP and GEM T&E testing
9. Discussion, recommendations and next steps for future CSEP experiments

Summaries of Presentations and Discussions

Session 1:
Overview of OEF and CSEP

Presentations

Max Werner provided an overview of CSEP and an update on recent results. He noted that 434 models (or variants of models) were now under testing. Selected results included the performance of multiplicative hybrid models formed from the Regional Earthquake Likelihood Models (RELM) in a study by Rhoades et al. (2014), which suggests that hybrids formed of models with different information and data types perform particularly well and retrospectively had greater predictive skill than the best performing single model (the Helmstetter adaptive smoothing model). A recent evaluation of 3-month models in California by Schneider et al. (2013) used residuals as an interesting visual diagnostic of model performance. Max showed initial results of an adaptive space-time smoothing model (Conan) applied to California and Japan. Jeremy Zechar analyzed the performance of available 1-day models in California and found that (1) new models tend to perform better, (2) the ETAS and non-parametric model K^3 by Helmstetter and Werner (2014) currently performed best, and (3) the Kagan-Jackson model underpredicts strongly. Max mentioned the new 30-minute testing class in California and noted the substantial computational challenges for the new high-resolution global experiments. Finally, he provided a status update on experiments of External Forecasts and Predictions (EFPs) including M8 and QuakeFinder predictions.

Warner Marzocchi presented an update on the liaison between OEF and CSEP in Italy. CSEP-Italy has been live since August 1, 2009 and will directly serve as the source of credible information for the OEF system being developed by the Civil Protection Agency and INGV. Ensemble modeling is used to construct optimal forecasts. Warner argued that consistency tests might be misleading and should be reconsidered. Other open issues include that CSEP uses 1-day and 3-month forecast horizons, while OEF requires 1-week forecasts. He noted that testing should consider the entire distribution rather than just the mean.

Matt Gerstenberger provided an update on OEF in New Zealand (NZ). Prior to Canterbury, no regular forecasts were issued, but forecasts have been operational and public since then. He discussed GNS' experience during and after the sequence and communication issues with engineers and the public. He noted that CSEP testing is not yet credible for some practitioners.

Yosi Ogata provided an update on OEF and CSEP in Japan. Michael Blanpied summarized the current UGSG strategy for OEF, including network requirements, development and testing of algorithms and communication. Ned Field described the status quo of the short-term component of the Uniform California Earthquake Rupture Forecast (UCERF3), which includes finite faults and elastic rebound. Ned argued that CSEP could help determine whether sequence-specific parameters are required in California. He also noted issues with respect to real-time data and access to the network and wondered how simulation-based forecasts should best be tested within CSEP.

Discussion: How adequate is CSEP's infrastructure for evaluating OEF models?

Moderators: P. Maechling and M. Werner; Reporter: M. Liukis

Warner Marzocchi noted that it is helpful to have miniCSEP distributions to test models for researchers before submitting/engaging with CSEP. Matt Gerstenberger summarized current OEF requirements and needs in NZ: longer forecast periods, longer lag-times, more retrospective testing, better results communication strategies, improved methods for dealing with real-time catalog issues. Attendees noted that forecast frequency updates and realtime data problems needed to be viewed from the perspective of the end-user. Bill Ellsworth noted that ground-motion testing should be emphasized in CSEP as it is important for OEF.

Phil Maechling listed UCERF's OEF requirements, including versioned catalogs including uncertainties, probabilities of missed events and associations between earthquakes and model faults. He summarized USGS OEF strategy by focusing on user needs and uses and the roles of different organizations. Ned Field noted that elastic rebound is an important problem for UCERF, science and OEF.

Presentations

Takahiro Omi outlined a Bayesian method to estimate the time-dependent completeness magnitude after large earthquakes, and its use for forecasting aftershocks after a strong earthquake. Harley Benz provided an overview of ComCat and real-time data access, including PDL (Product Distribution Layer), which is a distribution method for all types of earthquake-related content including time tags. He mentioned available Python code (https://github.com/usgs/libcomcat) to generate other formats of earthquake source parameters from the standard format available via the web service, including GeoJSON feeds and an epicentral iPhone app. He noted that the NEIC PDE catalog is currently available from 1983 to the present, but not 'fully validated'. NEIC is currently loading the SCSN catalog and will load the NCSN catalog next.

Panel Discussion: *H. Benz / E. Hauksson / E. Field / Y. Ogata*
How should real-time data uncertainties be handled in OEF models and their evaluation? What are the data requirements for OEF?
Moderator: L. Jones Reporter: Nicholas van der Elst

OEF requires accurate estimates of earthquake parameters to feed into the model. We do not have a good idea how much this will limit the usefulness of the models currently in testing through CSEP (Werner). The OEF models based on clustering are particularly sensitive to errors in magnitude, because the forecast rates depend exponentially on magnitude. Some models are also sensitive to finite fault extent. We could use information from multiple estimates of magnitude to try to project to a more consistent magnitude estimate (Ogata).

The comprehensive catalog is useful, but without versioning and archiving of the versions, we cannot really investigate the effect of the poorer-quality earthquake parameters available in real-time on the model. Outside of OEF, versioning is still a necessary condition for scientific reproducibility (Page). If we design new techniques and want to see if they give the old answers, we cannot do that if the dataset has been changed. Will ComCat take on versioning as their responsibility? It seems like this is a solved problem in programming/IT with tools to track changes in files. Would these tools be applicable to ComCat? ComCat also contains ShakeMap and DYFI information that could be useful in evaluating ground motion forecasts. Although versioning tools for software development and databases are not identical, modern databases should provide tools to implement versioning on the single dataset level.

Session 2:
GEM Overview, Global Experiments and GEM's GEAR models

Discussion: How should global experiments be conducted?
How can CSEP's testing methods be improved?
Moderator: Warner Marzocchi Reporter: Morgan Page

Dave Jackson gave a short presentation arguing for a few simple tests that are easy for others to reproduce. The L-test could be abolished, as this conflates information that is already in the N, S, and M tests. Specificity and information gain tests should be added. We should allow modelers to mask cells that should not be included in the forecast. In addition, simulations could be based only on the functional form of the probability distribution, eliminating the needs for lat/long/mag binning. David Rhoades presented an alternative view for the direction of the CSEP tests, advocating efficiency improvements that would make the global tests more numerically tractable. The uniform gridding currently used oversamples polar regions, creating a large number of cells that is problematic. Efficient alternatives exist for nearly all of the current tests. David advocated abolishing simulated catalogs, using the large-number approximation when applicable, the point-process likelihood formulation, and moving to event-based testing, which removes deviations from Poisson statistics. During the discussion period, Yosihiko Ogata asked for the tests to include more magnitude resolution, to help elucidate the ability of the models to forecast b-value differences. Warner Marzocchi suggested that CSEP compare the global models in regions to the corresponding regional models. Finally, Bruce Shaw suggested that CSEP mask the tests in regions of time/magnitude where short-term aftershock incompleteness is a problem, so that modelers do not have to model short-term catalog incompleteness in order to do well in the tests.

Presentations

Sam Mak of GFZ Potsdam presented a study whose goal is to establish whether non-instrumental earthquake records (from the Did-You-Feel-It? and ShakeMap databases) can be useful for validation hazard forecasts.

Discussion: How should GEM and USGS ground motion forecasts be evaluated?
Is CSEP ready to evaluate seismic hazard maps?
Moderator: B. Ellsworth Reporter: M. Gerstenberger

The most vigorous debate was about the value in testing the combined outputs from an earthquake rupture forecast and ground-motion prediction equation(s). Testing the combined output makes it very difficult to interpret the result and to understand what each of the various components are contributing to the final result; however, because the final product that a PSHA model produces is a hazard curve or some variation, it was decided that it is necessary to test the final result, but that it must be done along side testing of the ERF and GMPE(s) independently.

Other questions were raised about if it is reasonable to test using aftershocks given that current PSHA models explicitly exclude aftershocks. The point was raised that aftershocks can contribute substantially to hazard and, by including them in the testing, we can better understand how significant their contribution is. It was recommended that hazard models provide an algorithm to exclude what the models define as aftershocks. Otherwise, testing will require defining aftershocks and choosing a declustering algorithm, which may or may not be consistent with the hazard model.

The usefulness in testing long-term models (e.g., 50 years) was also questioned due to the paucity of large events and the short time periods we have available to test. There was no satisfactory answer to this question but the societal impact of these models was discussed with the implication that a focus on understanding how to test such models is necessary (and we will not better learn how to test them until we begin the effort).

Participants noted that testing hazard is qualitatively different from testing earthquake probability forecasts. A major difficulty concerns open data availability: some agencies openly provide data, many others do not. In addition, it was noted that data standards are currently lacking for testing ground motions, which require much more information (e.g., meta-data) than the level of information required for testing earthquake forecasts. The creation of the NGA flatfile is an attempt to standardize data, but it is not updated regularly.

Session 3: Retrospective Canterbury Experiment

Discussion: How are physics-based models performing?
How can retrospective experiments help evaluate OEF models?
Moderator: J. Hardebeck Reporter: A. Strader

The Canterbury Experiment evaluated and compared a series of statistical and physics-based earthquake forecasts in the area surrounding the 2010 Mw7.1 Darfield earthquake. Statistics-based forecasts included ETAS, EEPAS and STEP models, whereas physics-based forecasts relied upon quasi-static (rate-and-state) Coulomb stress evolution. Forecasts were evaluated from the time of the Darfield mainshock using best available data. An ongoing evaluation using real-time data will assess the impact of data deficiencies on forecast quality.

Due to limited information value from short (five-year period) timescales, prospective tests have limited usefulness for informed decision making and operational earthquake forecasting (OEF) that is required now. Although retrospective tests are subject to bias due to prior knowledge of seismicity distributions, they can indicate poorly formulated models, provide confidence in subjective model building, and inform model skill over longer periods compared to prospective tests.

Coulomb/rate-and-state forecasts performed better than the statistical model in the one-year forecast group. Coulomb models' performance also improved for one-year forecasts, compared to previous retrospective tests using 1-day intervals. Multiple one-month forecasts, with start periods adjusted to major earthquake times, outperformed the Poissonian model consistently, as expected. While CRS1 performed decently, ETAS was not a significant improvement over physics-based models. Some Coulomb stress models were competitive for one-month forecasts as well as one-year tests.

During the panel discussion, Yosihiko Ogata questioned whether the Coulomb model explained aftershock distributions following the Darfield earthquake, given that a significant number of aftershocks nucleated within stress shadows. Camilla Cattania, who contributed to Coulomb rate-and-state forecasts, suggested that stress heterogeneity may account for earthquakes occurring in stress shadows, as varying the receiver planes near ruptured faults resulted in a stronger association between stress sign and earthquake location. The Coulomb stress models have been shown in other regions to outperform statistical models, particularly away from faults (and away from most of the influence of stress uncertainties). For example, Margarita Segou tested rate-and-state Coulomb forecasts in northern California, and found that the Coulomb model was a

significant improvement over statistical models. Coulomb stress models also displayed significant information gain following the Landers earthquake for long- term (20 years) retrospective tests.

Session 4: New Directions

CSEP & Induced Seismicity
Panel Discussion: *Emily Brodsky / B. Ellsworth / A. Llenos*
(Reporter: M. Segou)

At the beginning of the panel discussion Emily Brodsky pointed out that predicting induced seismicity is an important question from the industry's point of view. She presented results suggesting good correlation between seismicity and injection rates in the Salton Sea geothermal field. However, it is impossible at this time to support real-time experiments of seismicity rate changes since usually there is a 2-month lag in providing the data logs from industry.

W. Ellsworth's presentation focused on the increase in seismicity in the mid-continent the last few years. In order to describe the phenomenon he presented a basic model supporting that "there is no long-term increase in the background seismicity rates (BR)" and 4 alternative approaches supporting: 1) important short-term variability in the long-term BR, 2) the new seismicity rate is the "New Normal", 3) new earthquakes come from a new population, yet to be understood, and 4) the observed increase is transient in nature.

A. Llenos in her presentation focused on the statistical modeling of the seismicity rates in the mid-continent. She noted that induced seismicity in the mid-continent deviates from the standard Omori law decay and that successful models will be required to incorporate parameter variability in space. She also noted firstly, that sometimes seismicity expands outside of the exploitation area and secondly, that both background rates and aftershock productivity changes. The estimation of maximum expected magnitude in induced seismicity sites ranges between M5.0-5.7.

CSEP & paleo- and simulator-based earthquake rupture probabilities
Panel Discussion: Dave Jackson and Ned Field
(Reporter: D. Harte)

Paleo-seismic event determinations over hundreds or thousands of years appear to produce very regular inter-event times. David Jackson showed a cumulative plot of event times from the UCERF3 data set in California that is remarkably linear, indicating a fairly consistent recurrence time. His plot was also consistent with Berryman et al (2012) who did a very similar plot for paleo times of some events on the Alpine Fault of NZ. A better determination of these historical major events would be very beneficial to future seismic risk assessment. If we take the results at face value, it appears that the interevent times have less variability than one would even expect from a simple Poisson process. In particular, the current open interval since the last earthquake at the combination of the chosen sites is much greater than expected at a very high confidence level, suggesting potential issues in the modeling. The discussion in this session related to various possible explanations, and also how one tests the validity of long term forecasts based on such data. These results seem inconsistent with recurrence times observed on major faults using catalogue data spanning the last few hundred years. To what extent do the paleo analytical methods used, sample selection criteria, and data presentation explain these results? For example, if the determined times were essentially discretized by the analytical method (say to the closest 100th year), this could make it rather regular looking.

Recommendations

OEF

Workshop participants agreed that CSEP is ideally situated as an independent entity to evaluate the predictive skill of OEF models and candidates, especially because of the testing of models in a variety of tectonic settings with correspondingly greater data points that in any single national region of interest to a government agency. To improve the service to the OEF community and the impact of CSEP results, participants recommended a stronger focus on activities that directly relate to scientific (and to some extent technical) questions arising from OEF efforts. For example, further retrospective testing of models, as in the retrospective Canterbury experiment, was noted as very helpful. In addition, forecast horizons and updating intervals should be more flexible, for example to allow for 1-week forecasts, and the immediate updating of 1-week forecasts after earthquakes.

Attendees noted that the real-time data problems deserved greater attention. In particular, OEF requires immediate access to data, but modeling strategies of catalog incompleteness, such as those presented by Omi et al., are required to address the potential adverse effects of real-time data issues on the quality of real-time forecasts. The ongoing discussion with USGS ComCat developers to develop priorities for real-time products, such as providing uncertainties and versioning, were encouraged to deepen.

Global experiments and testing methods

Participants concluded that a change in the status quo of CSEP's testing methods is required for high-resolution global experiments, such as the upcoming testing of the GEM's Global Earthquake Activity Rate (GEAR) models. Recommendations by Dave Jackson, David Rhoades and others were embraced for the simplification of tests and for greater efficiency.

In addition, a greater emphasis on the role of epistemic uncertainty was recommended. For example, current likelihood-based tests make reasonable but strong assumptions on the aleatory variability of models, and leave little room for single-model epistemic uncertainties such as parameter uncertainties or input data uncertainties. More flexible frameworks were recommended. In addition, ensemble modeling techniques, whether Bayesian or multiplicative, were seen as promising options for constructing optimal models and for tracking the relative performance of models over time. Automating such methods would lead to a substantial reduction in the wealth of results that are based on classical frequentist hypothesis testing.

Ground Motion and Intensity Prediction Equation Testing

Participants recommended testing final hazard models in addition to testing hazard model components such as GMPEs and earthquake rupture forecasts. Testing components has greater power and is simpler to interpret, but only the final model will help place the significance of forecast performance in the realm of risk governance.

It was additionally recommended that aftershocks were included in the modeling and testing because there is no unique definition of aftershocks and they at least occasionally contribute substantially to hazard.

Participants recommended CSEP begin testing long-term models (e.g., 50 years) despite the obvious shortcoming that large quakes occur rarely while the test periods are relatively short. It

was argued that the societal impact of these models is substantial and therefore warrants *efforts* to test them that may eventually lead to better approaches.

Participants noted major challenges for extending CSEP's principles to the testing of hazard models and their components. For example, ground motion information is more complex, and no standard format exists that is widely used. Attendees agreed that improved data openness and availability needs to be prioritized.

Canterbury

Participants noted the relatively good performance of the Coulomb-based models in the retrospective Canterbury experiment (on certain timescales) as very encouraging for the future of physics-based forecasting. Further analysis was recommended, as well as the possible design of other retrospective experiments to complement this study.

Induced seismicity

Participants embraced the idea of focused CSEP testing regions to evaluate hypotheses of injection-induced seismicity and recommended proceeding with the Salton Sea experiment. Potential future testing regions might include Oklahoma, where various simple but competing hypotheses outlined by Ellsworth could be evaluated. In addition, it was noted that the (partial) knowledge of injection parameters might lead to an improved understanding of tectonic earthquakes.

Testing of Paleo-event rate

The implications of the current UCERF3-implied rates of ruptures at paleosites in California, as presented by Jackson, were hotly debated. Consensus emerged that the results deserved further study by the WGCEP because they appear to imply that the implied rates (and thus a component of UCERF3) may need to be revised. The study highlighted that the design of prospective tests can help identify poor model performance, and that CSEP could play a role in hosting blind evaluations of UCERF3 model components.

Next Steps

OEF

CSEP will work with USGS to develop strategy for testing real-time OEFs as external forecasts imported into CSEP. Questions to be addressed include the role of real-time data, event-based updating and overlapping forecast horizons. USGS and CSEP will provide data requirements for OEF to ComCat developers. In addition, CSEP will assess further retrospective experiments to help evaluate candidate OEF models. CSEP currently has two draft designs for event-based testing that need to be evaluated from an OEF perspective.

Global Experiments and Testing Methods

GEM will provide GEAR models to CSEP. GEM T&E group will develop new testing software based on this workshop's recommendations and, with the help of Liukis, integrate these codes

into the CSEP software stack. Joint GEM T&E and CSEP retrospective testing will be performed; prospective testing will begin as soon as possible.

Rhoades, Jackson, Ogata, Werner and others will develop further changes to the testing methods by (1) implementing tests that do not require simulations, and (2) considering a more flexible framework allowing for individual model epistemic and aleatory uncertainty.

Ground Motion and Intensity Prediction Equation Testing

The GEM T&E group received constructive feedback from the attendees and will work on several questions related to their study of validating hazard forecasts using Did-You-Feel-It? and ShakeMap data. These include the role of aftershocks on hazard, site effects, the quality of the DYFI data and others. The GEM T&E and the newly formed Dynamic Risk Quantification Potsdam group will work on combining CSEP rate tests with ground-motion and IPE testing.

The importance of lobbying for standardized data formats of ground motion information was noted. Participants hoped the USGS will begin to provide such information in a standardized manner.

Canterbury

CSEP will complete generation and evaluation of forecasts, including the 1-day forecast group and the test case using near-real-time data. In addition, CSEP will automate ensemble modeling methods for this experiment and transfer codes to other experiments.

Induced seismicity

CSEP will work with Emily Brodsky to design a prospective experiment to evaluate the hypothesis that seismicity rates in the geothermal Salton Sea field correlate with net fluid extraction rates. CSEP and others will assess opportunities for further new testing regions targeted at injection-induced seismicity.

Testing of Paleo-event rate

WGCEP participants will discuss Jackson's results and search for explanations. Future discussions between CSEP and WGCEP will be held to assess the feasibility of CSEP-based testing of implied paleo-rates.