2010 SCEC Annual Report

How Long Will it Take to Obtain Meaningful Test Results (and Distinguish Models) in CSEP?

Matthew C. Gerstenberger
and
David A. Rhoades

Institute of Geological & Nuclear Sciences, P.O. Box 30-368, Lower Hutt, New Zealand
m.gerstenberger@gns.cri.nz

J. Douglas Zechar

ETH-Zürich, Zürich, NO H 3, Sonneggstrasse 5, 8092 Zürich, Switzerland

This grant was to supply funding for M. Gerstenberger to travel to Zurich to work with Jeremy Zechar on providing insight into 1) how long it will take, using CSEP style tests, to detect significant differences between long term forecast models, and 2) what is the appropriate time-period over which a probabilistic seismic hazard model, such as the U.S. National Seismic Hazard Model (NSHM; Frankel et al, 2002), should be evaluated? David Rhoades was also actively involved in this research.

Gerstenberger spent one week in Zurich, during which time the methodology for answering the questions was specified in detail, simulations were run, and preliminary results were obtained.

To answer question one, there are two main considerations. The first being, how many earthquakes will it take to distinguish significant differences between two particular models? The second is, if seismicity is not stationary, how representative will any 5-year period be of the true performance of a model?

We are currently working with OpenSHA developers to obtain regional deaggregation results (i.e., not deaggregating for a single point, but for a large region) from the UCERF model to assist us in answering question two.

Declustering

Our aim is to understand the results of long-term model forecasts, which is defined by CSEP to be a static 5 year forecast. Several important models which belong to this class are created using declustered catalogs. For this reason we have focused our initial work on declustered data. All results presented below have been calculated with both Gardner Knopoff declustering and Reasenberg declustering. In both cases we have used the default parameters for the techniques.

Statistical Power

Using methods introduced by Zechar et al. (2010), we have examined the statistical power of the CSEP M-Test and S-Test to determine how many earthquakes are required before we can distinguish between two models. Statistical power depends on the models being compared, and in this approach, one must assume that one model is “true” and simulate catalogs consistent with the true model. We also note that power is not symmetric under these conditions, meaning that it may be harder to distinguish Model A from Model B if Model A is assumed to be true than if Model B is assumed to be true.

For this analysis, we considered three spatial distributions in the RELM testing area: that of the 1996 USGS national hazard map (Frankel et al.), that of the Helmstetter et al. RELM model (Helmstetter et al., 2007), and a uniform

distribution. We simulated earthquakes with magnitudes greater than or equal to 4.95. If we take a Type II error rate of 0.2 to indicate satisfactory power, and NSHM is the true model, one earthquake is likely to distinguish the uniform model, while five earthquakes are required to distinguish the Helmstetter model. If the Helmstetter model is the true model, 13 earthquakes are required to distinguish the NSHM model. We are currently working to see how the power of the S-test varies with different minimum magnitudes—for example, how many M6+ earthquakes would be required to distinguish NSHM and Helmstetter?

To explore the power of the M-Test, we compared the NSHM magnitude distribution of M4.95+ with the empirical magnitude distribution from the ANSS catalog. As one might expect, far more earthquakes are required to distinguish these distributions. If NSHM were the true magnitude distribution model, 56 M4.95+ events would be required to reach a Type II error rate of 0.2; if the magnitude distribution were stationary (i.e., the observed empirical distribution were true), only 14 earthquakes would be required.

We are still actively pursuing this work; we intend to recompute similar scenarios using the UCERF2 model.

Stationarity of Seismicity

While we can never fully anticipate what the changes in stationarity of seismicity will be in the future, we can use the existing seismicity catalog to gain an understanding of how the non-stationarity present in the catalog will effect CSEP style tests that are evaluated for a short time period (i.e., 5 years). Using the catalog as a starting point, we have created a series of simulated catalogs using different bootstrapping techniques; each one of these simulated catalogs is then used to evaluate forecast models. By examining the variance in the results of the tests, we can gain some insight into how representative any five year period is of the overall performance of the model.

Bootstrapping

We have used two methods of bootstrapping to creating the synthetic catalogs. Block bootstrapping is a method where a continuous subset of data is selected using randomly chosen starting points in time; this results in simulated catalogs of differing lengths. For example, we have created 10,000 simulated catalogs each for the time periods of 90 days, 1 year, 5 years, 10 years, 20 years and 50 years. In this case, the model forecast are evaluated only over the length of time of the bootstrapped catalog.

The second method we used, which we call synthetic catalogs is similar in that we randomly select continuous blocks of data from that catalog; however, in each case we string the sub-catalogs together to form total synthetic catalogs of five years. In this case we use smaller blocks of data: 1 day, 10

days, 30 days, 90 days, 1 year and 2.5 years. Again, 10,000 synthetic catalogs were created for each period and models were evaluated over the full five year time period.

In interpreting the results, it must be kept in mind that as the length of the time period grows longer, the independence of the simulated catalogs declines due to the length of observations available. This particularly effects the 20 year and 50 year simulations.

The Models and Tests

For the initial work we have done, we have evaluated the Helmstetter, Kagan and Jackson model (HKJ). We have used three standard CSEP tests: The N-Test, which examines the total number of events predicted for consistency with the observations; The S-Test, which evaluates the spatial distribution of the predicted earthquakes; and the M-Test, which evaluates the predicted magnitude distribution compared to the observations.

Preliminary Results

We are currently in the process of evaluating and understanding the test results. Some preliminary results of interest, which currently only pertain to the model investigated, include

- The power of the N-Test appears to be lacking. In the 20 and 50 year tests, the model is rejected; however in any five year period, the model is most likely to not be rejected.
- For a “good” model the expected result is a 5% chance of rejection (using an alpha of 5%). For the M- and S-Tests there is a systematic increase in rejection rate as the block size increase that goes beyond this 5% level.
- The HKJ model fails the S-Test using the true observations from 1932-2005. Because the percentage of rejections grow with increasing time length, it is likely that the eventual rejection comes from overall “slightly wrong” information, rather than from individual events. In other words, the model is not particularly bad at any location, but as more events are observed the discrepancy becomes clear.
- The results from the synthetic catalog tests are similar across all time scales. This indicates that the scale of clustering in the data is very small and on the order of < 1 day.

Overall, the preliminary results are potentially worrying for the value of a 5 year CSEP style tests and indicate that the results can change significantly depending on the length of the period investigated and the time period in which it occurs. However, in order to better understand these results we must evaluate several other models using our simulated catalogs. Future tested models will include a recently obtained gridded version of the UCERF2

model, the ALM model, a spatially uniform Poisson model and the Bird-a model

Finally, once we are able to produce regional deaggregation results, we will better be able to understand the time period that a typical PSHA model needs to be tested for.