Statistical Science
2013, Vol. 28, No. 4, 510–520
DOI: 10.1214/13-STS440

Assessment of Point Process Models for Earthquake Forecasting

Andrew Bray and Frederic Paik Schoenberg

Abstract. Models for forecasting earthquakes are currently tested prospectively in well-organized testing centers, using data collected after the models and their parameters are completely specified. The extent to which these models agree with the data is typically assessed using a variety of numerical tests, which unfortunately have low power and may be misleading for model comparison purposes. Promising alternatives exist, especially residual methods such as super-thinning and Voronoi residuals. This article reviews some of these tests and residual methods for determining the goodness of fit of earthquake forecasting models.

Key words and phrases: Earthquakes, model assessment, point process, residual analysis, spatial–temporal statistics, super-thinning.

1. INTRODUCTION

A major goal in seismology is the ability to accurately anticipate future earthquakes before they occur (Bolt, 2003). Anticipating major earthquakes is especially important, not only for short-term response such as preparation of emergency personnel and disaster relief, but also for longer-term preparation in the form of building codes, urban planning and earthquake insurance (Jordan and Jones, 2010). In seismology, the phrase *earthquake prediction* has a specific definition: it is the identification of a meaningfully small geographic region and time window in which a major earthquake will occur with very high probability. An example of earthquake predictions are those generated by the M8 method (Keilis-Borok and Kossobokov, 1990), which issues an alarm whenever there is a suitably large increase in the background seismicity of a region. Such alarms could potentially be very valuable for short-term disaster preparedness, but unfortunately examples of M8-type alarms, including the notable Reverse Tracing of Precursors (RTP) algorithm, have generally exhibited low reliability when tested prospectively, typically failing to outperform naive methods based simply on smoothed historical seismicity (Geller et al., 1997; Zechar and Jordan, 2008).

Earthquake prediction can be contrasted with the related *earthquake forecasting*, which means the assignment of probabilities of earthquakes occurring in broader space–time-magnitude regions. The temporal scale of an earthquake forecast is more on par with climate forecasts and may be over intervals that range from decades to centuries (Hough, 2010).

Many models have been proposed for forecasting earthquakes, and since different models often result in very different forecasts, the question of how to assess which models seem most consistent with observed seismicity becomes increasingly important. Concerns with retrospective analyses, especially regarding data selection, overfitting and lack of reproducibility, have motivated seismologists recently to focus on prospective assessments of forecasting models. This has led to the development of the Regional Earthquake Likelihood Models (RELM) and Collaborative Study of Earthquake Predictability (CSEP) testing centers, which are designed to evaluate and compare the goodness of fit of various earthquake forecasting models. This paper surveys methods for assessing the models in these RELM and CSEP experiments, including methods currently used by RELM and CSEP and some others not yet in use but which seem promising.

Andrew Bray is Ph.D. Student and Frederic Paik Schoenberg is Professor, UCLA, Department of Statistics, 8125 Math Sciences Building, Los Angeles, California 90095-1554, USA.

2. A FRAMEWORK FOR PROSPECTIVE TESTING

The current paradigm for building and testing earthquake models emerged from the working group for the development of Regional Earthquake Likelihood Models (RELM) in 2001. As described in Field (2007), the participants were encouraged to submit differing models, in the hopes that the competition between models would prove more useful than trying to build a single consensus model. The competition took place within the framework of a prospective test of their seismicity forecasts. Working from a standardized data set of historical seismicity, scientists fit their models and submit to RELM a forecast of the number of events expected within each of many pre-specified spatial–temporal-magnitude bins. The first predictive experiment required models to forecast seismicity in California between 2006 to 2011 using only data from before 2006.

This paradigm has many benefits from a statistical perspective. The prospective nature of the experiments effectively eliminates concerns about overfitting. Furthermore, the standardized nature of the data and forecasts facilitates the comparison among different models. RELM has since expanded into the Collaborative Study of Earthquake Predictability (CSEP), a global-scale project to coordinate model development and conduct prospective testing according to community standards (Jordan, 2006). CSEP serves as an independent entity that provides standardized seismicity data, inventories proposed models and publishes the standards by which the models will be assessed.

3. SOME EXAMPLES OF MODELS FOR EARTHQUAKE OCCURRENCES

The first predictive experiment coordinated through RELM considered time-independent spatial point process models, which can be specified by their Papangelou intensity $\lambda(s)$, a function of spatial location s. A representative example is the model specified by Helmstetter, Kagan and Jackson (2007) that is based on smoothing previous seismicity. The intensity function is estimated with an isotropic adaptive kernel

$$\lambda(s) = \sum_{i=1}^{N} K_d(s - s_i),$$

where N is the total number of observed points, and K_d is a power-law kernel

$$K_d(s - s_i) = \frac{C(d)}{(|s - s_i|^2 + d^2)^{1.5}},$$

where d is the smoothing distance, $C(d)$ is a normalizing factor so that the integral of $K_d(\cdot)$ over an infinite area equals 1, and $|\cdot|$ is the Euclidean norm. The estimated number of points within the pre-specified grid cells is obtained by integrating $\lambda(s)$ over each cell.

Models of earthquake occurrence that consider it to be a time-dependent process are commonly variants of the epidemic-type aftershock sequence (ETAS) model of Ogata (1988, 1998) (see, e.g., Helmstetter and Sornette, 2003; Ogata, Jones and Toda, 2003; Sornette, 2005; Vere-Jones and Zhuang, 2008; Console, Murru and Falcone, 2010; Chu et al., 2011; Wang, Jackson and Kagan, 2011; Werner et al., 2011; Zhuang, 2011; Tiampo and Shcherbakov, 2012). According to the ETAS model, earthquakes cause aftershocks, which in turn cause more aftershocks, and so on. ETAS is a point process model specified by its conditional intensity, $\lambda(s, t)$, which represents the infinitesimal expected rate at which events are expected to occur around time t and location s, given the history H_t of the process up to time t. ETAS is a special case of the linear, self-exciting Hawkes' point process (Hawkes, 1971), where the conditional intensity is of the form

$$\lambda(s, t|H_t) = \mu(s, t) + \sum_{t_i<t} g(s - s_i, t - t_i; M_i),$$

where $\mu(s, t)$ is the mean rate of a Poisson-distributed background process that may in general vary with time and space, g is a *triggering function* which indicates how previous occurrences contribute, depending on their spatial and temporal distances and marks, to the conditional intensity λ at the location and time of interest, and (s_i, t_i, M_i) are the origin times, epicentral locations and moment magnitudes of observed earthquakes.

Ogata (1998) proposed various forms for the triggering function, g, such as the following:

$$g(s, t, M) = K(t + c)^{-p} e^{a(M - M_0)} \big(|s|^2 + d\big)^{-q},$$

where M_0 is the lower magnitude cutoff for the observed catalog.

The parameters in ETAS models and other spatial–temporal point process models may be estimated by maximizing the log-likelihood,

$$\sum_{i=1}^{n} \log\{\lambda(s_i, t_i)\} - \int_S \int \lambda(s, t)\, \mathrm{d}s\, \mathrm{d}t.$$

The maximum likelihood estimator (MLE) of a point process is, under quite general conditions, asymptotically unbiased, consistent, asymptotically normal and

asymptotically efficient (Ogata, 1978). Finding the parameter vector that maximizes the log-likelihood can be achieved using any of the various standard optimization routines, such as the quasi-Newton methods implemented in the function optim(·) in *R*. The spatial background rate μ in the ETAS model can be estimated in various ways, such as via kernel smoothing seismicity from prior to the observation window or kernel smoothing the largest events in the catalog, as in Ogata (1998) or Schoenberg (2003). Note that the integral term in the loglikelihood function can be cumbersome to estimate, and an approximation method recommended in Schoenberg (2013) can be used to accelerate computation of the MLE.

There are of course many other earthquake forecasting models quite distinct from the two point process models above. Perhaps most important among these are the Uniform California Earthquake Rupture Forecast (UCERF) models, which are consulted when setting insurance rates and crafting building codes (Field et al., 2009). They are constructed by soliciting *expert opinion* from leading seismologists on which components should enter the model, how they should be weighted, and how they should interact (Marzocchi and Zechar, 2011). Examples of the components include slip rate, geodetic strain rates and paleoseismic data. Note that some seismologists have argued that evaluating some earthquake forecasting models such as UCERF using model validation experiments such as RELM and CSEP may be inappropriate, though such a conclusion seems to run counter to basic statistical and scientific principles.

Although the UCERF models draw upon diverse information related to the geophysics of earthquake etiology, commonly used models such as ETAS and its variants rely solely on previous seismicity for forecasting future events. Many attempts have been made to include covariates, but when assessed rigorously, most predictors other than the locations and times of previous earthquakes have been shown not to offer any noticeable improvement in forecasting. Recent examples of such covariates include electromagnetic signals (Jackson, 1996; Kagan, 1997), radon (Hauksson and Goddard, 1981) and water levels (Bakun et al., 2005; Manga and Wang, 2007). A promising exception is moment tensor information, which is now routinely recorded with each earthquake and seems to give potentially useful information regarding the directionality of the release of stress in each earthquake. However, this information appears not to be explicitly used presently in models in the CSEP or RELM forecasts.

4. NUMERICAL TESTS

Several numerical tests were initially proposed to serve as the metrics by which RELM models would be evaluated (Schorlemmer et al., 2007). For these numerical tests, each model consists of the estimated number of earthquakes in each of the spatial–temporal-magnitude bins, where the number of events in each bin is assumed to follow a Poisson distribution with an intensity parameter equivalent to the forecasted rate.

The *L*-test (or Likelihood test) evaluates the probability of the observed data under the proposed model. The numbers of observed earthquakes in each spatial–temporal-magnitude bin are treated as independent random variables, so the joint probability is calculated simply as the product of their corresponding Poisson probabilities. This observed joint probability is then considered with respect to the distribution of joint probabilities generated by simulating many synthetic data sets from the model. If the observed probability is unusually low in the context of this distribution, the data are considered inconsistent with the model.

The *N*-test (Number) ignores the spatial and magnitude component and focuses on the total number of earthquakes summed across all bins. If the proposed model provides estimates $\hat{\lambda}_i$ for i corresponding to each of B bins, then according to this model, the total number of observed earthquakes should be Poisson distributed with mean $(\sum_{i=1}^{B} \hat{\lambda}_i)$. If the number of observed earthquakes is unusually large or small relative to this distribution, the data are considered inconsistent with the model.

The *L*-test is considered more comprehensive in that it evaluates the forecast in terms of magnitude, spatial location and number of events, while the *N*-test restricts its attention to the number of events. Two additional data consistency tests were proposed to assess the magnitude and spatial components of the forecasts, respectively: the *M*-test and the *S*-test (Zechar, Gerstenberger and Rhoades, 2010). The *M*-test (Magnitude) isolates the forecasted magnitude distribution by counting the observed number of events in each magnitude bin without regard to their temporal or spatial locations, standardized so that the observed and expected total number of events under the model agree, and computing the joint (Poisson) likelihood of the observed numbers of events in each magnitude bin. As with the *L*-test, the distribution of this statistic under the forecast is generated via simulation.

The *S*-test (Spatial) follows the same inferential procedure but isolates the forecasted spatial distribution by

summing the numbers of observed events over all times and over all magnitude ranges. These counts within each of the spatial bins are again standardized so that the observed and expected total number of events under the model agree, and then one computes the joint (Poisson) likelihood of the observed numbers of events in the spatial bins.

The above tests measure the degree to which the observations agree with a particular model, in terms of the probability of these observations under the given model. As noted in Zechar et al. (2013), tests such as the L-test and N-test are really tests of the consistency between the data and a particular model, and are not ideal for comparing two models. Schorlemmer et al. (2007) proposed an additional test to allow for the direct comparison of the performance of two models: the Ratio test (R-test). For a comparison of models A and B, and given the numbers of observed events in each bin, the test statistic R is defined as the log-likelihood of the data according to model A minus the corresponding log-likelihood for model B. Under the null hypothesis that model A is correct, the distribution of the test statistic is constructed by simulating from model A and calculating R for each realization. The resulting test is one-sided and is supplemented with the corresponding test using model B as the null hypothesis. The T-test and W-test of Rhoades et al. (2011) are very similar to the R-test, except that instead of using simulations to find the null distribution of the difference between log-likelihoods, with the T-test and W-test, the differences between log-likelihoods within each space–time-magnitude bin for models A and B are treated as independent normal or symmetric random variables, respectively, and a t-test or Wilcoxon signed rank test, respectively, is performed.

Unfortunately, when used to compare various models, such likelihood-based tests suffer from the problem of variable null hypotheses and can lead to highly misleading and even seemingly contradictory results. For instance, suppose model A has a higher likelihood than model B. It is nevertheless quite possible for model A to be rejected according to the L-test and model B not to be rejected using the L-test. Similarly, the R-test with model A as the null might indicate that model A performs statistically significantly better than model B, while the R-test with model B as the null hypothesis may indicate that the difference in likelihoods is not statistically significant. Seemingly paradoxical results like these occur frequently, and at a recent meeting of the Seismological Society of America, much confusion was expressed over such results; even some seismologists quite well versed in statistics referred to results in such circumstances as "somewhat mixed," even though model A clearly fit better according to the likelihood criterion than model B.

The explanation for such results is that the null hypotheses of the two tests are different: when model A is tested using the L-test, the null hypothesis is model A, and when model B is tested, the null hypothesis is model B. The test statistic may have very different distributions under these different hypotheses.

Unfortunately, these types of discrepancies seem to occur frequently and hence, the results of these numerical tests may not only be uninformative for model comparison, but in fact highly misleading. A striking example is given in Figure 4 of Zechar et al. (2013), where the Shen, Jackson and Kagan (2007) model produces the highest likelihood of the five models considered in this portion of the analysis, and yet under the L-test has the lowest corresponding p-value of the five models.

5. FUNCTIONAL SUMMARIES

Functional summaries, that is, those producing a function of one variable, such as the weighted K-function and error diagrams, can also be useful measures of goodness of fit. However, such summaries typically provide little more information than numerical tests in terms of indicating where and when the model and the data fail to agree or how a model may be improved.

The weighted K-function is a generalized version of the K-function of Ripley (1976), which has been widely used to detect clustering or inhibition for spatial point processes. The ordinary K function, $K(h)$, counts, for each h, the total number of observed pairs of points within distance h of one another, per observed point, standardized by dividing by the estimated overall mean rate of the process, and the result is compared to what would be expected for a homogeneous Poisson process. The weighted version, $K_w(h)$, was introduced for the inhomogeneous spatial point process case by Baddeley, Møller and Waagepetersen (2000), and is defined similarly to $K(h)$, except that each pair of points (s_i, s_j) is weighted by $1/[\hat{\lambda}(s_i)\hat{\lambda}(s_j)]$, the inverse of the product of the modeled unconditional intensities at the points s_i and s_j. This was extended to spatial–temporal point processes by Veen and Schoenberg (2006) and Adelfio and Schoenberg (2009).

Whereas the null hypothesis for the ordinary K-function is a homogeneous Poisson process, in the case

of K_w, the weighting allows one to assess whether the degree of clustering or inhibition in the observations is consistent with what would be expected under the null hypothesis corresponding to the model for $\hat{\lambda}$. While weighted K-functions may be useful for indicating whether the degree of clustering in the model agrees with that in the observations, such summaries unfortunately do not appear to be useful for comparisons between multiple competing models, nor do they accurately indicate in which spatial–temporal-magnitude regions there may be particular inconsistencies between a model and the observations.

Error diagrams, which are also sometimes called receiver operating characteristic (ROC) curves (Swets, 1973) or Molchan diagrams (Molchan, 1991, 2010; Zaliapin and Molchan, 2004; Kagan, 2009), plot the (normalized) number of alarms versus the (normalized) number of false negatives (failures to predict), for each possible alarm, where in the case of earthquake forecasting models an *alarm* is defined as any value of the modeled conditional rate, $\hat{\lambda}$, exceeding some threshold. Figure 1 presents error diagrams for two RELM models, Helmstetter, Kagan and Jackson (2007) and Shen, Jackson and Kagan (2007) (see Sections 3 and 7 for model details).

The ease of interpretation of such diagrams is an attractive feature, and plotting error diagrams with multiple models on the same plot can be a useful way to compare the models' overall forecasting efficacy. In Figure 1 we learn that Shen, Jackson and Kagan (2007) slightly outperforms Helmstetter, Kagan and Jackson (2007) when the threshold for the alarm is high, but as the threshold is lowered Helmstetter, Kagan and Jackson (2007) performs noticeably better. For the purpose of comparing models, one may even consider normalizing the error diagram so that the false negative rates are considered relative to one of the given models in consideration as in Kagan (2009). This tends to alleviate a common problem with error diagrams as applied to earthquake forecasts, which is that most of the relevant focus is typically very near the axes and thus it can be difficult to inspect differences between the models graphically. A more fundamental problem with error diagrams, however, is that while they can be useful overall summaries of goodness of fit, such diagrams unfortunately provide little information as to where models are fitting poorly or how they may be improved.

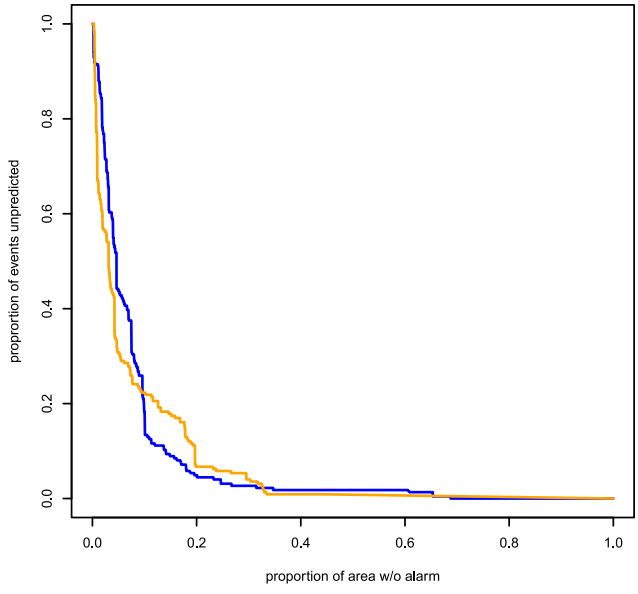

FIG. 1. *Error diagrams for Helmstetter, Kagan and Jackson (2007) in blue and Shen, Jackson and Kagan (2007) in orange. Model details are in Sections 3 and 7, respectively.*

6. RESIDUAL METHODS

Residual analysis methods for spatial–temporal point process models produce graphical displays which may highlight where one model outperforms another or where a particular model does not ideally agree with the data. Some residual methods, such as thinning, rescaling and superposition, involve transforming the point process using a model for the conditional intensity λ and then inspecting the uniformity of the result, thus reducing the difficult problem of evaluating the agreement between a possibly complex spatial–temporal point process model and data to the simpler matter of assessing the homogeneity of the residual point process. Often, departures from homogeneity in the residual process can be inspected by eye, and many standard tests are also available. Other residual methods, such as pixel residuals, Voronoi residuals and deviance residuals, result in graphical displays that can quite directly indicate locations where a model appears to depart from the observations or where one model appears to outperform another in terms of agreement with the data.

6.1 Thinned, Superposed and Super-Thinned Residuals

Thinned residuals are based on the technique of random thinning, which was first introduced by Lewis and Shedler (1979) and Ogata (1981) for the purpose of simulating spatial–temporal point processes and extended for the purpose of model evaluation in

Schoenberg (2003). The method involves keeping each observed point (earthquake) independently with probability $b/\hat{\lambda}(s_i, t_i)$, where $b = \inf_{(s,t)\in\mathcal{S}}\{\hat{\lambda}(s,t)\}$ and $\hat{\lambda}$ is the modeled conditional intensity. If the model is correct, that is, if the estimate $\hat{\lambda}(s,t) = \lambda(s,t)$ almost everywhere, then the residual process will be homogeneous Poisson with rate b (Schoenberg, 2003). Because the thinning is random, each thinning is distinct, and one may inspect several realizations of thinned residuals and analyze the entire collection to get an overall assessment of goodness of fit, as in Schoenberg (2003).

An antithetical approach was proposed by Bremaud (1981), who suggested superposing a simulated point process onto an observed point process realization so as to yield a homogeneous Poisson process. As indicated in Clements, Schoenberg and Veen (2012), tests based on thinned or superposed residuals tend to have low power when the model $\hat{\lambda}$ for the conditional intensity is volatile, which is typically the case with earthquake forecasts since earthquakes tend to be clustered in particular spatial–temporal regions. Thinning a point process will lead to very few points remaining if the infimum of $\hat{\lambda}$ over the observed space is small (Schoenberg, 2003), while in superposition, the simulated points, which are by construction approximately homogeneous, will form the vast majority of residual points if the supremum of $\hat{\lambda}$ is large.

A hybrid approach called super-thinning was introduced in Clements, Schoenberg and Veen (2012). With super-thinning, a tuning parameter k is chosen, and one thins (deletes) the observed points in locations of space–time where $\hat{\lambda} > k$, keeping each point independently with probability $k/\hat{\lambda}(s,t)$, and superposes a Poisson process with rate $\hat{\lambda}(s,t)/k$ where $\hat{\lambda} < k$. When the tuning parameter k is chosen wisely, the method appears to be more powerful than thinning or superposing in isolation.

6.2 Rescaled Residuals

An alternative method for residual analysis is rescaling. The idea behind rescaled residuals dates back to Meyer (1971), who investigated rescaling temporal point processes according to their conditional intensities, moving each point t_i to a new time $\int_0^{t_i}\hat{\lambda}(t)\,\mathrm{d}t$, creating a transformed space in which the rescaled points are homogeneous Poisson of unit rate. Heuristically, the space is essentially compressed when $\hat{\lambda}$ is small and stretched when $\hat{\lambda}$ is large, so that the points are ultimately uniformly distributed in the resulting transformed space, if the model for $\hat{\lambda}$ is correct. This method was used in Ogata (1988) to assess a temporal ETAS model and extended in Merzbach and Nualart (1986), Nair (1990), Schoenberg (1999) and Vere-Jones and Schoenberg (2004) to the spatial and spatial–temporal cases. Rescaling may result in a transformed space that is difficult to inspect if $\hat{\lambda}$ varies widely over the observation region, and in such cases standard tests of homogeneity such as Ripley's K-function may be dominated by boundary effects, as illustrated in Schoenberg (2003).

6.3 Pixel Residuals

A different type of residual analysis which is more closely analogous to standard residual methods in regression or spatial statistics is to consider the (standardized) differences between the observed and expected numbers of points in each of various spatial or spatial–temporal pixels or grids, producing what might be called *pixel residuals*. These types of residuals were described in great detail by Baddeley et al. (2005) and Baddeley, Møller and Pakes (2008). More precisely, the *raw* pixel residual on each pixel A_i is defined as $N(A_i) - \int\hat{\lambda}(s,t)\,\mathrm{d}t\,\mathrm{d}s$, where $N(A_i)$ is simply the number of points (earthquakes) observed in pixel A_i (Baddeley et al., 2005). Baddeley et al. (2005) also proposed various standardizations including Pearson residuals, which are scaled in relation to the standard deviation of the raw residuals: $r_i = \frac{N(A_i)-\int\hat{\lambda}(s,t)\,\mathrm{d}t\,\mathrm{d}s}{\sqrt{\int\hat{\lambda}(s,t)\,\mathrm{d}t\,\mathrm{d}s}}$.

A problem expressed in Bray et al. (2014) is that if the pixels are too large, then the method is not powerful to detect local inconsistencies between the model and data, and places in the interior of a pixel where the model overestimates seismicity may cancel out with places where the model underestimates seismicity. On the other hand, if the pixels are small, then the majority of the raw residuals are close to zero while those few that correspond to pixels with an earthquake are close to one. In these situations where the residuals have a highly skewed distribution, the skew is only intensified by the standardization to Pearson residuals. As a result, plots of both the raw and the Pearson residuals are not informative and merely highlight the pixels where earthquakes occur regardless of the fit of the model. The raw or Pearson residuals may be smoothed, as in Baddeley et al. (2005), but such smoothing typically only reveals gross, large-scale inconsistencies between the model and data.

If one is primarily interested in comparing competing models, then instead one may plot, in each pixel, the difference between log-likelihoods for the two models, as in Clements, Schoenberg and Schorlemmer (2011). The resulting residuals may be called

deviance residuals, in analogy with residuals from logistic regression and other generalized linear models. Deviance residuals appear to be useful for comparing models on grid cells and inspecting where one model appears to fit the observed earthquakes better than the other. It remains unclear how these residuals may be used or extended to enable comparisons of more than two competing models, other than by comparing two at a time.

6.4 Voronoi Residuals

One method of addressing the problem of pixel size specification is to use a data-driven, spatially adaptive partition such as the Voronoi tessellation, as suggested in Bray et al. (2014). Given n observed earthquakes, one may obtain a collection of n Voronoi cells $A_1, \ldots, A_n$, where A_i is defined as the collection of spatial–temporal locations closer to the particular point (earthquake) i than to any of the other observed points (Okabe et al., 2000). Thus, $N(A_i) = 1$ for each cell A_i. One may then compute the corresponding standardized residuals $r_i = \frac{1-\int \hat{\lambda}(s,t)\,dt\,ds}{\sqrt{\int \hat{\lambda}(s,t)\,dt\,ds}}$ over the Voronoi cells A_i. As with pixel residuals, for each Voronoi cell one may choose to plot the raw residual, or the residual deviance if one is interested in comparing competing models. Voronoi residuals are shown in Bray et al. (2014) to be generally less skewed than pixel residuals and are approximately Gamma distributed under quite general regularity conditions.

7. EXAMPLES

In the present section we apply some of the residual methods discussed above to models and seismicity data from the 5-year RELM prediction experiment that ran from 2006 to 2011. The original experiment called for modelers to estimate the number of earthquakes above magnitude 4.95 that would occur in many pre-specified spatial bins in California. During this time period only 23 earthquakes that fit these criteria were recorded, a fairly small data set from which to assess a model. In order to better demonstrate the methods available in residual analysis, the models that we consider were recalibrated using their specified magnitude distributions to forecast earthquakes of greater than magnitude 4.0, of which there are 232 on record.

The first model under consideration is one that was submitted to RELM by Helmstetter, Kagan and Jackson (2007) and is described in Section 3. The left panel of Figure 2 shows the estimated number of earthquakes in every pixel in the greater California region that were part of the prediction experiment. Pixels shaded very light gray have a forecast of near zero earthquakes, while pixels shaded black forecast much greater seismicity. The tan circles are the epicenters of the 232 earthquakes in the catalog, many of which are concentrated just South of the Salton Sea, near the border between California and Mexico.

The extent to which the observed seismicity is in agreement with the forecast can be visualized in the raw pixel residual plot (center panel). The pixels are those established by the RELM experiment. Pixels where the model predicted more events than were observed are shaded in red; pixels where there was underprediction are shown in blue. The degree of color saturation indicates the p-value of the observed residual in the context of the forecasted Poisson distribution. Thus, while the Helmstetter, Kagan and Jackson (2007) model greatly underpredicted the number of events in the Salton Sea trough (dark blue), it also forecasted a high level of seismicity in several isolated pixels that experienced no earthquakes (dark red). The majority of the pixels are shaded very light red, indicating regions where the model forecast a very low rate of seismicity and no earthquakes were recorded.

The Voronoi residual plot for the Helmstetter, Kagan and Jackson (2007) model is shown in the right panel of Figure 2. The spatial adaptivity of this partition is evidenced by the small tiles in regions of high point density and larger tiles in low density regions. The region of consistent underprediction in the Salton Sea trough is easily identified. Unlike the raw pixel residual plot, the Voronoi plot appears to distinguish between areas where the high isolated rates can be considered substantial overprediction (dark red) and areas where, considered in the context of the larger tile, the overprediction is less extreme (light red).

In Figure 3 we assess how well the Helmstetter, Kagan and Jackson (2007) model performs relative to another model in RELM using deviance residuals. The Shen, Jackson and Kagan (2007) model is notable for utilizing geodetic strain-rate information from past earthquakes as a proxy for the density (intensity) of the process. $\mu(\cdot)$ is then an interpolation of this data catalog. The result is a forecast that is generally much smoother than the Helmstetter, Kagan and Jackson (2007) forecast, as seen in the left panel of Figure 3. The center panel displays the deviance residuals for the Helmstetter, Kagan and Jackson (2007) model relative to the Shen, Jackson and Kagan (2007) model. The color scale is mapped to a measure of the comparative performance of the two models ranging from 1

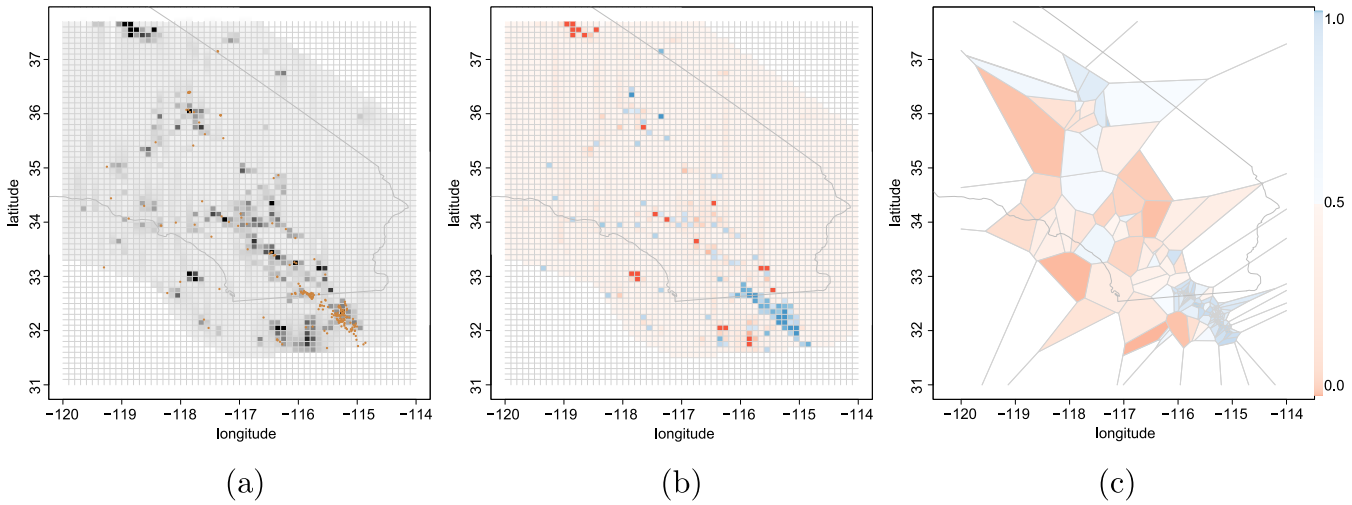

FIG. 2. (a) *Estimated rates under the Helmstetter, Kagan and Jackson* (2007) *model, with epicentral locations of observed earthquakes with* $M \geq 4.0$ *in Southern California between January* 1, 2006 *and January* 1, 2011 *overlaid.* (b) *Raw pixel residuals for Helmstetter, Kagan and Jackson* (2007) *with pixels colored according to their corresponding p-values.* (c) *Voronol residuals for Helmstetter, Kagan and Jackson* (2007) *with pixels colored according to their corresponding p-values.*

(dark blue) indicating better performance of the Helmstetter, Kagan and Jackson (2007) model to −1 (dark red) indicating better performance of the Shen, Jackson and Kagan (2007) model. This deviance residual plot reveals that the Helmstetter, Kagan and Jackson (2007) model's relative advantage is in broad areas off of the main fault lines where the forecast was lower and there were no recorded earthquakes. It appeared to fit worse than the Shen, Jackson and Kagan (2007) model, however, just West of the Salton Sea trough region of high seismicity, in a swath off the coast, and in isolated pixels in central California.

The Voronoi deviance plot (right panel) identifies the same relative underperformance of the Helmstetter, Kagan and Jackson (2007) model relative to the Shen, Jackson and Kagan (2007) model in the central California region and off the coast and is a bit more informative in the areas of higher recorded seismicity.

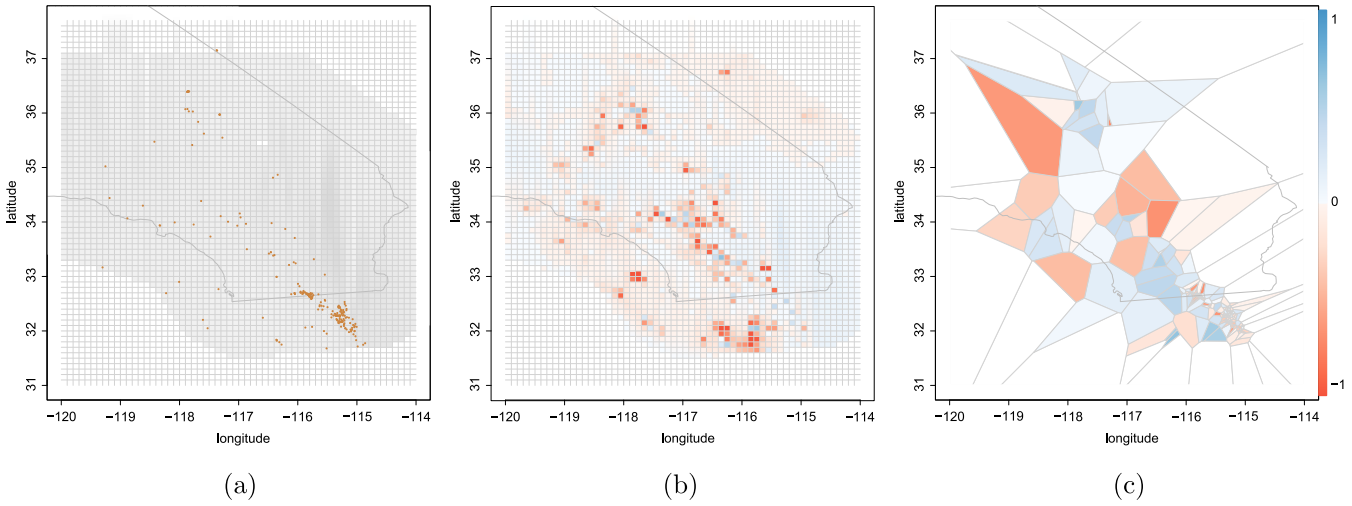

FIG. 3. (a) *Estimated rates under the Shen, Jackson and Kagan* (2007) *model, with epicentral locations of observed earthquakes with* $M \geq 4.0$ *in Southern California between January* 1, 2006 *and January* 1, 2011 *overlaid.* (b) *Pixel deviance plot with blue favoring model A, Helmstetter, Kagan and Jackson* (2007), *versus model B, Shen, Jackson and Kagan* (2007). *Coloration is on a linear scale.* (c) *Voronoi deviance plot with blue favoring model A, Helmstetter, Kagan and Jackson* (2007), *versus model B, Shen, Jackson and Kagan* (2007). *Coloration is on a linear scale.*

In the Salton Sea trough region, just south of the border of California with Mexico, the Helmstetter, Kagan and Jackson (2007) model appears to outperform the Shen, Jackson and Kagan (2007) model in a vertical swath on the Western side of the seismicity, while the results on the Eastern side are more mixed. While these regions appear nearly white in the pixel deviance residual plot, suggesting roughly equivalent performance of the models, the aggregation of many of those pixels in the Voronoi plot allows for a stronger comparison of the two models.

The utility of residual methods can be seen by contrasting the residual plots with the error diagram of these same two models (Figure 1 in Section 5). While the error diagram and other functional summaries collapse the model and the observations into a new measure (such as the false negative rate), residual methods preserve the spatial referencing, which can help inform subsequent model generation.

8. DISCUSSION

The paradigm established by RELM and CSEP is a very promising direction for earthquake model development. In addition to requiring the full transparent specification of earthquake forecasts before the beginning of the experiment, the criteria on which these models would be evaluated, namely, the L, N and R tests, was also established. As the first RELM experiment proceeded, it became apparent that these tests can be useful summaries of the degree to which one model appears to agree with observed seismicity, but that they leave much to be desired. They are not well-suited to the purpose of comparing the goodness of fit of competing models or to suggesting where models may be improved. It is worth noting that numerical tests such as the L-test, can be viewed as examples of scoring rules (see Gneiting and Raftery, 2007), and developing research on scoring rules may result in numerical tests of improved power and efficiency.

Future prediction experiments will allow for the implementation of more useful assessment tools. Residuals methods, including super-thinned, pixel and Voronoi residuals, seem ideal for comparison and to see where a particular model appears to overpredict or underpredict seismicity. Deviance residuals are useful for comparing two competing models and seeing where one appears to outperform another in terms of agreement with the observed seismicity. These methods are particularly useful in the CSEP paradigm, as insight gained during one prediction experiment can inform the building of models for subsequent experiments.

A note of caution should be made concerning the use of these model assessment tools. It is common to estimate the intensity function nonparametrically, for example using a kernel smoother. If the selection of the tuning parameter is done while simultaneously assessing the fit of the resulting models, this will likely lead to a model that is overfitted. A simple way to avoid this danger is to have a clear separation between the model fitting stage and the model assessment stage, as occurs when models are developed for prospective experiments.

Although the best fitting models for forecasting earthquake occurrences involve clustering and are thus highly non-Poissonian, it is unclear whether the Poisson assumption implicit in the *evaluation* of these models in CSEP or RELM has anything more than a negligible impact on the results. Since the quadrats used in these forecast evaluations are rather large, the dependence between the numbers of events occurring in adjacent pixels may be slight after accounting for inhomogeneity. Further, a departure from the Poisson distribution for the number of events occurring within a given cell would typically have similar impacts on competing forecast models and thus have little noticeable effect when it comes to evaluation of the relative performance of competing models. Nonetheless, further study is needed to clarify the importance of this assumption in the CSEP model evaluation framework. An alternative approach to the Poisson model would be to require that modelers provide not only the expected number of earthquakes within each bin, but also the joint probability distribution of counts within the bins.

Although this paper has focused on assessment tools for earthquake models, there is a wide range of point process models to which these methods can be applied. Super-thinned residuals and the K-function have been useful in assessing models of invasive species (Balderama et al., 2012). Other recent examples, such as the use of functional summaries in a study of infectious disease, can be found in Gelfand et al. (2010).

ACKNOWLEDGMENTS

We thank the Editor, Associate Editor and referees for very thoughtful remarks which substantially improved this paper.

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

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., MØLLER, J. and PAKES, A. G. (2008). Properties of residuals for spatial point processes. *Ann. Inst. Statist. Math.* **60** 627–649. MR2434415

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

BAKUN, W. H., AAGAARD, B., DOST, B., ELLSWORTH, W. L., HARDEBECK, J. L., HARRIS, R. A., JI, C., JOHNSTON, M. J. S., LANGBEIN, J., LIENKAEMPER, J. J., MICHAEL, A. J., MURRAY, J. R., NADEAU, R. M., REASENBERG, P. A., REICHLE, M. S., ROELOFFS, E. A., SHAKAL, A., SIMPSON, R. W. and WALDHAUSER, F. (2005). Implications for prediction and hazard assessment from the 2004 Parkfield earthquake. *Nature* **437** 969–974.

BALDERAMA, E., SCHOENBERG, F. P., MURRAY, E. and RUNDEL, P. W. (2012). Application of branching models in the study of invasive species. *J. Amer. Statist. Assoc.* **107** 467–476. MR2980058

BOLT, B. (2003). *Earthquakes*, 5th ed. Freeman, New York.

BRAY, A., WONG, K., BARR, C. and SCHOENBERG, F. P. (2014). Residuals for spatial point processes based on Voronoi tessellations. *Ann. Appl. Stat.* To appear.

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

CHU, A., SCHOENBERG, F. P., BIRD, P., JACKSON, D. D. and KAGAN, Y. Y. (2011). Comparison of ETAS parameter estimates across different global tectonic zones. *Bull. Seismol. Soc. Amer.* **101** 2323–2339.

CLEMENTS, R. A., SCHOENBERG, F. P. and SCHORLEMMER, D. (2011). Residual analysis methods for space–time point processes with applications to earthquake forecast models in California. *Ann. Appl. Stat.* **5** 2549–2571. MR2907126

CLEMENTS, R. A., SCHOENBERG, F. P. and VEEN, A. (2012). Evaluation of space-time point process models using superthinning. *Environmetrics* **23** 606–616. MR3020078

CONSOLE, R., MURRU, M. and FALCONE, G. (2010). Probability gains of an epidemic-type aftershock sequence model in retrospective forecasting of $M \geq 5$ earthquakes in Italy. *J. Seismology* **14** 9–26.

FIELD, E. H. (2007). Overview of the working group for the development of regional earthquake models (RELM). *Seismological Research Letters* **78** 7–16.

FIELD, E. H., DAWSON, T. E., FELZER, K. R., FRANKEL, A. D., GUPTA, V., JORDAN, T. H., PARSONS, T., PETERSEN, M. D., STEIN, R. S., WELDON, R. J. and WILLS, C. J. (2009). Uniform California Earthquake Rupture Forecast, Version 2 (UCERF 2). *Bull. Seismol. Soc. Amer.* **99** 2053–2107.

GELFAND, A., DIGGLE, P., GUTTORP, P. and FUENTES, M., eds. (2010). *Handbook of Spatial Statistics*. CRC Press, Boca Raton, FL. MR2761512

GELLER, R. J., JACKSON, D. D., KAGAN, Y. Y. and MULARGIA, F. (1997). Earthquakes cannot be predicted. *Science* **275** 1616–1617.

GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. MR2345548

HAUKSSON, E. and GODDARD, J. G. (1981). Radon earthquake precursor studies in Iceland. *J. Geophys. Res.* **86** 7037–7054.

HAWKES, A. G. (1971). Point spectra of some mutually exciting point processes. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **33** 438–443. MR0358976

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.

HELMSTETTER, A. and SORNETTE, D. (2003). Predictability in the Epidemic-Type Aftershock Sequence model of interacting triggered seismicity. *J. Geophys. Res.* **108** 2482–2499.

HOUGH, S. (2010). *Predicting the Unpredictable*: *The Tumultuous Science of Earthquake Prediction*. Princeton Univ. Press, Princeton, NJ.

JACKSON, D. D. (1996). Earthquake prediction evaluation standards applied to the VAN method. *Geophys. Res. Lett.* **23** 1363–1366.

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

JORDAN, T. H. and JONES, L. M. (2010). Operational earthquake forecasting: Some thoughts on why and how. *Seismological Research Letters* **81** 571–574.

KAGAN, Y. Y. (1997). Are earthquakes predictable? *Geophys. J. Int.* **131** 505–525.

KAGAN, Y. Y. (2009). Testing long-term earthquake forecasts: Likelihood methods and error diagrams. *Geophys. J. Int.* **177** 532–542.

KEILIS-BOROK, V. and KOSSOBOKOV, V. G. (1990). Premonitory activation of earthquake flow: Algorithm M8. *Physics of the Earth and Planetary Interiors* **6** 73–83.

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

MANGA, M. and WANG, C. Y. (2007). Earthquake hydrology. In *Treatise on Geophysics* (G. Schubert, ed.) **4** 293–320. Elsevier, Amsterdam.

MARZOCCHI, W. and ZECHAR, J. D. (2011). Earthquake forecasting and earthquake prediction: Different approaches for obtaining the best model. *Seismological Research Letters* **82** 442–448.

MERZBACH, E. and NUALART, D. (1986). A characterization of the spatial Poisson process and changing time. *Ann. Probab.* **14** 1380–1390. MR0866358

MEYER, P. A. (1971). Démonstration simplifiée d'un théorème de Knight. In *Séminaire de Probabilités*, *V* (*Univ. Strasbourg*, *Année Universitaire* 1969–1970) *Lecture Notes in Math.* **191** 191–195. Springer, Berlin. MR0380972

MOLCHAN, G. M. (1991). Structure of optimal strategies in earthquake prediction. *Tectonophysics* **193** 267–276.

MOLCHAN, G. (2010). Space-time earthquake prediction: The error diagrams. *Pure and Applied Geophysics* **167** 907–917.

NAIR, M. G. (1990). Random space change for multiparameter point processes. *Ann. Probab.* **18** 1222–1231. MR1062066

OGATA, Y. (1978). The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Ann. Inst. Statist. Math.* **30** 243–261. MR0514494

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. (1998). Space–time point process models for earthquake occurrences. *Ann. Inst. Statist. Math.* **50** 379–402.

OGATA, Y., JONES, L. M. and TODA, S. (2003). When and where the aftershock activity was depressed: Contrasting decay patterns of the proximate large earthquakes in southern California. *Journal of Geophysical Research* **108** 2318.

OKABE, A., BOOTS, B., SUGIHARA, K. and CHIU, S. N. (2000). *Spatial Tessellations*: *Concepts and Applications of Voronoi Diagrams*, 2nd ed. Wiley, Chichester. MR1770006

RHOADES, D. A., SCHORLEMMER, D., GERSTENBERGER, M. C., CHRISTOPHERSEN, A., ZECHAR, J. D. and IMOTO, M. (2011). Efficient testing of earthquake forecasting models. *Acta Geophysica* **59** 728–747.

RIPLEY, B. D. (1976). The second-order analysis of stationary point processes. *J. Appl. Probab.* **13** 255–266. MR0402918

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

SCHOENBERG, F. P. (2013). Facilitated estimation of ETAS. *Bulletin of the Seismological Society of America* **103** 1–7.

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.

SHEN, Z. K., 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.

SORNETTE, D. (2005). Apparent clustering and apparent background earthquakes biased by undetected seismicity. *J. Geophys. Res.* **110** B09303.

SWETS, J. A. (1973). The relative operating characteristic in psychology. *Science* **182** 990–1000.

TIAMPO, K. R. and SHCHERBAKOV, R. (2012). Seismicity-based earthquake forecasting techniques: Ten years of progress. *Tectonophysics* **522** 89–121.

VEEN, A. and SCHOENBERG, F. P. (2006). Assessing spatial point process models using weighted K-functions: Analysis of California earthquakes. In *Case Studies in Spatial Point Process Modeling* (A. Baddeley, P. Gregori, J. Mateu, R. Stoica and D. Stoyan, eds.). *Lecture Notes in Statist.* **185** 293–306. Springer, New York. MR2232135

VERE-JONES, D. and SCHOENBERG, F. P. (2004). Rescaling marked point processes. *Aust. N. Z. J. Stat.* **46** 133–143. MR2055791

VERE-JONES, D. and ZHUANG, J. (2008). On the distribution of the largest event in the critical ETAS model. *Phys. Rev. E* (3) **78** 047102.

WANG, Q., JACKSON, D. D. and KAGAN, Y. Y. (2011). California earthquake forecasts based on smoothed seismicity: Model choices. *Bull. Seismol. Soc. Amer.* **101** 1422–1430.

WERNER, M. J., HELMSTETTER, A., JACKSON, D. D. and KAGAN, Y. Y. (2011). High-Resolution Long-Term and Short-Term Earthquake Forecasts for California. *Bull. Seismol. Soc. Amer.* **101** 1630–1648.

ZALIAPIN, I. and MOLCHAN, G. (2004). Tossing the earth: How to reliably test earthquake prediction methods. *Eos. Trans. AGU* **85** 47. S23A–0302.

ZECHAR, J. D., GERSTENBERGER, M. C. and RHOADES, D. A. (2010). Likelihood-based tests for evaluating space–rate-magnitude earthquake forecasts. *Bull. Seismol. Soc. Amer.* **100** 1184–1195.

ZECHAR, J. D. and JORDAN, T. H. (2008). Testing alarm-based earthquake predictions. *Geophys. J. Int.* **172** 715–724.

ZECHAR, J. D., SCHORLEMMER, D., WERNER, M. J., GERSTENBERGER, M. C., RHOADES, D. A. and JORDAN, T. H. (2013). Regional earthquake likelihood models I: First-order results. Unpublished manuscript.

ZHUANG, J. (2011). Next-day earthquake forecasts for the Japan region generated by the ETAS model. *Earth Planets Space* **63** 207–216.