Ensemble Earthquake Forecasting Models for the 2010 Darfield, New Zealand, Earthquake Sequence

Report for SCEC Award #15169
Submitted November 15, 2016

Investigators: Max Werner (Bristol), Matt Gerstenberger (GNS Science), Maria Liukis (USC) & Warner Marzocchi (INGV Rome)

I. Project Overview

A. Abstract

In the box below, describe the project objectives, methodology, and results obtained and their significance. If this work is a continuation of a multi-year SCEC-funded project, please include major research findings for all previous years in the abstract. (Maximum 250 words.)

The project's objective was to investigate the ability of ensemble models to forecast the 2010 Canterbury, New Zealand, earthquake sequence. Ensemble modeling refers to methodologies for merging forecasts from multiple models according to different criteria. Ensemble models offer two main advantages. First, no single model has to be selected a priori for making decisions. Second, merged forecasts may perform better than individual forecasts. Ensemble models may therefore contribute to Operational Earthquake Forecasting (OEF) systems. The Darfield earthquake sequence offers a unique opportunity for studying the performance of ensembles. The Collaboratory for the Study of Earthquake Predictability (CSEP) had previously assembled fourteen individual forecast models for the sequence. We developed four different ensemble models for the 1-day and 1-month forecasting experiments, using both real-time and best-available datasets. The weights of models in our ensembles are based on a measure of past performance. The weights reveal temporal fluctuations that show model performance evolves substantially over time. We ranked the ensembles against individual models using the likelihood metric. All ensemble models perform almost as well as the best individual model. This suggests that ensemble models provide a robust OEF approach for merging forecast without selecting one model a priori. However, the current approaches do not merge information in a complimentary manner to improve on the single best forecast (which is, however, only known a posteriori). Two ensemble models are now implemented within the California testing region and are running live.

B. SCEC Annual Science Highlights

Each year, the Science Planning Committee reviews and summarizes SCEC research accomplishments, and presents the results to the SCEC community and funding agencies. Rank (in order of preference) the sections in which you would like your project results to appear. Choose up to 3 working groups from below and re-order them according to your preference ranking.

Collaboratory for the Study of Earthquake Predictability (CSEP)
Earthquake Forecasting and Predictability (EFP)
Working Group on California Earthquake Probabilities (WGCEP)

C. Exemplary Figure

Select one figure from your project report that best exemplifies the significance of the results. The figure may be used in the SCEC Annual Science Highlights and chosen for the cover of the Annual Meeting Proceedings Volume. In the box below, enter the figure number from the project report, figure caption and figure credits.

Figure 1: Evolution of weights in the 1-month ensemble model forecasts over the 20 forecast horizons of the 2010-12 Canterbury earthquake sequence. Weights are calculated from the cumulative performance of individual models; performance metrics vary between ensemble models. Left panel: best-available data is used to generate forecasts and ensembles. Right panel: real-time data is used. The BMA and gSMA show strong sensitivity to past performance and interesting fluctuations. The SMA and PGSMA ensembles are more robust and conservative. [from Taroni, Marzocchi, Werner & Zechar, 2015, in prep.]

D. SCEC Science Priorities

In the box below, please list (in rank order) the SCEC priorities this project has achieved. See https://www.scec.org/research/priorities for list of SCEC research priorities. *For example: 6a, 6b, 6c*

2b, 2e, 2d

E. Intellectual Merit

How does the project contribute to the overall intellectual merit of SCEC? *For example: How does the research contribute to advancing knowledge and understanding in the field and, more specifically, SCEC research objectives? To what extent has the activity developed creative and original concepts?*

Our results contribute to SCEC's goals of improving the science of Operational Earthquake Forecasting (OEF). This project advanced our understanding of how forecasts can be merged for optimal performance. Bayesian Model Averaging (BMA) techniques are well known, but many other ensembles are possible and indeed perform comparatively well. Ensemble models also provide new tools for visualizing the performance of models over time.

F. Broader Impacts

How does the project contribute to the broader impacts of SCEC as a whole? *For example: How well has the activity promoted or supported teaching, training, and learning at your institution or across SCEC? If your project included a SCEC intern, what was his/her contribution? How has your project broadened the participation of underrepresented groups? To what extent has the project enhanced the infrastructure for research and education (e.g., facilities, instrumentation, networks, and partnerships)? What are some possible benefits of the activity to society?*

The INGV uses an ensemble model of CSEP-tested individual models to provide real-time information about the time-dependence of seismic hazards to the Italian Civil Protection Agency. Other government agencies, including the USGS, are making plans to deploy OEF systems. Our results show that the tested ensembles perform well: they are robust and suitable for OEF systems.

G. Project Publications

All publications and presentations of the work funded must be entered in the SCEC Publications database. Log in at *http://www.scec.org/user/login* and select the Publications button to enter the SCEC Pubications System. Please either (a) update a publication record you previously submitted or (b) add new publication record(s) as needed. If you have any problems, please email *web@scec.org* for assistance.

II. Technical Report

The technical report should describe the project objectives, methodology, and results obtained and their significance. If this work is a continuation of a multi-year SCEC-funded project, please include major research findings for all previous years in the report. (Maximum 5 pages, 1-3 figures with captions, references and publications do not count against limit.)

A. Project Objectives

The 2010 M_W7.1 Darfield, New Zealand, earthquake set off a complex and devastating earthquake cascade that has drastically increased seismic hazard estimates over the coming years and decades in the Christchurch and surrounding Canterbury region (Gerstenberger et al., 2014). The sequence provides a wealth of new scientific data to study earthquake clustering and to evaluate the predictive skills of time-dependent forecasting models. To this end, the Collaboratory for the Study of Earthquake Predictability (CSEP) is conducting a retrospective evaluation of fifteen short-term statistical and physics-based forecasting models that were developed by groups in New Zealand, Europe and the US. Our results may eventually contribute to operational earthquake forecasting systems that seek to disseminate credible information about time-dependent seismic hazards and risks to the public.

This assessment pitches well-known models such as the Epidemic Type Aftershock Sequence (ETAS) models and Coulomb/rate-state models against newly developed hybrid physical/statistical models that use the static Coulomb stress hypothesis to replace isotropic spatial aftershock footprints of statistical clustering models. Despite much research, the predictive power of the Coulomb hypothesis remains controversial. One goal of the retrospective evaluation is to assess whether hybrid statistical/Coulomb models improve on purely statistical models.

Another goal is to understand the influence of near-real-time data quality on the predictive skills of forecasts. The experiment therefore includes two testing modes. In the retrospective mode, models can access the best available data; in the pseudo-real-time mode, models can only access data available shortly after the earthquakes (e.g., preliminary slip models).

The objective of this specific project was to understand the performance of multiple ensemble modeling techniques during the Canterbury sequence, using both best-available and near-real-time datasets as model inputs. Ensemble models provide two advantages. First, they provide a framework for merging forecasts when multiple models are available and no individual model is known to be the best (a priori). This is particularly important for operational forecasts, where the objective is to provide the best possible forecast rather than understand the scientific basis. Second, merging model forecasts offers the opportunity to surpass the performance of individual models if the models provide complimentary predictive information.

B. Methodology

The CSEP experiment for the 2010-12 Canterbury sequence includes 15 models that start their forecasts on 4th September 2010 (right after the M7.1 mainshock of the sequence) and finish on 29th February 2012. Each forecast consists of a space-magnitude grid (0.1 by 0.1 spatial degree, 0.1 magnitude bins) that specifies for each spatial cell the number of events with magnitude bigger than 3.95 Ml forecasted for the next time window (1-month or 1-day) in the region of Canterbury. 394 events with magnitude Ml >3.95 are listed in the best available catalog for the Canterbury sequence, while the real time catalog contains only 217. This is a big difference, and we investigate also how results change by using one or the other catalog.

All forecasts were updated after the big shocks in the sequence in addition to regular 1-day or 1-month updates. The 1-month models provide 20 forecasts over the 18-month period (the additional two are a result of the regeneration of forecasts after the large quakes). The 1-day models provide 546 time-windows. We focus our analysis on forecasts and ensembles that are generated with real-time data to mimic an operational real-time setting. We used the best-available catalog to evaluate forecasts and ensembles because the best-available catalog reflects experienced earthquakes better. However, to understand the importance of improved data for forecasting skill, we repeat the analysis using the best-available catalog to generate forecasts and build ensembles.

Models and references are provided in Table 1.

Table 1: models participating in the CSEP Canterbury experiment

Model Name	Description	Authors/Reference
CRS-0	Basic Coulomb/rate-state	Cattania et al., 2014
CRS-1	Incl. uncertainties	Cattania et al., 2014
CRS-2	Incl. aftershock focal mechanisms as stress sources	Cattania et al., 2014
CRS-3	Incl. all aftershocks as stress sources	Cattania et al., 2014
CRS-4	Incl. hetereogeneous background stressing rate	Cattania et al., 2014
ETAS-0	Basic ETAS model	Bach & Hainzl, 2012
ETAS-1	With Coulomb-based spatial kernel	Bach & Hainzl, 2012
ETAS-2	With fault geometry instread of epicenter	Bach & Hainzl, 2012
ETAS-HW	ETAS model with early-aftershock smoothing	Helmstetter & Werner, 2014
K2	Space-time smoothing kernels	Helmstetter & Werner, 2014
K3	Space-time-magnitude smoothing kernels	Helmstetter & Werner, 2014
R-ETAS-0	ETAS-0 with Coulomb-based productivity	Hainzl et al., 2010
R-ETAS-1	ETAS-1 with Coulomb-based productivity	Hainzl et al., 2010
R-ETAS-2	ETAS-2 with Coulomb-based productivity	Hainzl et al., 2010
STEPCoulomb	STEP model with spatial Coulomb kernel	Steacy et al., 2014
SUP	Stationary uniform Poisson model	Rhoades

Ranking methods are based on the likelihood, that is the probability of observing the data by using a given earthquake model (see e.g.,. Zechar et al., 2010).

We build four types of ensembles: the Bayesian Model Averaging (BMA, Hoeting et al. 1999), the Score Model Averaging (SMA, Good 1952), the generalized SMA (gSMA, Marzocchi et al. 2012) and the Parimutuel Gambling Score Model Averaging (PGSMA, Taroni et al. 2013) by taking into account the correlation between forecasts (we use the same procedure of Marzocchi et al. 2012, the capped eigenvalue method of Garthwaite and Mubwandarikwa 2010).

Here, BMA weights are proportional to the cumulative likelihood; SMA weights are proportional to the inverse of the cumulative log-likelihood; gSMA weights are proportional to the inverse of the difference between the cumulative log-likelihood of one model and the model with largest log-likelihood; finally the PGSMA is proportional to the Parimutuel Gambling Score (PGS, Zechar and Zhuang 2010) obtained by the model. This PGS use a completely different approach (see Zechar and Zhuang 2010). The PGS is also different from the likelihood regarding the maximum loss: the likelihood can have a value equal to zero, and the log-likelihood can approach and equal minus infinity. There is no limit for poor performance of a model. The PGS instead is more conservative: we have both a minimum and maximum that a model can win or lose. In all four ensembles, past performance of individual models determines their weights in the average.

C. Results

The evolution of weights over the course of the earthquake sequence reflects the ensemble's sensitivity to past performance (Figure 1). The BMA and gSMA are both sensitive to past performance (weights fluctuate), while the SMA and PGSMA are much more robust or conservative (weights remain relatively constant). In the BMA framework, the data-generating ("true") model is often assumed to participate in the ensemble and the framework therefore quickly converges to the most likely model as the "true" model.

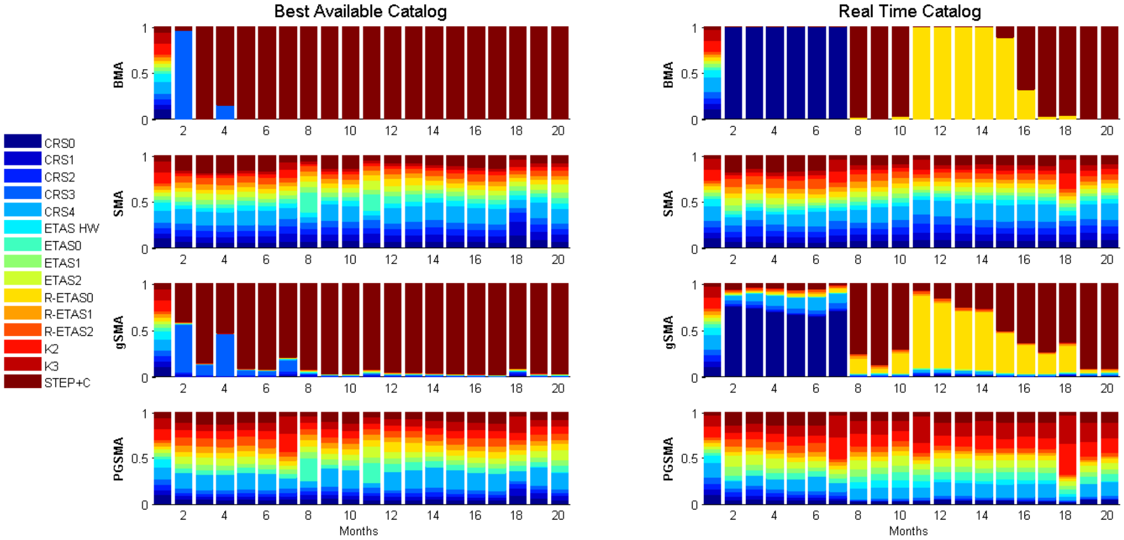

Figure 2: Evolution of weights in the 1-month ensemble model forecasts over the 20 forecast horizons of the 2010-12 Canterbury earthquake sequence. Weights are calculated from the cumulative performance of individual models; performance metrics vary between ensemble models. Left panel: best-available data is used to generate forecasts and ensembles. Right panel: real-time data is used. The BMA and gSMA show strong sensitivity to past performance and interesting fluctuations. The SMA and PGSMA ensembles are more robust and conservative.

The ensembles that are more sensitive to past performance (BMA, gSMA) show interesting fluctuations in relative performance. If the best-available catalog is used to construct forecasts and ensembles, then the STEPCoulomb model dominates the ensemble after only a few forecast periods. The CRS-3 model is the only model to also contribute substantially. However, these contributions are limited to early periods. In contrast, the ensembles constructed from real-time data show some significant differences. Here, the dominating model fluctuates from CRS0 to STEPCoulomb to R-ETAS0 and back to STEPCoulomb.

The changes in the ensemble-dominating models in the real-time data case may result from the poor data quality. The real-time catalog is of much poorer data quality, and we see that model performance and the composition of ensembles are strongly affected (compared to the best-available data case).

The BMA approach reveals its flaws in the real-time data case. The data-generating model is certainly not amongst our available models, and the relative performance of the models changes over time. This effective non-stationarity in model performance appears to violate the BMA assumption that the "correct" model is amongst the available model, or at least that one model is closest to the "true" model. Here, however, the model closest to the "true" model changes over time.

Irrespective of philosophical questions about the BMA approach, the fluctuations in the weights provide a neat summary of cumulative relative model performance. The next steps are to identify the causes of the fluctuations.

The ensemble models rank highly amongst the individual models. Figure 2 shows the cumulative probability gains per earthquake relative to the best individual model (STEPCoulomb) using the real-time dataset (wide bars) and the best-available datasets (narrow bars). The red ensemble models are ranked 2nd through 4th and 7th amongst the 19 models in the real-time data set. Using best-available data, the ensemble models perform even better: three ensembles outperform all individual models, while the PGSMA ensemble outperforms all but the STEPCoulomb model.

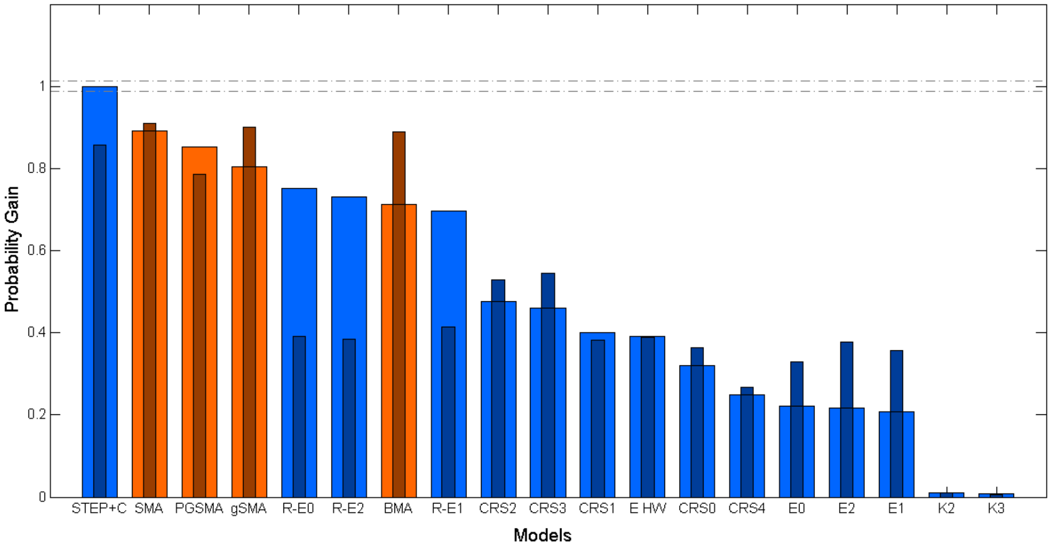

Figure 3: Probability gains of individual models in blue and ensemble models in red, ranked from left to right using the real-time dataset. Thin bars show gains when forecasts and ensembles are constructed using best-available data.

D. Significance

All tested ensemble models perform well compared to individual models. Ensembles tend to rank highest or perform slightly worse than the best individual model. These results provide evidence that ensemble modeling is an effective tool to merge forecasts when multiple models are available and no a priori model selection is desired.

Robust/conservative ensemble models have similar predictive skill as ensembles (e.g. BMA) that depend sensitively on the performance of individual models. The fluctuations in the weights may not be required from a predictive point of view, and more robust ensembles are therefore desirable. Additional ensemble methods should be tested for their performance.

High-quality datasets improve the relative skill of most ensembles. The relative performance of only one ensemble decreased when best-available data was used. However, the cumulative performance of the best individual model (STEPCoulomb) actually decreased when best available data is used to generate the forecasts. This highlights the role of potential statistical fluctuations in the probability gain that require further analysis.

The evolution of weights in an ensemble provides a succinct summary of relative cumulative model performance. CSEP has implemented the BMA ensemble for prospective testing in the California region.

E. References

Cattania, C., S. Hainzl, L. Wang, F. Roth, and B. Enescu (2014), Propagation of coulomb stress uncertainties in physics-based aftershock models, Journal of Geophysical Research: Solid Earth, doi:10.1002/2014JB011183.

Hainzl, S., G. B. Brietzke, and G. Zoller (2010), Quantitative earthquake forecasts result- ing from static stress triggering, J. Geophys. Res. (Solid Earth), 115(B14), B11,311, doi: 10.1029/2010JB007473.

Helmstetter, A., and M. J. Werner (2014), Adaptive smoothing of seismicity in time, space, and magnitude for time-dependent earthquake forecasts for California, Bulletin of the Seismological Society of America, 104(2), 809–822, doi:10.1785/0120130105.

Steacy, S., M. Gerstenberger, C. Williams, D. Rhoades, and A. Christophersen (2014), A new hybrid Coulomb/statistical model for forecasting aftershock rates, Geophysical Journal International, 196(2), 918–923, doi:10.1093/gji/ggt404.

Taroni, M., W. Marzocchi, M. J. Werner and J. D. Zechar (2015, in preparation), Operational Ensemble Model Earthquake Forecasting: Case Study 2010-12 Canterbury, New Zealand, Sequence.

Zechar, J. D., M. C. Gerstenberger, and D. Rhoades (2010), Likelihood-based tests for evaluat- ing space-rate-magnitude earthquake forecasts, Bull. Seismol. Soc. Am., 100(3), doi:10.1785/ 0120090192.