Evaluation of ETAS and STEP forecasting models for California seismicity using point process residuals.

Joshua Ward[1], Maximilian Werner[2], William Savran[3], Frederic Schoenberg[1,4]

August 26, 2024

[1] Department of Statistics, University of California, Los Angeles, 90095-1554, USA.
[2] Southern California Earthquake Center, University of Southern California, USA.
[3] School of Earth Sciences, University of Bristol, UK.
[4] Corresponding author, frederic@stat.UCLA.edu.

Abstract

Variants of the Epidemic-Type Aftershock Sequence (ETAS) and Short-Term Earthquake Probabilities (STEP) models have been used for earthquake forecasting and are entered as forecast models in the purely prospective Collaboratory Study for Earthquake Predictability (CSEP) experiment. Previous analyses have suggested the ETAS model offered the best forecast skill for the first several years of CSEP. Here, we evaluate the prospective forecasting ability of the ETAS and STEP one-day forecast models for California from 2013-2017, using super-thinned residuals and Voronoi residuals. We find very comparable performance of the two models, with slightly superior performance of the STEP model compared to ETAS according to most metrics.

Keywords: CSEP, earthquakes, seismology, self-exciting point process, super-thinning, Voronoi deviance residuals.

1 Introduction

Since 2007, the Collaboratory for the Study of Earthquake Predictability (CSEP) has hosted a variety of models in prospective experiments that forecast the rate of earthquakes in particular magnitude ranges occurring in each given spatial-temporal grid cell [3, 27, 37, 34]. In order to remedy the problems of overfitting, publication bias, and other problems common in retrospective analyses of earthquake forecasts, the forecasts in CSEP are made on a purely prospective basis: the computer code for calculating forecasts must be made and submitted completely in advance, and no changes or retrospective adjustments may be made by humans after the fact [13]. Thus, CSEP offers seismologists a unique perspective into the abilities of models for earthquake occurrences to forecast seismicity in a purely prospective way, and thus a statistically sound platform for evaluating the relative performances of these models.

In particular, variants of the Epidemic-Type Aftershock Sequence (ETAS) ([17], [18]) and Short-Term Earthquake Probabilities (STEP) models [10] have been used for earthquake forecasting and have been entered as forecast models in the CSEP experiments. Several studies have suggested that, for the first several years of CSEP's operation, the ETAS model has appeared to offer one of the best forecasts among the models in CSEP ([6],[32],[36]). In this project, we evaluate and compare in detail the prospective forecast skill of the ETAS and STEP one-day forecast models for California in a one year interval of 2016-2017 and a five year interval of 2013-2017 using super-thinned and Voronoi residual based methods. Overall, we find very comparable performance of the two models, with slightly superior performance of the STEP model compared to ETAS according to most metrics.

Short-term forecasts of earthquake occurrences lead to unique challenges that make traditional evaluation and comparison techniques difficult. For one, very few earthquakes exceeding a minimum magnitude of 3.95 are recorded in California on a day-to-day or even annual basis. Therefore, for most models in CSEP, the estimated conditional intensities are very close to zero, and can be quite volatile. Under such conditions, traditional residual testing techniques can suffer from low power, numerical instability, and high variance [5]. To resolve these difficulties, some binning or smoothing techniques can be useful, to facilitate a more robust evaluation of the long-term success of these models with likelihood-based metrics.

Our research builds on recent efforts to rigorously assess the space-time models used to

forecast earthquake occurrences (e.g. [1], [2], [3], [5], [6], [7], [8], [9], [11], [12], [15], [16], [21], [22], [23], [25], [29], [32], [33], [35], [36], and [37]). The model evaluation methods used in some of these efforts have relied on tools such as the N-test, CL-test and S-test that are generally not ideally powerful at discerning between closely competing models and generally do not indicate where and when one model might be performing relatively poorly. Indeed, the N-test and CL-test examine the quantiles of the total numbers of events in the pixels or likelihoods over all pixels, in comparison with those expected under the given model, and the resulting tests typically have limited power [1]. Further, even when these tests do reject a model, they do not typically indicate where or when the model performed poorly, or how it could be improved. An additional problem with such tests is the reliance on the assumption of independence across cells and a Poisson distribution for the number of points in each cell, which is particularly troublesome for short-term forecasts [30].

Recent statistical developments in the assessment of space-time point process models have resulted in new, powerful model evaluation tools, and we apply these techniques to assist in the comparison and improvement of models for earthquake occurrences. Specifically, we apply residual point process methods including super-thinned residuals [8] and Voronoi deviances [5], which are useful to help detect inconsistencies between data and models and to suggest areas where models can be improved.

The structure of the remainder of this paper is as follows. After a brief description of the data and the binning techniques used in Section 2, evaluation and comparison techniques are reviewed in Section 3, and Section 4 summarizes the results. Finally, a discussion is given in Section 5.

2 Data

The CSEP modeling and testing region was designed to include all recorded shallow earthquakes of magnitude 3.95 and higher within a region covering California and extending approximately 1^o in longitude and 1^o in latitude in each direction around the state of California. This spatial region is divided into square cells with sizes of 0.1^o longitude by 0.1^o latitude, and the observations are further divided into 100 magnitude bins of length .10

each, ranging from 3.95 to 8.95. Thus, a model seeking to make a prediction must output a forecasted expected number of earthquakes occurring in each of the 100 bins of magnitude ranges for each $.1^o$ x $.1^o$ spatial cell. Since STEP was one of the first models to be introduced into the CSEP experiment, and since some changes to CSEP were made since its inception in 2007, STEP only makes predictions on a reduced subset of this spatial region (see Figure 1). We therefore restrict the evaluation of both STEP and ETAS to be the regions they both share in common for the current evaluation. Prospective forecasts for STEP and ETAS were provided via CSEP for every day over the 5 year period from January 1, 2013, to December 31, 2017. Additionally, California earthquake data was provided from the ANSS Comprehensive Earthquake Catalog (ComCat) [31]; queried from CSEP's python package pyCSEP [20]. The fields included in this were the longitudes, latitudes, magnitudes, depths, and times of each earthquake through the above 5 year period.

3 Methods

For each of the spatial seismicity forecast models, in order to address the high variance inherent in residual analyses with very small amounts of data, we focus on the aggregated conditional intensity estimates summed over all 100 magnitude bins and over all days in the 5-day period, so that the focus here is on the purely spatial distribution of the observed earthquakes and their corresponding forecasts.

Voronoi residuals [5] and Voronoi deviances ([5], [8]) are useful for evaluating gridded forecasts especially when a substantial proportion of pixels have very small integrated conditional intensities. For any point in a point pattern, one defines its corresponding Voronoi cell as the region consisting of all locations that are closer to the observed event than to any of the other events, and a Voronoi tessellation is the collection of such Voronoi cells. Voronoi residuals, defined as the difference between the integrated conditional intensity and the observed number of points in each Voronoi cell, are not only spatially adaptive and entirely data-driven, but in addition have been shown to be considerably less skewed than pixel residuals ([5], [28]). Choices for appropriate color scales when plotting Voronoi residuals have been proposed in [5] and [11].

Two competing point process models can also be compared using Voronoi deviances,

which are the differences between the log-likelihoods of the two point process models, integrated over each Voronoi cell. If the deviance in a particular cell is close to 0, then the two models forecast about equally well in the cell. Large Voronoi deviance residuals indicate locations where one model substantially outperforms the other in terms of forecast skill, and the sign of the residual indicates which model had superior performance. Voronoi residuals can also be used for hypothesis testing. Here, in order to see if different models have different forecast skill in regions of varying degrees of seismic activity, we conduct a 2-way repeated measure ANOVA to test if the difference in mean Voronoi residuals for the two models, STEP and ETAS, is statistically significant across different Voronoi cell sizes.

Super-thinned residuals [8] are also useful to compare the performance for two models. In super-thinning, given a model with estimated conditional intensity $\hat{\lambda}(x, y, t)$, one first chooses an appropriate value of k, thins the point process N by keeping each point $\tau_i = (x_i, y_i, t_i)$ independently with probability $\min\{k/\hat{\lambda}, 1\}$, and then superposes points simulated according to a Cox process directed by $\max\{k - \hat{\lambda}, 0\}$. The resulting points are called super-thinned residuals, and should look uniformly distributed with rate k if and only if the modeled conditional intensity is correct almost everywhere [8]. Thus one may inspect these residual points for uniformity as a way of assessing the performance ability of the model. For instance, one may use the centered, variance-stabilized L-function [4], where the L-function is defined as $L(r) = \sqrt{K(r)/\pi} - r$, and K is Ripley's K-function [19], indicating the normalized average number of other residual points within distance r of any given residual point, so that values of L greater than zero indicate clustering in the residuals and values less than zero indicate inhibition.

Since estimates of the conditional intensity λ at each observed earthquake is required in order to compute Voronoi deviance residuals and super-thinned residuals, we estimate λ at each such point by dividing the modeled expected number of earthquakes per cell by the cell size. Essentially this is assuming that, according the model, λ is constant within each grid cell.

4 Results

Figure 2 shows Voronoi residuals for STEP and ETAS for the California data over the 5-year period from 1/1/2013 to 12/31/2017. One sees that overall, STEP appears to forecast seismicity more accurately than ETAS, preferring to slightly over-estimate whereas ETAS usually under-estimates. While STEP dramatically under-predicts for the earthquake around the San Francisco Metropolitan Area, most cells are better specified. Indeed, Figure 3 shows a plot of the same cells colored by which model's predictions are closer to 0. STEP outperforms ETAS here, especially in the interior of California. The overall log-likelihoods for the STEP and ETAS models are -308.7 and -350.2, respectively, again indicating slightly superior overall forecast skill for the STEP model.

Figure 4 shows the super-thinned residuals corresponding to STEP and ETAS, and Figure 5 shows the centered L-functions of the super-thinned residuals, along with 95% confidence bounds generated from 1000 simulations of homogeneous Poisson processes. For both STEP and ETAS, the super-thinned residuals exhibit statistically significant clustering across most distances, indicating that a very substantial amount of the clustering in the observations was not adequately forecast by the models. Indeed, the original earthquakes that are not thinned are concentrated along the major fault lines where large amounts of space not filled in by super-imposed earthquakes. This suggests that these areas around the fault lines were over-estimated by both models which causes this clustering in the centered L-functions. However, for smaller distances of approximately 0-100km, the super-thinned residuals for the ETAS model exhibit less clustering than that of STEP, indicating superior performance of ETAS relative to STEP at this distance scale.

Finally, the observed earthquakes were plotted to examine if there were any spatial relationships to model performance. Here, the deviance defined as the difference in log-likelihoods between STEP and ETAS was calculated and plotted against various attributes of the corresponding earthquake such as longitude, latitude, location, and magnitude. In the first three plots of Figure 6, there appears to be a trend of ETAS performing slightly better in the Northwestern region of California (Longitude = -124, Latitude= 40.5) whereas STEP performs better in most other areas of the state. Comparing the deviance to the magnitude of the earthquakes, there does not appear to be a trend which is further supported where

the Pearson correlation of deviation and magnitude is -.05.

5 Discussion

The fact that STEP appears to outperform ETAS overall in prospective CSEP testing for the 5 year period of 1/12013 to 12/31/2017 is rather surprising. A possible advantage of STEP over other models is that STEP uses observed seismological heuristics to create rules in which to classify spatial aftershock zones. These zone classifications then influence time dependent components of the model including a generic region, sequence specific, and spatially varying element which are then added to a background seismicity element. This differs tremendously from ETAS, in that ETAS makes no difference in triggering seismicity among foreshocks, mainshocks and aftershocks, instead allowing each event to have the potential to trigger seismicity. ETAS forecasts more seismic activity around more recent events compared to STEP, potentially allowing STEP an advantage in forecasting over longer time frames such as five years. Indeed, Figure 7 shows that STEP is much more conservative in its predictions, focusing on 'hot spots' that are very concentrated on faults that have experienced earthquakes in their immediate vicinity. While ETAS captures these hot spots, a larger surrounding area than STEP is given higher intensities where with such few earthquakes, actually decreases model performance.

While the statistical properties of the ETAS approach make it an appealing model choice for short term predictions, the focus of STEP taking an opposite approach and using heuristics to differentiate foreshocks and aftershocks based on spatial zones may give it an advantage in forecasting areas of relatively high and low seismicity. 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. Similarly, these results indicate there may be room for the development of an ETAS-style model that seeks to incorporate elements of STEP and differentiate between different zones or forecast different aftershock activity for different types of earthquakes.

References

[1] Asim, K. M., Schorlemmer, D., Hainzl, S., Iturrieta, P., Savran, W. H., Bayona, J. A., and Werner, M. J. Multi-Resolution Grids in Earthquake Forecasting: The Quadtree Approach. *Bulletin of the Seismological Society of America 113*, 1 (12 2022), 333–347.

[2] Baddeley, A., Turner, R., MÃ¸ller, J., and Hazelton, M. Residual analysis for spatial point processes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*, 5 (2005), 617–666.

[3] Bayona, J. A., Savran, W. H., Rhoades, D. A., and Werner, M. J. Prospective evaluation of multiplicative hybrid earthquake forecasting models in California. *Geophysical Journal International 229*, 3 (01 2022), 1736–1753.

[4] Besag, J. Comments on 'Modelling spatial patterns' by B.D. Ripley. *Journal of the Royal Statistical Society B39*, 2 (1977), 193–195.

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

[6] Cattania, C., Werner, M. J., Marzocchi, W., Hainzl, S., Rhoades, D., Gerstenberger, M., Liukis, M., Savran, W., Christophersen, A., Helmstetter, A., Jimenez, A., Steacy, S., and Jordan, T. H. The Forecasting Skill of Physics-Based Seismicity Models during the 2010-2012 Canterbury, New Zealand, Earthquake Sequence. *Seismological Research Letters 89*, 4 (06 2018), 1238–1250.

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

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

[9] Fox, E., Schoenberg, F., and Gordon, J. Spatially inhomogeneous background rate estimators and uncertainty quantification for nonparametric Hawkes point process models of earthquake occurrences. *Annals of Applied Statistics 10*, 3 (2016), 1725–1756.

[10] Gerstenberger, M., Wiemer, S., Jones, L., and Reasenberg, P. Real-time forecasts of tomorrow's earthquakes in California. *Nature 435*, 7040 (2011), 328–331.

[11] Gordon, J., Clements, R., Schoenberg, F., and Schorlemmer, D. Voronoi residuals and other residual analyses applied to CSEP earthquake forecasts. *Spatial Statistics 14b* (2015), 133–150.

[12] Gordon, J., Fox, E., and Schoenberg, F. A nonparametric Hawkes model for forecasting California seismicity. *BSSA 111*, 4 (2021), 2216–2234.

[13] Iturrieta, P., Bayona, J. A., Werner, M. J., Schorlemmer, D., Taroni, M., Falcone, G., Cotton, F., Khawaja, A. M., Savran, W. H., and Marzocchi, W. Evaluation of a Decade-Long Prospective Earthquake Forecasting Experiment in Italy. *Seismological Research Letters* (04 2024).

[14] Jordan, T. H. Earthquake predictability, brick by brick. *Seismological Research Letters 77*, 4 (2006), 3–6.

[15] Kagan, Y. Testing long-term earthquake forecasts: Likelihood methods and error diagrams. *Geophysical Journal International 177*, 2 (2009), 532–542.

[16] Mancini, S., Segou, M., Werner, M. J., Parsons, T., Beroza, G., and Chiaraluce, L. On the use of high-resolution and deep-learning seismic catalogs for short-term earthquake forecasts: Potential benefits and current limitations. *Journal of Geophysical Research: Solid Earth 127*, 11 (2022), e2022JB025202.

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

[18] Ogata, Y., and Zhuang, J. Space-time ETAS models and an improved extension. *Tectonophysics 413*, 1-2 (2006), 13–23.

[19] Ripley, B. D. *Spatial statistics*. Wiley New York, 1981.

[20] Savran, W. H., Bayona, J. A., Iturrieta, P., Asim, K. M., Bao, H., Bayliss, K., Herrmann, M., Schorlemmer, D., Maechling, P. J., and Werner, M. J. pyCSEP: A Python Toolkit for Earthquake Forecast Developers. *Seismological Research Letters 93*, 5 (07 2022), 2858–2870.

[21] Savran, W. H., Werner, M. J., Marzocchi, W., Rhoades, D. A., Jackson, D. D., Milner, K., Field, E., and Michael, A. Pseudoprospective Evaluation of UCERF3-ETAS Forecasts during the 2019 Ridgecrest Sequence. *Bulletin of the Seismological Society of America 110*, 4 (07 2020), 1799–1817.

[22] Schneider, M., Clements, R., Rhoades, D., and Schorlemmer, D. Likelihood- and residual-based evaluation of medium-term earthquake forecast models for California. *Geophysical Journal International 198*, 3 (06 2014), 1307–1318.

[23] Schoenberg, F. Multi-dimensional residual analysis of point process models for earthquake occurrences. *J. Amer. Statist. Assoc. 98*, 464 (2003), 789–795.

[24] Schoenberg, F., Gordon, J., and Harrigan, R. Analytic computation of non-parametric Marsan-Lengliné estimates for Hawkes point processes. *Journal of Nonparametric Statistics 30*, 3 (2018), 742–757.

[25] Schorlemmer, D., Gerstenberger, M., Wiemer S., Jackson, D. D., and Rhoades, D. A. Earthquake likelihood model testing. *Seism. Res. Lett. 78*, 1 (2007), 17–29.

[26] Schorlemmer, D., and Gerstenberger, M. C. RELM testing Center. *Seism. Res. Lett. 78*, 1 (2007), 30–35.

[27] Schorlemmer, D., Mele, F., and Marzocchi, W. A completeness analysis of the National Seismic Network of Italy. *Journal of Geophysical Research: Solid Earth 115*, B4 (2010).

[28] Schorlemmer, D., Zechar, J. D., Werner, M., Jackson, D. D., Field, E. H., Jordan, T. H., and the RELM Working Group. First results of the Regional Earthquake Likelihood Models Experiment. *Pure and Applied Geophysics 167* (2007), 17–29.

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

[30] Stark, P. B. Earthquake prediction: the null hypothesis. *Geophysical Journal International 131*, 3 (1997), 495–499.

[31] Survey, U. G. Earthquake Hazards Program. *Advanced National Seismic System (ANSS) Comprehensive Catalog of Earthquake Events and Products: Various* (2017).

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

[33] Vere-Jones, D., and Schoenberg, F. Rescaling marked point processes. *Aust. N. Z. J. Stat. 46* (2004), 133–143.

[34] Werner, M., Gerstenberger, M., Liukis, M., Marzocchi, W., Rhoades, D., Taroni, M., Zechar, J., Cattania, C., Christophersen, A., Hainzl, S., Helmstetter, A., Jiménez, A., Steacy, S., and Jordan, T. Retrospective evaluation of time-dependent earthquake forecast models during the 2010-12 canterbury, new zealand, earthquake sequence (abstract). *Seismological Research Letters 86* (01 2015), 587–588.

[35] Werner, M., Zechar, J., Marzocchi, W., and Wiemer, S. Retrospective evaluation of the five-year and ten-year CSEP-Italy earthquake forecasts. *Annals of Geophysics 53*, 3 (2010), 11–30.

[36] Werner, M. J., Helmstetter, A., Jackson, D. D., and Kagan, Y. Y. High-Resolution Long-Term and Short-Term Earthquake Forecasts for California. *Bulletin of the Seismological Society of America 101*, 4 (08 2011), 1630–1648.

[37] Zechar, J.D. and Schorlemmer, D. and Werner, M.J. and Gerstenberger, M.C. and Rhoades, D.A. and Jordan, T.H. Regional earthquake likelihood mod-

els i: First order results. *Bulletin of the Seismological Society of America 103*, 2A (2013), 787–798.

Appendix

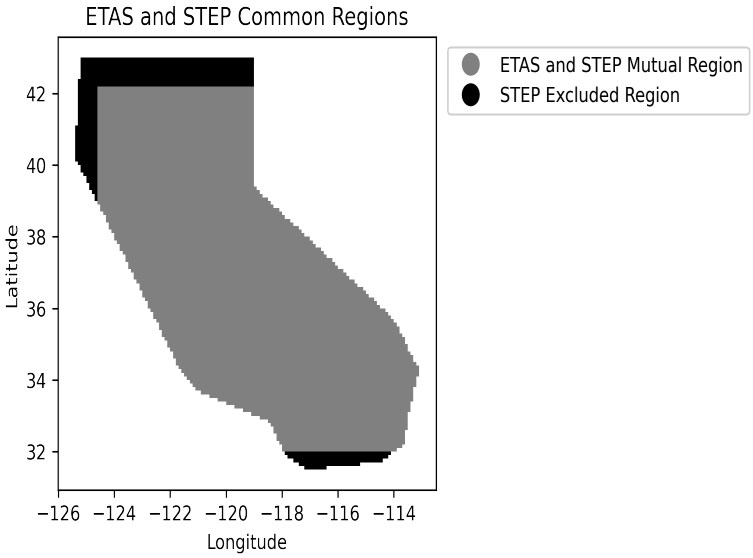

Figure 1: STEP and ETAS differ in regions for their predictions. For this analysis we only analyze the intersection of area where both models have made forecasts (the mutual region).

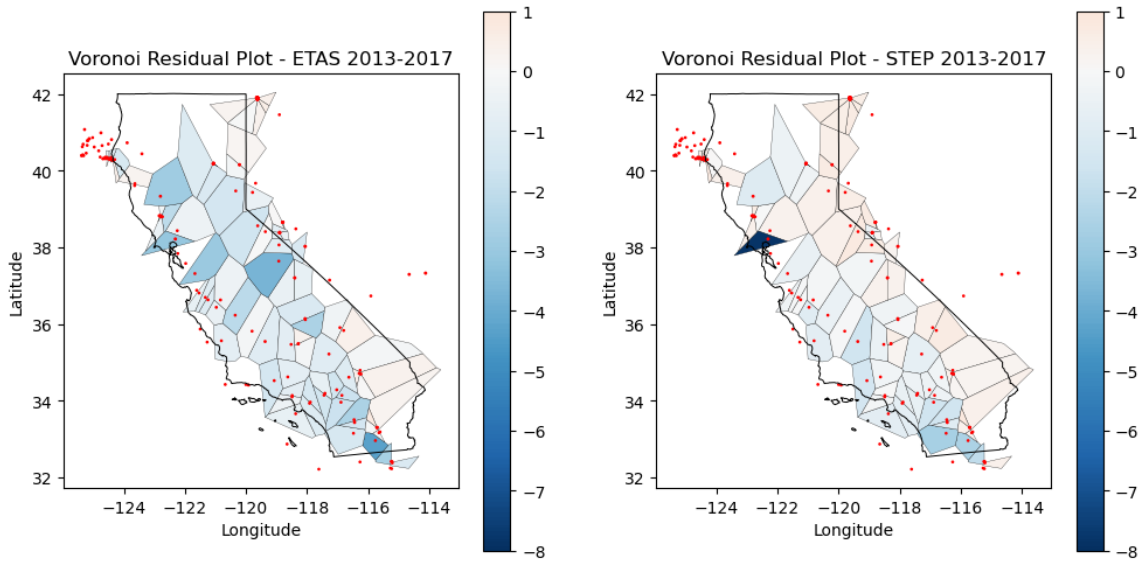

Figure 2: Voronoi residuals for STEP and ETAS for 1/1/2013 to 12/31/2017. Here, Voronoi residuals are the integrated rate given from a model over all area minus the observed number of events in the cell. Cells around the 'boundary' of the area of interest were removed to prevent skewed results from large areas being outside of model prediction zones.

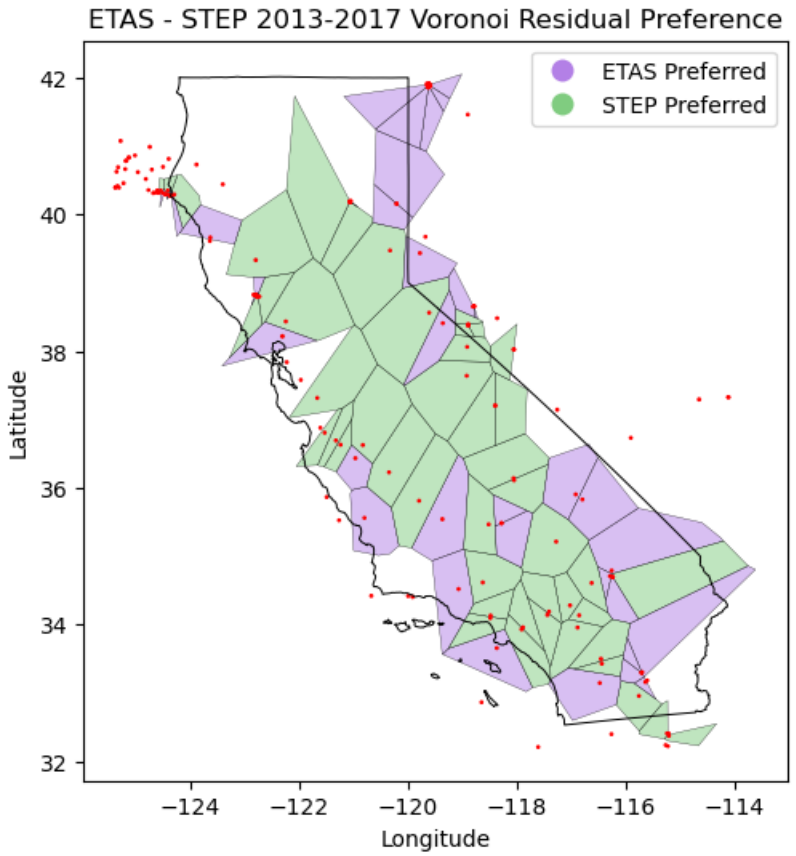

Figure 3: Comparison between ETAS and STEP for which model's prediction was closer to 0 for each cell.

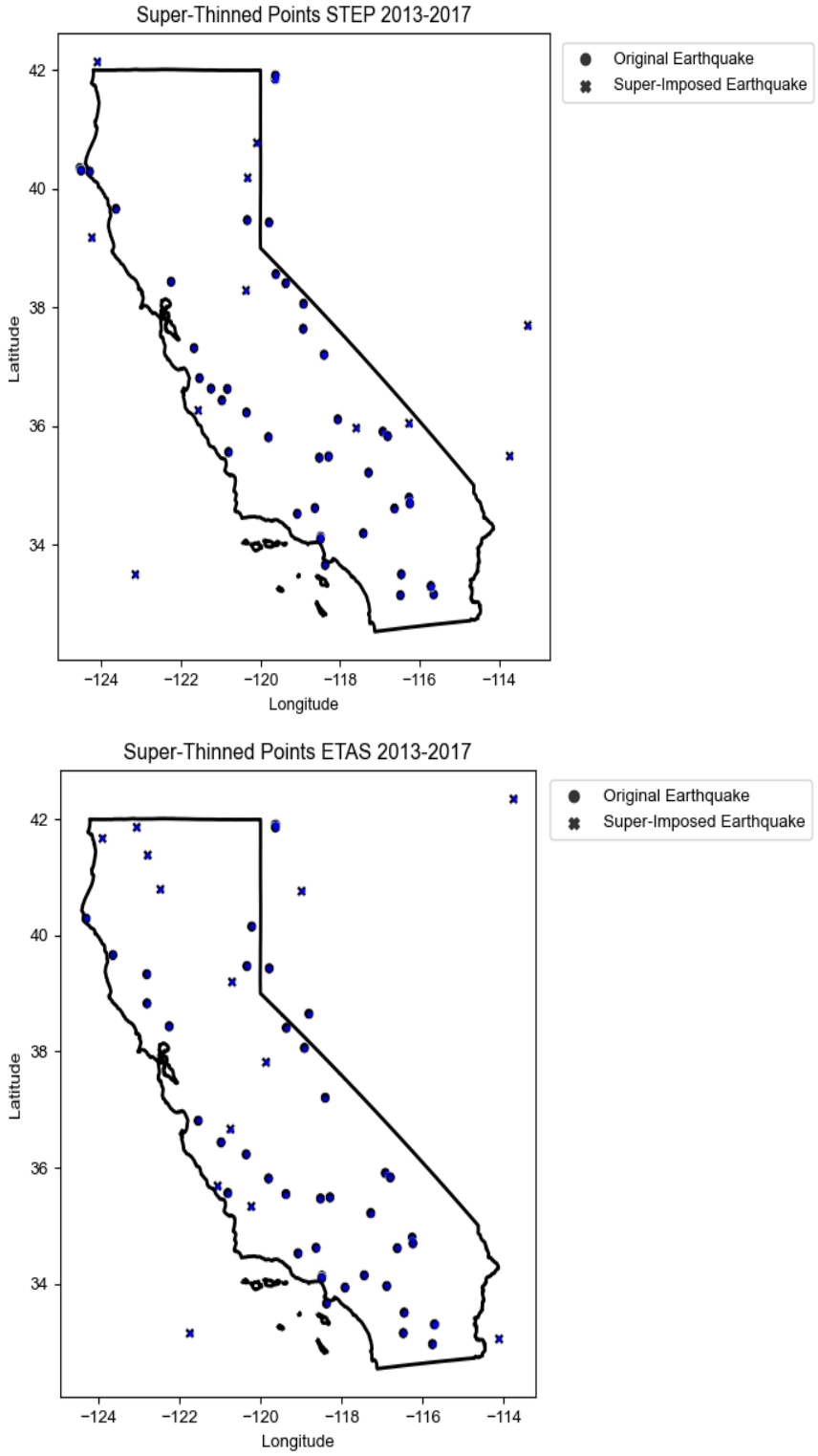

Figure 4: Super-thinned residuals for (a) STEP and (b) ETAS, for 1/1/2013 to 12/31/2017.

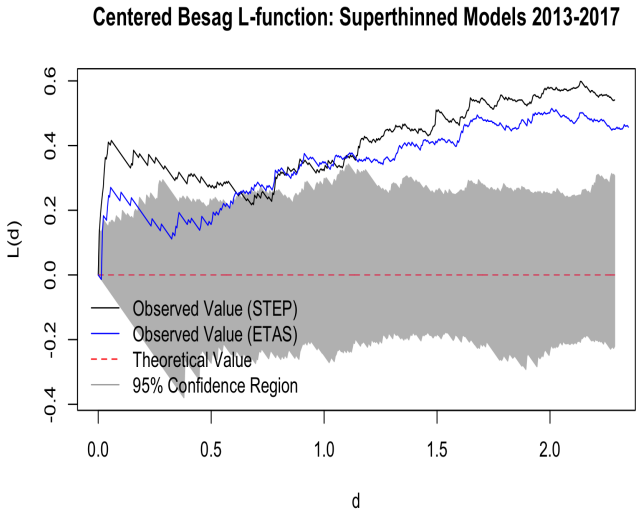

Figure 5: Centered L-functions with 95% confidence regions for STEP and ETAS.

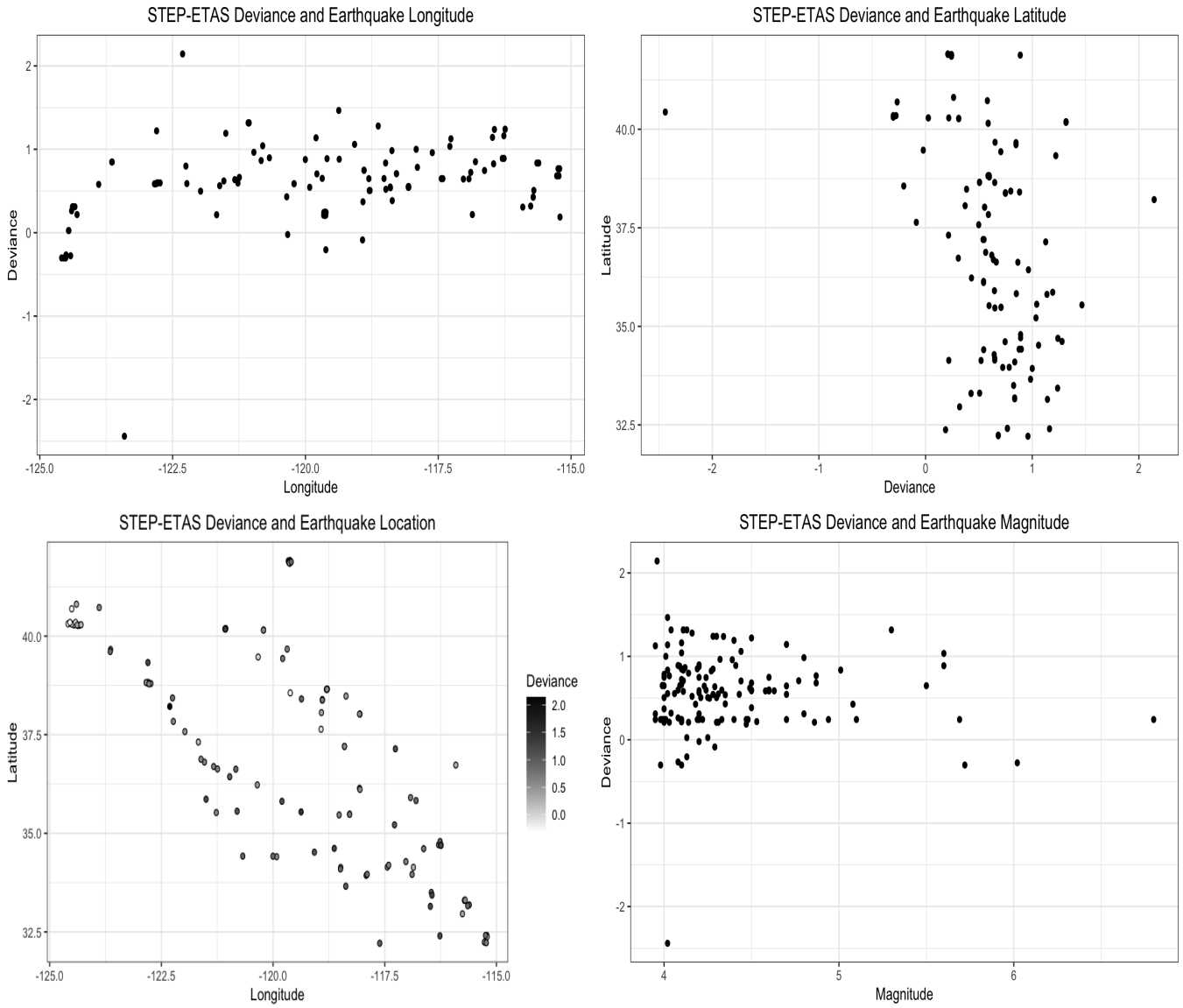

Figure 6: The STEP-ETAS Deviances (the log-likelihood for STEP subtracted from the log-likelihood for ETAS) for each significant earthquake in California from 2013-2017 plotted against longitude, latitude, and magnitude.

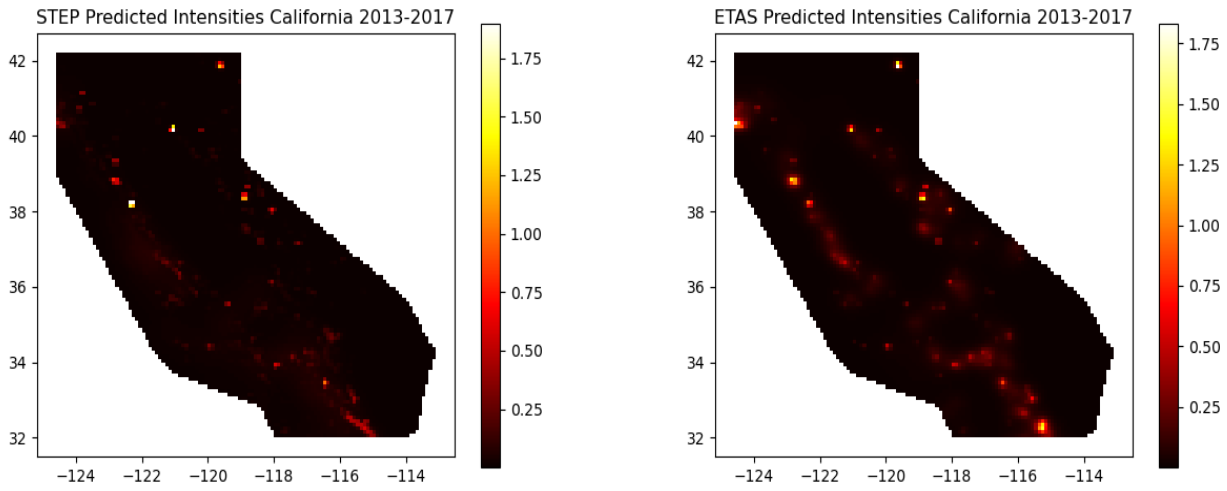

Figure 7: The aggregated predicted intensities for STEP and ETAS from 2013 to 2017.