The Annals of Applied Statistics
2011, Vol. 5, No. 4, 2549–2571
DOI: 10.1214/11-AOAS487

RESIDUAL ANALYSIS METHODS FOR SPACE–TIME POINT PROCESSES WITH APPLICATIONS TO EARTHQUAKE FORECAST MODELS IN CALIFORNIA[1]

BY ROBERT ALAN CLEMENTS, FREDERIC PAIK SCHOENBERG AND DANIJEL SCHORLEMMER

University of California, Los Angeles, University of California, Los Angeles, and University of Southern California and GFZ Potsdam

Modern, powerful techniques for the residual analysis of spatial-temporal point process models are reviewed and compared. These methods are applied to California earthquake forecast models used in the Collaboratory for the Study of Earthquake Predictability (CSEP). Assessments of these earthquake forecasting models have previously been performed using simple, low-power means such as the L-test and N-test. We instead propose residual methods based on rescaling, thinning, superposition, weighted K-functions and deviance residuals. Rescaled residuals can be useful for assessing the overall fit of a model, but as with thinning and superposition, rescaling is generally impractical when the conditional intensity λ is volatile. While residual thinning and superposition may be useful for identifying spatial locations where a model fits poorly, these methods have limited power when the modeled conditional intensity assumes extremely low or high values somewhere in the observation region, and this is commonly the case for earthquake forecasting models. A recently proposed hybrid method of thinning and superposition, called super-thinning, is a more powerful alternative. The weighted K-function is powerful for evaluating the degree of clustering or inhibition in a model. Competing models are also compared using pixel-based approaches, such as Pearson residuals and deviance residuals. The different residual analysis techniques are demonstrated using the CSEP models and are used to highlight certain deficiencies in the models, such as the overprediction of seismicity in inter-fault zones for the model proposed by Helmstetter, Kagan and Jackson [*Seismological Research Letters* **78** (2007) 78–86], the underprediction of the model proposed by Kagan, Jackson and Rong [*Seismological Research Letters* **78** (2007) 94–98] in forecasting seismicity around the Imperial, Laguna Salada, and Panamint clusters, and the underprediction of the model proposed by Shen, Jackson and Kagan [*Seismological Research Letters* **78** (2007) 116–120] in forecasting seismicity around the Laguna Salada, Baja, and Panamint clusters.

Received September 2010; revised May 2011.

[1]Supported by the Southern California Earthquake Center. SCEC is funded by NSF Cooperative Agreement EAR-0106924 and USGS Cooperative Agreement 02HQAG0008. The SCEC contribution number for this paper is 1495.

Key words and phrases. Thinned residuals, rescaled residuals, superposition, super-thinned residuals, Pearson residuals, deviance residuals.

1. Introduction. Recent statistical developments in the assessment of space–time point process models have resulted in new, powerful model evaluation tools. These tools include residual point process methods such as thinning, superposition and rescaling, comparative quadrat methods such as Pearson residuals and deviance residuals, and weighted second-order statistics for assessing particular features of a model such as its background rate or the degree of spatial clustering.

Unfortunately, these methods have not yet become widely used in seismology. Indeed, recent efforts to assess and compare different space–time models for earthquake occurrences have led to developments such as the Regional Earthquake Likelihood Models (RELM) project [Field (2007)] and its successor, the Collaboratory for the Study of Earthquake Predictability (CSEP) [Jordan (2006)]. The RELM project was initiated to create a variety of earthquake forecast models for seismic hazard assessment in California. Unlike previous projects that were addressing earthquake forecast modeling for seismic hazard assessment, the RELM participants decided to develop a multitude of competing forecasting models and to rigorously and *prospectively* test their performance in a dedicated testing center [Schorlemmer and Gerstenberger (2007)]. With the end of the RELM project, the forecast models became available and the development of the testing center was done within the scope of CSEP. CSEP inherited not only all models developed for RELM and is testing them for the previously defined period of 5 years, but also a suite of forecast performance tests that was developed during the RELM project. In RELM, a community consensus was reached that all models will be tested with these tests [Jackson and Kagan (1999), Schorlemmer et al. (2007)]. The tests include the Number or N-Test that compares the total forecasted rate with the observation, the Likelihood or L-Test that assesses the quality of a forecast in the likelihood space, and the Likelihood-Ratio or R-Test that compares the performance of two forecast models. However, over time several drawbacks of these tests were discovered [Schorlemmer et al. (2010)] and the need for more and powerful tests became clear to better discern between closely competing models. The N-test and L-test simply compare the quantiles of the total numbers of events in each bin or likelihood within each bin to those expected under the given model, and the resulting low-power tests are typically unable to discern significant lack of fit unless the overall rate of the model fits extremely poorly. Further, even when the tests do reject a model, they do not typically indicate *where* or *when* the model fits poorly, or how it could be improved.

The purpose of the current paper is to review modern model evaluation techniques for space–time point processes and to demonstrate their use and practicality on earthquake forecasting models for California. The RELM project represents an ideal test case for this purpose, as a variety of relevant, competing space–time models are included, and these models yield genuinely prospective forecasts of earthquake rates based solely on prior data. The rates are specified per bins which are spatial-magnitude-temporal volumes (called pixels in the statistical domain). These bins have been predefined in a community consensus process in order to

have the model forecast rates in the exact same bins. The models' forecasts translate into strongly different estimates of seismic hazard. Its accurate estimation is important for seismic hazard assessment, urban planning, disaster preparation efforts and in the pricing of earthquake insurance premiums [Bolt (2006)], so distinguishing among competing models is an extremely important task.

In Section 2 we describe a group of earthquake forecast models to be evaluated, along with the observed earthquake occurrences used to assess the fit of the models. The methods currently used by seismologists for model evaluation are briefly reviewed in Section 3. Pixel-based residuals for model comparison are discussed in Section 4. In Section 5 weighted second-order statistics, primarily the weighted K-function, are investigated. Section 6 reviews various residual methods based on rescaling, thinning and superposition, and introduces and applies the method of super-thinning. Section 7 summarizes some of the benefits and weaknesses of these tools.

2. CSEP earthquake forecast models and earthquake occurrence catalogs. CSEP expanded and now collects and evaluates space–time earthquake forecasts for different regions around the world, including California, Japan, New Zealand, Italy, the Northwest Pacific, the Southwest Pacific and the entire globe. The forecasts are evaluated in testing centers in Japan, Switzerland, New Zealand and the United States. The U.S. testing center is located at the Southern California Earthquake Center (SCEC) and hosts forecast experiments for California, the Northwest and Southwest Pacific, and the global experiments. We have chosen to apply a variety of measures to assess the fit of a collection of the California forecast models currently being tested at SCEC.

The forecast models are arranged in classes according to their forecast time period: five-year, three-month and one-day. There are two types of forecasts, rate-based and alarm-based. Within the five-year group are a set of rate-based models developed as part of the RELM project. In this paper we evaluate the RELM project rate-based one-day and five-year models, and will be ignoring the three-month models due to their very recent introduction to the CSEP testing center.

All CSEP forecasts are grid-based, providing a forecast in each spatial-magnitude bin within a given time window. For the one-day models, each bin is of size 0.1° longitude (lon) by 0.1° latitude (lat) by 0.1 units magnitude for earthquake magnitudes ranging from 3.95 to 8.95. For magnitudes 8.95–10, there is a single bin of size 0.1° by 0.1° by 1.05 units of magnitude. The RELM forecasts are identical, except with a lower magnitude bound of 4.95 instead of 3.95. For each bin, an expected number of earthquakes in the forecast period is forecasted.

There are five models in the RELM project that are considered mainshock+aftershock models. These models forecast both mainshocks and aftershocks with a single forecast for a period of five years. Models proposed in Helmstetter, Kagan and Jackson (2007) and Kagan, Jackson and Rong (2007), which we will call models A and B, respectively, base their forecasts exclusively on previous seismicity.

The model proposed in Shen, Jackson and Kagan (2007), denoted model C here, is based on other geodetic or geological data. All RELM models are five-year forecasts, beginning 1 January 2006, 00:00 UTC and ending 1 January 2011, 00:00 UTC. CSEP is also testing two one-day forecast models: The Epidemic-Type Aftershock Sequences (ETAS) model [Zhuang, Ogata and Vere-Jones (2004), Ogata and Zhuang (2006)] and the Short-Term Earthquake Probabilities (STEP) model [Gerstenberger et al. (2005)] since September of 2007. Both of these models produce forecasts based exclusively on prior seismicity.

CSEP evaluates the RELM models using a lower magnitude cutoff of 4.95. Because there are so few earthquakes of magnitude 4.95 and higher in the catalog over the observed period we use a lower magnitude cutoff of 3.95 instead. The forecasts for models A, B and C were extrapolated using each model's fitted magnitude distribution. Models A and B assume the magnitude distribution follows a tapered Gutenberg–Richter law [Gutenberg and Richter (1944)] with a b-value of 0.95 and a corner magnitude of 8.0. Model C uses a b-value of 0.975 and the same corner magnitude. Model A adjusts the magnitude distribution in a small region in northern California influenced by geothermal activity ($122.9°\text{W} < \text{lon} < 122.7°\text{W}$ and $38.7°\text{N} < \text{lat} < 38.9°\text{N}$) by using a b-value of 1.94 instead of 0.95.

Earthquake catalogs containing the estimated earthquake hypocenter locations and magnitudes were obtained from the Advanced National Seismic System (ANSS). From 1 January 2006 to 1 September 2009 there were 142 shallow earthquakes with a magnitude of 3.95 or larger which occurred in RELM's spatial-temporal window (see Figure 1). Note that each RELM model does not necessarily produce a forecasted seismicity rate for every pixel in the space–time region.

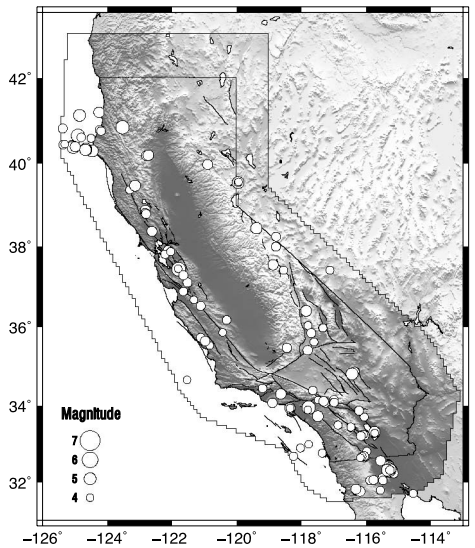

FIG. 1. *Locations of earthquakes with magnitude $M \geq 3.95$ in the RELM testing region.*

Hence, each model essentially has its own relevant spatial-temporal observation region, and thus we may have different numbers of observed earthquakes corresponding to different models. For instance, all 142 recorded earthquakes from 1 January 2006 to 1 September 2009 corresponded to pixels where model A made forecasts, but only 81 corresponded to pixels where model B made forecasts, and 86 where model C made forecasts. 85 earthquakes of magnitude 3.95 or greater occurred since 1 September of 2007, all of which corresponded to forecasts made by ETAS but only 83 of which corresponded to forecasts made by STEP.

3. L-test and N-test. CSEP initially implemented two numerical summary tests, called the Likelihood-test (L-test) and the Number-test (N-test), to evaluate the fit of the earthquake forecast models they collect. A full description of these methods can be found in Schorlemmer et al. (2007). These goodness-of-fit tests are similar to other numerical goodness-of-fit summaries such as the Akaike Information Criterion [Akaike (1974)] and the Bayesian Information Criterion [Schwarz (1978)] in that they provide a score for the overall fit of the model without indicating where the model may be fitting poorly.

The L-test, described in Schorlemmer et al. (2007), works by first simulating some fixed number s of realizations from the forecast model. The log-likelihood (ℓ) is computed for the observed earthquake catalog (ℓ_{obs}) and each simulation (ℓ_j, for $j = 1, 2, \ldots, s$). The quantile score, γ, is defined as the fraction of simulated likelihoods that are less than the observed catalog likelihood:

$$\gamma = \frac{\sum_{j=1}^{s} \mathbf{1}_{\{\ell_j < \ell_{\text{obs}}\}}}{s},$$

where $\mathbf{1}$ denotes the indicator function. If γ is close to zero, then the model is considered to be inconsistent with the data, and can be rejected. Otherwise, the model is not rejected and further tests are necessary.

The N-test is similar to the L-test, except that the quantile score examined is instead the fraction of simulations that contain fewer points than the actual observed number of points in the catalog, N_{obs}. That is,

$$\delta = \frac{\sum_{j=1}^{s} \mathbf{1}_{\{N_j < N_{\text{obs}}\}}}{s},$$

where N_j is the number of points in the jth simulation of the model. With the N-test, the model is rejected if δ is close to 0 or 1. If a model is underpredicting or overpredicting the total number of earthquakes, then $\delta \sim 1$ or 0, respectively, and the model will likely be rejected with the N-test.

Table 1 shows results for the L- and N-test for selected models. The L-test would lead to rejection of models A, B, C and STEP as seen by the very low γ scores. The ETAS model would not be rejected based on the γ score alone, requiring the application of the N-test for a final decision. At the 5% level of significance, the δ scores indicate that the STEP model is underpredicting the total number

TABLE 1

Results of the L and N-test. Listed are the observed log-likelihoods, ℓ_{obs}, the L-test γ scores, the observed number of events, N_{obs} and the N-test δ scores. δ scores that are bold-faced are significant at the 5% level leading to rejection of the forecast

Model	ℓ_{obs}	γ	N_{obs}	δ
		Mainshock+Aftershock		
A. Helmstetter	−22881.46	0.000	142	**0.000**
B. Kagan	−10765.43	0.008	81	**0.001**
C. Shen	−10265.20	0.002	86	**0.043**
		Daily		
ETAS	−387.69	1.00	85	**0.00**
STEP	−50.43	0.00	83	**0.99**

of earthquakes, while models A, B, C and ETAS are significantly overpredicting earthquake rates.

Unfortunately, in practice, both statistics γ and δ test essentially the same thing, namely, the agreement between the observed and modeled *total* number of points. Indeed, for a typical model, the likelihood for a given simulated earthquake catalog depends critically on the number of points in the simulation.

4. Pixel-based methods. Baddeley et al. (2005) introduced methods for residual analysis of purely spatial point processes, based on comparing the total number of points within predetermined bins to the number forecast by the model. Such methods extend readily to the spatial-temporal case, and are quite natural for evaluating the CSEP forecasts since the models are constrained to have a constant conditional intensity within prespecified bins. The differences between observed and expected numbers of events within bins can be standardized in various ways, as described in what follows.

4.1. *Preliminaries*. Earthquake occurrence times and locations are typically modeled as space–time point processes, with the estimated epicenter or hypocenter of each earthquake representing its spatial location. Along with each observation, one may also record several *marks* which may be used in the model to help forecast future events; an important example of a mark is the magnitude of the event. Space–time point process models are often characterized by their associated conditional intensity, $\lambda(t, \mathbf{x})$, that is, the infinitesimal rate at which one expects points to occur around time t and location $\mathbf{x}$, given full information on the occurrences of points prior to time t, and given the marks and possibly other covariate information observed before time t. Note that due to the lack of a natural ordering of points in the plane, purely spatial point processes are typically characterized by their Papangelou intensities [Papangelou (1972)], which may be thought of as the

limiting rate at which points are expected to accumulate within balls centered at location **x** given what *other* points have occurred at all locations outside of these balls, as the size of the balls shrink to zero. For a review of point processes and conditional intensities, see Daley and Vere-Jones (2003).

An aggregate conditional intensity is derived for each spatial bin for all models by summing the forecast rates over all magnitude bins and then dividing the sum by the area of each pixel. Since we are evaluating the five-year models A, B and C after only 44 of the 60 months of the forecast period have elapsed, their conditional intensities are scaled by a factor of $44/60$.

4.2. *Raw and Pearson residuals.* Consider a model $\hat{\lambda}(t, x, y)$ for the conditional intensity at any time t and location (x, y). *Raw residuals* may be defined following Baddeley et al. (2005) as simply the number of observed points minus the number of expected points in each pixel, that is,

$$R(B_i) = N(B_i) - \int_{B_i} \hat{\lambda}(t, x, y)\, dt\, dx\, dy, \tag{1}$$

where $N(B_i)$ is the number of points in bin i. Note that Baddeley et al. (2005) consider only the case of purely spatial point processes characterized by their Papangelou intensities; Zhuang (2006) showed that one may nevertheless extend the definition to the spatial-temporal case using the conventional conditional intensity as in (1).

One may wish to rescale the raw residuals in such a way that they have mean 0 and variance approximately equal to 1. The *Pearson residuals* are defined as

$$R_{\mathrm{P}}(B_i) = \sum_{(t_j, x_j, y_j) \in B_i} \frac{1}{\sqrt{\hat{\lambda}(t_j, x_j, y_j)}} - \int_{B_i} \sqrt{\hat{\lambda}(t, x, y)}\, dt\, dx\, dy$$

for all $\hat{\lambda}(t_i, x_i, y_i) > 0$. These are analogous to the Pearson residuals in Poisson log-linear regression.

Both STEP and model C have several pixels with forecasted conditional intensities of 0, which complicates the standardization of the corresponding residuals for these two models. Pearson residuals were obtained for each of the remaining models. For instance, Figure 2 shows that the largest Pearson residual for model B is 2.817 located in a pixel in Mexico, just south of the California border near the Imperial Valley fault zone (lon $\approx 115.3°$W and lat $\approx 32.4°$N), which is the location of a large cluster of earthquakes. Another very large residual for model B can be seen just above the San Bernardino and Inyo county border near the Panamint Valley fault zone (lon $\approx 117.0°$W and lat $\approx 36.0°$N). This is also the location of the largest ETAS Pearson residual (2.221). The largest Pearson residual for model A (4.068) is located at a small earthquake cluster near the Peterson Mountain fault northwest of Reno, Nevada (lon $\approx 199.9°$W and lat $\approx 39.5°$N).

Note that when spatial-temporal bins are very small and/or the estimated conditional intensity in some bins is very low, as in this example, the raw and especially

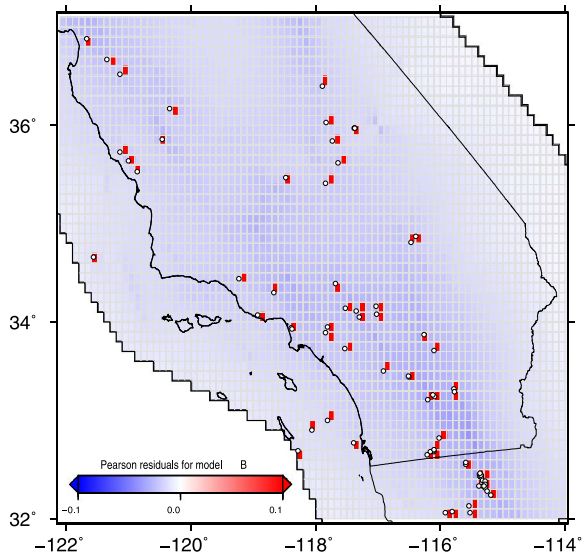

FIG. 2. *Pearson residuals for model B. The maximum observed Pearson residual is* 2.817.

the standardized residuals are highly skewed. In such cases, the residuals in such pixels where points happen to occur tend to dominate, and the skew may complicate the analysis. Indeed, Pearson residuals fail to provide much useful information about the model's fit in the other pixels where earthquakes did not happen to occur, and graphical displays of the Pearson residuals tend to highlight little more than the locations of the earthquakes themselves. Therefore, while Pearson and raw residuals may help to identify individual bins containing earthquakes that require an adjustment in their forecasted rates, Pearson and raw residuals generally fail to identify other locations where the models may fit relatively well or poorly.

4.3. *Deviance residuals.* A useful method for comparing models is using the deviance residuals proposed by Wong and Schoenberg (2009), in analogy with deviances defined for generalized linear models in the regression framework. As with Pearson residuals, S is divided into evenly spaced bins, and the differences between the log-likelihoods within each bin for the two competing models are examined. Given two models for the conditional intensity, $\hat{\lambda}_1$ and $\hat{\lambda}_2$, the deviance residual in each bin, B_i, of $\hat{\lambda}_1$ against $\hat{\lambda}_2$ is given by

$$R_{\mathrm{D}}(B_i) = \sum_{i\,:\,(t_i,x_i,y_i)\in B_i} \log\left(\hat{\lambda}_1(t_i,x_i,y_i)\right) - \int_{B_i} \hat{\lambda}_1(t,x,y)\,dt\,dx\,dy$$
$$- \left(\sum_{i\,:\,(t_i,x_i,y_i)\in B_i} \log\left(\hat{\lambda}_2(t_i,x_i,y_i)\right) - \int_{B_i} \hat{\lambda}_2(t,x,y)\,dt\,dx\,dy \right).$$

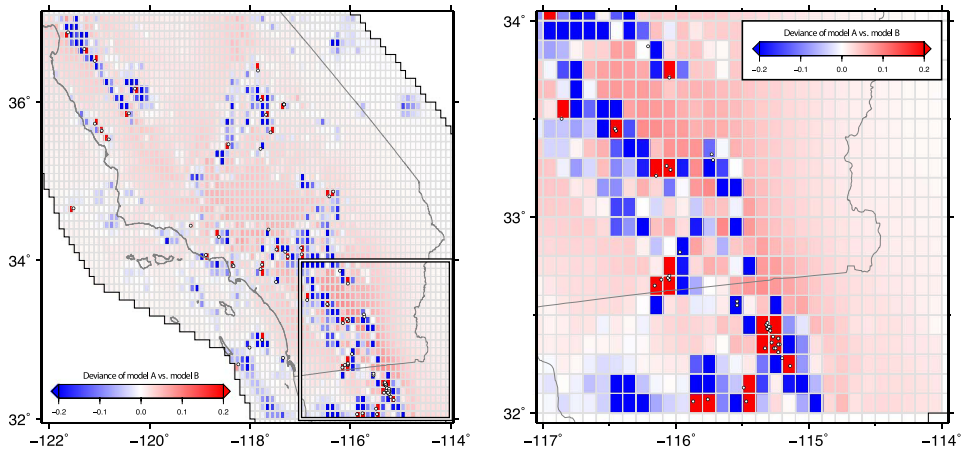

FIG. 3. *Left panel* (a): *deviance residuals for model A versus B. Sum of deviance residuals is* 84.393. *Right panel* (b): *close-up of deviance residuals for model A versus B near the Imperial fault.*

Positive residuals imply that the model $\hat{\lambda}_1$ fits better in the given pixel and negative residuals imply that $\hat{\lambda}_2$ provides better fit. By simply taking the sum of the deviance residuals, $\sum_i R_D(B_i)$, we obtain a log-likelihood ratio score, giving us an overall impression of the improvement in fit from the better fitting model. If $\hat{\lambda}_1$ or $\hat{\lambda}_2$ is estimated, then one may use this estimate in computing the deviance residuals, and similarly if $\hat{\lambda}_1$ or $\hat{\lambda}_2$ is given, that is, not estimated, then one would simply use this given model in computing the residuals.

Figure 3(a) shows the deviance residuals for model A versus model B. Model A outperforms model B in almost all locations where earthquakes actually occurred, and, in particular, model A forecasts the Imperial earthquake cluster and another cluster near the Laguna Salada and Yuha Wells faults just north of the California–Mexico border (lon $\approx$ 116.0°W and lat $\approx$ 32.7°N) much better than model B. The pixel with the largest residual, highlighted in Figure 3(b), is located in the Imperial cluster. Model B seems to fit better in several selected areas, mostly regions close to known faults but where earthquakes did not happen to occur in the time span considered. In most locations, however, including the vast majority of locations far from seismicity, model A offers better fit, as model B tends to overpredict events in these locations more than model A. Overall, the log-likelihood ratio score is 84.393, indicating a significant improvement from model A compared to model B.

Results are largely similar for model A versus model C, as seen in Figure 4(a), with model A forecasting the rate at all observed earthquake clusters, including a cluster at the extreme southern end of the observation region on the Baja, Mexico peninsula (lon $\approx$ 116.3W and lat $\approx$ 31.8N), more accurately than model C. Overall, model A offers substantial improvement over model C with a likelihood ratio score of 86.427. Residuals for model B versus model C can be seen in Figure 4(b). Model C forecasts the rate near the Imperial cluster better, and model B forecasts

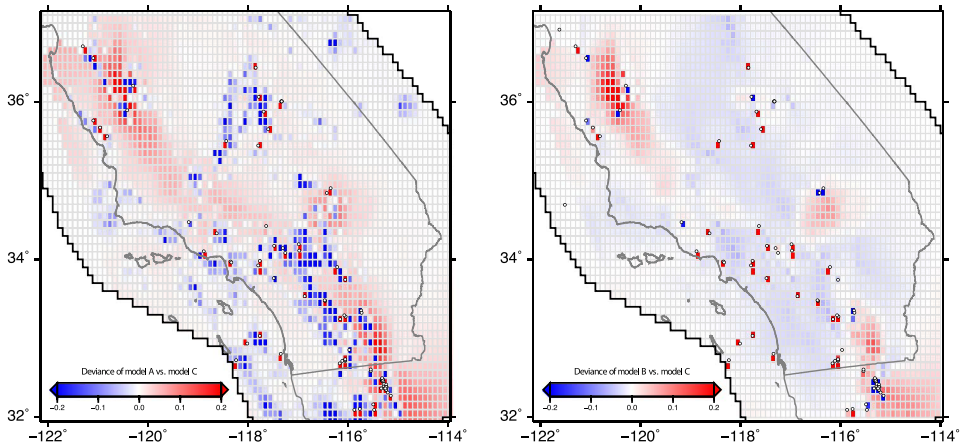

FIG. 4. *Left panel* (a): *deviance residuals for model A versus C. Sum of deviance residuals is* 86.427. *Right panel* (b): *deviance residuals for model B versus C. Sum of deviance residuals is* −7.468.

more accurately around the Laguna Salada cluster. There are vast regions where model B outperforms model C and vice versa. Overall, model C fits slightly better than model B, with a likelihood ratio score of −7.468. Deviance residuals for ETAS versus STEP (not shown) reveal that the ETAS model performs somewhat better for this data set overall, with a log-likelihood ratio score of 76.261, providing substantially more accurate forecasts in nearly all locations, especially where earthquakes occur.

5. Weighted second-order statistics. A common model assessment tool used for detecting clustering or inhibition in a point process is Ripley's K-function [Ripley (1981)], defined as the average number of points within r of any given point divided by the overall rate λ, and is typically estimated via

$$\hat{\mathrm{K}}(r) = AN^{-2} \sum_{i<j, \|x_i - x_j\| < r} s(\mathbf{x}_i, \mathbf{x}_j),$$

where A is the area of the observation region, N is the total number of observed points, and $s(\mathbf{x}_i, \mathbf{x}_j)^{-1}$ is the proportion of area of the ball centered at $\mathbf{x}_i$ and passing through $\mathbf{x}_j$ that falls within the observation region [see Ripley (1981), Cressie (1993)]. For a homogeneous Poisson process in $\mathbf{R}^2$, $\mathrm{K}(r) = \pi r^2$, Besag (1977) suggested a variance stabilized version of the K-function, called the L-function, given by $\mathrm{L}(r) = \sqrt{\mathrm{K}(r)/\pi}$.

The null hypothesis for most second-order tests such as Ripley's K-function is that the point process is a homogeneous Poisson process. Stark (1997) argues that this is a poor null hypothesis for the case of earthquake occurrences because a homogeneous Poisson model fits so poorly to actual data. Adelfio and Schoenberg (2009) described a variety of weighted analogues of second-order tests that

are useful when the null hypothesis in question is more general. Most useful among these is the weighted analogue of Ripley's K-function, first introduced by Baddeley, Møller and Waagepetersen (2000). They discussed the case where the null model $\hat{\lambda}_0$, can be any inhomogeneous Poisson process, and this was extended by Veen and Schoenberg (2005) to the case of non-Poisson processes as well. The weighted K-function is useful for testing the degree of clustering in the model, and was used by Veen and Schoenberg (2005) to assess a spatial point process model fitted to Southern California earthquake data. The standard estimate of the weighted K-function is given by

$$\mathrm{K_W}(r) = \frac{b}{\int_S \hat{\lambda}_0(\mathbf{x})\,d\mathbf{x}} \sum_i \hat{\lambda}_0(\mathbf{x}_i)^{-1} \sum_{j \neq i} \hat{\lambda}_0(\mathbf{x}_j)^{-1} \mathbf{1}_{\{|\mathbf{x}_j - \mathbf{x}_i| \leq r\}},$$

where $b = \min(\hat{\lambda})$, $\mathbf{1}$ is the indicator function, and $\hat{\lambda}_0(\mathbf{x}_i)$ is the conditional intensity at point $\mathbf{x}_i$ under the null hypothesis. Edge-corrected modifications can also be used, especially when the observed space is irregular. Guan (2009) proposed a local empirical K-function which can assess lack-of-fit in subsets of S and can be compared to the weighted K-function applied globally to S. Here, we apply the weighted K-function globally to derive an overall impression of each model's lack of fit.

As with Ripley's K-function, under the null hypothesis, for a spatial point process with intensity λ_0, $\mathrm{K_W}(r) = \pi r^2$ [Veen and Schoenberg (2005)]. To obtain a centered and standardized version, one can also transform the weighted K-function into a weighted L-function as before, and plot $\mathrm{L_W}(r) - r = \sqrt{\mathrm{K_W}(r)/\pi} - r$ versus r.

Space–time versions of the L-function have been proposed, but for the purpose of examining, in particular, the range and degree of purely spatial clustering in each model, it seems preferable to apply the purely spatial weighted L-function previously described, after first integrating the conditional intensities of the ETAS and STEP models over time. Figure 5 shows the estimated centered weighted L-functions for the five models considered here, along with 95% confidence bounds based on the normal approximation in Veen and Schoenberg (2005), who showed that asymptotically, the distribution of the weighted K-function should generally obey

$$\mathrm{K_W}(r) \sim N\left(\pi r^2, \frac{2\pi r^2 A}{[\int_S \hat{\lambda}_0(\mathbf{x})\,d\mathbf{x}]^2}\right). \tag{2}$$

The catalog of observed earthquakes is significantly more clustered than would be expected according to model A, especially within distances of 0.2 degrees of longitude/latitude, or approximately 22.2 km. However, at distances greater than 0.3°, or approximately 33.3 km, the observed data exhibit greater inhibition than one would expect according to model A. This suggests that model A is underpredicting

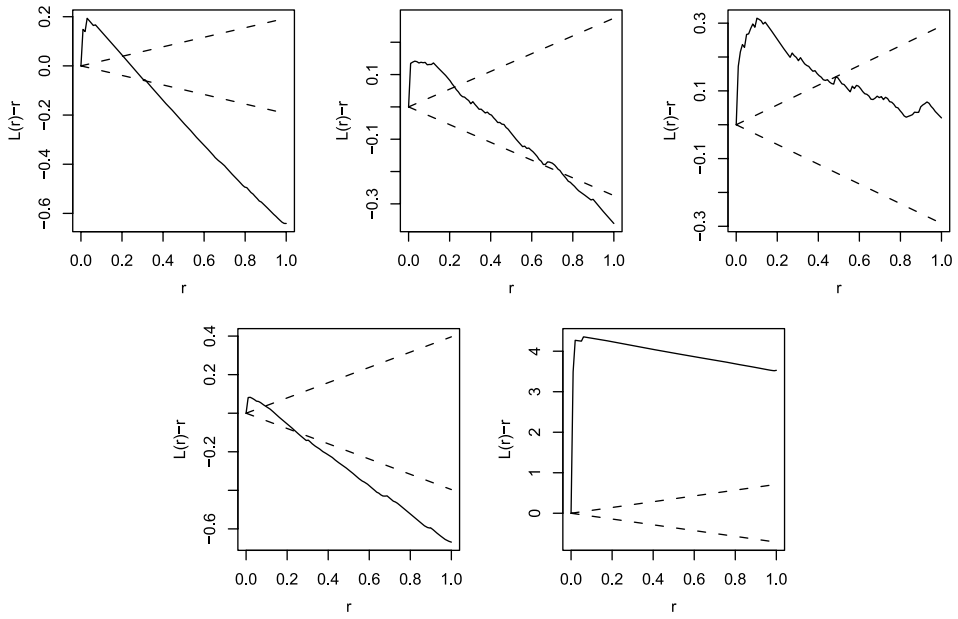

FIG. 5. *Estimated centered weighted L-function* (*solid curve*) *and* 95% *confidence bands* (*dashed curves*). *Top-left panel*: (a) *model A*. *Top-center panel*: (b) *model B*. *Top-right panel*: (c) *model C*. *Bottom-left panel*: (d) *ETAS*. *Bottom-right panel*: (e) *STEP*.

the degree of clustering in the observed seismicity and may be generally underpredicting the seismicity rate within highly active seismic areas, and may be overpredicting seismicity elsewhere. Results are similar for model B and the ETAS model. The estimated L-function for model C shows significantly more clustering of the (weighted) seismicity than one would expect within distances of 0.4° or 44.4 km, that is, model C is significantly underpredicting the degree of clustering within this range, but seems consistent with the data outside of this range. The estimated L-function shows clear discrepancies between the STEP model and the data, as the (weighted) seismicity is significantly more clustered than one would expect according to the model at both small and large distances. These results are not surprising considering that STEP tends to underpredict seismicity overall: according to the STEP forecasts, one would expect only 63 earthquakes in total during the period in which 85 occurred. By contrast, ETAS tends to overpredict the overall rate, forecasting more than 114 earthquakes in this same period.

6. Residual point process methods. As shown in Section 4.2, when the spatial-temporal pixels are small, the distribution of raw and Pearson residuals tend to be highly skewed, and this limits their utility. When pixels are larger, however, a drawback of pixel-based residuals is that considerable information is lost in aggregating over the pixels. Instead, one may wish to examine the extent to which the

data and model agree, without relying on such aggregation. One way to perform such an assessment is to transform the points of the process, by rescaling, thinning, superposition or superthinning, to form a new point process that should be a homogeneous Poisson process if and only if the model used to govern this transformation is correct. The residual points can then be assessed for inhomogeneity as a means of evaluating the goodness of fit of the underlying model.

6.1. *Rescaled residuals*. Meyer (1971) observed that the temporal coordinates of a multivariate point process can be rescaled according to the integrated conditional intensity in order to form a sequence of stationary Poisson processes. For a space–time point process, one may thus rescale one axis, for example, the x-axis, moving each observation (t_i, x_i, y_i) to the new rescaled position $(t_i, \int_0^{x_i} \hat{\lambda}(t, x, y)\, dx, y_i)$, and assess the space–time homogeneity of the resulting process. This sort of method was used by Ogata (1988) for model evaluation for the purely temporal case and by Schoenberg (2003) for the spatial-temporal case. The spatial homogeneity of these residual points may be assessed, for instance using Ripley's K-function.

If λ is spatially volatile, the transformed space bounding the rescaled residuals can be highly irregular, which makes it difficult to detect uniformity using the K-function. In this case, one can rescale the points along a different axis as in Schoenberg (1999) and see if there is any improvement. Unfortunately, most CSEP forecast models have volatile conditional intensities, resulting in a highly irregular boundary regardless of which axis is chosen for rescaling. In such cases, the K-function is dominated by boundary effects and has little power to detect excessive clustering or inhibition in the residuals. Figure 6 shows the rescaled residuals for models B and C, which had the most well behaved of the rescaled residuals for the five models we considered. There is significant clustering in both the vertically and horizontally rescaled residuals for all five models, apparently due to clustering in the observations not adequately accounted for by the models, the most noticeable of which is the very large Imperial cluster. One must be somewhat cautious, however, in interpreting rescaled residuals, because patterns observed in the points in the rescaled coordinates may be difficult to interpret.

6.2. *Thinned residuals*. Thinned residuals are a modification to the simulation techniques used by Lewis and Shedler (1979) and Ogata (1981), and, as shown in Schoenberg (2003), are useful for assessing the spatial fit of a space–time point process model and revealing locations where the model is fitting poorly. Unlike rescaled residuals, thinned residuals have the advantage that the coordinates of the points are not transformed and, thus, the resulting residuals may be easier to interpret. To obtain thinned residuals, each point (t_i, x_i, y_i) is kept independently with probability

$$\frac{b}{\hat{\lambda}(t_i, x_i, y_i)},$$

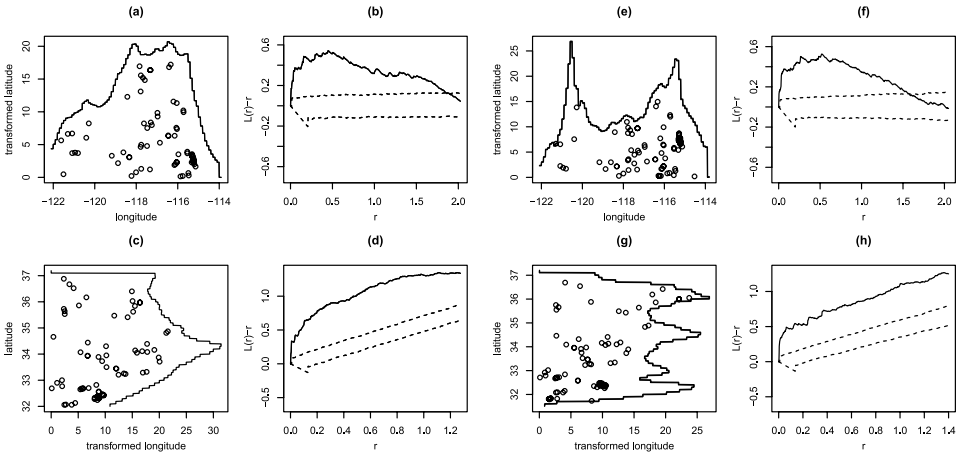

FIG. 6. *Rescaled residuals and transformed space for models B and C*. (a): *vertically rescaled residuals for model B*. (b): *estimated centered L-function for vertically rescaled residuals* (*solid line*) *and middle* 95% *ranges of estimated centered L-functions for* 1,000 *simulated homogeneous Poisson processes* (*dashed lines*). (c): *horizontally rescaled residuals for model B*. (d): *estimated centered L-function for horizontally rescaled residuals* (*solid line*) *and middle* 95% *ranges of estimated centered L-functions for* 1,000 *simulated homogeneous Poisson processes* (*dashed lines*). (e): *vertically rescaled residuals for model C*. (f): *estimated centered L-function for vertically rescaled residuals* (*solid line*) *and middle* 95% *ranges of estimated centered L-functions for* 1,000 *simulated homogeneous Poisson processes* (*dashed lines*). (g): *horizontally rescaled residuals for model B*. (h): *estimated centered L-function for horizontally rescaled residuals* (*solid line*) *and middle* 95% *ranges of estimated centered L-functions for* 1,000 *simulated homogeneous Poisson processes* (*dashed lines*).

where $b = \inf\{\hat{\lambda}(t, x, y) : (t, x, y) \in S\}$ is the infimum of the estimated intensity over the entire observed space–time window, S. The remaining points, called *thinned residual points*, should be homogeneous Poisson with rate b if and only if the fitted model for λ is correct [Schoenberg (2003)]. For this method to have sufficient power, several realizations of thinned residuals can be collected, each realization being tested for uniformity using the K-function, and then all K-functions may be examined together to get the best overall assessment of the model's fit.

When applied to the CSEP earthquake forecasts, b tends to be so small that thinning results in very few points (often zero) being retained. One can instead obtain *approximate thinned residuals* by forcing the thinning procedure to keep, on average, a certain number, k, of points by keeping each point with probability

$$k \Big/ \left(\hat{\lambda}(t_i, x_i, y_i) \sum_{i=1}^{N(S)} \hat{\lambda}(t_i, x_i, y_i)^{-1} \right)$$

as in Schoenberg (2003).

Typical examples of approximate thinned residuals for the five models we consider, using $k = 25, 15, 15, 25$ and 25 for models A, B, C, ETAS and STEP, respectively, are shown in Figure 7. Excessive clustering or inhibition in the residual

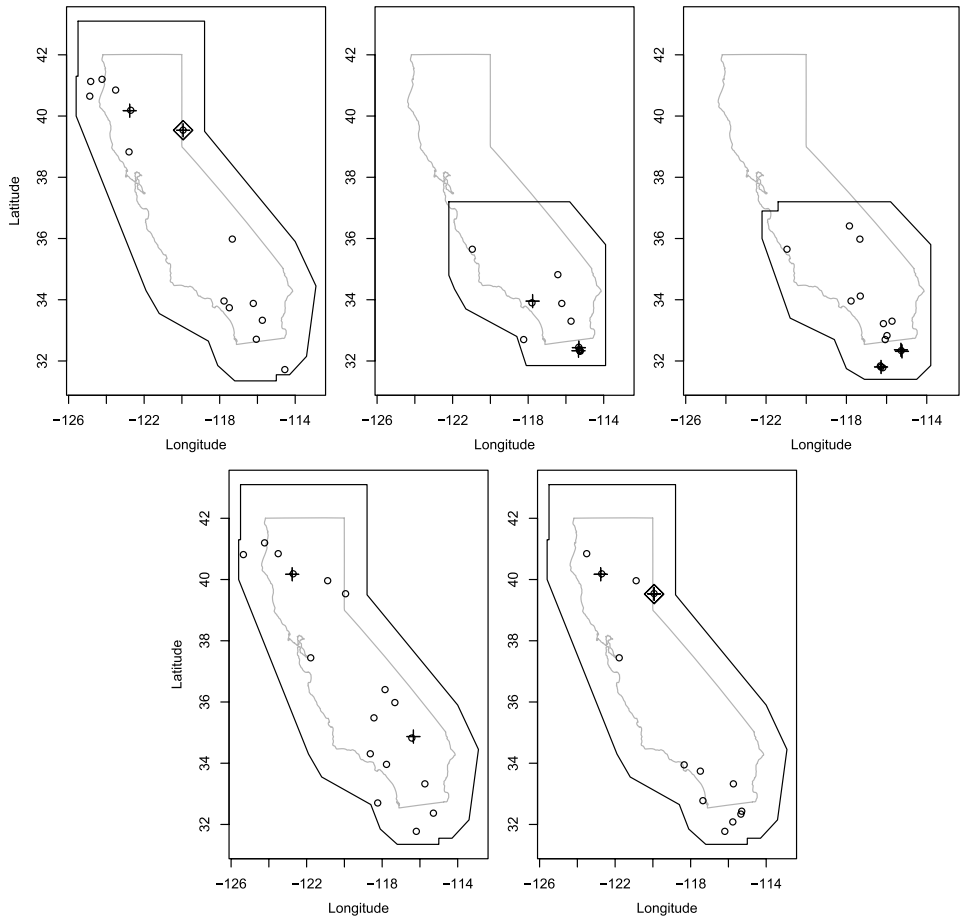

FIG. 7. *One realization of thinned residuals for each of the five models considered* (*nearby points are plotted with different symbols so they can be differentiated*). *Top-left panel* (a): *model A* ($k = 25$). *Top-center panel* (b): *model B* ($k = 15$). *Top-right panel* (c): *model C* ($k = 15$). *Bottom-left panel* (d): *ETAS* ($k = 25$). *Bottom-right panel* (e): *STEP* ($k = 25$).

process, compared with what would be expected from a homogeneous Poisson process with overall rate k, indicates lack of fit. To test the residuals for homogeneity, one may apply the weighted K-function to the residuals, with $\hat{\lambda}_0(\mathbf{x}_i) = k$ for all points $\mathbf{x}_i$. This is equivalent to using the unweighted version of the K-function on the residuals, except that here the overall rate is k, whereas with the conventional unweighted K-function, the overall rate is typically estimated as $N(S)/|S|$. The estimated centered weighted L-functions for each model, along with the 95%-confidence bands based on 2, are shown in Figure 8. Models A and STEP most noticeably fail to thin out the small cluster near the Peterson Mountain fault northwest of Reno, Nevada, and another small cluster in northern California that occurs

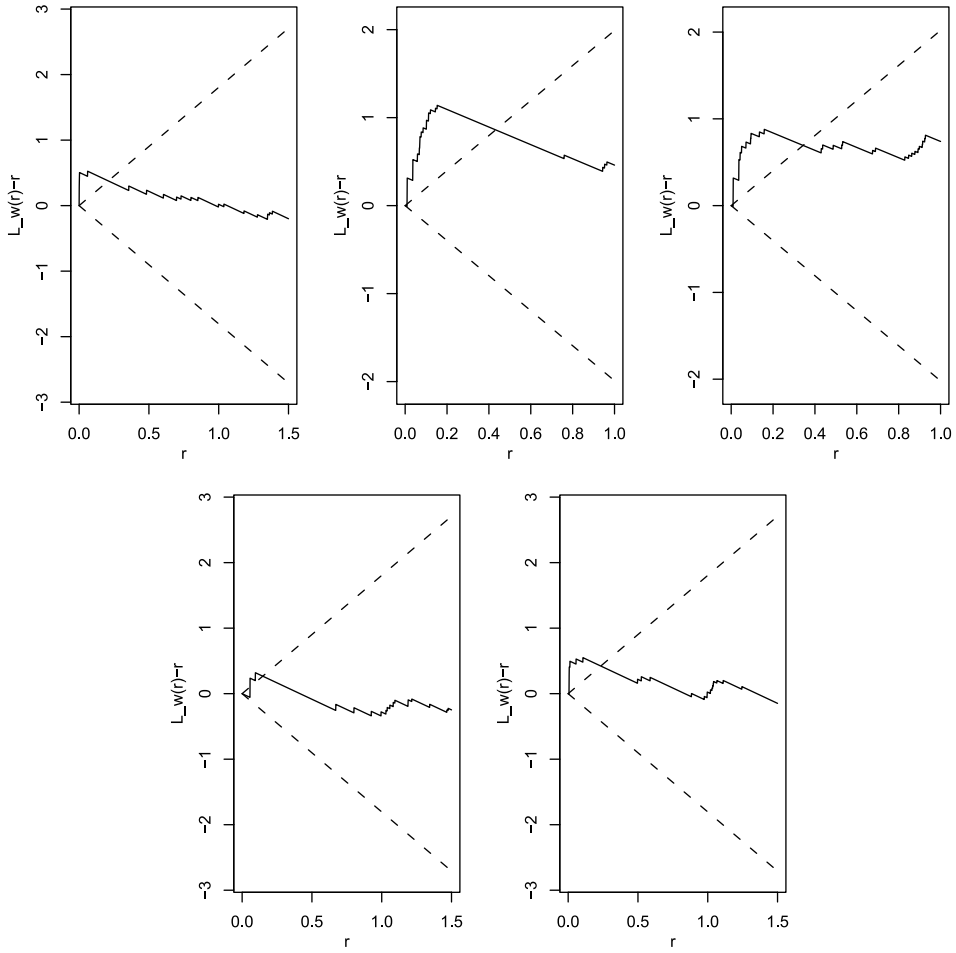

FIG. 8. *Estimated centered weighted L-function* (*solid line*) *for one realization of super-thinned residuals and* 95% *bounds* (*dashed lines*). *Top-left panel* (a): *model A* ($\hat{\lambda}_0 = 0.296$). *Top-center panel* (b): *model B* ($\hat{\lambda}_0 = 0.406$). *Top-right panel* (c): *model C* ($\hat{\lambda}_0 = 0.394$). *Bottom-left panel* (d): *ETAS* ($\hat{\lambda}_0 = 0.296$). *Bottom-right panel* (e): *STEP* ($\hat{\lambda}_0 = 0.296$).

approximately 35 kilometers south of the Battle Creek fault (lon $\approx 122.7°$W and lat $\approx 40.2°$N). This residual clustering is significant, as shown by the weighted L-functions in Figures 8(a) and (e). Model B has trouble forecasting the Imperial cluster, as evidenced by the significant clustering at distances up to 0.6°. The residuals for both models C and ETAS appear to be closer to uniformly distributed throughout the space, though further investigation of several realizations of thinned residuals reveals that model C has trouble thinning out the Baja, California cluster, which leads to some significant clustering in the residuals at very small distances.

6.3. *Superposition*. Superposition is a residual analysis technique similar to thinned residuals, but instead of removing points, one simulates new points to be added to the data and examines the result for uniformity. This procedure was proposed by Brémaud (1981), but examples of its use have been elusive. Points are simulated at each location (t, x, y) according to a Cox process with intensity $c - \hat{\lambda}(t_i, x_i, y_i)$, where $c = \sup_S\{\hat{\lambda}(t, x, y)\}$. As with thinning and rescaling, if the model for λ is correct, the union of the superimposed residuals and observed points will be homogeneous Poisson. Any patterns of inhomogeneity in the residuals aid us in identifying spots where the model fits poorly.

Superposition helps solve one of the biggest disadvantages of thinned residuals: the lack of information on the goodness of fit of the model in locations where no events occur. However, if c is large, then there is a possibility that too many points will be simulated, meaning that the behavior of the K-function will be primarily influenced by simulated points rather than actually observed data points. For models A and STEP, for example, simulated points comprise $\geq 99\%$ of the total points after superposition. For models C and ETAS, simulated points comprise $\geq 90\%$ of the superposed residual points. See Figure 9 for an example of superposed residuals for model C. Since the test for uniformity is based almost entirely on the simulated points, which are by construction approximately homogeneous for large c, the test has low power for model evaluation in such situations.

A realization of superposed residuals for model B can be seen in Figure 10, along with the corresponding centered weighted L-function as a test for homogeneity of the residuals. 95%-confidence bands for the L-function are constructed under the null hypothesis $\hat{\lambda}_0(\mathbf{x}_i) = c$ for all points $\mathbf{x}_i$. The superposed residuals are significantly more clustered than would be expected, up to distances of 0.4°,

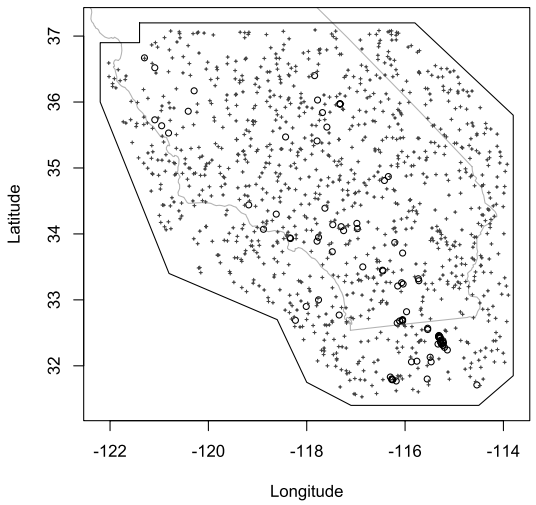

FIG. 9. *Superposed residuals for model C. Simulated points make up* 90.7% *of all points.*

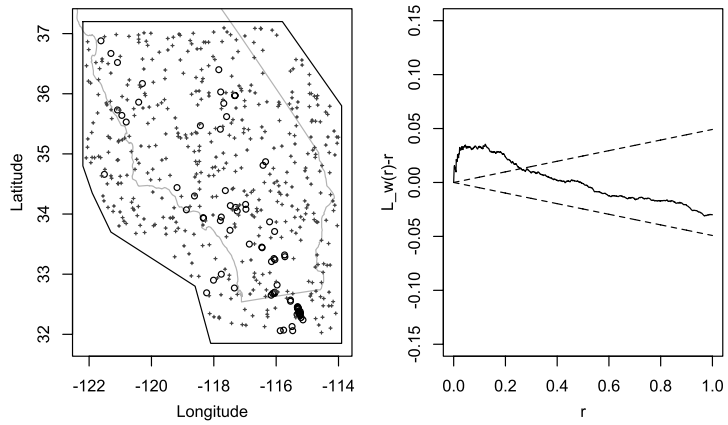

FIG. 10. *Superposed residuals for model B. Left panel* (a): *one realization of superposed residuals* (*circles* = *observed earthquakes*; *plus signs* = *simulated points*). *Right panel* (b): *estimated centered weighted L-function for superposed residuals* (*solid line*) *and* 95%*-confidence bounds* (*dashed lines*).

or approximately 44.4 km. This is likely the result of the underprediction of the seismicity rate in the Imperial cluster. One also observes significantly more inhibition in the superposed residuals than would be expected at distances greater than 0.5°, or approximately 55.5 km. This inhibition can most likely be attributed to the model's overprediction of the seismicity rate in areas devoid of earthquakes, which can be seen in the portions of Figure 10(a) in various regions lacking both simulated and observed points.

6.4. *Super-thinning*. A more powerful approach than thinning or superposition individually is a hybrid approach where one thins in areas of high intensity and superposes simulated points in areas of low intensity, resulting in a homogeneous point process if the model for λ used in the thinning and superposition is correct. The benefit of this method, called super-thinning by Clements, Schoenberg and Veen (2010), is that the user may specify the overall rate of the resulting residual point process, Z, so that it contains neither too few or too many points.

In super-thinning, one first keeps each observed point (t, x, y) in the catalog independently with probability $\min\{1, k/\hat{\lambda}(t, x, y)\}$ and subsequently superposes points generated according to a simulated Cox process with rate $\max\{0, k - \hat{\lambda}(t, x, y)\}$. The result is a homogeneous Poisson process with rate k if and only if the model $\hat{\lambda}$ for the conditional intensity is correct [Clements, Schoenberg and Veen (2010)] and, hence, the resulting super-thinned residuals can be assessed for homogeneity as a way of evaluating the model. In particular, any clustering or inhibition in the residual points indicates a lack of fit.

In the application to earthquake forecasts, a natural choice for k is the total number of expected earthquakes according to each forecast. Figure 11 shows one

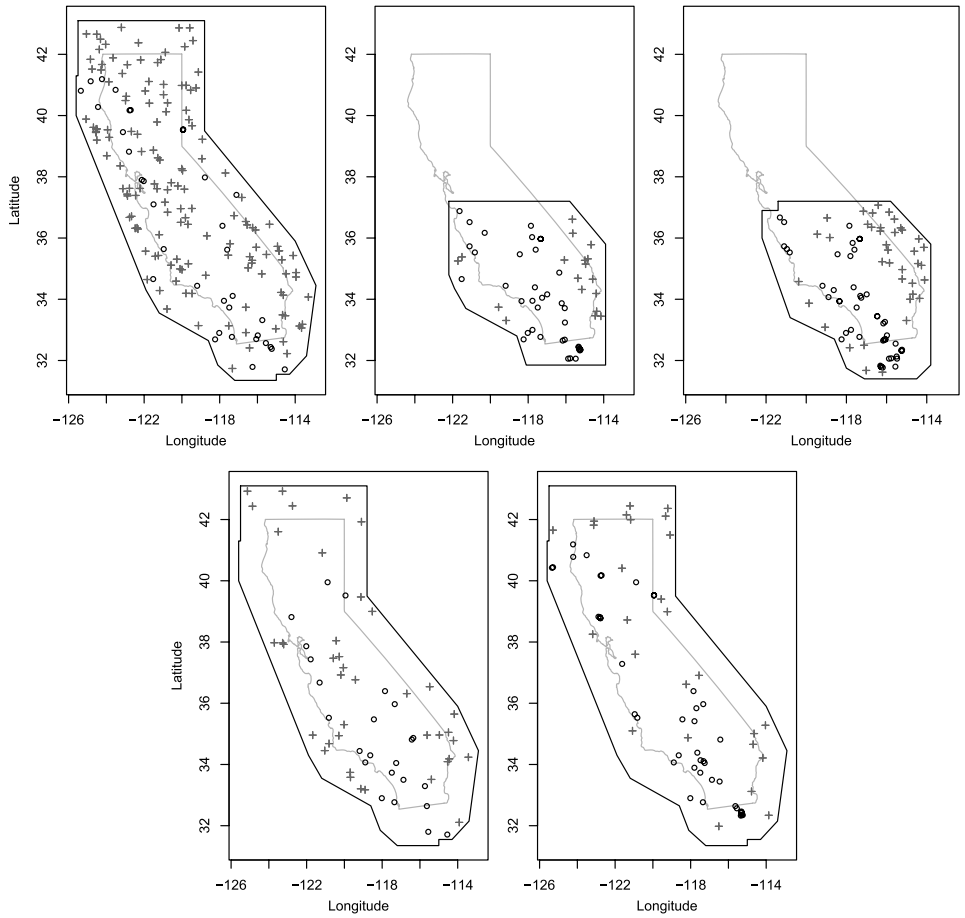

FIG. 11. *One realization of super-thinned residuals for the five models considered* (*circles = observed earthquakes*; *plus signs = simulated points*). *Top-left panel* (a): *model A* ($k = 2.76$). *Top-center panel* (b): *model B* ($k = 2.95$). *Top-right panel* (c): *model C* ($k = 2.73$). *Bottom-left panel* (d): *ETAS* ($k = 1.35$). *Bottom-right panel* (e): *STEP* ($k = 0.75$).

realization of super-thinned residuals for each model, and Figure 12 shows the estimated centered weighted L-functions for the corresponding residuals, with $\hat{\lambda}_0(\mathbf{x}_i) = k$ for all points $\mathbf{x}_i$, along with 95%-confidence bands. Model A appears to fit rather well overall, with some significant clustering in the residuals at very small distances (from 0° to 0.1°) most likely attributable to the same small clusters that remained in the thinned residuals. However, the L-function in Figure 12(a) reveals that there is somewhat more inhibition in the residual process than we would expect. This is likely attributable to model A's overprediction of the seismicity rate especially in inter-fault zones. The super-thinned residuals for model B contain a few significant clusters (Imperial, Laguna Salada and Panamint) and some slight inhibition due to overprediction of seismicity in two regions devoid of any

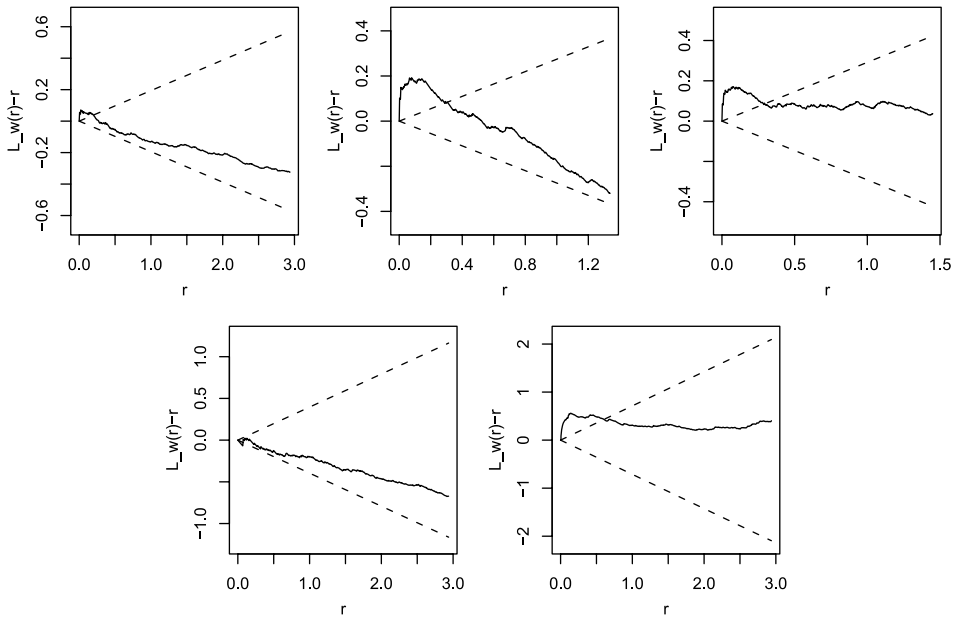

FIG. 12. *Estimated centered weighted L-function* (*solid line*) *and* 95% *bands* (*dashed lines*) *for the super-thinned residuals in Figure* 11. *Top-left panel* (a): *model A* ($\hat{\lambda}_0 = 2.76$). *Top-center panel* (b): *model B* ($\hat{\lambda}_0 = 2.95$). *Top-right panel* (c): *model C* ($\hat{\lambda}_0 = 2.73$). *Bottom-left panel* (d): *ETAS* ($\hat{\lambda}_0 = 1.35$). *Bottom-right panel* (e): *STEP* ($\hat{\lambda}_0 = 0.75$).

simulated points or retained earthquakes: the San Diego-Imperial County areas and the Los Angeles–San Bernardino areas. There is also significant clustering for model C up to distances of 0.2°, particularly the Laguna Salada, Baja and Panamint clusters. The ETAS residuals contain significant clustering at distances up to 0.1°, and this is largely attributable to the Imperial cluster and to clusters in Peterson Mountain and the Mt. Konocti area near Clearlake, California at lon $\approx 122.1°$W and lat $\approx 38.8°$N. The STEP residuals exhibit significant clustering at distances up to 0.4°, with obvious clustering at Imperial, Peterson Mountain, Battle Creek, Mt. Konocti and the Mendocino fault zone off the coast of Northwest California.

7. Summary. A litany of residual analysis methods for spatial point processes can be implemented to assess the fit and reveal weaknesses in point process models, and many of these methods provide more reliable estimates of the overall fit and more detailed information than the L-test and N-test. Rescaled residuals can assist in the evaluation of the overall spatial fit, but are not easily interpretable due to the transformed spatial window. Thinned residuals are much more easily interpretable, but suffer from variability in the thinned residual point pattern and low power if b is too small. Superposition is similar to thinning in that it also suffers from sampling variability and low power in the case of a very large supremum of $\hat{\lambda}$.

Super-thinning appears to be a promising alternative, but, like superposition, may have low power if the modeled intensity is extremely volatile. Deviance residuals and weighted second-order statistics appear to be quite powerful, especially for comparisons of competing models.

Clearly, the availability of a larger number of observed earthquakes in the tests would lead to more detailed and more meaningful results, and this suggests further decreasing the lower magnitude threshold. However, considerations of catalog incompleteness at lower magnitudes, as well as the fact that not all forecast models in the study are capable of forecasting small events and their spatial-temporal fluctuations, lead to limits on how low one may place the lower magnitude threshold for the catalog. Indeed, lowering the threshold requires stronger time-dependence of the models to account for the short-term fluctuations of microseismicity. Due to these considerations, CSEP sets the lower magnitude threshold in most cases to 3.95 for the time-varying models like STEP and ETAS.

Overall, model A seems to be overpredicting seismicity at the time of testing, but this may change once the forecast period is complete if there is a greater amount of seismic activity. Models B and C appear to be significantly underpredicting seismicity in many locations, and unless the seismic activity in these regions slows down considerably, these models will continue to underpredict for the remainder of the forecast period. The spatial distribution of model A is quite accurate, coupling forecasts of high conditional intensity in areas along active faults with very low intensity forecasts in areas adjacent to these faults which typically are devoid of earthquakes. Models B and C have smooth spatial distributions yielding erroneously high forecasts at distances far from any faults.

The question of what choice of k is optimal in thinning or super-thinning remains open for future research. Ideally, k should be chosen such that a poorly fitting model is rejected with high probability, while a "correct" or satisfactorily fitting model is rejected with low probability (i.e., the Type I error probability, α, is small). When thinning, we lose information when points are removed, so we prefer to keep as many points as possible, while keeping α low. With super-thinning, we would also ideally want to retain many of the original points while simulating few points, so that any assessment of the homogeneity of the residuals is not highly dependent on the simulations. Simulation and theoretical studies are needed in the future to compare the power of these goodness-of-fit measures under various hypotheses.

Acknowledgments. We thank Yan Kagan and Alejandro Veen for helpful comments, the Advanced National Seismic System for the earthquake catalog data, and the Collaboratory for the Study of Earthquake Predictability and the Southern California Earthquake Center for supplying the earthquake forecasts.

REFERENCES

ADELFIO, G. and SCHOENBERG, F. P. (2009). Point process diagnostics based on weighted second-order statistics and their asymptotic properties. *Ann. Inst. Statist. Math.* **61** 929–948. MR2556772

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **AC-19** 716–723. MR0423716

BADDELEY, A. J., MØLLER, J. and WAAGEPETERSEN, R. (2000). Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. *Stat. Neerl.* **54** 329–350. MR1804002

BADDELEY, A., TURNER, R., MØLLER, J. and HAZELTON, M. (2005). Residual analysis for spatial point processes. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 617–666. MR2210685

BESAG, J. E. (1977). Comments on "Modeling spatial patterns" by B. D. Ripley. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **B39** 193–195.

BOLT, B. (2006). *Earthquakes*, 5th ed. W.H. Freeman, New York.

BRÉMAUD, P. (1981). *Point Processes and Queues: Martingale Dynamics*. Springer, New York. MR0636252

CLEMENTS, R. A., SCHOENBERG, F. P. and VEEN, A. (2010). Evaluation of space–time point process models using super-thinning. *UCLA Statistics Preprints* **579** 1–18.

CRESSIE, N. A. C. (1993). *Statistics for Spatial Data*. Wiley, New York. MR1239641

DALEY, D. and VERE-JONES, D. (2003). *An Introduction to the Theory of Point Processes, Vol.* 1. Springer, New York.

FIELD, E. H. (2007). Overview of the working group for the development of Regional Earthquake Likelihood Models (RELM). *Seismological Research Letters* **78** 7–16.

GERSTENBERGER, M. C., WIEMER, S., JONES, L. M. and REASENBERG, P. A. (2005). Real-time forecasts of tomorrow's earthquakes in California. *Nature* **435** 328–331.

GUAN, Y. (2009). On nonparametric variance estimation for second-order statistics of inhomogeneous spatial point processes with a known parametric intensity form. *J. Amer. Statist. Assoc.* **104** 1482–1491.

GUTENBERG, B. and RICHTER, C. F. (1944). Frequency of earthquakes in California. *Bull. Seismol. Soc. Amer.* **142** 185–188.

HELMSTETTER, A., KAGAN, Y. Y. and JACKSON, D. D. (2007). High-resolution time-independent grid-based forecast $M \geq 5$ earthquakes in California. *Seismological Research Letters* **78** 78–86.

JACKSON, D. D. and KAGAN, Y. Y. (1999). Testable earthquake forecasts for 1999. *Seismological Research Letters* **70** 393–403.

JORDAN, T. (2006). Earthquake predictability, brick by brick. *Seismological Research Letters* **77** 3–6.

KAGAN, Y. Y., JACKSON, D. D. and RONG, Y. (2007). A testable five-year forecast of moderate and large earthquakes in southern California based on smoothed seismicity. *Seismological Research Letters* **78** 94–98.

LEWIS, P. A. W. and SHEDLER, G. S. (1979). Simulation of nonhomogeneous Poisson processes by thinning. *Naval Res. Logist. Quart.* **26** 403–413. MR0546120

MEYER, P. (1971). Demonstration simplifiée d'un théorème de Knight. In *Séminaire de Probabilités V. Lecture Notes in Math.* **191** 191–195. Springer, Berlin. MR0380972

OGATA, Y. (1981). On Lewis' simulation method for point processes. *IEEE Trans. Inform. Theory* **IT-27** 23–31.

OGATA, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *J. Amer. Statist. Assoc.* **83** 9–27.

OGATA, Y. and ZHUANG, J. (2006). Space–time ETAS models and an improved extension. *Tectonophysics* **413** 13–23.

PAPANGELOU, F. (1972). Integrability of expected increments of point processes and a related random change of scale. *Trans. Amer. Math. Soc.* **165** 483–506. MR0314102

RIPLEY, B. D. (1981). *Spatial Statistics*. Wiley, New York. MR0624436

SCHOENBERG, F. (1999). Transforming spatial point processes into Poisson processes. *Stochastic Process. Appl.* **81** 155–164. MR1694573

SCHOENBERG, F. P. (2003). Multidimensional residual analysis of point process models for earthquake occurrences. *J. Amer. Statist. Assoc.* **98** 789–795. MR2055487

SCHORLEMMER, D. and GERSTENBERGER, M. C. (2007). RELM testing center. *Seismological Research Letters* **78** 30–35.

SCHORLEMMER, D., GERSTENBERGER, M. C., WIEMER, S., JACKSON, D. D. and RHOADES, D. A. (2007). Earthquake likelihood model testing. *Seismological Research Letters* **78** 17–27.

SCHORLEMMER, D., ZECHAR, J. D., WERNER, M. J., FIELD, E. H., JACKSON, D. D., JORDAN, T. H. and THE RELM WORKING GROUP (2010). First results of the Regional Earthquake Likelihood Models Experiment. *Pure and Applied Geophysics* **167** 859–876.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

SHEN, Z., JACKSON, D. D. and KAGAN, Y. Y. (2007). Implications of geodetic strain rate for future earthquakes, with a five-year forecast of M5 earthquakes in southern California. *Seismological Research Letters* **78** 116–120.

STARK, P. B. (1997). Earthquake prediction: The null hypothesis. *Geophysical Journal International* **131** 495–499.

VEEN, A. and SCHOENBERG, F. P. (2005). Assessing spatial point process models using weighted K-functions: analysis of California earthquakes. In *Case Studies in Spatial Point Process Models* (A. Baddeley, P. Gregori, J. Mateu, R. Stoica and D. Stoyan, eds.) 293–306. Springer, New York. MR2232135

WONG, K. and SCHOENBERG, F. P. (2009). On mainshock focal mechanisms and the spatial distribution of aftershocks. *Bull. Seismol. Soc. Amer.* **99** 3402–3412.

ZHUANG, J. (2006). Second-order residual analysis of spatiotemporal point processes and applications in model evaluation. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 635–653. MR2301012

ZHUANG, J., OGATA, Y. and VERE-JONES, D. (2004). Analyzing earthquake clustering features by using stochastic reconstruction. *Journal of Geophysical Research* **109** B05301-17.

R. A. CLEMENTS
F. P. SCHOENBERG
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
8125 MATH SCIENCES BUILDING
LOS ANGELES, CALIFORNIA 90095-1554
USA
E-MAIL: clements@stat.ucla.edu
frederic@stat.ucla.edu

D. SCHORLEMMER
SOUTHERN CALIFORNIA EARTHQUAKE CENTER
UNIVERSITY OF SOUTHERN CALIFORNIA
3651 TROUSDALE PARKWAY
LOS ANGELES, CALIFORNIA 90089-0740
USA
E-MAIL: ds@usc.edu
AND
GERMAN RESEARCH CENTER FOR GEOSCIENCES
TELEGRAFENBERG
14473 POTSDAM
GERMANY
E-MAIL: ds@gfz-potsdam.de