Final Report

2015 CSEP/USGS/GEM Workshop: Epistemic Uncertainty

October 12, 2015

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

Date: September 13, 2015
Location: Hilton Palm Springs Resort, Palm Springs, California, USA
Attendees: by invitation only.
Website: http://www.scec.org/workshops/2015/csep/index.html
(including a full list of attendees and links to presentations)

OVERVIEW: This 1-day workshop is organized jointly by the Collaboratory for the Study of Earthquake Predictability (CSEP) and representatives of the U.S. Geological Survey (USGS) and the Global Earthquake Model (GEM) Foundation. The goals of the workshop are to strengthen collaborations on the evaluation of earthquake and ground-motion forecasting methods, to assess new CSEP results and to continue reshaping CSEP's modus operandi to better address the evolving needs of the community. Several major themes will be covered. First, Operational Earthquake Forecasting (OEF) and its evaluation will be a major focus in light of on-going efforts in the US, in New Zealand, and in Italy. Second, the implications of epistemic uncertainty for forecasting (including automated ensemble modeling) and CSEP testing methods will be addressed. A third theme will be the predictability of fluid injection-induced seismicity. Finally, we will focus on recent research on hazard models testing and evaluation.

The workshop will bring 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 will include sessions on the following topics:

1. Overview of CSEP, OEF and hazard models:
   - Overview of CSEP and recent results
   - USGS perspectives on OEF, hazard and CSEP
   - GEM overview and perspectives on earthquake and ground motion forecasting and evaluating
   - CSEP and OEF in New Zealand, Japan, Italy and China
2. Evaluating earthquake forecast models:
   - Evaluating USGS models: UCERF2/3, the NSHMP and U3-ETAS
   - Evaluating GEM's GEAR1 models and other global experiments
   - Evaluating physics-based and statistical OEF candidate models: the retrospective Canterbury, New Zealand, experiment

    - Status of prototype External Earthquake Forecast and Prediction (EFP) experiments: M8 and QuakeFinder algorithms
3. Seismic hazard models testing:
    - Methods for testing and evaluating hazard models
    - Ground-motion forecasts of USGS and GEM models
    - Results from GEM testing of GMPEs and IPEs
4. Injection-induced seismicity
    - USGS perspectives on induced hazard estimation
    - Overview of statistical and geomechanical forecast models of induced seismicity
    - The Salton Sea CSEP experiment prototype
    - Developing new experiments to evaluate induced seismicity forecasts

Summaries of Presentations and Discussions

Session 1: Overview of OEF and CSEP

Moderator: D. Schorlemmer Reporter: M. Liukis

Presentations

Max Werner provided an overview of CSEP and an update on recent results. He noted that 436 models (or variants of models) were now under testing. Current software developments as of release v15.10 includes (i) the GEAR1 model by Bird et al. (2015) developed in collaboration with GEM, (ii) major OS and Python upgrades, (iii) miniCSEP as a virtual image, (iv) raw catalog retrieval with time-lag and version-control, (iv) enhancements to test routines. Selected results from California experiments included the 1-day, 3-month and 5-yr RELM forecast group. Preliminary Bayesian ensemble modelling of 8 CA 1-day models shows (i) two models dominating the Bayesian model averaging weights, (ii) irrespective of whether or not weights are corrected for correlations between forecasts. Hybrid RELM forecasts for CA show that some hybrids attain a greater information score than the most informative RELM forecast, namely the Helmstetter et al model. Max presented global forecast results as well as the substantial computational challenges for the new high-resolution global experiments. The Ocean Transform Faults experiment shows no earthquakes since January 2012 but none of three models can be rejected (thus far). The Salton Sea geothermal field experiment is being developed: Brodsky's student is testing their model on the CSEP development server.

David Rhoades presented current CSEP and OEF activities in New Zealand. The software was updated and issues with downloading the catalog were fixed. An ETAS model by Harte is being considered for installation. Harte's model would provide simulations of seismicity rather than expected rates, requiring an enhancement to CSEP testing capabilities. Rhoades et al are setting up an experiment to compare real-time versus delayed earthquake forecasts using real-time and processed catalogs. David showed test results of the expert elicitation model developed for OEF for the Canterbury region. The EE hybrid outperformed most individual models at all time lags. David showed model developments involving multiplicative hybrids using both fault and earthquake data. High-order hybrids perform well. The next step is to include strain rate maps into long-term as well as medium term models. Recent OEF activities include forecasting during the Wilberforce earthquake (January 2015).

Naoshi Hirata reported on the CSEP activities in Japan. He showed results from the Five-year Japanese earthquake predictability experiment with multiple runs since 2009 including the 2011 Tohoku-oki earthquake and the 2014 Northern Nagano earthquake. In all 3 testing regions, 251 models are installed. Results from the 3-month group show that overall models underestimated the Tohoku sequence and that Ogata's HISTETAS model performed best. He mentioned that time-variant components are required even for 3-month models. Plans involve generating real-time forecasts based on automated catalogs and a 3D experiment in Kanto.

Changsheng Jiang presented the status and challenges for implementing CSEP in China. The proposed testing region is the North South Seismic Belt. The completeness of the region has been evaluated. Tests of models include the Receiver Operating Characteristic (ROC) and the Molchan diagram. Models include four 5yr models, three 3yr and 2 1day models, including a model based on gravity. Challenges include involving more contributors, soliciting more models and coordinating CSEP activities with the existing earthquake predictions efforts in China.

Warner Marzocchi reported on OEF and Operational Earthquake Loss Forecasting in Italy. He provided an overview of CSEP activities in Italy and the current status of results and models in the European CSEP testing center at ETH Zurich. The official catalog is only available until 31 December 2012, thus delaying calculation of recent CSEP results. The OEF system in Italy, aka Cassandra, uses an ensemble model of 1-day models. Challenges include the fact that CSEP models produce 1-day forecasts while Cassandra produces 1-week forecasts tailored to end users. Warner remarked that the testing method of CSEP must be strengthened to be able to deal with epistemic uncertainties. Calculations of risk of dying show that levels deemed acceptable were breached during a particular earthquake sequence in Italy, suggesting that low OEF probabilities may translate to actionable risk assessments. The OEF model uses real-time catalogs, collects weights from past performance, and submits forecasts to testing center.

Mike Blanpied (USGS) reported on ongoing OEF activities at the USGS. He stated the USGS already did OEF but that it was not tailored to user needs. For example, aftershock probabilities are published after large earthquakes in California and, more recently, on a global scale. These are done on an ad hoc basis. Goals for the program include (i) a broader scope for global OEF using (ii) modernized and verified calculation methods. The OEFs should be (iii) tailored to specific end users and (iv) well coordinated and communicated within the USGS, to the public and end users. OEF needs to be (v) added to the range of USGS products. Further efforts should be guided by end users. Mike reported on a recent NEPEC meeting on OEF as well as the USGS Powell Center meetings on OEF. He showed an example of OEF during the Nepal Gorka earthquake that required a two-source Reasenberg Jones (RJ) model that was supplanted by an ETAS model. The simple RJ model uses parameters that have been estimated in various regions of the globe (Garcia zones). Mike mentioned that the USGS would like to evaluate forecasts within CSEP. NEPEC will advise on how to proceed with advanced methods for modelling and evaluation.

Danijel Schorlemmer summarized results of evaluating UCERF2 and plans for testing UCERF3. Likelihood-based results show that the UCERF2 model passes all consistency tests in a prospective 5yr test, suggesting consistency with observed earthquakes. In a comparison, only the Helmstetter et al outperformed the UCERF2 and NSHM forecasts for California. A 40yr retrospective test revealed that UCERF2 was inconsistent with past data. Danijel mentioned upcoming objectives for optimizing CSEP codes for global experiments as well as testing the time-dependent UCERF model.

Session 2: OEF, Aftershocks and Retrospective Experiments

Moderator: *P. Maechling* Reporter: *A. Llenos*

Presentations

Ned Field presented on the testing of UCERF3-ETAS, which is the component that includes spatiotemporal clustering for use in California OEF. His presentation touched on how to gauge the "usefulness" of a model, which will vary with its particular use, and asked whether the CSEP tests are necessary or sufficient for this, if there are other tests that can be used, and how to test models in the context of their specific uses. The UCERF3-ETAS model doesn't include segmentation, but does include elastic rebound which is required by the inclusion of both clustering and finite-fault ruptures, and this seems to get rid of the problem of having non-GR magnitude-frequency distributions. Other issues include how to apply tight spatial clustering statistics to faults with large uncertainties around their spatial locations and dimensions, and how to tune and test models at magnitudes that are relevant for hazard and loss. The main leap of faith required is that the statistics of small earthquakes applies to big earthquakes as well. Some of the main testing issues are how to deal with over 1000 logic tree branches, how to test the synthetic catalogs the model produces in CSEP, and how to test the forecasts when they depend on the interplay between elastic rebound and characteristic MFD's. The model can be used to simulate specific earthquakes and scenarios. To help determine the usefulness of the model, it is important to identify potential early adaptors and formalize tests for these users.

Max Werner showed some results from the Canterbury retrospective experiment, which tested 15 models (including physics-based, statistics-based and hybrids) in two modes (retrospective using best available data, and pseudo-prospective using archived preliminary raw catalogs) over several evaluation periods (1 year, 1 month, 1 day). He presented the results of the T/W tests for 2 1-yr forecasts, and the surprising results are that the physics-based models are now outperforming the statistical models, likely because of improvements in the Coulomb modeling that account for uncertainties in parameters and calculations. The other surprising result is that the importance of real-time data is strongly model-dependent; some models actually performed better using the raw catalogs instead of the official catalogs. This exercise was a useful example of how CSEP can do retrospective experiments, and some questions that were put forward are whether the data, model forecasts and results should be provided for other researchers to use in benchmarking; whether the Canterbury models can be used in other experiments; and what other specific sequences/regions might be good targets for future retrospective experiments. Retrospective experiments around other sequences would be useful to investigate whether there is something special about Canterbury that allows the physics-based models to perform better. Yan Kagan raised the question of possible biases in the retrospective test, because the number of degrees of freedom in each model varies and ideally you want a "zero degree of freedom" model.

Matteo Taroni discussed 1-month ensemble model forecasts from the Canterbury experiment, comparing them using a ranking measure based on the loglikelihood of the observed data given each model. The best single model was the STEP-Coulomb model. Then he showed ensemble models built from the weighted average of the best 4 single models, with weights proportional to either the likelihood or its reciprocal, which accounts for correlation between the model forecasts. The ensemble models are comparable to the best single model, and are useful because you don't know a priori which model will perform the best overall.

Tom Jordan briefly introduced a new opportunity at SCEC – the Collaboratory for Interseismic Simulation and Modeling (CISM), which is oriented to constructing system-specific forecasts and building infrastructure to test these increasingly complex models and forecasts.

Anne Strader discussed testing ensemble models over time using dynamic risk quantification (DRQ). The main goal is to start with earthquake forecasts, combine these with GMPEs to get seismic hazard, and ultimately provide a seismic risk forecast. The main idea is to replace expert-opinion decision-making with an algorithm-driven, data-driven framework. DRQ ensemble forecasts combine weighted individual forecasts according to information gain and are then evaluated against an appropriate null hypothesis. The mixing parameters are then repeatedly optimized over a series of time intervals. Ultimately this will reveal which forecasts have the greatest weights and how stable are they.

Panel discussion: Epistemic uncertainties in CSEP

Warner Marzocchi led off by introducing the general problem that current models being tested simply produce a single rate in a single space-time-magnitude bin, ignoring the epistemic uncertainties completely. He emphasized the need to start thinking of ways to describe these epistemic uncertainties in CSEP, particularly because some models might not be rejected if their uncertainties were accounted for. Also not all models are Poissonian, and the target earthquakes of the forecasts are not independent in time and space, so using the loglikelihood may not make much sense.

Continuing on this theme, David Rhoades pointed out that this is a particular problem for ETAS models and other short-term models that contain a higher degree of clustering, and the CSEP consistency tests, which rely on the Poisson assumption, often wrongly reject these types of models. Therefore he proposed a pilot experiment, where a new NZ model class would be developed to accept specifically non-Poissonian models. Modelers would provide the grid cell forecasts and simulated catalogs conforming to the model, from which the distribution of expected numbers, statistics and joint distribution of expected numbers in different grid cells. For the N-test, the negative binomial distribution (NBD) could potentially be used to replace the Poisson N-tests, since the NBD is a better fit to clustered data, although the 2nd parameter of the NBD would need to be supplied by the modelers, or determined from the simulated catalogs by the testing center, as Yan Kagan suggests. There was some concern later in the discussion about the large amount of computational resources and potentially data transmission between modelers and testing centers that would be required to generate and/or provide the simulated catalogs necessary to characterize the epistemic uncertainties, and whether it might be difficult to get modelers to produce epistemic uncertainties for their models, but some modelers are already doing this (for example, Takahiro Omi's Bayesian forecasting approach).

Dave Jackson opened the discussion a bit, wanting to move away from these sorts of model validation questions, towards the broader motivation for these kinds of experiments that starts with: what is the scientific question that needs to be answered? What is a model that articulates this? How do we go about testing this model? He emphasized the importance of designing experiments that address scientific questions and specific areas/sequences. Max pointed to UCERF3-ETAS as an interesting example. Potential runaway sequences would have massive variability over Poisson, which would affect the likelihood of triggering something on a larger fault in the future. For example, in the UCERF model, the likelihood of triggering something big on the San Andreas is small but the variability is large, so there is a need to go beyond the Poisson distribution to get the likelihood comparisons right. Simulated catalogs would be useful to characterize the spatial correlation between cells.

Session 3: Evaluating Seismic Hazard Models

Moderator: *W. Marzocchi* Reporter: *N. van der Elst*

Presentations

Celine Beauval provided a summary of evaluations of the French and Turkish hazard models against observations. Marco Pagani provided an overview of GEM, its hazard component and thoughts about how to learn from testing to improve hazard models. Sam Mak summarized evaluations of the US hazard model against Did-You-Feel-It (DYFI) data.

Presenters agreed that since Probabilistic Seismic Hazard Assessement (PSHA) is a forecast, we need to test the predictive power of the model. Preliminary studies have been carried out. Strong motion data is of course preferred, but such observations are sparse, so intensity data and DYFI responses are used as proxies. Preliminary data suggest the models are not doing great at the small return times (the range that can be tested over the short term).

However, there is also some concern that the data sets (intensity and DYFI) contained characteristics that are problematic, such as spatial/temporal biases. For instance, some felt that it does not seem correct to 'reject' a PSHA forecast for the entire US based on data that emphasize small areas within that total forecast. Data being used are strongly correlated and focused on very small parts of the model. We really need independent measurements spread over the entire model domain to test the entire model.

Other issues were raised. The training dataset for PSHA is generally declustered, but there was disagreement over whether this is appropriate, especially if the testing dataset has not been declustered. To obtain data at large accelerations, participants noted that global datasets are required. For data at long return times, other indicators, such as precariously balanced rocks, might be used to test PSHA, because other data won't be available any time soon.

Participants discussed the utility of testing the final model ouput/forecast rather than the components. If the data don't match the PSHA predictions, what does this tell us about how the model is failing? PSHA (unlike EQ rates) is a composite forecast. We need to test individual components (source model, ground motion model, other branches of the logic tree) to guide development/improvements. An end-result test is an important exercise, but has limited feedback into development. Participants agreed that the useful evaluation results should suggest targeted model improvements.

To really understand why a model performed the way it did, it is necessary to pinpoint performance to individual data points and to work with modelers to assess whether the individual data point performance was significant in some way or unexpected. This interaction drives model improvements. Techniques that allow assessing this performance include individual likelihood scores and spatial residuals.

Ogata suggested that a focus should be on improving forecasts of the magnitude distribution as a function of time – because this is what drives the hazard.

There was discussion about the definition of usefulness. Is usefulness more valuable than correctness? Is there a tradeoff between these two? The use changes the scientific question, so it's important to test PSHA taking into account its utility. How do you measure utility -- via risk, monetary losses? It may be difficult or even impossible to measure or test usefulness. A better path may be to just solicit user requests from early adopters and find out what people want. But knowing the use is key. For instance, if your intended use allows a week of data gathering before the forecast is issued, the model we build will be very different from one that needs to make a prediction on day zero. This is a controversial proposal. Others in the audience want to see a separation between science and use (scientists should deliver knowledge to applied fields because there's no way to know what people will come up with.) Let's not let uncertainty about the use stall our scientific investigations.

Session 4: Induced Seismicity

Moderator: *M. Page* Reporter: *M. Taroni*

Presentations

A. Llenos showed results about induced seismicity hazard for the U.S. national hazard map. Some open questions about induced earthquakes included: where and how often they occur? How much higher is the ground motion that they cause? Which is the best way to forecast it? Is the Poisson assumption still valid? In this work Llenos and co-authors produce 1-year earthquake forecasts, by looking at the rate of the induced earthquakes. They investigate 17 areas in the U.S. and use as annual rate the maximum between long-term, 1-year and 2-year rates. In this manner they take into account the induced earthquakes (if there are any). They finally compare the rate obtained with the NSHM14 rates: the new rates are on average much higher.

Q&A included the reason for choosing the maximum of (long-term, 1-yr, 2-yr) rates, which is simplicity in a method that captures the 1-yr variation of the seismicity. Another topic of conversation included the availability of industrial (injection/withdrawal) data and its use in statistical models. Forecasts are updated annually and cover a 1-yr period, based on end-user input. Retrospective tests had not yet been performed. A suggestion was made to assess the assumption of iid Gutenberg-Richter magnitudes.

Bill Ellsworth started his presentation with a figure that shows the annual number of magnitude above 3 earthquakes in California and Oklahoma: there is a clear growth for earthquakes in OK. In the last year there are even more earthquakes in OK than in CA! According to Bill it is an open question how well we can forecast next year's seismicity. For seismicity in OK, there is a clear correlation with fluid injection. Bill illustrated an ARIMA method to forecast seismicity, which can be modified to include monthly injection volumes. USGS hazard maps that account for induced seismic hazard should be available in December 2015. Tom Jordan asked about the b-value, which Bill explained was a big problem: they assume a value of one, but smaller regions show substantial variations (up to 1.5).

Participants discussed the role and importance of declustering and characteristics of particular declustering methods. Apparently the Gardner-Knopoff declustering method deletes most earthquakes and leaves little except the M5.6 Prague earthquake. A stochastic declustering method was mentioned as an alternative. Marco Pagani asked which GMPEs will be used for induced earthquake hazard; these will be built ad hoc, as induced quakes are very shallow. But eventually it will be necessary to build specific and validated GMPEs.

Max Werner provided a brief update on the model installation by Brodsky et al for forecasting induced seismicity in the Salton Sea geothermal field. One issue is the availability of injection and withdrawal data from the DOGGR website.

Ogata suggested to use strainmeters in regions of induced seismicity, to capture potential aseismic deformation. Ellsworth stated that the strain rate is often indistinguishable from zero. Blanpied urged collecting InSAR data that are more accurate than point-wise GPS measurements.

Ogata suggested that induced seismicity may be similar to background seismicity variation due to magma intrusion, in the sense that similar modelling approaches may (should) be pursued.

Shaw asked whether focal mechanisms of induced earthquakes show any anomalies, which does not appear to be the case, as far as we can tell.

Recommendations

CSEP & OEF

Attendees agreed that CSEP evaluations could provide credible evidence about the predictive skills of operational earthquake forecasting models that will be or are deployed in real time. UCERF3-ETAS was identified as a prime target for independent testing with CSEP. UCERF3-ETAS forecasts are of a different type than typical CSEP forecasts, and testing them will require extending CSEP capabilities to deal with (many) stochastic event sets (simulations) as well as fault-based forecasts of finite ruptures. Additionally, the influence of real-time on the evaluation of forecasts needs to be assessed in more detail, in collaboration with real-time earthquake data product developers such as ComCat. In this context, retrospective experiments such as the Canterbury experiment were highlighted as providing useful information. 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.

Epistemic Uncertainty

Several recommendations emerged from the discussions. First, CSEP must address epistemic uncertainty by either assessing the potential impact on current test results or by developing ways to include epistemic uncertainties in forecasts. Assessing the potential impact could be performed via simulations, while including epistemic uncertainties will require relaxing the Poisson distribution. Second, it was recommended to address the more philosophical question of how to assess epistemic uncertainty and to test models with epistemic uncertainty. Third, epistemic uncertainty is represented in several ways, including in ensemble models, in logic trees and in Bayesian forecasts. How to specify forecasts and evaluate them appears to depend on the specific representation quite significantly. These were identified as important avenues for future research.

Evaluating Hazard Models

Attendees generally recommended further developing the evaluation of hazard models and their components but noted the challenges of data openness/availability, as well as the limited 'return times' currently available. Participants noted the importance of the feedback loop of using information from model evaluations to drive model improvements. Sanity checks were recommended as useful, as well as focussing on how to learn from these.

There continues to be debate about the importance of aftershocks in seismic hazard modelling and assessment. They are non-uniquely defined and solely retrospectively identifiable and in addition contribute to hazard, too. Nonetheless, extent models often assume aftershocks have been removed. Participants noted that the USGS plans to provide OK hazard maps that do not distinguish between mainshocks and aftershocks.

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.

Next Steps

CSEP & OEF

CSEP will continue to work with the USGS to test current and future versions of UCERF. Issues to be resolved include: testing forecasts of space-time correlated stochastic event sets; testing forecasts of finite ruptures; uniquely attributing earthquakes to faults; assessing the uncertainty of forecast evaluations on data incompleteness. CSEP has already installed GEM's GEAR1 model and is implementing new information scores (Kagan's) to evaluate this global high resolution forecast.

Epistemic Uncertainties

An important next step is to provide simple examples of how to specify and treat epistemic uncertainty in probabilistic forecasts and their evaluation. Several CSEP members will work on developing simple scenarios with shared data and forecasts to facilitate progress. CSEP will focus on both additive and multiplicative ensemble modelling of forecasts in California and Canterbury.

Evaluating Hazard Models

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.

Induced seismicity

CSEP will continue to 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. Current data availability and the most recently available date of open data access need to be followed up on.