Summary of Results from SCEC grant 25166 during February 1, 2025 - January 31, 2026, Potency-weighted scoring metrics for assessing CSEP Forecasts.

October 11, 2025

1 Overview.

The Collaboratory for the Study of Earthquake Predictability (CSEP) hosts a variety of models in prospective experiments that predict the rate of earthquakes in particular magnitude ranges occurring in a given spatial-temporal grid cell. According to analyses by Schorlemmer et al. (2010) and Zechar (2013) the ETAS model appeared to offer the best fit for the first several years of CSEP. However, most metrics used to evaluate these forecasts weight all earthquakes equally, essentially, even though forecasting the very few largest events is of primary importance in practice. This research extends standard scoring metrics such as the log-likelihood score and Brier score so that they properly weight earthquakes according to their potency. We evaluate short-term (daily) models in CSEP, to determine which models have the highest efficacy at forecasting according to these metrics. Weighting earthquakes by their potency in the model evaluation enables the identification of models that best succeed at forecasting the largest events. In particular, we found very comparable performance of the STEP and ETAS, with slightly superior performance of the STEP model compared to ETAS using potency-weighted metrics.

2 Data

The CSEP modeling and testing region was designed to include all earthquakes with a magnitude of 3.95-8.95 in California and approximately 1^o of longitude and latitude around it. The space of this region is divided into square cells with sizes of 0.1^o longitude by 0.1^o latitude and each cell is binned with each bin representing a .05 increase in magnitude. Thus, a model seeking to make a prediction for the conditional intensity of an earthquake must make predictions for each of the 100 bins of magnitude ranges for each $.1^o$ x $.1^o$ cell.

From CSEP, files for STEP and ETAS were provided for every day's prediction from 2013, the first year of live predictions from both models, to 2017, a 5 year period. Additionally, the ground truth data of earthquakes in California was provided through CSEP's API.

In order to combat the skewness problem in Pearson Residuals and allow for a large enough sample of ground truth data to be used for computing Voronoi Residuals, the magnitude bins for each cell were added together such that for each cell, the predicted conditional intensity could be interpreted as the intensity corresponding to an earthquake occurring with a magnitude $M \geq 3.95$ and $M \leq 8.95$. Next, these intensities were summed within each spatial grid cell, for each day over the 5 year period. This results in a a single predicted conditional intensity value for each cell that represents the overall intensity of an earthquake occurring within the cell from 2013-2017.

3 Methods

Our research builds on recent efforts to assess rigorously the space-time models used to forecast earthquake occurrences (e.g. Schoenberg 2003, Vere-Jones and Schoenberg 2004, Baddeley et al. 2005, Schorlemmer et al. 2007, Schorlemmer et al. 2010, Clements et al. 2011, Clements et al. 2012, Bray et al. 2014, Gordon et al. 2015, Fox et al. 2016, Catania et al. 2018, Taroni et al. 2018, Gordon and Schoenberg 2020). The model evaluation methods used in these efforts have often relied on tools such as the N-test and L-test that lack the requisite power to discern between closely competing models or to indicate where and when one model might be fitting relatively poorly. Further, such methods as well as other tests and residual methods typically assign essentially equal weight to each earthquake, and thus do not properly reward models that more accurately forecast the largest and most important events.

Instead, we proposed potency-weighted measures for evaluating the goodness-of-fit of such models, and evaluated the performance of these models in CSEP. In particular, the two most promising measures we proposed were the potency-weighted log-likelihood score and the Q-score, a quotient indicating the extent to which the model forecasts the locations and times of the largest 5% of events relative to the other events. Specifically, the Q-score takes the form

$$Q = \frac{mean\{\lambda_i : m_i > m^{[.95]}\}}{mean\{\lambda(t,x,y)\}},$$

where $m^{[.95]}$ is the 95th percentile of the magnitude distribution, estimated e.g. based on prior seismicity, and the denominator $mean\{\lambda(t,x,y)\}$ is simply the mean forecast rate according to the model, over the entire space-time region. The higher the value of Q, the better the model appears to be forecasting the spatial-temporal locations of the largest 5% of events. For a homogeneous Poisson model with uniform magnitude density, Q will be close to 1. Any model that adequately accounts for the spatial inhomogeneity of seismicity will have $Q > 1$, as it should since such a model will tend to vastly outperform a homogeneous Poisson model at forecasting the largest events. Among competing models, the model that forecasts the larger events more accurately will tend to be the model with higher Q, especially if all the models are similarly calibrated overall [i.e. $mean\{\lambda(t,x,y)\}$ is close to the overall rate of seismicity] which can readily be checked via other methods, such as the N-test.

We also investigated the statistical properties of these measures, especially the quotient Q, to determine its anticipated mean and variance under various conditions.

4 Main results

We found that ETASOneDay models typically performed best among CSEP models in terms of Q score and potency-weighted log-likelihood. The STEP-JAVA model's results were highly variable, as STEP performed best in 2012 and 2nd best in 2014 among all models in CSEP, but worst among all CSEP models in 2017. It also became clear during this research that the models in CSEP are as a whole too homogeneous, and more alternative models that are fundamentally different from ETAS should be included for comparison.

The fact that STEP appears to outperform ETAS in prospective CSEP testing during many of the years of CSEP, both in terms of root-mean-square residual size and in terms of potency-weighted metrics such as the Q score, is rather surprising. Our results suggest that further modification, improvement, calibration, and improved estimation of parameters of STEP models may deserve the kind of attention that similar methods for ETAS has generated over the past two decades.

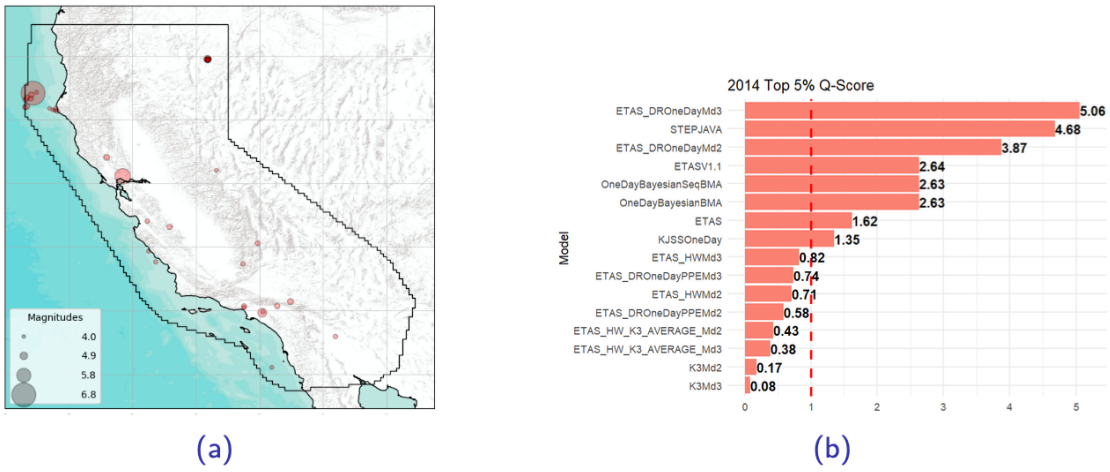

Figure 1: (a) 44 events above magnitude 3.95 in California in 2014, with three in the top 5% of magnitudes (M5.08). (b) Q-scores of various CSEP models for the year 2014. STEP-JAVA has the 2nd highest Q score among all CSEP models for 2014.

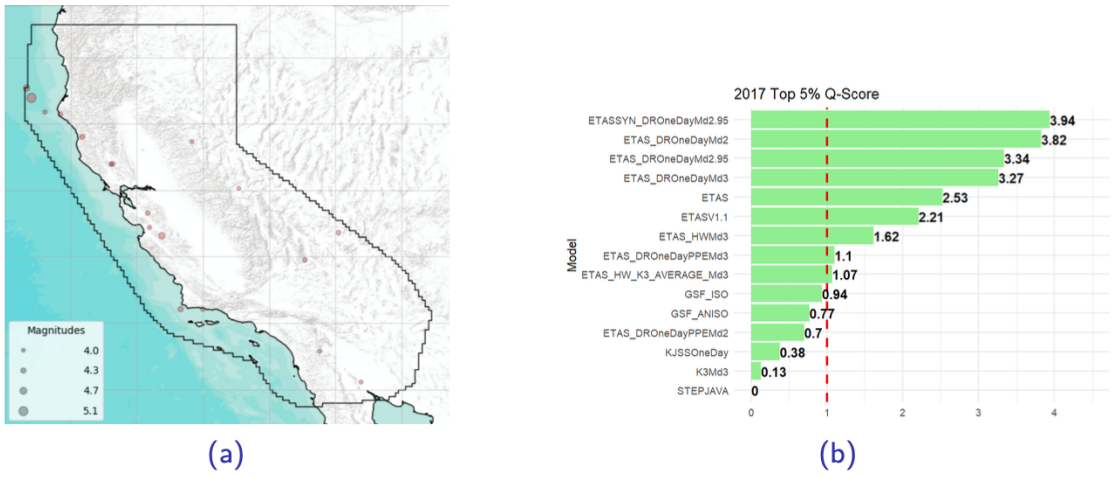

Figure 2: (a) 18 events above magnitude 3.95 in California in 2017, with one in the top 5% of magnitudes (M5.01, 6.02, 6.80). (b) Q-scores of various CSEP models for the year 2017. STEP-JAVA lowest Q score among all CSEP models for 2017.

We were also able to prove theoretical results regarding the mean and variance of Q, including showing that Q is asymptotically ratio-unbiased for magnitude-separable models under quite general conditions, and the variance of Q is extremely high relative to its mean.

References

Baddeley, A., Turner, R., Moeller, J., and Hazelton, M., 2005. Residual analysis for spatial point processes. Journal of the Royal Statistical Society B, 67(5), 617?666.

Bray, A., Wong, K., Barr, C.D., and Schoenberg, F.P., 2014. Voronoi cell based residual analysis of spatial point process models with applications to Southern California earthquake forecasts. Annals of Applied Statistics, 8(4), 2247-2267.

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

Clements, R.A., Schoenberg, F.P., and Veen, A., 2012. Evaluation of space-time point process models using super-thinning. Environmetrics, 23(7), 606-616.

Fox, E.W., Schoenberg, F.P., and Gordon, J.S., 2016. Spatially inhomogeneous background rate estimators and uncertainty quantification for nonparametric Hawkes point process models of earthquake occurrences. Annals of Applied Statistics 10(3), 1725-1756.

Gordon, J.S., Clements, R.A., Schoenberg, F.P., and Schorlemmer, D., 2015. Voronoi residuals and other residual analyses applied to CSEP earthquake forecasts. Spatial Statistics , 14b, 133-150.

Gordon, J.S., and Schoenberg, F.P., 2020. A nonparametric Hawkes model for forecasting California seismicity . BSSA , in review.

Jordan, T. H., 2006. Earthquake predictability, brick by brick. Seismological Research Letters 77, 3-6.

Ogata, Y., 1998. Space-time point-process models for earthquake occurrences, Ann. Inst. Statist. Math., 50(2), 379-402.

Schoenberg, F.P., 2003. Multi-dimensional residual analysis of point process models for earthquake occurrences. J. Amer. Statist. Assoc. 98(464), 789?795.

Schoenberg, F.P., Gordon, J.S., and Harrigan, R., 2018. Analytic computation of nonparametric Marsan-Lengliné estimates for Hawkes point processes. Journal of Nonparametric Statistics , 30(3), 742-757.

Schoenberg, F., and Schorlemmer, D. (2024). Critical Questions About CSEP, in the Spirit of Dave, Yan, and Ilya. Seismological Research Letters 95 (6), 3617-3625.

Schorlemmer, D., and M. C. Gerstenberger, 2007. RELM testing Center, Seism. Res. Lett., 78(1), 30-35.

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

Schorlemmer, D., J. D. Zechar, M. Werner, D. D. Jackson, E. H. Field, T. H. Jordan, and the RELM Working Group, 2009. First results of the Regional Earthquake Likelihood Models Experiment, Pure and Applied Geophysics, 167, 859-876.

Taroni, M, Marzocchi, W, Schorlemmer, D, Werner, M, Wiemer, S, Zechar, JD, Heiniger, L and Euchner, F, 2018. Prospective CSEP evaluation of 1-day, 3-month, and 5-yr earthquake forecasts for Italy. Seismological Research Letters, 89, 1251-1261.

Vere-Jones, D. and Schoenberg, F.P., 2004. Rescaling marked point processes. Aust. N. Z. J. Stat. 46, 133-143.