Summary of Results from SCEC grant 21051 during 2021-2022, Statistical evaluation of CSEP forecasts.

March 15, 2022

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. In particular, variants of the Epidemic-Type Aftershock Sequence (ETAS) (Ogata 1988) and Short-Term Earthquake Probabilities (STEP) models (Jordan 2006) have been used for earthquake forecasting and are entered as forecast models in the CSEP experiments. 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. In this project, we evaluated the prospective fit of the ETAS and STEP one-day forecast models for California from 2013-2017, using super-thinned residuals and Voronoi residuals. We found very comparable performance of the two models, with slightly superior performance of the STEP model compared to ETAS according to most 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. Indeed, the N-test and L-test simply examine the quantiles of the total numbers of events in each pixel or likelihood within each pixel, in comparison with those expected under the given model, and the resulting low-power tests are typically unable to discern significant lack of fit unless a model fits extremely poorly. Further, even when the tests do reject a model, they do not typically indicate where or when the model fit poorly, or how it could be improved.

Recent statistical developments in the assessment of space-time point process models have resulted in new, powerful model evaluation tools, and we applied these techniques to assist in the comparison and improvement of models for earthquake occurrences. These tools include residual point process methods such as super-thinned residuals and Voronoi deviances, which can be used to help detect inconsistencies between data and models and to suggest areas where models can be improved.

Voronoi residuals (Bray et al. 2014) and Voronoi deviances (Clements et al. 2011) are useful for evaluating gridded forecasts especially when a substantial proportion of pixels have very small integrated conditional intensities. Furthermore, Voronoi based residual methods offer advantages over grid based residuals in that with the former type of residuals, the spatial partition is data-driven and spatially adaptive, and the resulting distribution of residuals is usually far less skewed in such situations than residuals integrals over fixed rectangular grid cells (Bray et al. 2014). For any point in a point pattern, one may define its corresponding Voronoi cell as the region consisting of all locations that are closer to the observed event than to any of the other points. A Voronoi tessellation is the collection of such Voronoi cells. Voronoi residuals, meaning the difference between the integrated conditional intensity and the observed number of points in each Voronoi cell, were shown to be considerably less skewed than pixel residuals in Bray et al. (2014). There are issues needing further exploration, however, such as the determination of an appropriate color scale, and partial solutions have been proposed in Bray et al. (2014) and Gordon et al. (2015). Competing point process models can be compared using Voronoi deviance analysis, where one considers the difference between the log-likelihoods of the two point process models over Voronoi cells. If the difference is close to 0, then the two models fit about equally well in the given cell. Large Voronoi deviance residuals indicate places where one model fit substantially better than the other, and the sign of the residual indicates which model had superior fit.

Super-thinned residuals (Clements et al. 2012) are also useful to compare the goodness of fit for two models. In super-thinning a given model with estimated conditional intensity λ, one first chooses an appropriate value of k, thins the point process N, keeping each point τ_i independently with probability $\min\{k/\lambda, 1\}$, and then adds on points simulated according to a Cox process directed by $\max\{k - \lambda, 0\}$. The resulting points are called super-thinned residuals, and should look like uniformly distributed with rate k if and only if the modeled conditional intensity is correct almost everywhere (Clements et al. 2012).

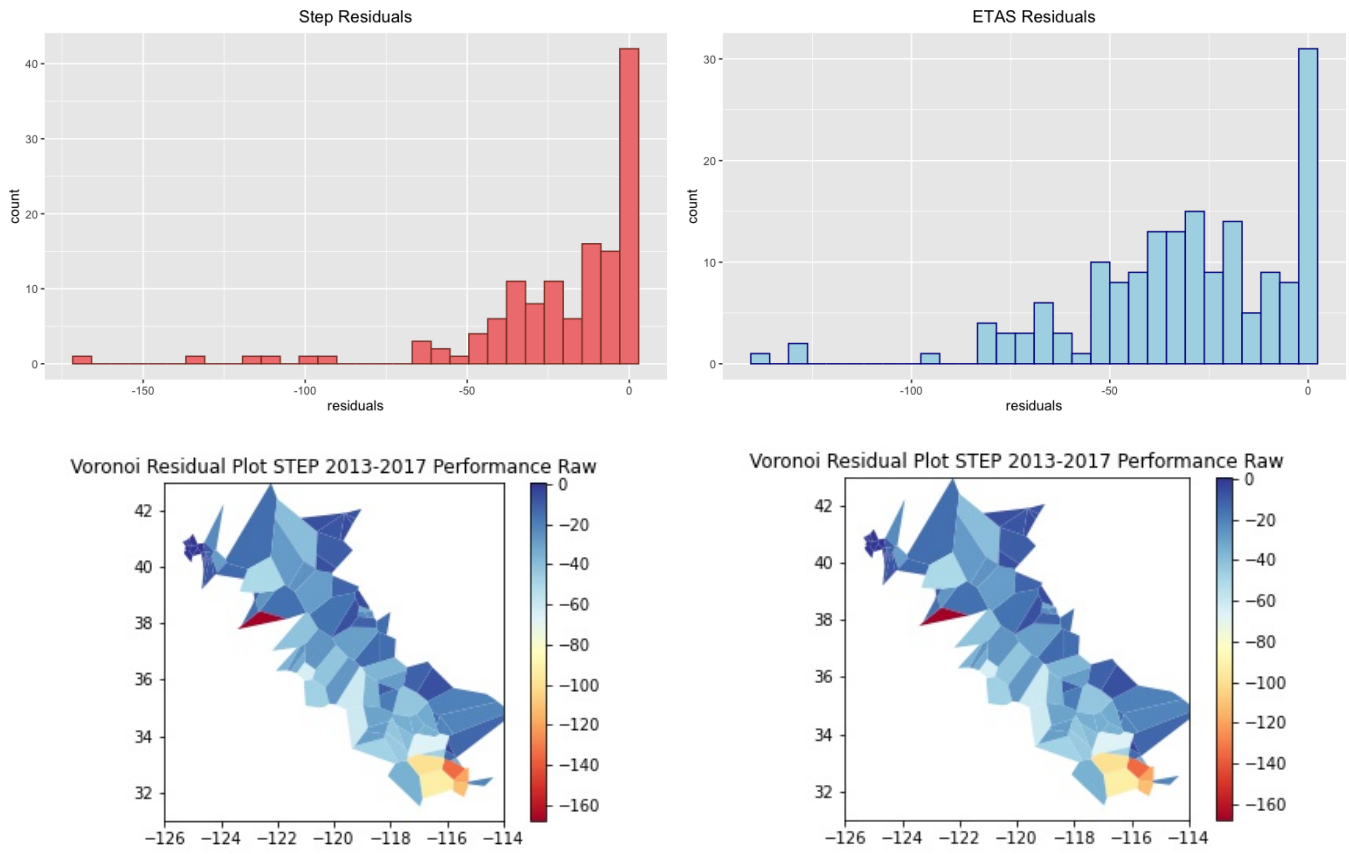

Figure 1: Top row: histogram of raw Voronoi Residuals for STEP (left) and ETAS (right). Bottom row: Spatial plot of Voronoi residuals for STEP (left) and ETAS (right). Blue indicates better fit.

4 Main results

The Voronoi residuals for STEP were generally smaller in absolute value than the corresponding residuals for ETAS, indicating overall better forecasting performance during this time period. The Voronoi residuals for each model are shown in Figure 1. The log likelihood for STEP was higher (-3367 for STEP versus -4243 for ETAS), again indicating better fit.

From the 2-Way Repeated Measures ANOVA, both the effects of the models and of the cell size binning along with an interaction of the two were shown to be significant at the $\alpha = 0.05$ significance level. Additionally, post-hoc Bonferoni corrected testing showed that there was a statistically significant difference at the $\alpha = 0.01$ significance level between the model difference in means for cells ranging from areas of 0 to 0.25 and 1 to 3. In other words, the residual mean for STEP was significantly smaller than that of ETAS in these cell area ranges, indicating that STEP fit better than ETAS in areas that have both the lowest and highest frequencies of earthquakes. Differences in all other ranges were found not to be statistically significant. A histogram of the ANOVA design can be seen in Figure 2 and qq-plots of the residuals of the ANOVA can be seen in Figure 3. The results suggest the approximate normality of the ANOVA residuals in this comparison.

The fact that STEP appears to outperform ETAS in prospective CSEP testing from 2013-2017, both in terms of root-mean-square residual size and in terms of overall behavior of the Voronoi residuals is rather surprising. Our results suggest that further modification, improvement, calibra-

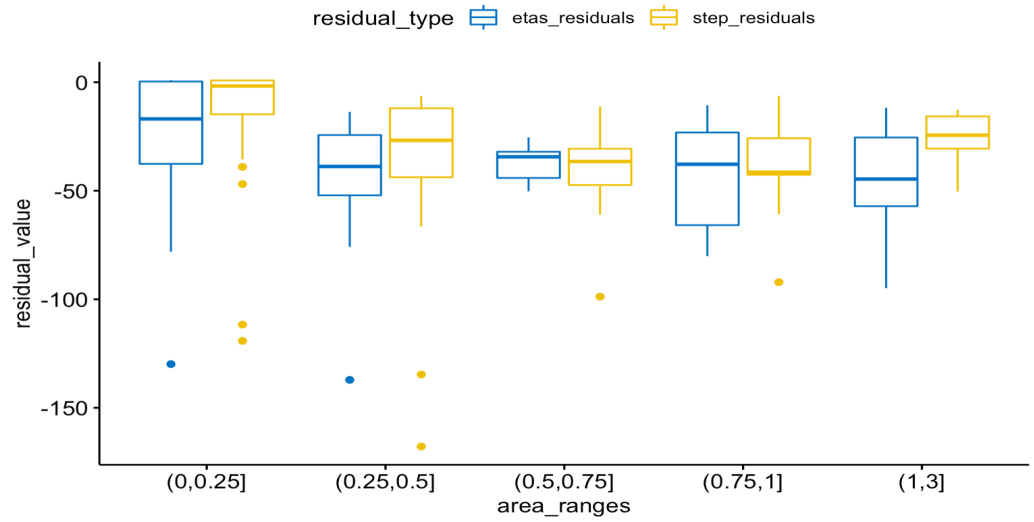

Figure 2: Post-hoc Bonferroni tests showing that differences in means in the ranges [0, 0.25] and [1,3] were statistically significant at the $\alpha = .0001$ and $\alpha = .01$ significance levels, respectively.

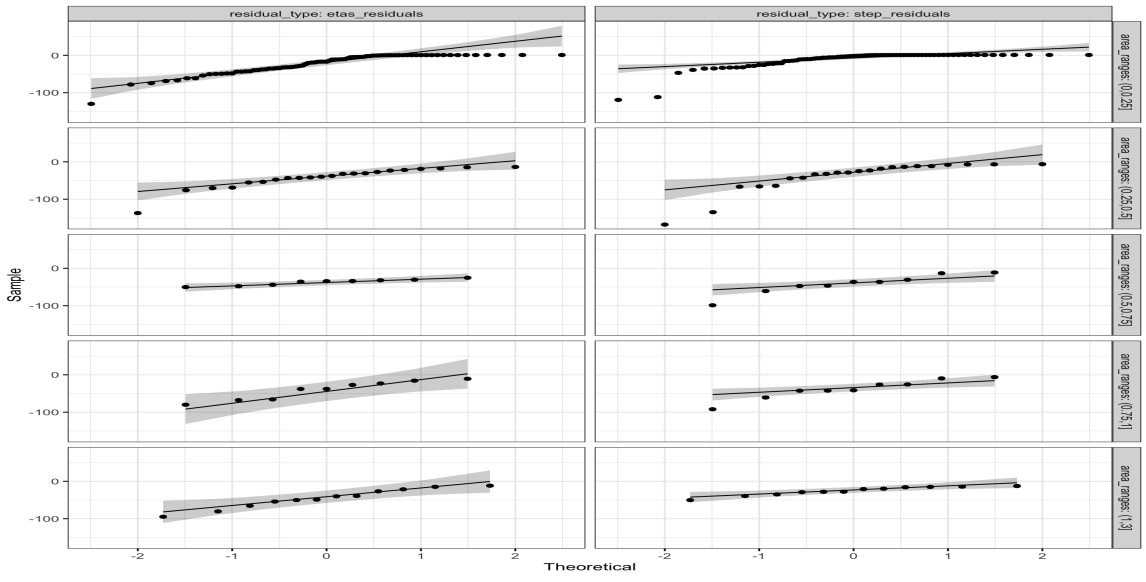

Figure 3: QQ-plots for the residuals of the two-way repeated measures ANOVA.

tion, 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.

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.

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.