Pseudo-prospective Evaluation of UCERF3-ETAS Forecasts During the 2019 Ridgecrest Sequence

William H. Savran, Maximilian J. Werner, Warner Marzocchi, David A. Rhoades, David D. Jackson, Kevin Milner, Edward Field, Andrew Michael

William Savran, University of Southern California, Southern California Earthquake Center, 3651 Trousdale Parkway, Los Angeles, CA 90089

Abstract

The 2019 Ridgecrest sequence provides the first opportunity to evaluate Uniform California Earthquake Rupture Forecast Version 3 with Epidemic Type Aftershock Sequences (UCERF3-ETAS) in a pseudo-prospective sense. For comparison, we include a version of the model without explicit faults more closely mimicking traditional ETAS models (UCERF3-NoFaults). We evaluate the forecasts with new metrics developed within the Collaboratory for the Study of Earthquake Predictability (CSEP). The metrics consider synthetic catalogs simulated by the models rather than synoptic probability maps, thereby relaxing the Poisson assumption of previous CSEP tests. Our approach compares statistics from the synthetic catalogs directly against observations, providing a flexible approach that can account for dependencies and uncertainties encoded in the models. We find that, to first order, both UCERF3-ETAS and UCERF3-NoFaults approximately capture the spatiotemporal evolution of the Ridgecrest sequence, adding to the growing body of evidence that ETAS models can be informative forecasting tools. However, we also find that both models mildly overpredict the seismicity rate, on average, aggregated over the evaluation period. More severe testing indicates the overpredictions occur too often for observations to be statistically indistinguishable from the model. Magnitude tests indicate that the models do not include enough variability in forecasted magnitude-number distributions to match the data. Spatial tests highlight discrepancies between the forecasts and observations, but the greatest differences between the two models appear when aftershocks occur on modeled UCERF3-ETAS faults. Therefore, any predictability associated with embedding earthquake triggering on the (modeled) fault network may only crystalize during the presumably rare sequences with aftershocks on these faults. Accounting for uncertainty in the model parameters could improve test results during future experiments.

Introduction

A fundamental question in seismology is: What is the probability of observing an earthquake within some predefined space-time-magnitude region? Earthquake forecasting models try to answer this question by incorporating ideas of varying complexity about the earthquake process, including both empirical statistical relations, such as the Omori-Utsu and Gutenberg-Richter relations (Gutenberg and Richter, 1944; Utsu, 1961), and physical modeling, such as Coulomb stress calculations (Oppenheimer et al., 1988; King et al., 1994; Stein, 1999; Woessner et al., 2011; Cattania et al., 2018). The simplest models use locations of previous earthquakes to forecast locations of future earthquakes via smoothing (Kagan and Jackson, 1994; Frankel, 1995; Werner et al., 2010; Zechar and Jordan, 2010; Werner et al., 2011; Helmstetter and Werner, 2014). By contrast, UCERF3-ETAS (hereafter U3ETAS) combines long-term earthquake probabilities on faults based on elastic rebound statistics with short-term earthquake clustering as epidemic type aftershock sequences (Ogata, 1998) into a single model with fault-specific magnitude distributions (Field et al., 2017a; Field et al., 2017b). Most notably, U3ETAS provides probabilities of triggering ruptures on known faults, such as the Garlock and San Andreas faults. U3ETAS is a candidate model for Operational Earthquake Forecasting (OEF) issued by the US Geological Survey, motivating model evaluations also from a practical perspective.

The Collaboratory for the Study of Earthquake Predictability (CSEP) has established the philosophy and cyber-infrastructure required to conduct earthquake forecasting experiments in an unbiased and transparent fashion (Jordan, 2006; Schorlemmer and Gerstenberger, 2007; Jordan et al., 2011; Michael and Werner, 2018; Schorlemmer et al., 2018). Since its inception, CSEP has been using likelihood-based consistency tests (Schorlemmer et al., 2007; Zechar et al., 2010; Rhoades et al., 2011; Werner et al., 2011) that are rooted in the concepts that 1) earthquakes occur in space-time-magnitude bins independently, 2) earthquakes follow the Poisson distribution in each bin, and 3) modelers provide the 'true' parameter of the distribution in each bin. Thus, CSEP required that modelers provide forecasts giving the expected number of earthquakes in discrete space-time-magnitude bins. This pragmatic simplification allows multiple types of models, including those without explicit likelihood functions, to participate in the experiments.

However, Poisson likelihood-based evaluations of gridded forecasts can incorrectly report discrepancies between forecasts and observations when the true likelihood function of a forecast does not match a Poisson distribution or when strong dependencies exist between events within a forecast period. For example, the ETAS model is overdispersed with respect to a Poisson process, causing forecasts to be more frequently rejected than expected (Werner and Sornette 2008, Lombardi and Marzocchi 2010, Nandan et al., 2019). This is particularly noticeable when evaluating forecasts over multiple forecasting periods.

Evaluating gridded forecasts over multiple time periods exploits the property that the sum of N Poisson random variables each with parameter λ_i is a Poisson random variable with parameter $\sum \lambda_i$. The same convenience does not hold for catalog-based forecasts, because, in general, simulated events in catalogs from later time periods are not consistent with simulated events from earlier catalogs. Thus, catalog-based forecasting models should be evaluated for consistency by comparing realizations from their predictive distributions against observations. This approach is formally referred to as calibration, which is based on the idea that observations should be indistinguishable from realizations drawn from the predictive distributions of the model (Gneiting et al., 2006; Gneiting et al., 2007; Gneiting and Katzfuss, 2014). In other words, if the model were the data generator, we would expect observations to uniformly sample the forecasted distribution over independent trials.

Fundamentally, calibration is a different type of evaluation approach that can be potentially more severe than previously used cumulative evaluations. For example, evaluations over individual periods might indicate that observations consistently fall within the forecasted distribution, but instead of sampling the forecasted distribution uniformly they are concentrated towards one end. Thus, the model would fail calibration, but potentially pass a cumulative test. Understanding the overall performance of these models is more important than ‘rejecting’ a particular forecast; therefore, we focus on characteristics of the models and differences between models that potentially uncover new insights that might lead to model improvements.

Page and van der Elst (2018) introduced Turing-style evaluations for assessing forecasting models that produce synthetic catalogs. The tests evaluate important features of the simulated catalogs such as: aftershock productivity, seismicity rate, magnitude distribution, and clustering behavior. The Turing tests provide useful insights into the behavior of the forecasts, and can help to inform modeling decisions and identify discrepancies between the model and observations. However, they are not well suited for consistency testing or calibration, because they do not formally score forecasts against observations.

Here, we introduce new validation methods (consistency tests) for catalog-based earthquake forecasting models. Most notably, these methods relax the assumption that earthquakes follow independent Poisson distributions in discrete space-time-magnitude bins (Schorlemmer et al., 2007). Catalog-based forecasts differ from gridded forecasts in that they can capture the full aleatory variability of the model and can also account for epistemic uncertainty (such as in parameter estimates). Exhaustive sets of simulated catalogs retain the full spatiotemporal dependencies amongst modeled earthquakes, i.e., they can reflect the full complexity of the model through simulations. We build predictive distributions from the forecasts, empirically, by defining statistics that emphasize important characteristics of seismicity. This enables hypothesis testing and calibration of probabilistic forecasts over multiple evaluation periods.

We organize this manuscript as follows. First, we introduce the evaluation metrics for catalog-based forecasts. We then apply the metrics to forecasts made during the Ridgecrest sequence for an eleven-week period following the M_w 7.1 mainshock. To benchmark the fault-based triggering component of U3ETAS, we also generate and evaluate forecasts from a simpler version of the model, named UCERF3-NoFaults (hereafter NoFaults), which removes the fault component of U3ETAS. We discuss the primary differences between U3ETAS and NoFaults in the Methods section. Finally, we discuss the evaluation results with respect to U3ETAS and NoFaults and comment on the evaluation metrics.

Methods: Evaluations

Definitions and Notation

We introduce some notation to help us define evaluations in the context of earthquake forecasts that are specified as synthetic earthquake catalogs. First, we define a testing region $\boldsymbol{\mathcal{R}}$, as the combination of a magnitude range $\boldsymbol{\mathcal{M}}$, spatial domain $\boldsymbol{\mathcal{S}}$, and time period $\boldsymbol{\mathcal{T}}$:

$$\boldsymbol{\mathcal{R}} = \boldsymbol{\mathcal{M}} \times \boldsymbol{\mathcal{S}} \times \boldsymbol{\mathcal{T}}. \tag{1}$$

These individual components can be regarded as filters that operate on a catalog which retain only the events within $\boldsymbol{\mathcal{R}}$.

Let us consider an event, $e = (t, \boldsymbol{x}, m)$. Each e can be specified exactly by its origin time, t, geographic location, $\boldsymbol{x}$, and magnitude, m. The spatial coordinate, $\boldsymbol{x}$, typically refers to a latitude and longitude pair, but can also include depth. Thus, an earthquake catalog is simply a collection of events.

We define an observed catalog as

$$\Omega = \{e_i \mid i = 1, \dots, N_{obs}; e_i \in \boldsymbol{\mathcal{R}}\}. \tag{2}$$

Here, Ω is the observed catalog containing N_{obs} observed events, e_i, within $\boldsymbol{\mathcal{R}}$. This catalog is used as the testing data set for the evaluations.

A forecast is a collection of synthetic catalogs whose events $\tilde{e}_{ij}$ in $\boldsymbol{\mathcal{R}}$ are defined as

$$\boldsymbol{\Lambda} \equiv \Lambda_j = \left\{\tilde{e}_{ij} \mid i = 1, \dots, N_j; j = 1, \dots, J;\ \tilde{e}_{ij} \in \boldsymbol{\mathcal{R}}\right\}. \tag{3}$$

The forecast, $\boldsymbol{\Lambda}$, contains J synthetic catalogs each with N_j events. Λ_j indicates the j^{th} catalog of the forecast $\boldsymbol{\Lambda}$, likewise $\tilde{e}_{ij}$ denotes the i^{th} event from the j^{th} synthetic catalog of $\boldsymbol{\Lambda}$. Each Λ_j is a synthetic catalog that represents a continuous space-time-magnitude realization of seismicity generated by the model. The synthetic catalogs from the forecast and the observed catalog share the same event definitions, therefore the same statistics can be readily applied to all catalogs.

$N_{obs} = |\Omega|,$

The testing methodology presented here follows three guiding principles: (1) statistics should be calculated directly from the simulated and observed catalogs to build test distributions empirically; (2) testing methods should be able to preserve space-time-magnitude dependencies between events that are encoded in the model and may exist within the earthquake process; and (3) these tests should reduce their reliance on approximate likelihood functions of models, whether parametric in the case of the Poisson assumption or non-parametric in the case of the spatial test and pseudo-likelihood tests presented here. The last principle requires compromise if (approximate) likelihood-based inference remains desirable for model comparison, especially if no analytical likelihood function is available. Models without explicit likelihood functions are also known as generative or simulator-based models (Gutmann and Corander, 2016), which is the case for U3ETAS. In the remainder of this section, we define a suite of evaluations that can be used to evaluate the consistency of earthquake forecasts specified as synthetic catalogs against observed seismicity. These evaluations by no means represent an exhaustive set of metrics that can be used to evaluate catalog-based forecast models.

Number Test

The number test asks whether the number of earthquakes observed in $\boldsymbol{\mathcal{R}}$ is inconsistent with the forecasted number distribution by assessing whether the observed number falls into the tails of the forecast distribution (Kagan and Jackson, 1995; Schorlemmer et al., 2007; Zechar et al., 2010). The test statistic for an arbitrary catalog, ξ, is $N = |\xi|$, where the bars denote the count of events in the catalog. Thus, the observed statistic is

$$N_{obs} = |\Omega|, \tag{4}$$

or simply the number of events in the observed catalog. To build the test distribution from the forecast $\mathbf{\Lambda}$ we calculate this statistic for every catalog forming the vector:

$$N_j = |\Lambda_j|; j = 1, \dots, J. \tag{5}$$

To identify potentially important discrepancies between the observation and the forecast distribution, we compute the quantiles of the observed number in the empirical cumulative distribution function (CDF) of the forecast distribution (Equation 5) according to

$$\delta_1 = 1 - F_N(N_{obs} - 1) = P(N_j \geq N_{obs}) \tag{6}$$

and

$$\gamma_N = \delta_2 = F_N(N_{obs}) = P(N_j \leq N_{obs}). \tag{7}$$

$F_N(n)$ denotes the empirical cumulative distribution function of N_j. For the number test, we should consider a two-sided test to assess the probabilities of observing (1) at least and (2) at most N_{obs} events, a distinction that becomes important when forecasted and observed numbers are small (Zechar et al., 2010). $F_N(n)$ denotes the empirical predictive CDF of N_j. For a probabilistically calibrated forecast, we expect the quantile scores, γ_N, to uniformly sample the forecasted number distribution over multiple independent trials.

Magnitude Test

The magnitude test evaluates whether an observed magnitude-frequency distribution (MFD) is inconsistent with the forecasted MFD. We base this statistic on a square metric computed from the difference in logarithms between the incremental MFDs of the so-called union catalog Λ_{U}, individual catalogs Λ_{j}, and the observed catalog Ω. This metric is loosely related to the quadratic Cramer von-Mises and Anderson tests (Anderson, 2006). Using the logarithm of bin-wise magnitude counts places greater weight on magnitude bins with relatively fewer observed (and predicted) earthquakes, which typically occur at larger magnitudes. Thus, each missed (or over-predicted) event at larger magnitudes should contribute more to the test statistic than the same absolute error between smaller magnitudes.

We first define the union catalog Λ_U as

$$\Lambda_U = \{\Lambda_1 \cup \Lambda_2 \cup \ldots \cup \Lambda_J\}. \tag{8}$$

The union catalog Λ_U contains all events from $\mathbf{\Lambda}$ totaling $N_U = \sum_{j=1}^{J} |\Lambda_j|$ events. We compute the standard histograms of (1) $\Lambda_U^{(m)}$, the magnitudes of the union catalog, (2) $\Lambda_j^{(m)}$, the magnitudes of each individual synthetic catalog, and (3) $\Omega^{(m)}$, the observed magnitudes, with all histograms discretized according to $\boldsymbol{\mathcal{M}}$ (say, in increments of 0.1 magnitude units). We normalize all histograms so that $\sum_k \xi^{(m)}(k) = N_{obs}$, where $\xi^{(m)}(k)$ represents the normalized number of events in the k^{th} bin of the incremental MFD for an arbitrary catalog. This ensures that differences in forecasted rates do not contribute directly to the bin-wise sum, although the earthquake rate may implicitly affect the shape of the MFD. We compute the observed statistic as the sum of squared logarithmic residuals between the normalized observed magnitude and union histograms following

$$d_{obs} = \sum_{k} \left(\log \left[\frac{N_{obs}}{N_U} \Lambda_U^{(m)}(k) + 1 \right] - \log\left[\Omega^{(m)}(k) + 1\right] \right)^2 . \tag{9}$$

$\Lambda_U^{(m)}(k)$ and $\Omega^{(m)}(k)$ represent the count in the k^{th} bin of the incremental MFDs from the union and observed catalogs, respectively. We add unity to each bin to prevent the singularity associated with $\log(0)$. Since we are only concerned with differences between two MFDs, this modification does not bias the statistic. Next, we build the test distribution from the catalogs in **Λ,** i.e., the distribution of test statistics if the forecast model were the data-generating model following

$$D_j = \sum_{k} \left(\log \left[\frac{N_{obs}}{N_U} \Lambda_U^{(m)}(k) + 1 \right] - \log \left[\frac{N_{obs}}{N_j} \Lambda_j^{(m)}(k) + 1 \right] \right)^2 ;\ j = 1, \dots J. \tag{10}$$

Here, $\Lambda_j^{(m)}(k)$ indicates the count of events in the k^{th} magnitude bin from the j^{th} synthetic catalog. Finally, we compute the quantile score of d_{obs} within the empirical cumulative distribution function defined as

$$\gamma_m = F_D(d_{obs}) = P\left(D_j \leq d_{obs}\right). \tag{11}$$

We expect the quantile scores, γ_m, should uniformly sample the test distribution D_j for either forecast.

Pseudo-Likelihood Test

Here, we introduce a statistic based on the continuous point-process likelihood function (Daley and Vere-Jones, 2004). While this statistic resembles the likelihood scores used by previous CSEP experiments (e.g., Schorlemmer et al., 2007), there are two differences. First, we do not compute an actual likelihood, whence the name pseudo-likelihood. Second, this pseudo-likelihood statistic is aggregated over target event likelihood scores as opposed to the Poisson likelihood scores computed over discrete cells (see also Rhoades et al., 2011). In the case of zero or one events the pseudo-likelihood and the Poisson likelihood scores are identical. Finally, and most importantly, we build test distributions of pseudo-likelihood scores using the simulated (non-Poissonian) catalogs provided by the forecasting model, thereby producing distributions that better represent models that are over-dispersed and more clustered than a Poisson process.

A continuous marked space-time point process can be represented by its conditional intensify function $\lambda(\boldsymbol{e} \mid H_t)$, where H_t denotes the history of all earthquake occurrences (and any other relevant input data) prior to time t. The log likelihood function of any point-process over a region $\boldsymbol{\mathcal{R}}$ is

$$L = \sum_{i=1}^{N} \ln \lambda(e_i \mid H_t) - \int_{\boldsymbol{\mathcal{R}}} \lambda(\boldsymbol{e} \mid H_t) d\boldsymbol{\mathcal{R}}. \tag{12}$$

CSEP seeks to accommodate a wide range of stochastic models, including generative or simulator-based models such as UCERF3-ETAS without explicit conditional intensity or likelihood functions. CSEP therefore does not require an explicit likelihood function for evaluation (although models that contain explicit likelihood functions can be evaluated using this idea, e.g., Ogata et al., 2013).

Instead, we approximate the expectation of $\lambda(\boldsymbol{e} \mid H_t)$ using the forecasted catalogs. To do this we introduce a discretization of $\boldsymbol{\mathcal{R}}$ similar to previous CSEP experiments. Heuristically, the approximate rate density is defined as the conditional expectation, given the discretized region, $\boldsymbol{\mathcal{R}_d}$, of its continuous rate density:

$$\hat{\lambda}(\boldsymbol{e} \mid H_t) = E[\lambda(\boldsymbol{e} \mid H_t) \mid \boldsymbol{\mathcal{R}_d}]. \tag{13}$$

Conceptually, we can still regard the model as continuous in space, time and magnitude, but its rate density is only approximated and takes a constant value within a given cell. The approximate rate density is readily derived from the standard CSEP forecast of gridded expected rates, by computing the mean event count from the forecast, $\boldsymbol{\Lambda}$, in each cell in $\boldsymbol{\mathcal{R}_d}$. The discrete grid cells are used only for approximation purposes; we use the synthetic catalogs from the full model to calculate the pseudo-likelihood statistic (rather than catalogs of the approximate model).

From the approximate rate density (Equation 13), we can define the pseudo log likelihood $\hat{L}$ by

$$\hat{L} = \sum_{i=1}^{N} \ln \hat{\lambda}(e_i \mid H_t) - \int_{\boldsymbol{\mathcal{R}}} \hat{\lambda}(\boldsymbol{e} \mid H_t) d\boldsymbol{\mathcal{R}}. \tag{14}$$

The pseudo-likelihood test applied here considers a discretized region in space to avoid introducing artifacts into the forecasts (such as minimum “water-levels” and smoothing operators that could bias the evaluations) to account for under-sampling in space-magnitude bins. Formally, we can write the spatial approximate rate density as

$$\hat{\lambda}_s(\boldsymbol{e} \mid H_t) = \sum_{\mathcal{M}} \hat{\lambda}(\boldsymbol{e} \mid H_t). \tag{15}$$

If $\hat{\lambda}_s(k)$ denotes the approximate rate density in the k^{th} spatial cell of the model, we can compute the observed pseudo-likelihood score using,

$$\hat{L}_{obs} = \sum_{i=1}^{N_{obs}} \ln \hat{\lambda}_s\,(k_i) - \bar{N}. \tag{16}$$

Here k_i denotes the spatial cell in which the i^{th} event occurs and $\bar{N}$ denotes the expected number of events in $\boldsymbol{\mathcal{R}_d}$. Following Equation 16, we compute the statistics for the test distribution as

$$\hat{L}_j = \left[\sum_{i=1}^{N_j} \ln \hat{\lambda}_s\big(k_{ij}\big) - \bar{N}\right]; j = 1, \ldots, J. \tag{17}$$

Here $\hat{\lambda}_s(k_{ij})$ denotes the approximate rate density of the i^{th} event of the j^{th} catalog from the forecast. We combine Equation 16 and Equation 17 to obtain the quantile score

$$\gamma_L = F_L(\hat{L}_{obs}) = P(\hat{L}_j \leq \hat{L}_{obs}). \tag{18}$$

The statistic captures simultaneously the spatial component and the rate component of the forecast. Thus, potential discrepancies in both rate and the spatial components of the forecasts should be reflected in this statistic. As with the magnitude test and the number test, we expect that the quantile scores γ_L should be uniformly distributed over multiple evaluation periods.

Spatial Test – Geometric Average of Target Event Rate Distribution

The spatial test isolates the spatial distribution of the forecast to evaluate whether the observed locations are consistent with the forecasted spatial distribution. This statistic utilizes the approximate rate density (Equation 15) with normalization $\hat{\lambda}_s^* = \hat{\lambda}_s / \sum_{\boldsymbol{\mathcal{R}}} \hat{\lambda}_s$ to isolate the spatial component of the forecast.

We define the observed spatial statistic according to

$$s_{obs} = \left[\sum_{i=1}^{N_{obs}} \ln \hat{\lambda}_s^*(k_i) \right] N_{obs}^{-1}, \tag{19}$$

where $\hat{\lambda}_s^*(k_i)$ denotes the normalized approximate rate density in the k^{th} cell corresponding to the i^{th} event in $\mathbf{\Omega}$. Likewise, we can define the test distribution for the statistic defined in Equation (19) using

$$S_j = \left[\sum_{i=1}^{N_j} \ln \hat{\lambda}_s^*(k_{ij}) \right] N_j^{-1} \,; j = 1, \dots, J. \tag{20}$$

As above, $\hat{\lambda}_s^*(k_{ij})$ denotes the approximate rate density in the k^{th} cell corresponding to the i^{th} event in the j^{th} simulated catalog. The observed spatial statistic (Equation 19) is scored by computing quantiles in the test distribution (Equation 20) using,

$$\gamma_S = F_S(\hat{s}_{obs}) = P(\hat{S}_j \le \hat{s}_{obs}). \tag{21}$$

We interpret this statistic as being the geometric mean of the target event rate distribution. Normalizing $\hat{\lambda}_s$ and computing the geometric mean of the target event rate distribution ensures that two catalogs (from the same forecast) with events occurring in identical bins will result in equivalent spatial test statistics irrespective of the number of events in either catalog. If the model were the data generator, we expect that γ_S will be uniformly distributed over multiple evaluation periods.

Testing Over Multiple Periods

To assess models over many periods, we exploit the following idea: quantile scores over multiple periods should be uniformly distributed if the model is the data generator (Gneiting and Katzfuss, 2014). Departures from a uniform distribution of the quantile scores flag discrepancies between the forecasting model and observation. Formally, we employ a Kolmogorov-Smirnov test between the quantile scores and the uniform distribution to test the hypothesis that the observed quantile scores are uniformly distributed. We calculate the p-value of this test and use a significance level $\alpha = 0.05$ to identify discrepancies.

Graphically, we consider different patterns of variation of the observed quantile scores from a uniform distribution. A model that under-predicts the test statistic produces a graph similar to that in Figure 1a. In this case, there is a small proportion of low quantile scores and a high proportion of high quantiles compared to the uniform distribution, because the observed test-statistic tends to be higher than the simulated test-statistic. Conversely, a model that tends to over-predict the test statistic produces a graph similar to Figure 1b, because in that case the actual test statistic tends to be lower than the simulated test statistics. If the model test statistics are under-dispersed relative to the observed test statistics, then the quantile scores will fall near the end-points 0 and 1 of the distribution. This produces the pattern seen in Figure 1c. Conversely, if the model test statistics are over-dispersed relative to the actual test statistic, the pattern seen in Figure 1d will be the result.

Methods: Pseudo-Prospective Experiment Design

The 2019 Ridgecrest sequence provides the first opportunity to evaluate operational aftershock forecasts in a pseudo-prospective sense. A pseudo-prospective experiment preserves the time-dependent causality of the data set by partitioning the dataset into a training set and a testing set (Schorlemmer et al., 2018), which happened naturally as these forecasts were computed in near real-time during the Ridgecrest sequence. Most of the forecasts produced in this study were computed in near-real-time using real-time data products with the exceptions listed in Table 1. The forecasts presented in this study use the *ShakeMap (v.14)* source model and default parameters described in Milner et al. (2020). We evaluate the forecasts starting at $t = 0$ and $t = 7$ days following the M_w 7.1 mainshock (along with the nine others) in this study.

Data

For this experiment, we use authoritative data from the Advanced National Seismic System (ANSS) provided by the United States Geological Survey (USGS) Comprehensive Catalog (ComCat). The evaluation data were accessed from ComCat on 11 November 2019, approximately 60 days following the date of the final forecast. We use the data directly provided by ComCat, and do not attempt to standardize magnitude types or manually relocate events. We apply the time-dependent magnitude of completeness model from Helmstetter et al. (2006) to account for missing events following the mainshock, modeled by a time-dependent magnitude of completeness $M_c(t)$. Therefore, the evaluation catalog has a threshold magnitude

$$M_t(t) = \max\big(M_{min}, M_c(t)\big). \tag{22}$$

Here, M_{min} , represents a minimum magnitude that is either defined to be $M_{min} = 2.5$ or $M_{min} = 3.5$, in the case of the number test. We apply the time-dependent magnitude of completeness model to both forecasted and observed catalogs. The inset in Figure 2a shows the events used for this study along with the time-dependent magnitude of completeness. In the 77 days following the M_w 7.1 Ridgecrest event, the catalog lists 1,362 events with $M \geq 2.5$ in the study region.

Finite-fault representations for trigger ruptures are based on surface field mapping and geodetic observations, and were provided by ShakeMap (Wald et al., 1999). These finite-fault models were made available on 11 July 2019 within six days after the M_w 7.1 mainshock. Milner et al. (2020) explains the various finite-fault representations available and the sensitivity of the forecasts to these.

Earthquake Forecasting Models

We consider two forecasting models, (1) UCERF3-ETAS (U3ETAS) and (2) UCERF3-NoFaults (NoFaults). The former model is explained in detail by Field et al. (2017b), so we summarize the important differences between U3ETAS and NoFaults here. Field et al. (2017a) provides a less technical overview of the UCERF3-ETAS model for the interested reader. The full mathematical description of these models can be found in the above manuscripts and their appendices.

U3ETAS is unique as compared with standard ETAS models, because the model includes finite faults that can host so-called supraseismogenic earthquakes. In U3ETAS, a supraseismogenic earthquake is defined as an earthquake with a rupture length at least as long as the seismogenic fault width. When a large earthquake in close enough proximity to a U3ETAS fault is sampled by ETAS, that earthquake is mapped onto the modeled fault-sections. Subsequently, the rates of all events that utilize the ruptured sections are modified according to Reid renewal statistics (Reid, 1910; Field et al., 2015). Therefore, U3ETAS provides stochastic event sets with ruptures on modeled finite faults in addition to 'off-fault' ruptures elsewhere, following a traditional ETAS model. U3ETAS makes no model-wide assumptions about magnitude-frequency distributions on faults, with most exhibiting non Gutenberg-Richter (GR) behavior depending on the relative rate of microseismicity versus inferred fault-based ruptures. On average the faults are slightly characteristic (elevated rates at higher magnitudes), which means off-fault areas are slightly anti-characteristic so that combined a GR *b*-value of 1.0 is maintained. However, the model assigns the regional faults surrounding the Ridgecrest sequence an anti-characteristic behavior (Field et al., 2017b; Milner et al., 2020) implying lower probabilities of triggering supraseismogenic aftershocks than under a pure GR model. In contrast, NoFaults applies the state-wide G-R relationship (*b*-value=1.0) throughout the entire model.

As the name suggests, NoFaults does not include information about modeled faults and behaves similar to a traditional space-time ETAS implementation (Ogata and Zhuang, 2006). Both U3ETAS and NoFaults explicitly model the depth distribution of seismicity. The computational requirements for the two models differ by approximately an order of magnitude, with U3ETAS being more expensive. Because model simplicity and computational efficiency are two desirable characteristics of robust operational forecasting tools (Jordan and Jones, 2010; Jordan et al., 2011), we seek to understand the relative predictive skills and usefulness of the models.

The forecasts issued by both U3ETAS and NoFaults consist of a family of 100,000 synthetic catalogs constrained to the bounding-box of the CSEP California testing region (Schorlemmer and Gerstenberger, 2007). As inputs to all forecasts, we include earthquakes with M2.5+ for seven days prior to the M_w 7.1 mainshock until the start-time of each forecast, including the M_w 6.4 Searles Valley event. We use identical input catalogs for both U3ETAS and NoFaults to maintain direct comparability between the two forecasts. Also, we do not include spontaneous (background) events in the conditioning data for the forecasts. Therefore, any discrepancies between the forecast and observations can be attributed to the implementation of the short-term components of the model and not the background seismicity model.

Spatial Region, Magnitude Bins, and Forecast Horizons

For this experiment, we choose magnitude bins

$$\boldsymbol{\mathcal{M}} = \{[2.5, 2.6), [2.7, 2.8), \dots, [8.4, 8.5), [8.5, \infty)\}. \tag{23}$$

The bins are uniformly spaced at $\Delta M = 0.1$ except for the right-most bin which extends to infinity. We remove events outside a spatial zone of three Wells and Coppersmith (1994) fault radii from the M7.1 epicenter (143 km) to isolate the Ridgecrest aftershocks from other seismicity. Each forecast horizon extends for seven non-overlapping days, which we treat as independent time intervals. Figure 2b shows the spatial extent of the circular region surrounding the hypocenter of the M_w 7.1 mainshock and the observed M2.5+ events during this time period.

These definitions completely define the extent of $\boldsymbol{\mathcal{R}}$ for our experiment. All forecasts are evaluated for seven days following the forecast start time to preserve effects of short-term clustering in the observed catalog. Table 1 contains the exact start and end times for all the forecasts considered in this study, which consist of eleven non-overlapping time periods following the M_w 7.1 mainshock. All but two U3ETAS forecasts were computed prospectively using real-time catalogs and data. The NoFaults simulations were run pseudo-prospectively, but using the same input catalogs and input finite-fault models as U3ETAS.

Results

Before we share the results of the quantitative evaluations of the forecasts, we show how differences between U3ETAS and NoFaults manifest in individual synthetic catalogs. Since the models are similar for events smaller than ~M_w 6.5, catalogs display similar characteristics for 'typical' realizations (Figure 3a,c), defined here as catalogs representing the median of the forecasted number distribution. The differences become obvious when viewing catalogs (Figure 3b,d) that sample the tails of the number distribution at the 99.9th percentile. We call these catalogs 'extreme' as they forecast rare, but possible, large aftershock sequences on potentially multiple faults. Extreme U3ETAS scenarios involve ruptures triggered on the Garlock fault and subsequently on the San Andreas fault. Their respective aftershocks are largely constrained within the fault zones. On the other hand, NoFaults assigns aftershock locations isotropically in space resulting in (nearly) isotropic catalogs that contain clusters of earthquakes (Figure 3d).

The differences illustrated in Figure 3, namely in the catalogs at the tails of the forecast, complicate robust model comparisons using typical California aftershock sequences, which only occasionally involve triggering of large aftershocks on (mapped) faults. This is because the models produce very similar (visually nearly indistinguishable) catalogs near the modes and medians of the number distributions. Sequences such as the 1992 Landers earthquake cascade and others that are thought to have triggered other large ruptures could help distinguish between these two models (Kisslinger and Jones, 1991; Hauksson et al., 1993; Freed and Lin, 2001).

We show test results as quantile scores for all evaluations in Table 2. The overall (aggregate) scores over all forecast periods are reported as p-values of Kolmogorov-Smirnov tests between a uniform distribution and the quantile scores of each test computed at the updating periods shown in Table 1.

Forecasted Seismicity Rates

Figure 4 shows the forecasted number distributions as a function of time during the aftershock sequence for both $M_t(t) = \max(2.5, M_c(t))$ and $M_t(t) = \max(3.5, M_c(t))$.

We observe the largest variability in the number distribution immediately following the mainshock, which decreases rapidly throughout the evaluation period. During the first evaluation period the median forecasted numbers are 925 and 956 for U3ETAS and NoFaults, respectively, with 829 observed events during this period. The median forecasted event counts are identical between the two models for the remaining forecasting periods.

We compute number test results for each forecast by reporting quantile scores for individual testing periods as a function of evaluation day (Figure 5a). Except for the first day, both forecasts produce nearly identical quantile scores. The difference in number distributions during the first forecasting period can potentially be explained by the anti-characteristic behavior of the U3ETAS faults surrounding the aftershock sequence. This behavior results in U3ETAS producing fewer large (M6.5+) events, along with their numerous aftershocks, and subsequently fewer catalogs with large numbers of aftershocks. During the first forecasting period, the 95-percentile range of the number distribution are (751, 2482) and (756, 3906) for U3ETAS and NoFaults respectively.

Figure 5b shows the number test quantile scores compared against standard uniform quantiles as a quantile-quantile plot. We assign the standard uniform quantiles following $U^{(k)} = k/(n+1)$, for $k = 1, \dots, n$, to space the quantiles equally along the distribution. We compute confidence intervals for the k^{th} order statistic of the standard uniform distribution using $U^{(k)} \sim \mathrm{B}(k, n+1-k)$ where n is the number of observations (Jones, 2004).

The distribution of quantile scores indicates the forecasts overpredict the observed seismicity (Figure 1b), as the observed numbers of earthquakes too frequently fall into the lower tails of the forecasted distributions. At both magnitude cutoffs, the Kolmogorov-Smirnov tests reject the hypothesis that the distribution of quantile scores from the number test are uniformly distributed. This suggests that, given this limited forecasting period, the observations are not indistinguishable from realizations from the forecast number distribution.

Magnitude Number Distribution

Figure 6a shows incremental MFDs aggregated over the eleven-week evaluation period. For the union MFD, $\Lambda_U^{(m)}$, and observed MFDs, $\Omega_U^{(m)}$, we sum bin-wise counts from each evaluation period to obtain aggregate counts. We estimate percentiles using an aggregate forecasted MFD (thin lines in Figure 6). We generate the aggregate forecasted MFD using a bootstrapped approach where we randomly sample one MFD per forecast per time-period and sum bin-wise counts over each evaluation period. This produces 100,000 aggregate MFDs approximating an MFD representative of the eleven-week evaluation period. Except between M3.0 and M4.0 the observations generally fall within the variability of the forecasted MFD. Above M6.5 we see differences in the tails of the magnitude frequency distributions that further show how the anticharacteristic MFDs assumed by U3ETAS manifest in the forecasts.

Figure 6b shows the bin-wise value of the magnitude test statistic over the full evaluation period to highlight the bin-wise contribution to the overall magnitude test statistic. From the bin-wise statistic, we can obtain the magnitude test statistic in Equation 9 by summing over all magnitude bins. This figure illustrates that discrepancies at larger magnitudes contribute more (per event) to the value magnitude test statistic, but this must be reconciled with statistics computed from simulated catalogs. We can identify bins whose values contribute most to the discrepancy between observations and forecasts by assessing the observed statistic with respect to the bin-wise distribution of magnitude test statistics.

The percentiles in Figure 6b (for both U3ETAS and NoFaults) are estimated from the bin-wise distributions of magnitude test statistics based on the bootstrapped aggregate MFD (explained above). We can associate the large peak observed near M4.7 in Figure 6b with catalogs from either model that contain zero events in that magnitude bin. This can be seen by comparing the square bin-wise difference with the union MFD and zero observed events in Figure 6a. The percentiles in Figure 6b indicate 2.5% of the catalogs contain no events at this magnitude, and 16% of the catalogs contain no events at M5.0. The largest discrepancies with respect to the forecasts occur from around M3.0 through M4.0 indicated by the observed bin-wise values falling outside the 95th percentile range of the bin-wise distribution. Generally, the observed values are frequently greater than the median from their respective bin-wise distributions, and this behavior is not confined to a particular magnitude range.

Figure 7a shows quantile scores for each evaluation period following the M_w 7.1 mainshock to assess the performance of the forecast over multiple updating periods. The shaded region in Figure 7a indicates the critical region assuming a 0.05 significance level for a right-tailed statistical test. (In this magnitude test, larger-than-expected values of the statistic, i.e. large quantile scores, indicate larger discrepancies). Figure 7b shows the quantile-quantile plot of the magnitude test scores against standard uniform quantiles. The quantile scores, γ_m, do not sample the test distribution uniformly and are instead concentrated near the upper end. The Kolmogorov-Smirnov test thus rejects the hypothesis of a uniform distribution of the quantile scores. The pattern in Figure 7b implies persistently greater-than-expected differences between the observed magnitude distribution and the forecast. The pattern of magnitude test quantile scores reflects the finding in Figure 6b that the bin-wise magnitude scores are typically greater than the median bin-wise values.

Spatial Distribution of Seismicity and Pseudo-likelihood Test

Figure 8a,b shows the approximate spatial rate density (Equation 15) for both U3ETAS and NoFaults during the first evaluation period following the M_w 7.1 mainshock. The expected cell-wise event counts clearly show differences between U3ETAS and NoFaults, specifically the increased expected rates along modeled faults in U3ETAS. The relatively high rates along the Garlock fault, for example, are dominated by catalogs containing supraseismogenic ruptures along these faults (which occur in about 7% of the catalogs). Thus, we should expect to see noticeable differences between these two models with observations of such aftershock sequences.

Figure 8c shows test distributions of spatial statistics for a single week-long forecast immediately following the M_w 7.1 mainshock. Likewise, Figure 8d shows test distributions for the pseudo-likelihood score. Positive values of the pseudo-likelihood scores can occur when multiple target events occur within the same spatial bin with $\hat{\lambda}_s \gg 1$ (the Poisson likelihood contains an explicit term to account this discretization artifact that does not appear in the pseudo-likelihood statistic), which can happen when scoring catalogs that sample upper tails of the number distribution. For this evaluation period, the observed statistic, $\hat{L}_{obs}$, lies in the lower tail of the test distribution $\hat{L}$.

The aggregate spatial test result in Figure 9a shows quantile scores and pseudo-likelihood quantiles for each evaluation period since the M_w 7.1 mainshock. In general, U3ETAS tends to have larger quantile scores, and thus, more favorable test statistics for a given forecast than NoFaults. We find that if differences are observed, they appear in both the pseudo-likelihood and spatial test statistics. Comparisons of quantile scores against the uniform distribution (Figure 9b) show the statistic from the observed catalog tends to fall in the lower tail of the spatial test distribution for most forecasts. Thus, according to the spatial test, random draws from the forecasted distribution are distinguishable from the observations; the latter more frequently occur in cells of lower rates than expected by the models.

The pseudo-likelihood quantiles γ_L shows seemingly better agreement with the standard uniform quantiles (we compute p=0.0280, p=0.0235 from the Kolmogorov-Smirnov test for U3ETAS and NoFaults, respectively); however, this observation must be analyzed in the context of both the number test and spatial test results. Since both models show inconsistencies in the number test and spatial test, we expect this to be reflected in the pseudo-likelihood test. Previous studies have shown that the Poisson-based likelihood test is anticorrelated with the number test results (Werner et al., 2011). The somewhat counterintuitive result causes forecasts that overpredict the seismicity rates to trivially pass the likelihood test. Therefore, this must be considered when interpreting the pseudo-likelihood test results. Specifically, the test results are probably better solely because the models overpredict.

Deconstructing the statistics helps to inform us about the behavior of the evaluation results. For the magnitude test, we showed the bin-wise value of the test statistics to identify problematic bins. Here, we show cell-wise spatial pseudo-likelihood ratios (U3ETAS – NoFaults) in Figure 10 to understand which cells contribute to the differences observed in the spatial test and the pseudo-likelihood tests. We represent the observed event rate distribution on the spatial grid as follows: spatial cells with no observed events show the difference in the approximate rate density between models, and cells containing observed events show the difference in the that cells' contribution to the pseudo-likelihood scores. Only cells containing observed events contribute to the spatial test statistic, therefore cells without observed events help to visualize differences in the spatial distributions of the forecast. These plots are similar to spatial deviance residuals (Schneider et al., 2014). We find that U3ETAS tends to show larger spatial test statistics, and thus quantile scores, when observed events occur along modeled U3ETAS faults.

Discussion

We have introduced a suite of evaluations for catalog-based earthquake forecasts that provide insight into the forecasted earthquake rates, magnitude-number distributions, and spatial distributions of seismicity. These evaluations are complementary to the Turing Tests introduced by Page and van der Elst (2018) and the comparative mean-information gain introduced by Nandan et al. (2019), which can also be used to evaluate generative or simulator models that produce synthetic catalogs. Importantly, these metrics begin to relax the independence and Poisson assumptions of previous forecast evaluations (Schorlemmer et al., 2007). Additionally, we introduced an approach, commonly applied to weather (and other) probabilistic forecasts (e.g., Gneiting et al., 2006), to calibrate probabilistic earthquake forecasting models. We apply these new methods to U3ETAS and NoFaults forecasts of the Ridgecrest sequence for eleven-weeks following the Mw 7.1 mainshock.

We find U3ETAS and NoFaults overpredict earthquake rates in 10 out of 11 evaluation periods for $M_t(t) = \max\big(2.5, M_c(t)\big)$ by comparing observed event counts against the mode of the forecasted number distribution (modal ratio), but 5 out of 11 modal ratios are within ±20% of the observed event count (with the maximum being a 140% overprediction). On average, from the modal ratio, the forecasts overpredict observed event counts by approximately 50%. NoFaults tends to produce larger variability in the number distribution than U3ETAS (e.g., Figure 4a,b), which is most noticeable during the first evaluation period. This likely occurs because the Airport Lake and Little Lake faults are both anti-characteristic in U3ETAS (Milner et al., 2020), which causes these faults to host fewer large magnitude events as compared with the GR (b=1.0) MFD implemented in NoFaults (Figure 6). Moreover, every event in NoFaults is treated as a point-source. In contrast, U3ETAS can assign large ruptures to faults (if the event occurs close enough to a modeled fault). This in turn activates the elastic-rebound model (Field et al., 2015), and this combined behavior effectively smooths the forecasted number of events in the vicinity of the rupture (Figure 8a,b). The anticharacteristic behavior of the Little Lake and Airport Lake faults is likely to have pronounced differences in the tails of the number distributions and the associated hazard and risk curves. In areas with anticharacteristic MFDs, U3ETAS produces lower expected rates of events except along the faults that host aftershock sequences. Visually, we see the larger rates along the faults for U3ETAS as compared with NoFaults (Figure 8a,b), but statistically the chance of damaging aftershocks is lower in U3ETAS.

On aggregate, the U3ETAS and NoFaults produce catalogs whose MFDs display lower variability with respect to the expected MFD than observations. By comparing the logarithms of bin-wise counts we find that observations are different, statistically, from realizations from the forecasts. Figure 6b shows contributions to this discrepancy across all magnitude ranges, but M3.0 through M4.0 show the largest discrepancy with respect to the forecasted bin-wise statistics. This can be interpreted in two ways: either significant discrepancies exist between U3ETAS (and NoFaults) and observations, or this magnitude test is too severe given the uncertainties in reported magnitudes and assumed b-values in the forecasting model. To address uncertainties in reported magnitudes, we recomputed the magnitude test with magnitude bins $\Delta M = 0.2$, and found consistent results with those presented in Figure 7. Moreover, using Monte Carlo simulations we find the magnitude test results are sensitive to changes in b-values of $\Delta b \leq 0.1$ units, which is on the order of the uncertainty in b-value estimates for U3ETAS (Felzer, 2013). Thus, including epistemic uncertainty in the assumed *b*-value could potentially improve calibration. Furthermore, we should consider explicitly accounting for uncertainties in observed magnitudes when evaluating earthquake forecasts.

Here, we discuss a potential reason for the inconsistencies in the spatial test results. ETAS models, due to their self-excitation property (Hawkes, 1971), have a particularly difficult time forecasting seismicity in areas that were not previously active. As a result, the approximate rate densities (Equation 13) and locations of events in the simulated catalogs are controlled by the events in the input catalog used to condition the forecast. For example, neither U3ETAS nor NoFaults forecast much seismicity off the northwest-end of the mainshock fault plane during the first forecasting period (Figure 8a,b), leading to the observations falling in the lower tail of the test distribution. This discrepancy can be reduced with more frequent updating of the ETAS intensity function, which would locally increase after each subsequent event. Ideally, the conditional intensity function would be updated continuously after each observed event; however, this might prove difficult in practice because of computational times and costs.

The spatial and pseudo-likelihood tests show the largest differences between U3ETAS and NoFaults amongst the statistics, which we expected because the spatial distribution of seismicity is the primary difference between these models. Figure 10 shows spatial (pseudo-) log-likelihood ratios (U3ETAS – NoFaults) to understand where differences in the spatial test statistic originate. Carefully looking at the cell-wise ratios where observed events occur in Figure 10, we find the differences manifest when aftershocks occur near modeled U3ETAS faults. This suggests that we should be able to identify differences between U3ETAS and NoFaults using the spatial test for sequences when aftershocks occur on modeled U3ETAS faults.

We draw counter-intuitive conclusions from the pseudo-likelihood test, when put in context of the spatial and number tests. We find that observations are inconsistent with both the rate and spatial forecasts from both models, and thus we expect the pseudo-likelihood scores to reflect this observation. Instead, the pseudo-likelihood test scores show more favorable agreement with the observations. Similar to the Poisson likelihood test (Schorlemmer et al., 2007), overpredictions in rates can result in artificially high pseudo-likelihood scores (e.g., Werner et al., 2011). From this, we conclude that the pseudo-likelihood test provides redundant information to the number and spatial tests, and the test is less severe than the spatial test when the forecast fails the number test.

U3ETAS uses ETAS parameters estimated from the state-wide California seismic catalog (Hardebeck, 2013). The moderate overprediction by U3ETAS (and NoFaults) suggests that the Ridgecrest sequence deviates from the state-wide average in aftershock productivity. Milner et al. (2020) found this behavior was due to high primary productivity of the mainshock, coupled with low secondary aftershock productivity. State-wide maximum-likelihood estimates (MLE) of ETAS parameters also result in over-predictions for this sequence when using traditional ETAS models (Mancini et al., 2020, In Press). These results suggest that accurate forecasting of aftershock rates requires proper treatment of intersequence variability or obtaining sequence specific parameters (Page et al., 2016).

MLE parameter estimates of a traditional ETAS model may well be different, however, from MLE estimates of U3ETAS parameters, because the models are different: non-GR behavior in U3ETAS is spatially variable, magnitude and spatial distributions are not separable, and ‘characteristic-ness’ impacts secondary triggering productivity (Milner et al., 2020). Milner et al. (2020) showed that adjustment of the ETAS *c*-value could improve the fit to the cumulative number of $M \geq 3.5$ events, but this required manual trial-and-error adjustments to optimize for a specific metric. If sequence specific parameters are not (yet) available, incorporating additional uncertainty in the ETAS parameters could make the model more general and perhaps calibrated, especially for the first forecasts following a large earthquake before sequence-specific information is available (Omi et al., 2015; Omi et al., 2019).

The discrepancies between the models and observations can potentially be explained by epistemic uncertainty in model parameters not accounted for by the model. Incorporating parameter uncertainty would broaden distribution functions (reduce sharpness) and potentially lead to calibrated probabilistic forecasts. Moreover, incorporating more sequences (and quiet periods) could uncover systematic discrepancies with observations that can lead to improvements in the models, and increase the robustness of the results. Retrospective as well as further prospective tests are required to understand the usefulness and accuracy of modeling decisions. In particular, the U3ETAS model will be most easily differentiated from standard ETAS models in the rare circumstances (of about 7%, assuming U3ETAS is correct) when supraseismogenic events are triggered. This relatively small percentage (which varies spatially in the model) implies that we expect to observe substantial differences between the models about once in 20 earthquake sequences. Future work should therefore evaluate the model retrospectively against all well-recorded aftershock sequences observed in California.

Conclusions

In this manuscript, we evaluate forecasts from UCERF3-ETAS and UCERF3-NoFaults during the Ridgecrest using new non-parametric evaluations developed for forecasts specified as simulated catalogs. We evaluate eleven week-long forecasts immediately following the Mw 7.1 mainshock using an idea, known as calibration, that suggests that random draws from the forecast should be indistinguishable from observations. Probabilistic calibration is severe, but is a useful approach to aggregate forecasts over multiple periods. Probabilistic forecasts should aim to maximize the sharpness of their predictive distributions, subject to calibration (Gneiting et al., 2007; Gneiting and Katzfuss, 2014). We introduce statistics that probe the forecasted earthquake rate, magnitude distributions, and spatial component of the forecast. Importantly, these evaluations relax the assumption that earthquakes occur in discrete Poissonian space-time-magnitude bins and better reflect the dependencies between earthquakes.

This pseudo-prospective evaluation of U3ETAS (and NoFaults) constitutes a milestone as it represents the first out-of-sample evaluation of a model under consideration for real-time operational earthquake forecasting by the US Geological Survey. (Pseudo-) Prospective model evaluation is a critical step in building confidence in the model outputs. To first order, both U3ETAS and NoFaults capture the temporal evolution and magnitude distributions of the earthquake sequence, notwithstanding the generic state-wide ETAS model parameters. For example, when considering the mode of the forecasted number distribution, the forecasts on average overpredict the observed number of events by approximately 50% with 5 out of 11 forecasts being within ±20% of the observed event count. This suggests that, in spite of the much more severe calibration test results, U3ETAS (and ETAS models in general) are effective tools to provide insight into the spatial and temporal distributions of seismicity, in real-time, during an aftershock sequence. As with any forecasting model, the usefulness depends on the specific use-case in mind (Field and Milner, 2018). For U3ETAS, in particular, estimates of probabilities of ruptures on nearby faults may provide valuable information for emergency planners and decision makers (Milner et al., 2020).

The results of the proposed tests lead to similar conclusions for both UCERF3-ETAS and NoFaults. For the number test, the forecasts systematically overpredict the observed seismicity. The overpredictions can be attributed to deviations in primary and secondary aftershock productivity during the Ridgecrest with respect to the state wide average. The observed MFDs show greater variability with respect to the expected MFD than predicted by the forecasts. We interpret this discrepancy as a result of unmodeled uncertainty in the magnitude data, highlighting a need to account for observational uncertainty in the tests. The spatial tests uncover an issue associated with the discrete updating of self-exciting ETAS models, that is, the models have difficulty forecasting seismicity in areas without previous seismicity. We find the largest differences between U3ETAS and NoFaults when observed aftershocks occur on modeled U3ETAS faults. In such cases, the pseudo-likelihood test provides redundant results to the number and spatial test.

Data and Resources

The evaluation results and data for individual simulations can be found at https://github.com/cseptesting/ridgecrest_evaluation_bssa. The UCERF3-ETAS and UCERF3-NoFaults simulations were generated using the UCERF3 model implemented in OpenSHA and can be found at https://github.com/opensha/ucerf3-etas-launcher/. The code used for the analysis can be found in development at https://github.com/SCECcode/csep2/. The finite-fault data was obtained from the ShakeMap accessed through the Comprehensive Catalog (ComCat) provided by the United States Geological Survey and can be access through the web at https://earthquake.usgs.gov/data/comcat/.

Acknowledgements

We would like to thank Morgan Page, Jeanne Hardebeck, and Nicholas van der Elst for their insight and helpful comments regarding forecast evaluations. M.J.W. and W.M. received funding from the European Union's Horizon 2020 research and innovation program (No 821115, RISE: Real-Time Earthquake Risk Reduction for a Resilient Europe). This research was supported by the Southern California Earthquake Center (Contribution No. 10082). SCEC is funded by NSF Cooperative Agreement EAR-1600087 & USGS Cooperative Agreement G17AC00047.

References

Anderson, T. W. (2006). On The Distribution of the Two-Sample Cramer von-Mises Criterion, Annals of Mathematical Statistics 1-12.

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

Daley, D. J., and D. Vere-Jones (2004). Scoring probability forecasts for point processes: the entropy score and information gain, Journal of Applied Probability **41** 297-312.

Felzer, K. R. (2013). Appendix L: Estimate of the Seismicity Rate and Magnitude-Frequency Distribution of Earthquakes in California from 1850 to 2011, 1-13.

Field, E. H., G. P. Biasi, P. Bird, T. E. Dawson, K. R. Felzer, D. D. Jackson, K. M. Johnson, T. H. Jordan, C. Madden, A. J. Michael, K. R. Milner, M. T. Page, T. Parsons, P. M. Powers, B. E. Shaw, W. R. Thatcher, R. J. Weldon II, and Y. Zeng (2015). Long-Term Time-Dependent Probabilities for the Third Uniform California Earthquake Rupture Forecast (UCERF3), Bulletin of the Seismological Society of America **105** 511-543.

Field, E. H., T. H. Jordan, M. T. Page, K. R. Milner, B. E. Shaw, T. E. Dawson, G. P. Biasi, T. Parsons, J. L. Hardebeck, A. J. Michael, R. J. Weldon, P. M. Powers, K. M. Johnson, Y. H. Zeng, K. R. Felzer, N. van der Elst, C. Madden, R. Arrowsmith, M. J. Werner, and W. R. Thatcher (2017a). A Synoptic View of the Third Uniform California Earthquake Rupture Forecast (UCERF3), Seismological Research Letters **88** 1259-1267.

Field, E. H., and K. R. Milner (2018). Candidate Products for Operational Earthquake Forecasting Illustrated Using the HayWired Planning Scenario, Including One Very Quick (and Not-So-Dirty) Hazard-Map Option, Seismological Research Letters **89** 1420-1434.

Field, E. H., K. R. Milner, J. L. Hardebeck, M. T. Page, N. J. van der Elst, T. H. Jordan, A. J. Michael, B. E. Shaw, and M. J. Werner (2017b). A Spatiotemporal Clustering Model for the Third Uniform California Earthquake Rupture Forecast (UCERF3-ETAS): Toward an Operational Earthquake Forecast, Bulletin of the Seismological Society of America **107** 1049-1081.

Frankel, A. (1995). Mapping Seismic Hazard in the Central and Eastern United States, **66** 8-21.

Freed, A. M., and J. Lin (2001). Delayed triggering of the 1999 Hector Mine earthquake by viscoelastic stress transfer, Nature **411** 180-183.

Gneiting, T., F. Balabdaoui, and A. E. Raftery (2007). Probabilistic forecasts, calibration and sharpness, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **69** 243-268.

Gneiting, T., and M. Katzfuss (2014). Probabilistic forecasting, Annual Review of Statistics and Its Application **1** 125-151.

Gneiting, T., K. Larson, K. Westrick, M. G. Genton, and E. Aldrich (2006). Calibrated Probabilistic Forecasting at the Stateline Wind Energy Center, **101** 968-979.

Gutenberg, B., and C. F. Richter (1944). Frequency of earthquakes in California, Bulletin of the Seismological society of America **34** 185-188.

Gutmann, M. U., and J. Corander (2016). Bayesian optimization for likelihood-free inference of simulator-based statistical models, The Journal of Machine Learning Research **17** 4256-4302.

Hardebeck, J. L. (2013). Appendix S: Constraining Epidemic Type Aftershock Sequence (ETAS) Parameters from the Uniform California Earthquake Rupture Forecast, Version 3

Catalog and Validating the ETAS Model for Magnitude 6.5 or Greater Earthquakes, U.S. Geol. Surv. Open-File Rept.

Hauksson, E., L. M. Jones, K. Hutton, and D. Eberhart-Phillips (1993). The 1992 Landers Earthquake Sequence: Seismological observations, **98** 19835.

Hawkes, A. G. (1971). Point spectra of some mutually exciting point processes, Journal of the Royal Statistical Society: Series B (Methodological) **33** 438-443.

Helmstetter, A., Y. Y. Kagan, and D. D. Jackson (2006). Comparison of short-term and time-independent earthquake forecast models for southern California, Bulletin of the Seismological Society of America **96** 90-106.

Helmstetter, A., and M. J. Werner (2014). Adaptive Smoothing of Seismicity in Time, Space, and Magnitude for Time-Dependent Earthquake Forecasts for California, Bulletin of the Seismological Society of America **104** 809-822.

Jones, M. (2004). Families of distributions arising from distributions of order statistics, Test **13** 1-43.

Jordan, T. H. (2006). Earthquake predictability, brick by brick, pubs.geoscienceworld.org **77** 3-6.

Jordan, T. H., Y. T. Chen, P. Gasparini, R. Madariaga, I. Main, W. Marzocchi, and G. Papadopoulos (2011). Operational earthquake forecasting. State of knowledge and guidelines for utilization, Annals of Geophysics.

Jordan, T. H., and L. M. Jones (2010). Operational Earthquake Forecasting: Some Thoughts on Why and How, Seismological Research Letters **81** 571-574.

Kagan, Y. Y., and D. D. Jackson (1994). Long-Term Probabilistic Forecasting of Earthquakes, Journal of Geophysical Research-Solid Earth **99** 13685-13700.

King, G. C., R. S. Stein, and J. Lin (1994). Static stress changes and the triggering of earthquakes, Bulletin of the Seismological Society of America **84** 935-953.

Kisslinger, C., and L. M. Jones (1991). Properties of aftershock sequences in southern California, Journal of Geophysical Research **96** 11947.

Mancini, S., M. J. Werner, M. Segou, and T. Parsons (2020, In Press). The Predictive Skills of Elastic Coulomb Rate-and-state Aftershock Forecasts During the 2019 Ridgecrest, California, Earthquake Sequence, Bulletin of the Seismological Society of America.

Michael, A. J., and M. J. Werner (2018). Preface to the Focus Section on the Collaboratory for the Study of Earthquake Predictability (CSEP): New Results and Future Directions, Seismological Research Letters **89** 1226-1228.

Milner, K. R., E. H. Field, W. H. Savran, M. T. Page, and T. H. Jordan (2020). Operational Earthquake Forecasting during the 2019 Ridgecrest, California, Earthquake Sequence with the UCERF3-ETAS Model, Seismological Research Letters.

Nandan, S., G. Ouillon, D. Sornette, and S. Wiemer (2019). Forecasting the Full Distribution of Earthquake Numbers Is Fair, Robust, and Better, Seismological Research Letters.

Ogata, Y. (1998). Space-time point-process models for earthquake occurrences, Ann I Stat Math **50** 379-402.

Ogata, Y., K. Katsura, G. Falcone, K. Nanjo, and J. Zhuang (2013). Comprehensive and Topical Evaluations of Earthquake Forecasts in Terms of Number, Time, Space, and Magnitude, Bulletin of the Seismological Society of America **103** 1692-1708.

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

Omi, T., Y. Ogata, Y. Hirata, and K. Aihara (2015). Intermediate-term forecasting of aftershocks from an early aftershock sequence: Bayesian and ensemble forecasting approaches, Journal of Geophysical Research: Solid Earth **120** 2561-2578.

Omi, T., Y. Ogata, K. Shiomi, B. Enescu, K. Sawazaki, and K. Aihara (2019). Implementation of a Real-Time System for Automatic Aftershock Forecasting in Japan, Seismological Research Letters **90** 242-250.

Oppenheimer, D. H., P. A. Reasenberg, and R. W. Simpson (1988). Fault plane solutions for the 1984 Morgan Hill, California, earthquake sequence: Evidence for the state of stress on the Calaveras fault, Journal of Geophysical Research: Solid Earth **93** 9007-9026.

Page, M. T., and N. J. van der Elst (2018). Turing-Style Tests for UCERF3 Synthetic Catalogs, Bulletin of the Seismological Society of America **108** 729-741.

Page, M. T., N. J. van der Elst, J. Hardebeck, K. Felzer, and A. J. Michael (2016). Three Ingredients for Improved Global Aftershock Forecasts: Tectonic Region, Time-Dependent Catalog Incompleteness, and Intersequence Variability, Bulletin of the Seismological Society of America **106** 2290-2301.

Reid, H. F. (1910). The California earthquake of April 18, 1906: Report of the State Earthquake Investigation Commission. 2. The mechanics of the earthquake, Carnegie Inst. of Washington.

Rhoades, D. A., D. Schorlemmer, M. C. Gerstenberger, A. Christophersen, J. D. Zechar, and M. Imoto (2011). Efficient testing of earthquake forecasting models, Acta Geophys **59** 728-747.

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

Schorlemmer, D., M. Gerstenberger, S. Wiemer, D. D. Jackson, and D. A. Rhoades (2007). Earthquake likelihood model testing, Seismological Research Letters **78**.

Schorlemmer, D., and M. C. Gerstenberger (2007). RELM Testing Center, Seismological Research Letters **78** 30-36.

Schorlemmer, D., M. J. Werner, W. Marzocchi, T. H. Jordan, Y. Ogata, D. D. Jackson, S. Mak, D. A. Rhoades, M. C. Gerstenberger, N. Hirata, M. Liukis, P. J. Maechling, A. Strader, M. Taroni, S. Wiemer, J. D. Zechar, and J. C. Zhuang (2018). The Collaboratory for the Study of Earthquake Predictability: Achievements and Priorities, Seismological Research Letters **89** 1305-1313.

Stein, R. S. (1999). The role of stress transfer in earthquake occurrence, Nature **402** 605-609.

Utsu, T. (1961). A statistical study on the occurrence of aftershocks, Geophys. Mag. **30** 521-605.

Wald, D. J., V. Quitoriano, T. H. Heaton, H. Kanamori, C. W. Scrivner, and C. B. Worden (1999). TriNet “ShakeMaps”: Rapid generation of peak ground motion and intensity maps for earthquakes in southern California, Earthquake Spectra **15** 537-555.

Wells, D. L., and K. J. Coppersmith (1994). New empirical relationships among magnitude, rupture length, rupture width, rupture area, and surface displacement, Bulletin of the Seismological Society of America **84** 974-1002.

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

Werner, M. J., A. Helmstetter, D. D. Jackson, Y. Y. Kagan, and S. Wiemer (2010). Adaptively smoothed seismicity earthquake forecasts for Italy, Annals of Geophysics **53** 107-116.

Woessner, J., S. Hainzl, W. Marzocchi, M. J. Werner, A. M. Lombardi, F. Catalli, B. Enescu, M. Cocco, M. C. Gerstenberger, and S. Wiemer (2011). A retrospective comparative forecast test on the 1992 Landers sequence, Journal of Geophysical Research-Solid Earth **116**.

Zechar, J. D., M. C. Gerstenberger, and D. A. Rhoades (2010). Likelihood-Based Tests for Evaluating Space-Rate-Magnitude Earthquake Forecasts, Bulletin of the Seismological Society of America **100** 1184-1195.

Zechar, J. D., and T. H. Jordan (2010). Simple smoothed seismicity earthquake forecasts for Italy, Annals of Geophysics 1-7.

Tables

Table 1. Start times for the forecasts considered in this study. UCERF3-NoFaults were computed pseudo-prospectively using the same input catalogs as their UCERF3-ETAS counterparts. UCERF3-ETAS forecasts were computed in near-real-time with real-time data products except as otherwise noted.

Mw 7.1 + Δ*T* (days)	**Start Time (GMT+0)**	**End Time (GMT+0)**
0.0[*]	2019-07-06 *03:19:54.04*	2019-07-13 *03:19:54.04*
7.0[†]	2019-07-13 *03:19:54.04*	2019-07-20 *03:19:54.04*
14.0[**]	2019-07-20 *03:19:54.04*	2019-07-27 *03:19:54.04*
21.0	2019-07-27 *03:19:54.04*	2019-08-03 *03:19:54.04*
28.0	2019-08-03 *03:19:54.04*	2019-08-10 *03:19:54.04*
35.0	2019-08-10 *03:19:54.04*	2019-08-17 *03:19:54.04*
42.0	2019-08-17 *03:19:54.04*	2019-08-24 *03:19:54.04*
49.0	2019-08-24 *03:19:54.04*	2019-08-31 *03:19:54.04*
56.0	2019-08-31 *03:19:54.04*	2019-09-07 *03:19:54.04*
63.0	2019-09-07 *03:19:54.04*	2019-09-14 *03:19:54.04*
70.0	2019-09-14 *03:19:54.04*	2019-09-21 *03:19:54.04*

[*] Calculated on 09/04/19, catalog input data accessed from ComCat 09/04/19

[†] Calculated on 07/16/19, catalog input data accessed from ComCat 07/16/19

[**] Calculated on 08/19/19, catalog input data accessed from ComCat 08/19/19

Table 2. Evaluation results for number, magnitude, pseudo-likelihood, and spatial tests results for UCERF3-ETAS and UCERF3-NoFaults for $M_t(t) = \max\left(2.5, M_c(t)\right)$.

Test day (since M7.1)	**U3ETAS (N-Test)***	**NoFaults (N-Test)***	**U3ETAS (M-Test)**	**NoFaults (M-Test)**	**U3ETAS (PL-Test)**	**NoFaults (PL-Test)**	**U3ETAS (S-Test)**	**NoFaults (S-Test)**
7	[0.818, 0.185]	[0.843, 0.160]	0.912	0.944	0.073	0.094	0.044	0.035
14	[0.688, 0.326]	[0.692, 0.322]	0.819	0.822	0.035	0.032	0.043	0.04
21	[0.995, 0.006]	[0.996, 0.006]	0.129	0.136	0.109	0.083	0.192	0.137
28	[0.958, 0.052]	[0.958, 0.051]	0.725	0.731	0.018	0.017	0.065	0.065
35	[0.999, 0.002]	[0.999, 0.002]	0.57	0.575	0.031	0.036	0.296	0.298
42	[0.907, 0.114]	[0.908, 0.113]	0.825	0.827	0.018	0.012	0.116	0.078
49	[0.399, 0.636]	[0.398, 0.636]	0.782	0.781	0.325	0.186	0.307	0.209
56	[0.998, 0.004]	[0.998, 0.004]	0.904	0.905	0.012	0.013	0.266	0.276
63	[0.999, 0.002]	[0.999, 0.002]	0.908	0.905	0.187	0.095	0.921	0.874
70	[0.995, 0.008]	[0.995, 0.008]	0.905	0.904	0.052	0.024	0.732	0.609
77	[1.000, 0.000]	[1.000, 0.001]	0.967	0.967	0.134	0.138	0.975	0.976
Overall	**8.450e-05**	**3.363e-05**	**1.425e-03**	**1.222e-03**	**2.796e-02**	**2.349e-02**	**2.432e-06**	**1.927e-08**

*(δ^1, δ^2)

Figure Captions

Figure 1. Schematic of cumulative distribution of quantile scores for a test statistic calculated over multiple test periods (points) as compared with the ideal uniform distribution (dashed line) expected for a well-calibrated model. Panels show instances of (a) under-prediction, and (b) over-prediction of the statistic by the model; (c) under-dispersion, and (d) over-dispersion of statistic in the model simulations.

Figure 2. (a) Ridgecrest sequence data beginning one week preceding the Mw 6.4 foreshock through the eleven-week evaluation period. Vertical gray dashed lines indicate the starting times of the forecasts. Brown data denote target (test) earthquakes. The forecasts are conditioned on all events until the start time of the forecast. The inset shows the Helmstetter et al. (2006) magnitude-completeness model for the first three days following the Mw 7.1 mainshock. (b) Distribution of spatial seismicity from ComCat during the period shown in (a). The circle shows the spatial region used for the evaluations based on an average Mw 7.1 fault length from Wells and Coppersmith (1994) with a radius of approximately 143 km.

Figure 3. Synthetic catalog realizations showing 7 days of aftershocks following the M_w 7.1 mainshock. (a) ‘Typical’ U3ETAS synthetic catalog, defined as the catalog whose event count lies along the median amongst all simulated catalogs. (b) ‘Extreme’ U3ETAS synthetic catalog, which is defined as the catalog whose event count falls in the uppermost 0.1 percentile of the forecasted number distribution. Notice the triggered ruptures on the Garlock and San Andreas faults that in turn generate aftershocks along these faults. (c) ‘Typical’ synthetic catalog generated by NoFaults and (d) an ‘extreme’ catalog from NoFaults, which lacks triggering of

ruptures on prescribed faults resulting in a nearly isotropic aftershock distribution. The 'extreme' catalogs highlight the predominant differences between these two models and suggest that differences will be most noticeable when large aftershocks occur on mapped faults in U3ETAS.

Figure 4. Forecasted number distributions and observed cumulative number over the eleven-week evaluation period. The forecasted event count distributions are offset by the number of observed events at the start of the forecast. Forecasted number distributions are plotted at the end of each evaluation period. The vertical extent of the lines indicates the 95-percentile range of the forecasted number distribution. The 'x' indicates evaluation periods with observed event counts that fall outside the 95-percentile range of the forecast. (a) Both observed and forecasted catalogs are filtered to threshold magnitudes $\mathrm{M_t}\ (\mathrm{t}) = \max(2.5, M_c(t))$ and (b) catalogs are filtered to $\mathrm{M_t(t)} = \max(3.5, M_c(t))$. During the first seven-day forecast period, the 95th percentile of the forecasted number distribution for M2.5+ events are 2,482 and 3,906 events for U3ETAS and NoFaults, respectively.

Figure 5. Aggregate number test results for $M_t(t) = \max(2.5, M_c(t))$ and $M_t(t) = \max(3.5, M_c(t))$ magnitude thresholds for U3ETAS and NoFaults for eleven weekly evaluation intervals following the $\mathrm{M_w}$ 7.1 mainshock. (a) Quantile scores δ_1 (top) and δ_2 (bottom) for individual weekly evaluation periods. (b) Quantile-quantile plot showing calibration of rate forecasts by comparing quantile scores, γ_N against standard uniform quantiles. The dashed lines indicate 95 percent confidence intervals around the standard uniform quantiles. Thus, U3ETAS and NoFaults overpredict the number of M2.5+ and M3.5+ events during this aftershock sequence.

Figure 6. (a) Magnitude frequency distribution in $\Delta M = 0.1$ bins aggregated over entire the eleven-week evaluation period. The thin lines approximate the 95% percentile range of the event counts in each magnitude bin. The U3ETAS magnitude frequency distribution shows anti-characteristic behavior through the lack of M6.5+ earthquakes as compared with NoFaults. (b) Bin-wise magnitude test statistic aggregated over the entire evaluation period. The circles depict the kernel of d_{obs} for both U3ETAS and NoFaults to show bin-wise contributions to d_{obs}. We find negligible differences between the two models. The solid lines show percentiles from the bin-wise value distribution, for both models.

Figure 7. Magnitude test results for events with $M_t(t) = (2.5, M_c(t))$ over the full eleven-week evaluation period. (a) Quantile scores are shown for individual week-long evaluation periods. Gray patch depicts the 0.05 significance level for the magnitude test. The largest differences between U3ETAS and NoFaults exist during the first week and become negligible over the remainder of the evaluation period. (b) Calibration of magnitude forecasts by comparing magnitude test quantile scores against standard uniform quantiles. The dashed lines depict 95 percent confidence intervals around the standard uniform quantiles.

Figure 8. Logarithm of the expected event counts per spatial bin per week for U3ETAS (a) and NoFaults (b) for the week-long forecast following the M_w 7.1. The relatively high expected counts along the faults in U3ETAS are controlled by scenarios whose aftershock sequences contain supraseismogenic ruptures along these faults. In both plots, target events during this period are shown as white circles. The color scale is manually saturated for comparison

purposes. The spatial bin with highest rate expects 64.24 and 65.76 events for U3ETAS and NoFaults, respectively. (c) Evaluation result for the spatial test for U3ETAS (top) and NoFaults (bottom) for the first evaluation period at seven days after the M_w 7.1 mainshock. $\hat{S}^{(95)}$ denotes the 95th percentile range of the test distribution of the spatial test statistic, $\hat{s}_{obs}$ is the observed statistic, and γ_S is the quantile score. (d) Same as (c) except for the pseudo-likelihood test statistics.

Figure 9. Spatial test and pseudo-likelihood results for events with $M_t(t) = \max(2.5, M_c(t))$ over the complete eleven-week evaluation period. The spatial test and likelihood tests show the greatest differences between U3ETAS and NoFaults. (a) Quantile scores shown for individual week-long evaluation periods. The patch depicts the 0.05 significance level for the spatial test. (b) Calibration of spatial forecasts by comparing quantile scores against standard uniform quantiles. The dashed lines depict 95 percent confidence intervals around the standard uniform quantiles.

Figure 10. Map of cell-wise spatial pseudo log-likelihood ratios between U3ETAS and NoFaults for individual evaluation periods ending on (a) day 35, (b) day 49, (c) day 56, and (d) day 63 following the Mw 7.1 mainshock. Maps show the higher rates along faults in U3ETAS. Evaluation periods at (b) 49 days and (d) 63 days show the largest differences in the observed spatial statistic, which is calculated only from spatial cells where events occur, while periods ending on days 35 and 56 show a negligible difference in the spatial statistic. This highlights how spatial test results are sensitive to events occurring on modeled U3ETAS faults and that such events are required to discern between the models. The color

scale is manually saturated between -0.05 and 0.05 to help comparisons; and dots show locations of target events

Figures

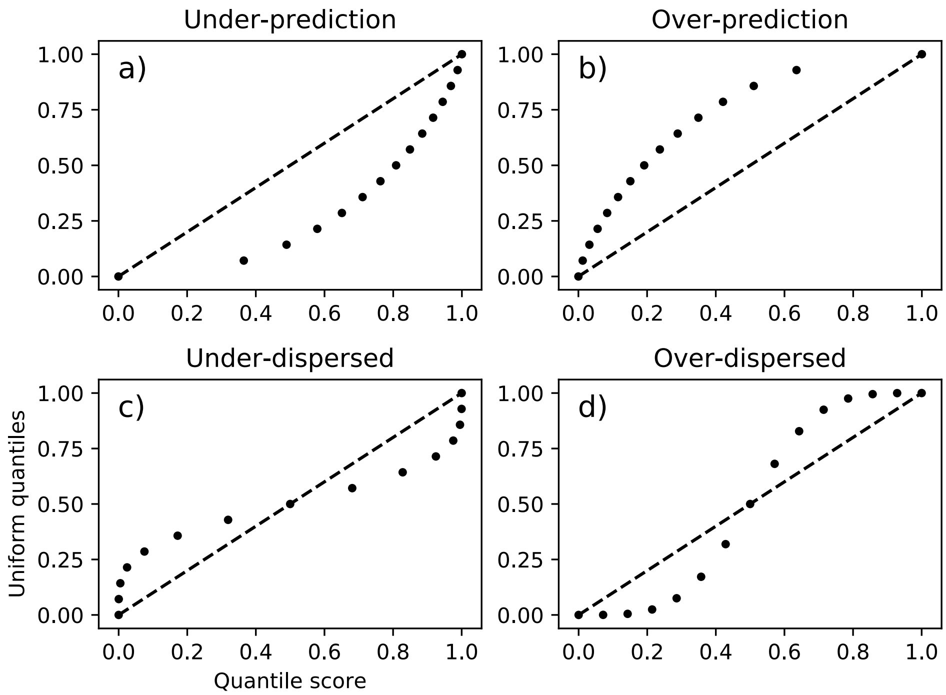

Figure 1. Schematic of cumulative distribution of quantile scores for a test statistic calculated over multiple test periods (points) as compared with the ideal uniform distribution (dashed line) expected for a well-calibrated model. Panels show instances of (a) under-prediction, and (b) over-prediction of the statistic by the model; (c) under-dispersion, and (d) over-dispersion of statistic in the model simulations.

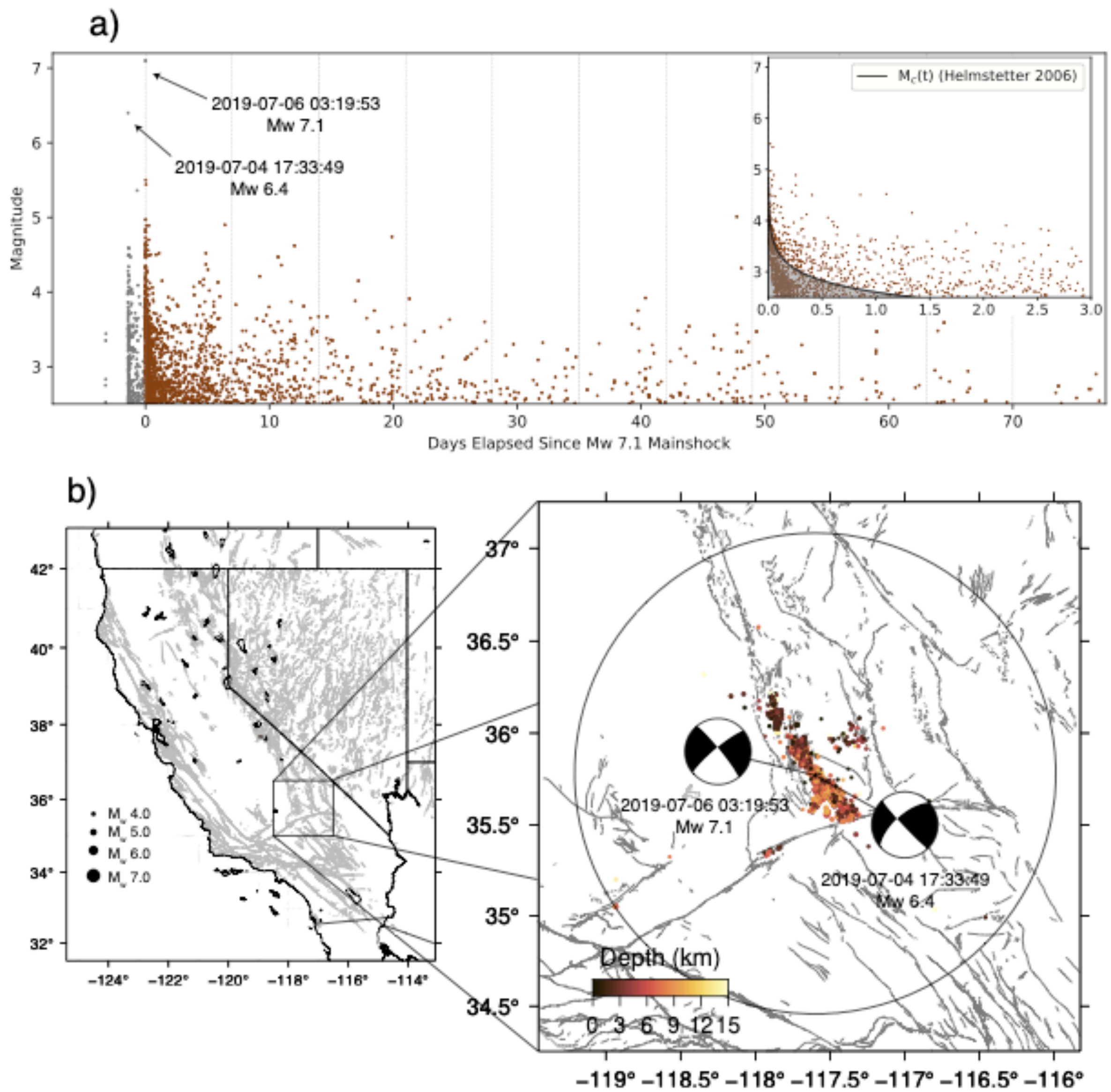

Figure 2. (a) Ridgecrest sequence data beginning one week preceding the M_w 6.4 foreshock through the eleven-week evaluation period. Vertical gray dashed lines indicate the starting times of the forecasts. Brown data denote target (test) earthquakes. The forecasts are conditioned on all events until the start time of the forecast. The inset shows the Helmstetter et al. (2006) magnitude-completeness model for the first three days following the M_w 7.1 mainshock. (b) Distribution of spatial seismicity from ComCat during the period shown in (a). The circle shows the spatial region used for the evaluations based on an average M_w 7.1 fault length from Wells and Coppersmith (1994) with a radius of approximately 143 km.

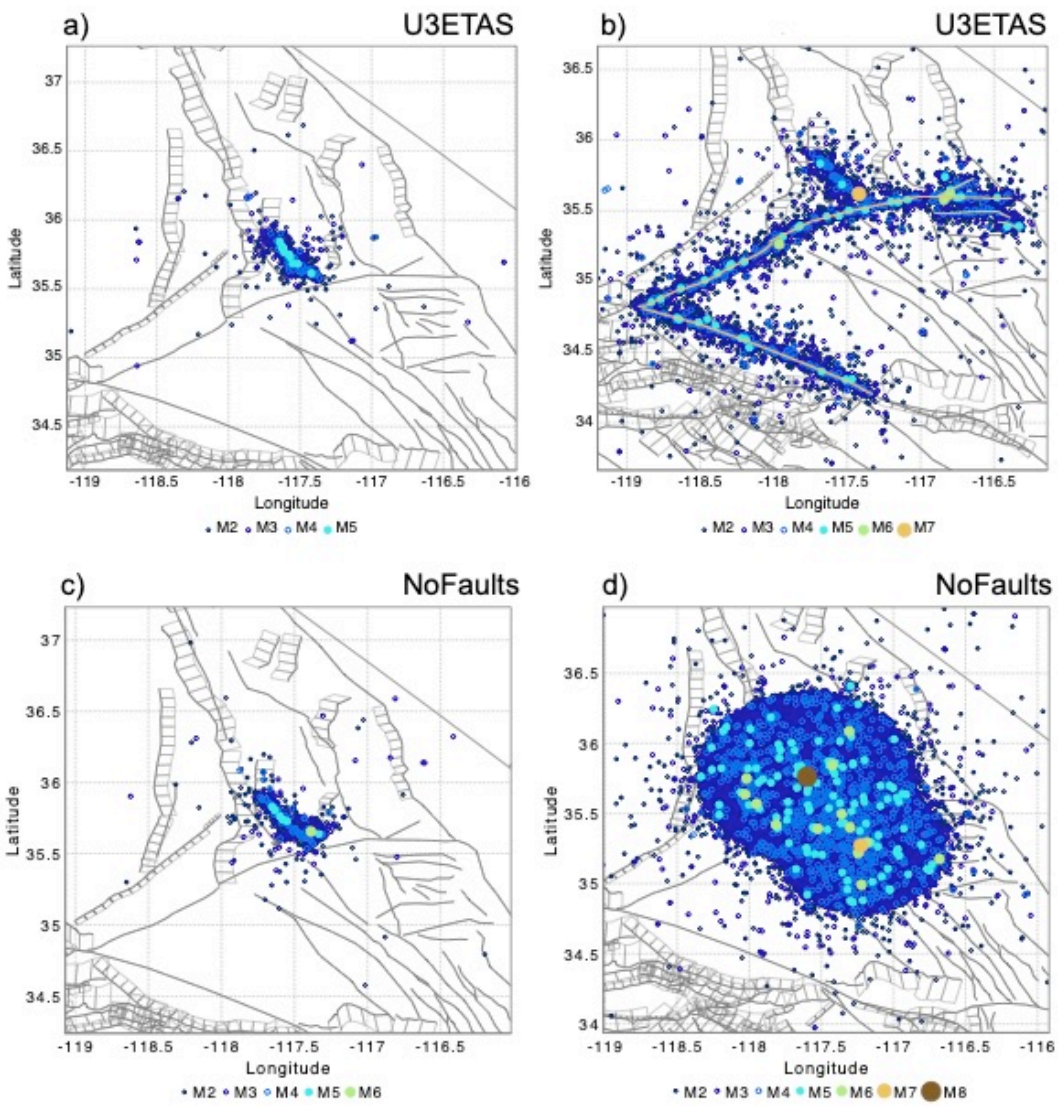

Figure 3. Synthetic catalog realizations showing 7 days of aftershocks following the M_w 7.1 mainshock. (a) ‘Typical’ U3ETAS synthetic catalog, defined as the catalog whose event count lies along the median amongst all simulated catalogs. (b) ‘Extreme’ U3ETAS synthetic catalog, which is defined as the catalog whose event count falls in the uppermost 0.1 percentile of the forecasted number distribution. Notice the triggered ruptures on the Garlock and San Andreas faults that in turn generate aftershocks along these faults. (c) ‘Typical’ synthetic catalog generated by NoFaults and (d) an ‘extreme’ catalog from NoFaults, which lacks triggering of ruptures on prescribed faults resulting in a nearly isotropic aftershock distribution. The ‘extreme’ catalogs highlight the predominant differences between these two models and suggest that differences will be most noticeable when large aftershocks occur on mapped faults in U3ETAS.

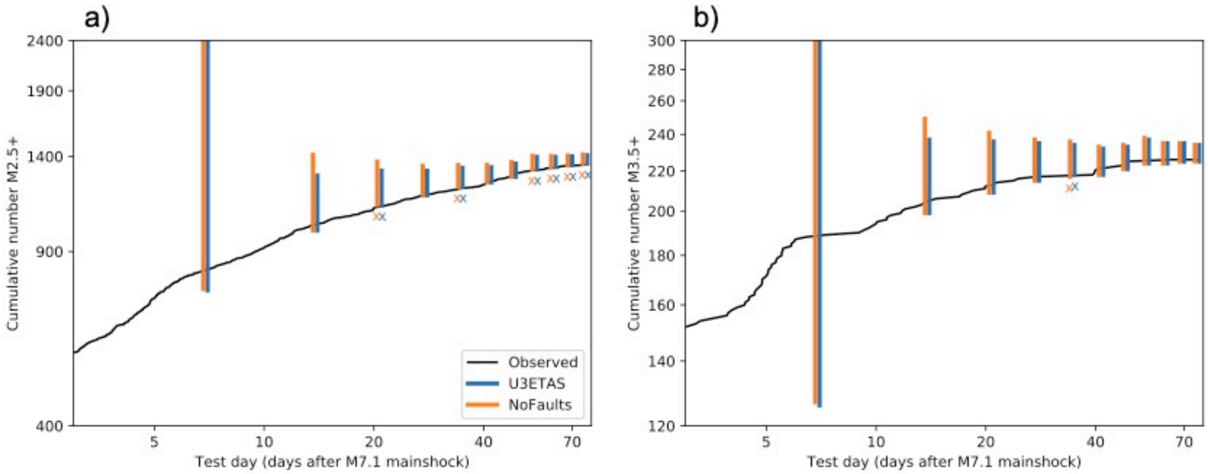

Figure 4. Forecasted number distributions and observed cumulative number over the eleven-week evaluation period. The forecasted event count distributions are offset by the number of observed events at the start of the forecast. Forecasted number distributions are plotted at the end of each evaluation period. The vertical extent of the lines indicates the 95-percentile range of the forecasted number distribution. The ‘x’ indicates evaluation periods with observed event counts that fall outside the 95-percentile range of the forecast. (a) Both observed and forecasted catalogs are filtered to threshold magnitudes $\mathrm{M_t}\ (\mathrm{t}) = \max(2.5, M_c(t))$ and (b) catalogs are filtered to $\mathrm{M_t}(\mathrm{t}) = \max(3.5, M_c(t))$. During the first seven-day forecast period, the 95th percentile of the forecasted number distribution for M2.5+ events are 2,482 and 3,906 events for U3ETAS and NoFaults, respectively.

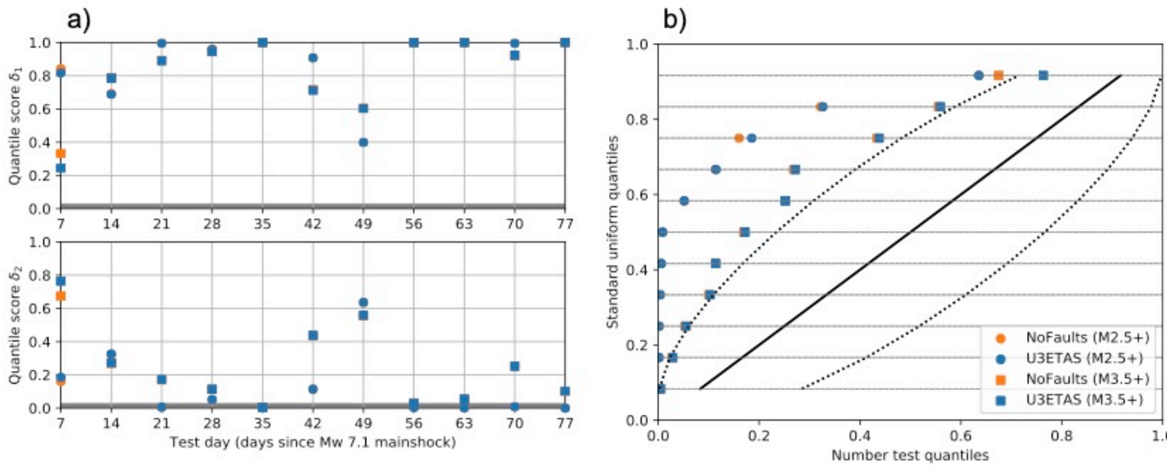

Figure 5. Aggregate number test results for $M_t(t) = \max(2.5, M_c(t))$ and $M_t(t) = \max(3.5, M_c(t))$ magnitude thresholds for U3ETAS and NoFaults for eleven weekly evaluation intervals following the M_w 7.1 mainshock. (a) Quantile scores δ_1 (top) and δ_2 (bottom) for individual weekly evaluation periods. (b) Quantile-quantile plot showing calibration of rate forecasts by comparing quantile scores, γ_N against standard uniform quantiles. The dashed lines indicate 95 percent confidence intervals around the standard uniform quantiles. Thus, U3ETAS and NoFaults overpredict the number of M2.5+ and M3.5+ events during this aftershock sequence.

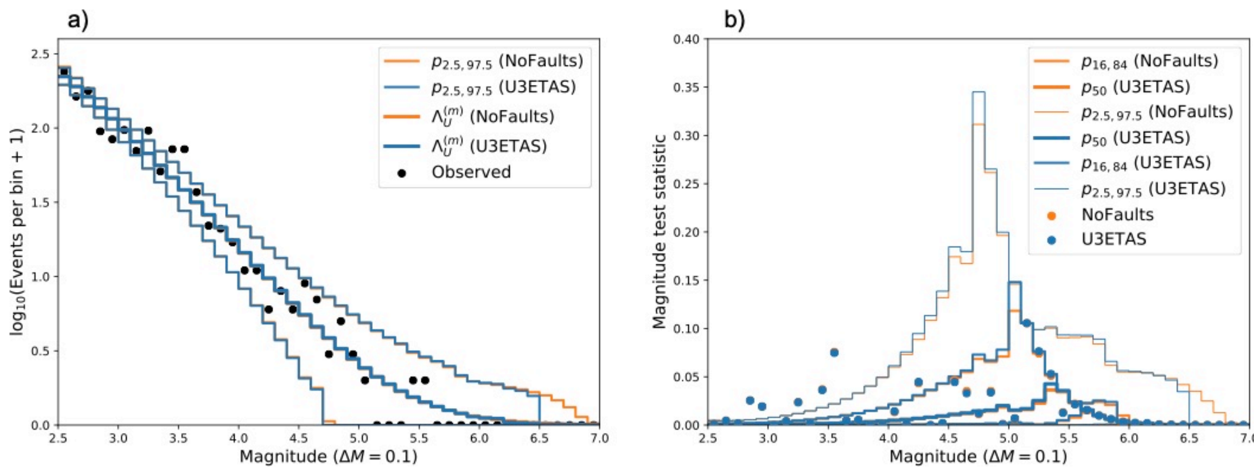

Figure 6. (a) Magnitude frequency distribution in $\Delta M = 0.1$ bins aggregated over entire the eleven-week evaluation period. The thin lines approximate the 95% percentile range of the event counts in each magnitude bin. The U3ETAS magnitude frequency distribution shows anti-characteristic behavior through the lack of M6.5+ earthquakes as compared with NoFaults. (b) Bin-wise magnitude test statistic aggregated over the entire evaluation period. The circles depict the kernel of d_{obs} for both U3ETAS and NoFaults to show bin-wise contributions to d_{obs}. We find negligible differences between the two models. The solid lines show percentiles from the bin-wise value distribution, for both models.

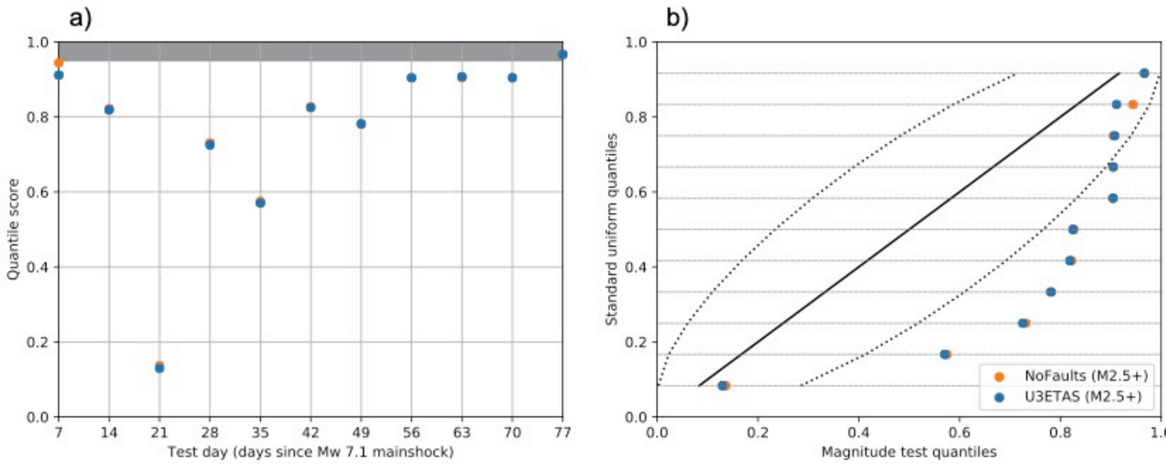

Figure 7. Magnitude test results for events with $M_t(t) = (2.5, M_c(t))$ over the full eleven-week evaluation period. (a) Quantile scores are shown for individual week-long evaluation periods. Gray patch depicts the 0.05 significance level for the magnitude test. The largest differences between U3ETAS and NoFaults exist during the first week and become negligible over the remainder of the evaluation period. (b) Calibration of magnitude forecasts by comparing magnitude test quantile scores against standard uniform quantiles. The dashed lines depict 95 percent confidence intervals around the standard uniform quantiles.

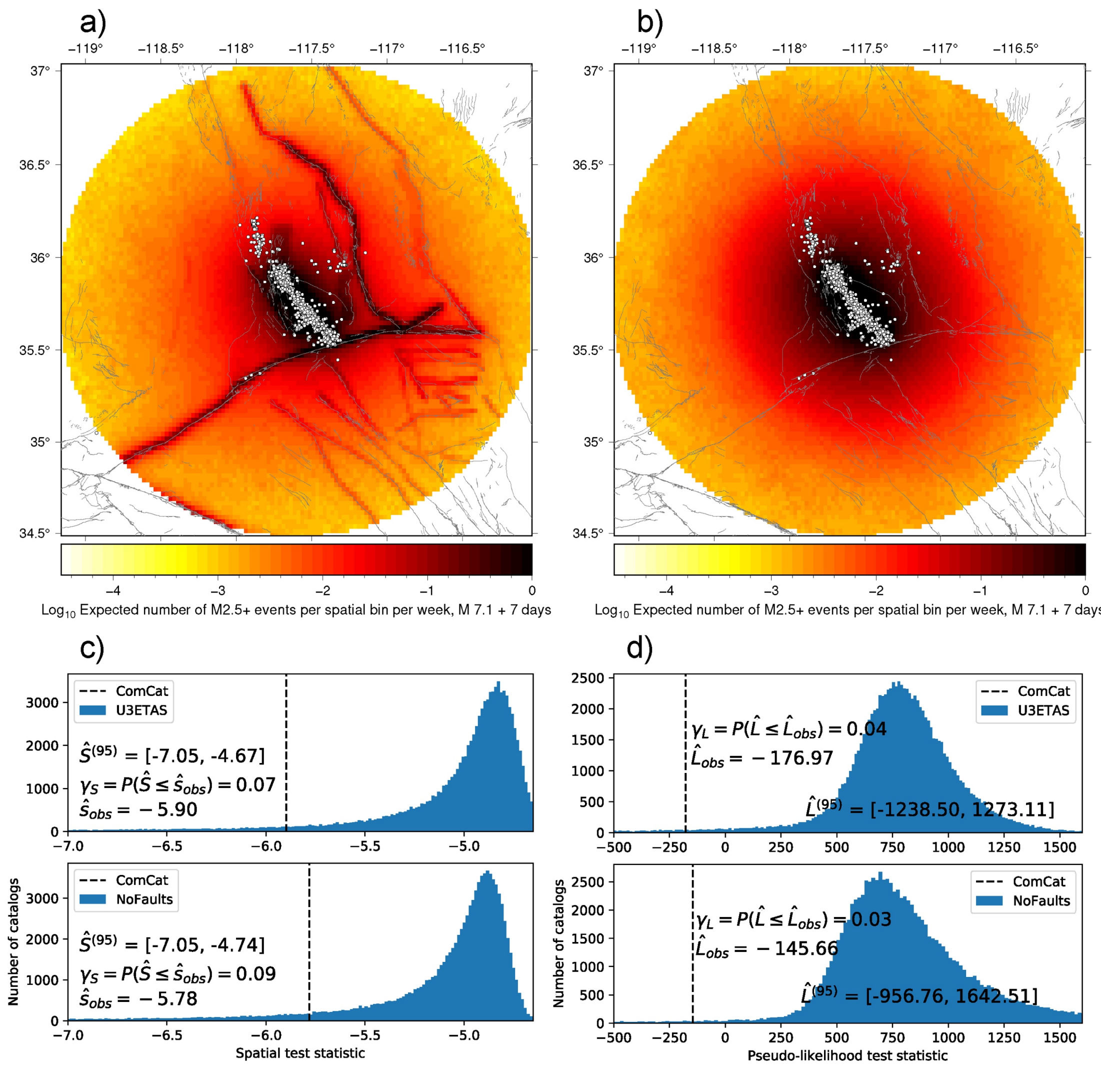

Figure 8. Logarithm of the expected event counts per spatial bin per week for U3ETAS (a) and NoFaults (b) for the week-long forecast following the M_w 7.1. The relatively high expected counts along the faults in U3ETAS are controlled by scenarios whose aftershock sequences contain supraseismogenic ruptures along these faults. In both plots, target events during this period are shown as white circles. The color scale is manually saturated for comparison purposes. The spatial bin with highest rate expects 64.24 and 65.76 events for U3ETAS and NoFaults, respectively. (c) Evaluation result for the spatial test for U3ETAS (top) and NoFaults (bottom) for the first evaluation period at seven days after the M_w 7.1 mainshock. $\hat{S}^{(95)}$ denotes the 95th percentile range of the test distribution of the spatial test statistic, $\hat{s}_{obs}$ is the observed statistic, and γ_S is the quantile score. (d) Same as (c) except for the pseudo-likelihood test statistics.

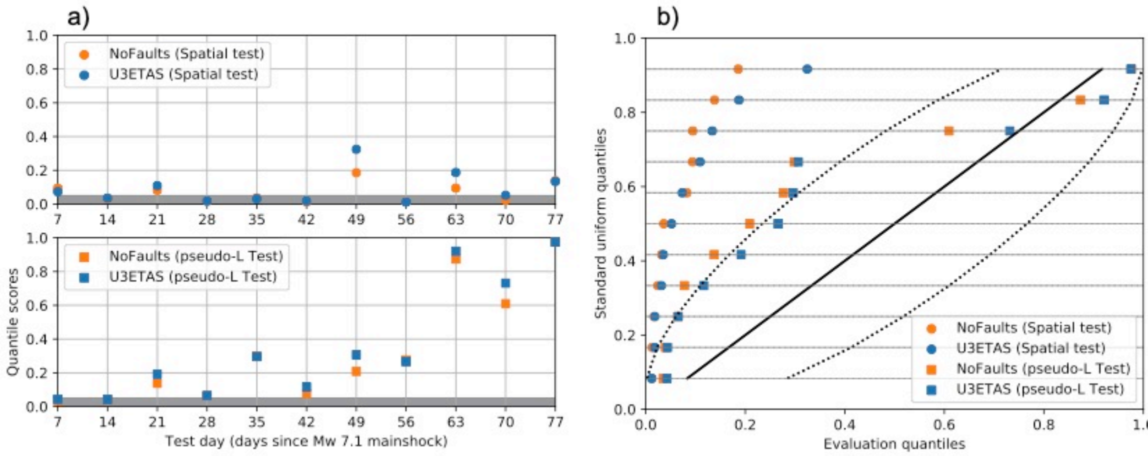

Figure 9. Spatial test and pseudo-likelihood results for events with $M_t(t) = \max(2.5, M_c(t))$ over the complete eleven-week evaluation period. The spatial test and likelihood tests show the greatest differences between U3ETAS and NoFaults. (a) Quantile scores shown for individual week-long evaluation periods. The patch depicts the 0.05 significance level for the spatial test. (b) Calibration of spatial forecasts by comparing quantile scores against standard uniform quantiles. The dashed lines depict 95 percent confidence intervals around the standard uniform quantiles.

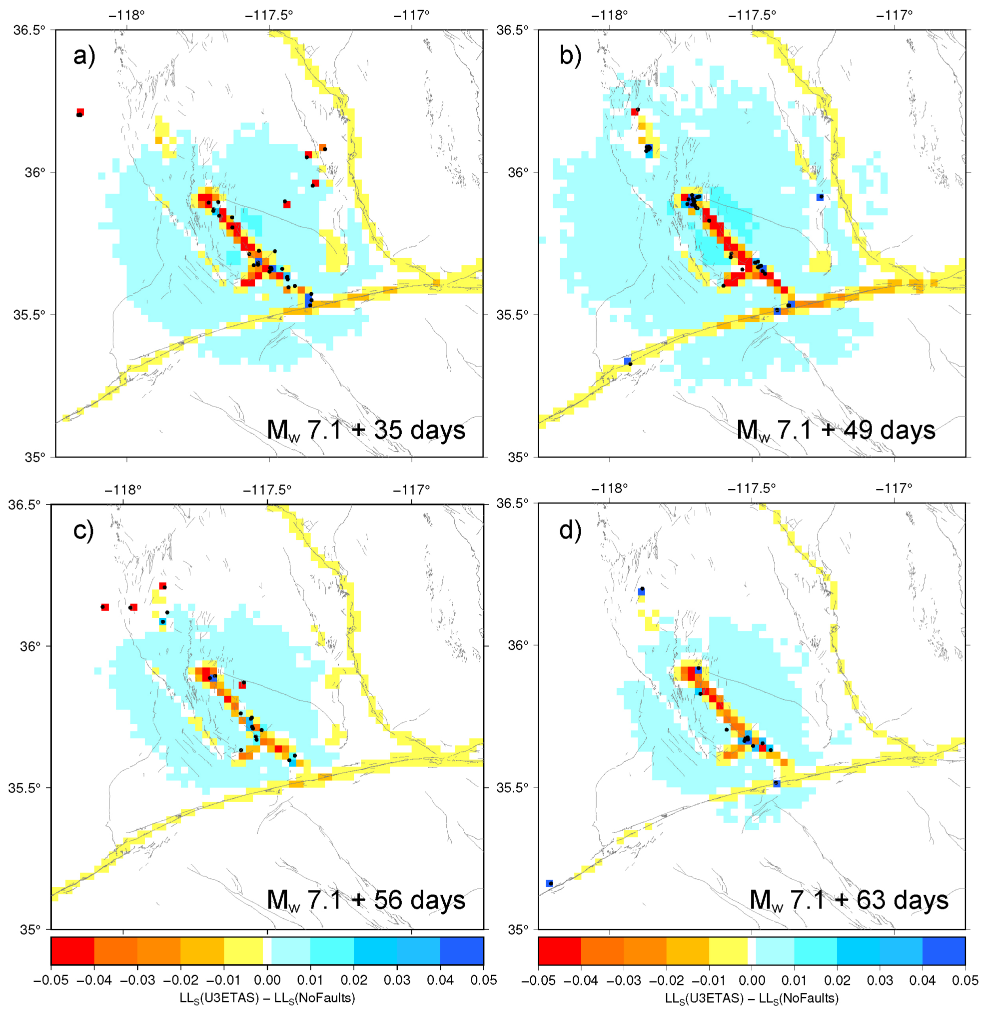

Figure 10. Map of cell-wise spatial pseudo log-likelihood ratios between U3ETAS and NoFaults for individual evaluation periods ending on (a) day 35, (b) day 49, (c) day 56, and (d) day 63 following the M_w 7.1 mainshock. Maps show the higher rates along faults in U3ETAS. Evaluation periods at (b) 49 days and (d) 63 days show the largest differences in the observed spatial statistic, which is calculated only from spatial cells where events occur, while periods ending on days 35 and 56 show a negligible difference in the spatial statistic. This highlights how spatial test results are sensitive to events occurring on modeled U3ETAS faults and that such events are required to discern between the models. The color scale is manually saturated between -0.05 and 0.05 to help comparisons; and dots show locations of target events.

Author Contact Information

Maximilian J. Werner, School of Earth Sciences, University of Bristol, Wills Memorial Building, Queens Road, Bristol BS8 1RJ, United Kingdom

Warner Marzocchi, Università degli Studi di Napoli Federico II, Corso Umberto I, 40, 80138 Napoli NA, Italy

David Rhoades, GNS Science, 1 Fairway Drive, Avalon, Lower Hutt 5011, New Zealand

David Jackson, Earth, Planetary, and Space Sciences University of California, Los Angeles 595 Charles Young Drive East Box 951567 Los Angeles, CA 90095-1567

Kevin Milner, University of Southern California, Southern California Earthquake Center, 3651 Trousdale Parkway, Los Angeles, CA 90089

Ned Field, United States Geological Survey, 1711 Illinois St Golden, CO 80401

Andrew Michael, United States Geological Survey, U.S. Geological Survey, 350 N. Akron Road Moffett Field, CA 94035