Bayesian Earthquake Forecasting Using Gaussian Process Modeling: GP-ETAS Applications

Christian Molkenthin[1], Gert Zöller[*1], Sebastian Hainzl[2,3], and Matthias Holschneider[1]

Abstract

Numerous seismicity models are known to simulate different observed characteristics of earthquake occurrence successfully. However, their ability of prospective forecasting future events is a priori not always known. The recently proposed semiparametric model, Gaussian process epidemic-type aftershock sequence (GP-ETAS) model, which combines the ETAS model with GP modeling of the background activity, has led to promising results when applied to synthetic seismicity. In this study, we focus on the ability of GP-ETAS for different forecasting experiments in two case studies: first, the Amatrice, Italy, sequence during 2016 and 2017, and second, long-term seismicity in Southern California. The results indicate that GP-ETAS performs well compared with selected benchmark models. The advantages become particularly visible in cases with sparse data, in which GP-ETAS shows in general a more robust behavior compared to other approaches.

Introduction

Earthquake modeling is crucial for various applications in seismology, for example, seismic hazard assessment or early warning. Because of the complexity and inherent randomness of earthquake occurrences, deterministic predictions are impossible. Consequently, the occurrence and distribution of earthquakes in a specific time window and predefined geographical region can only be modeled statistically as a realization of a stochastic process. More precisely it can be modeled as an outcome of a marked spatiotemporal point process, commonly described by an underlying intensity function (instantaneous rate of events), for which the earthquake magnitudes are the marks.

The benchmark model of a marked point process for spatiotemporal seismicity modeling is the epidemic-type aftershock sequence (ETAS) model introduced by Ogata (1988, 1998), which is built on empirical laws of the spatiotemporal clustering of earthquakes. ETAS is an example of a self-exiting point-process model, namely, a marked Hawkes process (Hawkes, 1971). Self-excitation means that each event causes an instantaneous increase in the intensity decaying with time and distance that may trigger future events and thus causing clustering in space and time. Although its main application is short-term seismicity modeling and forecasting, ETAS can also be used for long-term seismic forecasts because these can be derived from the background intensity part of the ETAS model or via simulations from the entire model, see for example, Ogata (2011) and Zhuang (2011).

To provide seismicity forecasts using the ETAS model, it is essential to infer its conditional intensity function from seismicity data. Traditional inference of the ETAS model relies on maximum-likelihood (ML) methods to derive point estimates for its model parameters (Ogata, 1998; Veen and Schoenberg, 2008; Lippiello *et al.*, 2014; Lombardi, 2015). However, achieving comprehensive uncertainty quantification of the inference results poses a challenge, commonly assessed by relying on standard errors of estimated model parameters typically based on the Hessian (Ogata, 1978; Rathbun, 1996; Wang *et al.*, 2010). This approach might be valuable in an asymptotic regime when the observational window is long enough. Other approaches to uncertainty quantification rely on bootstrap techniques (Fox *et al.*, 2016) or employing an ensemble of the best solutions to the ML optimization problem (Lombardi, 2015). However, when data are limited or scarce, ML estimation may not be robust and can encounter several problems, for example, a strong dependence on initial values, dependency on the optimization algorithm, and convergence problems. These problems may lead to underestimating the uncertainties of the inferred parameters, model estimates, and nonreliable forecasts.

By contrast, a Bayesian approach of the model inference offers several advantages compared to the traditional model

---

1. Institute of Mathematics, University of Potsdam, Potsdam, Germany, https://orcid.org/0000-0002-2384-2724 (GZ); https://orcid.org/0000-0003-3987-105X (MH); 2. GFZ German Research Centre for Geosciences, Potsdam, Germany, https://orcid.org/0000-0002-2875-0933 (SH); 3. Institute of Geosciences, University of Potsdam, Potsdam, Germany

*Corresponding author: zoeller@uni-potsdam.de

calibration and can complement it (see e.g., Ross, 2021). Bayesian inference starts with a prior assumption of model parameters, which can be informative, that is, based on specific knowledge, or noninformative. This prior distribution is combined with a likelihood function $\mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$, including a statistical model family, defined by a parameter vector $\boldsymbol{\theta}$, and the available data $\mathcal{D}$, which is in the present study an earthquake catalog. According to Bayes' theorem, a posterior distribution is obtained, which incorporates likelihoods of all possible models indicated by $\boldsymbol{\theta}$ for the given data $\mathcal{D}$. In contrast to point estimators such as the ML or the maximum a posteriori estimator, which select a single model in terms of $\boldsymbol{\theta}$, Bayesian inference returns a distribution of $\boldsymbol{\theta}$ and therefore allows for a comprehensive uncertainty assessment by construction. In this way, robust inference results can be obtained, even in situations with limited data. The inferred uncertainty of the model parameters, encoded in the posterior distribution, can be further propagated into seismicity forecasts, which is of great value. Bayesian seismicity forecasts can enhance the reliability and robustness of probabilistic seismic hazard analysis by explicitly addressing uncertainties, allowing for flexible modeling, integrating prior information, facilitating model updates, and accommodating the complexities of seismicity patterns. On the other hand, Bayesian inference of the ETAS model as we apply in this study is more complex and computationally more demanding than the traditional ML inference as one has to resort to Markov Chain Monte Carlo (MCMC) methods to obtain the posterior distribution. As a consequence, MCMC sampling of the exact posterior distribution can result in high-computational effort, if data sets are large. Serafini *et al.* (2023) have addressed this issue by decomposing the log-likelihood function into three parts and approximating each part linearly. Therefore, their R-package "inlabru" (Naylor *et al.*, 2023) is a suitable tool for applying Bayesian inference also to large data sets.

In this study, we demonstrate that a recently introduced Bayesian semiparametric ETAS model variant denoted as Gaussian process epidemic-type aftershock sequence (GP-ETAS; Molkenthin *et al.*, 2022) that works, in principle, without any approximation, is worthwhile pursuing for earthquake forecasting. A brief summary of GP-ETAS is provided in the subsequent Method section. First, it incorporates a flexible, fully Bayesian nonparametric model for the background intensity utilizing a Gaussian process prior. The idea of modeling spatial occurrences of seismicity using random fields has already been employed by Vere-Jones (1970), Bayliss *et al.* (2020, 2022), and others. Muir and Ross (2023) adopted the Gauss process modeling of background activity and modified it to a hierarchical version called "deep Gauss process." Second, it enables the joint Bayesian inference of model parameters for the triggering function as in Ross (2021) and Naylor *et al.* (2023). A similar approach has been proposed by Ross and Kolev (2022). They use a Dirichlet process rather than a Gaussian process for modeling the background. We conclude

that GP-ETAS is promising to provide a comprehensive uncertainty quantification that is valuable for time-dependent and time-independent earthquake forecasting.

For demonstration, we analyze two different cases in the Applications section. First, as a showcase for time-dependent forecasting in situations where data are limited, we apply GP-ETAS to the Amatrice seismic sequence during 2016 and 2017 in Italy, showing a better predictive power compared with standard ML inference results of the classical ETAS model. Second, as a showcase of the special importance of our nonparametric Bayesian approach for the inference of the background intensity, we compare long-term forecasts from GP-ETAS to the benchmark model of Helmstetter *et al.* (2007) for a test site in California. We show that GP-ETAS while using significantly less data seems to produce improved results and provides a comprehensive uncertainty quantification based on the posterior distribution.

Method

We consider seismicity data $\mathcal{D}$ in a specific time window $\mathcal{T}$ and in a predefined geographical region $\mathcal{X}$ and magnitudes $m \in \mathcal{M}$, structured as a time-ordered series $\mathcal{D} = \{(t_i, \boldsymbol{x}_i, m_i)\}_{i=1}^{N_\mathcal{D}}$. Here $N_\mathcal{D}$ is the number of recorded events, with $t_i \in \mathcal{T}$, $\boldsymbol{x}_i \in \mathcal{X}$, $m_i \geq m_0$ being the time, location (typically specified as longitude and latitude pairs; however, also depth can be incorporated), and magnitude of the events, for which the cutoff magnitude m_0 is larger equal to the completeness magnitude. The data $\mathcal{D}$ are statistically modeled by a marked point process and can be seen as a random point pattern in the space $\mathcal{T} \times \mathcal{X} \times \mathcal{M}$.

ETAS model

The ETAS model is characterized by a spatiotemporal-dependent conditional intensity function $\tilde{\lambda}(t, \boldsymbol{x}, m|H_t)$, which represents the instantaneous earthquake rate at time t, location $\boldsymbol{x}$, and magnitude m, conditioned on the history $H_t = \{(t_i, \boldsymbol{x}_i, m_i)\}_{i=1}^{N_H} : t_i < t\}$ of the N_H earthquakes up to t. The conditional intensity function factorizes $\tilde{\lambda}(t, \boldsymbol{x}, m|H_t) = \lambda(t, \boldsymbol{x}|H_t) f_M(m)$ in the intensity of a ground process $\lambda(t, \boldsymbol{x}|H_t)$ and the density of the magnitudes $f_M(m)$, for which we adopt the Gutenberg–Richter law $f_M(m) = \beta_m \exp(-\beta_m(m - m_0))$, a commonly used choice. The ground process is specified by

$$\lambda(t, \boldsymbol{x}|H_t) = \mu(\boldsymbol{x}) + \sum_{H_t} \varphi(t - t_i, \boldsymbol{x} - \boldsymbol{x}_i|H_t), \tag{1}$$

consisting of the sum of a Poisson intensity $\mu(\boldsymbol{x})$ of spontaneous earthquakes (background intensity) and a triggering part $\sum_{H_t} \varphi(t - t_i, \boldsymbol{x} - \boldsymbol{x}_i|H_t)$, which causes offspring events (aftershocks) depending on the history H_t. The triggering is characterized by function $\varphi(\cdot)$, which will be defined subsequently.

The background intensity $\mu(\boldsymbol{x})$ is commonly modeled either parametrically, typically employing a constant or a piecewise constant function (constant intensities for seismotectonic regions

or on a grid) or nonparametrically using a weighted kernel density estimator (KDE) with a Gaussian kernel and adaptive bandwidth, as introduced by Zhuang *et al.* (2002). In our ETAS applications, we use the latter method for which $\mu(x)$ is calculated by the weighted sum of Gaussian kernels $k_{d_i}(\cdot)$ with individual bandwidths, namely, $\mu_{\text{KDE}}(x) = (1/\mathcal{T}_{\mathcal{D}}) \sum_{i=1}^{N_{\mathcal{D}}} p_{i0} k_{d_i}(x - x_i)$, in which $\mathcal{T}_{\mathcal{D}}$ is the length of the observational time window, p_{i0} is the probability that event i is background, and $d_i = \max\{d_{\min}, r_{i,n}\}$ is the variable bandwidth. $r_{i,n}$ is the distance of event i from its n_pth nearest neighbor, and $d_{\min}$ is some minimum bandwidth. There are various recommendations for the choice of n_p. Zhuang *et al.* (2002) suggest opting for n_p within the range of 10–100. The minimal bandwidth $d_{\min}$ is commonly selected from the interval $d_{\min} \in [0.02, 0.05]°$. This range aligns with the usual localization error (Zhuang *et al.*, 2002). The probability p_{i0} that event i is background is given by $p_{i0} = \mu(x_i)/\lambda(t_i, x_i | H_{t_i})$.

The triggering function is given by (Ogata, 1998)

$$\varphi(t - t_i, x - x_i | H_t, \theta\varphi) = K_0 e^{\alpha_m(m_i - m_0)} (t - t_i + c)^{-p} s(x - x_i | m_i, \theta_s), \quad (2)$$

in which we state explicitly $\theta_\varphi = [K_0, c, p, \alpha_m, \theta_s]$, which are all parameters required for the function $\varphi(\cdot)$. The parameters K_0 and α_m determine the offspring (aftershocks) productivity of an event with magnitude m_i; c and p control the temporal decay of the offspring (aftershocks) intensity according to the empirical Omori–Utsu law (Omori, 1894; Utsu, 1961); and $s(x - x_i | m_i, \theta_s)$ represents a spatial density function characterizing how offspring events (aftershocks) triggered by a parent event with position x_i and magnitude m_i are distributed in space. Two commonly used forms are the short-range decay, often expressed by a Gaussian density, and the long-range decay, often expressed by a power law. Here, we utilize the latter form given by

$$s(x - x_i | m_i, \theta_s) = \frac{q - 1}{\pi \sigma_m(m_i)} \left(1 + \frac{(x - x_i)^{\mathsf{T}}(x - x_i)}{\sigma_m(m_i)} \right)^{-q}, \quad (3)$$

with $\sigma_m(m_i) = d^2 10^{2\gamma m_i}$ and $\theta_s = [d, \gamma, q]$ (Ogata, 1998; Zhuang *et al.*, 2002).

GP-ETAS model

The GP-ETAS, introduced by Molkenthin *et al.* (2022), is a Bayesian semiparametric spatiotemporal ETAS model variant for which the background intensity $\mu(x)$ is modeled in a very flexible Bayesian nonparametric way using a Gaussian process prior. A Gaussian process in space is defined as a set of random variables in space such that every finite sample of these random variables has a multivariate normal distribution. In the context of model design, Gaussian processes have the main advantage, compared to other stochastic processes, that several quantities can be derived explicitly without

approximation. In Molkenthin *et al.* (2022), it has been demonstrated that the use of Gaussian processes as a model for background activity in combination with data augmentation led to a conjugate structure in the Bayesian inference, that is, the Gaussian behavior propagates from the prior to the conditional posterior distribution, leading to a significant simplification in the evaluation of the posterior distribution.

Here, we assume that the background intensity is a priori defined by

$$\mu(x) = \bar{\lambda}\sigma(f(x)),$$

in which $\sigma(\cdot)$ is the logistic sigmoid function, and $\bar{\lambda}$ is the upper bound for the background intensity, such that $\mu(x)$ is bound between 0 and $\bar{\lambda}$. The function $f(x)$ is a Gaussian process prior meaning that the prior over any discrete set of K function values $f = \{f(x_i)\}_{i=1}^{K}$ at positions $\{x_1, x_2, ..., x_K\}$ is a K dimensional Gaussian distribution, which is defined by the prior mean and the $K \times K$ covariance matrix between function values at positions x_i. We note that no functional form of the background intensity has to be defined beforehand. The function $\mu(x)$ is directly learned from the data in a Bayesian way. More specifically, background intensity is treated as a random function and modeled as follows: $\mu(x) = \bar{\lambda}\sigma(f(x))$, in which $f(x) \sim \mathcal{GP}(m(x), k(x, x' | v))$ are latent function values with Gaussian process prior, $\sigma(z) = 1/(1 + \exp(-z))$ is the sigmoid function, and $\bar{\lambda} > 0$ is a scalar. The Gaussian process $f(x)$ captures the spatial correlation, and the sigmoid transformation ensures that the intensity function stays within the valid range for intensity values. A Gaussian process is defined by its mean function $m(x)$ and covariance function $k(x, x' | v)$, for details see Molkenthin *et al.* (2022) and Rasmussen (2013). In this study, a squared exponential covariance function is used. The background intensity, as modeled in GP-ETAS, provides a flexible framework for capturing smooth and localized variations across space. The triggering part of GP-ETAS is modeled by a parametric function, as in the classical ETAS model, which leads to the following functional form of the conditional intensity function:

$$\lambda_{\text{GP–ETAS}}(t, x | H_t) = \bar{\lambda}(\sigma(f(x))) + \sum_{t_i < t} \varphi(t - t_i, x - x_i | H_t, \theta_\varphi), \quad (4)$$

in which $\varphi(\cdot)$ is given by equation (2). The prior model specification of GP-ETAS, including the hyperparameters, is as follows:

$f \sim \mathcal{GP}$ prior with zero mean and a covariance function $k(x, x' | v))$
$v \sim p_v$, a prior on v (hyperparameters of covariance function)
$\bar{\lambda} \sim p_{\bar{\lambda}}$, a prior on $\bar{\lambda}$
$\mu | \bar{\lambda}, f, v \sim$ prior model on μ as defined in $\mu(x) = \bar{\lambda}\sigma(f(x))$
$\theta_\varphi \sim p_{\theta_\varphi}$, a prior on θ_φ of a triggering function $\varphi(\cdot)$.

The corresponding observational model is defined as $\mathcal{D} | \mu, \theta_\varphi$, a Hawkes process with GP-ETAS intensity function,

in which $\mathcal{D}$ denotes the observed data. The generative model for GP-ETAS, including a simulation algorithm, is comprehensively described in Molkenthin et al. (2022). Bayesian inference to obtain posterior density $p(\mu, \boldsymbol{\theta}_\varphi | \mathcal{D})$ is performed via data augmentation and MCMC sampling, that is, Gibbs sampling, as described in detail in Molkenthin et al. (2022). A particular advantage of these methods is that the likelihood function can, in principle, be evaluated exactly, that is, without approximation. However, the integral over the spatial domain related to the triggering function in the likelihood is approximated with the corresponding integral over $\mathbb{R}^2$, which is one because of the normalization. This simplification leads to a small bias, which is overall negligible. Details are discussed in Lippiello et al. (2014), Schoenberg (2013), and Petrillo and Zhuang (2024). The posterior density $p(\mu, \theta_\varphi | \mathcal{D})$ is given by K posterior samples $\{\mu_k, \boldsymbol{\theta}_{\varphi,k}\}_{k=1}^K \sim p(\mu, \boldsymbol{\theta}_\varphi | \mathcal{D})$, which encapsulate the inherent uncertainty after data observation and the incorporation of prior beliefs. This can be used to conduct a thorough assessment of uncertainty. Forecasts can be simulated directly from the posterior predictive distribution. This ability to propagate uncertainties into forecasts emanating from the posterior distribution is a noteworthy feature of GP-ETAS, enabling a nuanced understanding of the robustness and reliability of the model predictions in the face of both observed data and prior assumptions.

Forecast evaluation

To test and compare the forecast ability of seismicity models, the Collaboratory for the Study of Earthquake Predictability (CSEP) runs prospective tests of forecasts, specifying the mean number of expected earthquakes in predefined time–space–magnitude bins (Schorlemmer et al., 2007, 2018). CSEP provides a testing suite consisting of multiple tests designed to assess the consistency of different aspects of forecasts with observed data, assuming a homogeneous Poisson process in each bin. However, ETAS forecasts cannot be described by a homogeneous Poisson process and can be falsely rejected by those tests (Lombardi and Marzocchi, 2010). Thus, appropriate tests of the ETAS forecasts need to consider the full forecast distribution instead of only the mean value.

To assess the quality or reliability of probabilistic forecasts, different performance measures or metrics can be employed. Typically, proper or strictly proper scoring rules are utilized as outlined by Gneiting and Raftery (2007), and specifically for count data, as discussed by Czado et al. (2009) and Serafini et al. (2022). A systematic analysis of probabilistic forecasts based on MCMC simulations using scoring rules is given by Krüger et al. (2021). Scoring rules measure the quality of a forecast relative to observations and allow the comparison of different forecast-generating models simply by comparing the scores of their forecasts relative to observations. The score function denoted as $S(P, y_{\text{obs}})$ maps a forecast distribution P and an observation y_{obs} to a scalar value. P, which is the predictive distribution of the model, can be given either in parametric form or, if not analytically available, as in the case of ETAS, via K_{sim} simulated samples $\{F_i\}_{i=1}^{K_{\text{sim}}}$ with $F_i \sim P$. In this study, we follow the convention of employing negatively oriented scoring rules, which can be interpreted as error metrics to be minimized. Popular examples of suitable proper and strictly proper scoring rules for assessing discrete predictive distributions include the logarithmic score, the predictive deviance, the quadratic or Brier score, the spherical score, the ranked probability score (RPS), and the Dawid–Sebastiani score, as discussed in Czado et al. (2009).

In this study, we use the RPS, which is supposed to be more robust than the logarithmic score and which is particularly useful in the context of ensemble forecasts. In our simulation-based approach, the logarithmic or deviance scores can run into trouble when a single future realization has a zero probability under the predictive distribution—a situation not problematic for RPS. Specifically, we assume that forecast distribution P has a discrete probability mass $p(k)$ and a cumulative probability mass $P(k)$ for possible realizations $k = 0, 1, \ldots, \infty$. RPS, a strictly proper scoring rule, is defined as $S_{\text{RPS}}(P, y_{\text{obs}}) = \sum_{k=0}^\infty (P(k) - \mathbb{1}\{y_{\text{obs}} \leq k\})^2$. It measures the discrepancy between the cumulative distribution function (CDF) of the forecast $P(k)$ and the observed outcome y_{obs}, in which $\mathbb{1}\{\cdot\}$ is the indicator function giving one if the condition is fulfilled and zero otherwise. Using the empirical CDF, $\hat{P}(k) = 1/K_{\text{sim}} \sum_{i=1}^{K_{\text{sim}}} \mathbb{1}\{F_i \leq k\}$, instead of $P(k)$, RPS can be computed directly from the simulated samples $\{F_i\}_{i=1}^{K_{\text{sim}}}$ using the order statistics $F_1 < F_2 < \ldots < F_{K_{\text{sim}}}$,
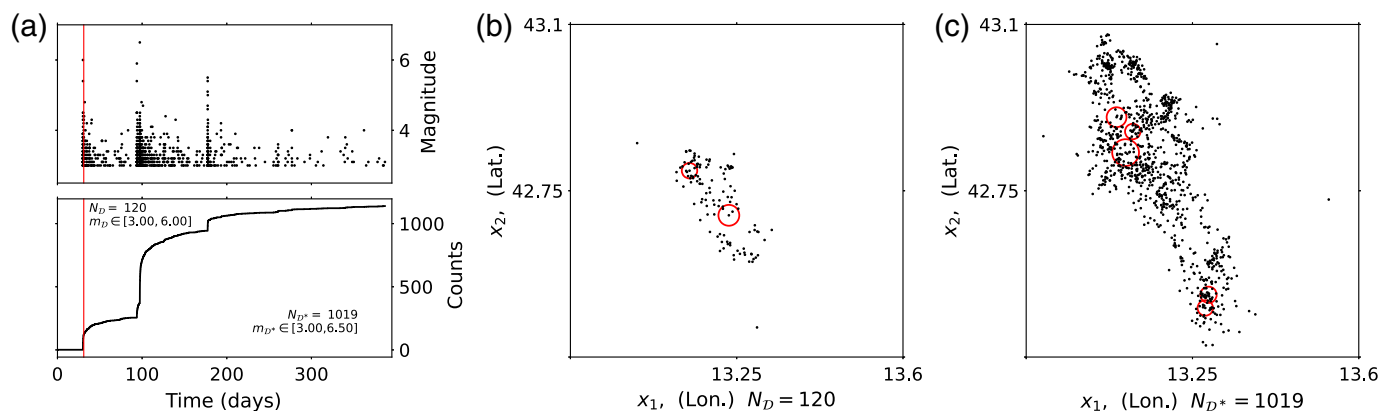
$$S_{\text{RPS}}(\hat{P}, y_{\text{obs}}) = \frac{2}{K_{\text{sim}}^2} \sum_{i=1}^{K_{\text{sim}}} (F_i - y_{\text{obs}}) \left(K_{\text{sim}} \mathbb{1}\{y_{\text{obs}} \leq F_i\} - i + \frac{1}{2} \right),$$

(6)

in which $F_i \sim P$ are sorted simulated samples from the predictive distribution P of the model, K_{sim} is the number of simulated samples, y_{obs} is the observed output, and i is an integer running from $i = 1, \ldots, K_{\text{sim}}$, see example, Reich and Cotter (2015). In this way, RPS can be considered as the error of the predictive CDF when compared to the ground truth (CDF based on observations).

Applications
Time-dependent earthquake forecasts
In this application, we undertake time-dependent earthquake forecasting by utilizing spatiotemporal GP-ETAS. Moreover, we assess its performance and outcomes, comparing them with those generated by the classical ETAS model. GP-ETAS adopts a Bayesian approach, while we employ ML estimation for the classical ETAS model. In addition, the two models differ in their representation of the background intensity. In GP-ETAS, the background intensity is characterized through Bayesian nonparametric Gaussian process modeling.

Figure 1. The Amatrice dataset consists of $N = 1139$ events, visualized as an earthquake sequence alongside its corresponding cumulative counts in panel (a). The temporal separation between the training data $\mathcal{D}$ and the test data $\mathcal{D}^*$ is indicated by a vertical solid red line for experiment 1, whereas the time window of the training data in experiment 2 is too small to be depicted in the figure. The spatial distribution of the training data $\mathcal{D}$, consisting of $N_{\mathcal{D}} = 120$ events, is illustrated as black dots in panel (b), and that of the test data $\mathcal{D}^*$ (unseen during inference) with $N_{\mathcal{D}^*} = 1019$ test events is shown in panel (c). Major shocks with magnitudes greater than or equal to 5.4 are highlighted by red circles, with the size of each circle proportional to the magnitude of the corresponding event. The color version of this figure is available only in the electronic edition.

In contrast, in the classical ETAS model, as implemented here, the background intensity is modeled using kernel density estimation with an adaptive step size. The parametric triggering function, denoted as $\varphi(\cdot)$, is the same for both models.

As our data set, we focus on the Amatrice seismic sequence during 2016 and 2017, which is thoroughly described in Marzocchi *et al.* (2017) and which has already been used for model testing in Serafini *et al.* (2023). The Amatrice data set comprises $N = 1139$ events with earthquake magnitudes $m \in \mathcal{M} = [3.0, 6.5]$. The time domain $\mathcal{T}$ covers the time period from 24 July 2016 to 15 August 2017, that is, $\mathcal{T} = [0, 395]$ days since 24 July 2016. The spatial domain $\mathcal{X}$ of the Amatrice data set spans from 12.9° to 13.6° E in longitudes and 42.4° to 43.1° N in latitudes, that is, $\mathcal{X} = [12.9°\text{E}, 13.6°\text{E}] \times [42.4°\text{N}, 43.1°\text{N}]$. The first major event of magnitude $m_1 = 6.0$ occurred at $t_1 = 30.067$ days. The data were obtained from the website of the National Institute of Geophysics and Vulcanology of Italy (Istituto Nazionale di Geofisica e Vulcanologia [INGV]; see Data and Resources), and are visualized in Figure 1.

To illustrate the benefit of employing the Bayesian GP-ETAS model in situations of data scarcity, we conduct two experiments with different training data windows. In the first case, experiment 1, we utilize a training data time window of about 31 days, denoted as $\mathcal{T}_{\mathcal{D}_1} = [0, 31.067]$ days, encompassing all data up to one day after the occurrence of the first major earthquake at t_1. In the second case, experiment 2, we use a significantly smaller training time window of just one day after t_1, that is, $\mathcal{T}_{\mathcal{D}_2} = [t_1, t_1 + 1]$ days. In both cases, we use the same standard priors and control parameters of the Gibbs sampler for the GP-ETAS model as given in Table 1. The model setup for the classical ETAS model corresponds to that described in the Method section with a kernel density estimation with variable bandwidths for the background intensity and a parametric triggering function. For the kernel density estimation of $\mu_{\text{kde}}(\mathbf{x})$, we adopt the conventional approach of using the minimal bandwidth (0.05°), with $n_p = 15$ representing the number of nearest neighbors used to determine the individual bandwidths (ETAS–classical; Zhuang *et al.*, 2002). We used the following starting values of $\boldsymbol{\theta}_{\varphi, \text{start}} = [K_0, c, p, d, \gamma, q]_{\text{start}} = [0.025, 0.01, 1.2, 1.8, 0.05, 0.5, 2]$ for

the ML estimation of the classical ETAS model. We remark that model parameters are estimated once in the training period and remain the same throughout the entire testing interval, whereas the earthquake history is updated on a daily basis. The robustness of the parameter estimations is shown in the supplemental material available to this article.

Various types of forecasts can be derived from the inference results of the models in the two experiments. Here are some examples: for instance, one approach involves generating forecasts in terms of the conditional intensity function $\lambda(t^*, \mathbf{x}|H_{t^*})$ at an arbitrary time point t^*, as illustrated in Figure 2 for experiment 1. We display the forecast of the spatially varying conditional intensity function evaluated at $t^* \approx 94.8$ days and all observed events in the following 24 hr. In the case of GP-ETAS, we show the mean of the posterior distribution of the intensity function in Figure 2, along with the uncertainty associated with the inferred intensity function in Figure 2b for a profile. In the case of the classical ETAS model, we display the ML point estimate in Figure 2b, which has no uncertainty information. Another possible forecast derived from the inference results is shown in Figure 3a,b. Here, we simulate samples from the predictive distribution of daily counts using the posterior samples in the case of GP-ETAS and the ML point estimate in the case of classical ETAS. This illustration only considers the predicted number of events for the entire region
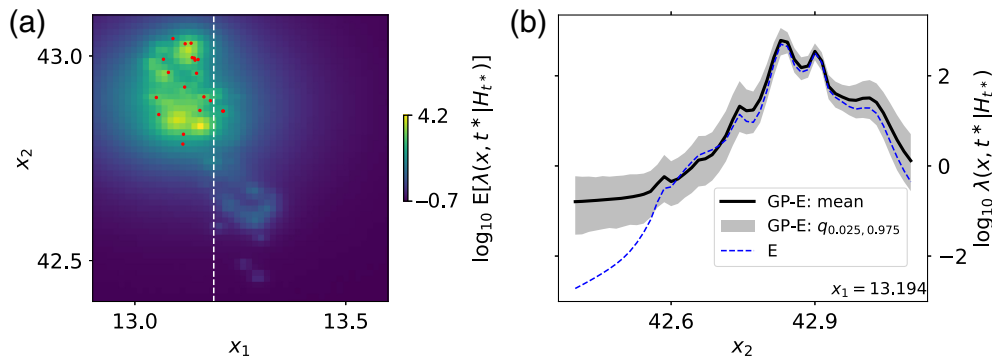
Figure 2. Forecast in terms of intensity $\lambda(\boldsymbol{x}, t^*|H_{t^*})$ at $t^* \approx 94.8$ days, arbitrarily chosen. (a) Color-coded Gaussian process epidemic-type aftershock sequence (GP-ETAS) mean intensity with red dots representing the 20 observed events that occurred in the next 24 hr, that is, in the range $[t^*, t^* + 1]$ days. (b) Specific result for the profile at $x_1 = 13.194$ (dashed line in panel a), for which the mean and quantiles of GP-ETAS are indicated by the solid black line and the gray shaded area, whereas the blue dashed line indicates the point estimate of the classical ETAS model. The color version of this figure is available only in the electronic edition.

$\mathcal{X}$. Although we apply the full spatiotemporal models, the spatial distribution of forecasted events in $\mathcal{X}$ is ignored here. The forecasts are updated daily and cover the prediction of events for the next 24 hr, following the procedure suggested in Serafini et al. (2023) for daily forecasts using a time-dependent seismicity model. The ith forecast time window starts at τ_i, and ends at $\tau_{i+1} = \tau_i + \Delta t$ with $\Delta t = 1$ day, assuming all events up to τ_i are known, thus the forecast for $[\tau_i, \tau_{i+1}]$ is delivered knowing H_{τ_i}. However, if a large earthquake with a magnitude $m^* \geq 5$ occurs at event time τ_{m^*} with $\tau_i < \tau_{m^*} < \tau_i + \Delta t$, the ith forecast time window ends at τ_{m^*} and is $[\tau_i, \tau_{m^*}]$. A new daily forecast is initiated starting from τ_{m^*} with a forecast time window of $[\tau_{m^*} + \epsilon, \tau_{m^*} + \epsilon + \Delta t]$, in which $\epsilon = 10^{-6}$ days is a

very small constant. Adding ϵ ensures the inclusion of the observed big shock at τ_{m^*} in the history, such that the subsequent forecast is based on $H_{\tau_{m^*}+\epsilon}$, which includes the shock at τ_{m^*}. Figure 3a,b shows the mean and the 0.025, 0.975 quantiles of forecasts of the total number of events for the entire region for GP-ETAS and classical ETAS. The spatial distribution of the events is neglected. Almost all observed earthquakes with $m \geq m_0$ (event counts) are within the quantiles $q_{0.025,0.975}$ of the daily GP-ETAS forecasts for both experiments, whereas for the classical ETAS model several

observed numbers for different forecast windows are not within the quantiles $q_{0.025,0.975}$ in the second experiment (experiment 2) based on a scarce data regime for the inference, see Figure 3b. We also note that in experiment 2 the training period includes one day only and is governed by an aftershock sequence with almost no background events. Consequently, parts of the aftershock sequence are misinterpreted as background events leading to an overestimation of the mean background activity in the testing period, which is visible in Figure 3b.

To rule out that the results might be dominated by the prior distributions of model parameters, we show in the supplemental material (Figs. S1, S2) the difference between prior and

TABLE 1
GP-ETAS Priors and Setup

Variable	Prior
Latent function f	Gaussian process prior with mean function $m(\boldsymbol{x}) = 0$, covariance function $k(\boldsymbol{x}, \boldsymbol{x}'|\boldsymbol{v})$ a squared-exponential kernel
Upper bound $\bar{\lambda}$	Gamma distribution $\Gamma(\alpha_0, \beta_0)$ with $\alpha_0 = 1/c_s^2, \beta_0 = 1/c_s^2/\mu_{\bar{\lambda}}$ with $c_s = 1$, $\mu_{\bar{\lambda}} = 2N_\mathcal{D}|\mathcal{X}|$
Hyperparameters $\boldsymbol{v}$	Exponential distribution with $\beta_{v_0} = 1/5$, $\beta_{v_l} = 5/2$, and $\boldsymbol{v} = [v_0, v_l]$
Triggering parameters $\boldsymbol{\theta}_\varphi$	Gamma distribution $\Gamma(\cdot, \cdot)$, $\boldsymbol{\theta}_\varphi = [K_0, c, p, \alpha_m, d, \gamma, q]$ $K_0 \sim \Gamma(1,10)$, $c \sim \Gamma(1,10)$, $p - 1 \sim \Gamma(1,2)$ $\alpha_m \sim \Gamma(1,0.5)$, $d \sim \Gamma(1,2)$, $\gamma \sim \Gamma(1,2)$, $q - 1 \sim \Gamma(1,1)$
Number of posterior samples K	1000
Burn-in iterations B	5000
Thinning	20, that is, each 20th sample after burn-in is accepted
Proposal distribution $\boldsymbol{v}$	Gaussian with $\sigma = 0.05$ in log units
Proposal distribution $\boldsymbol{\theta}_\varphi$	Gaussian with $\sigma = 0.01$ in log units

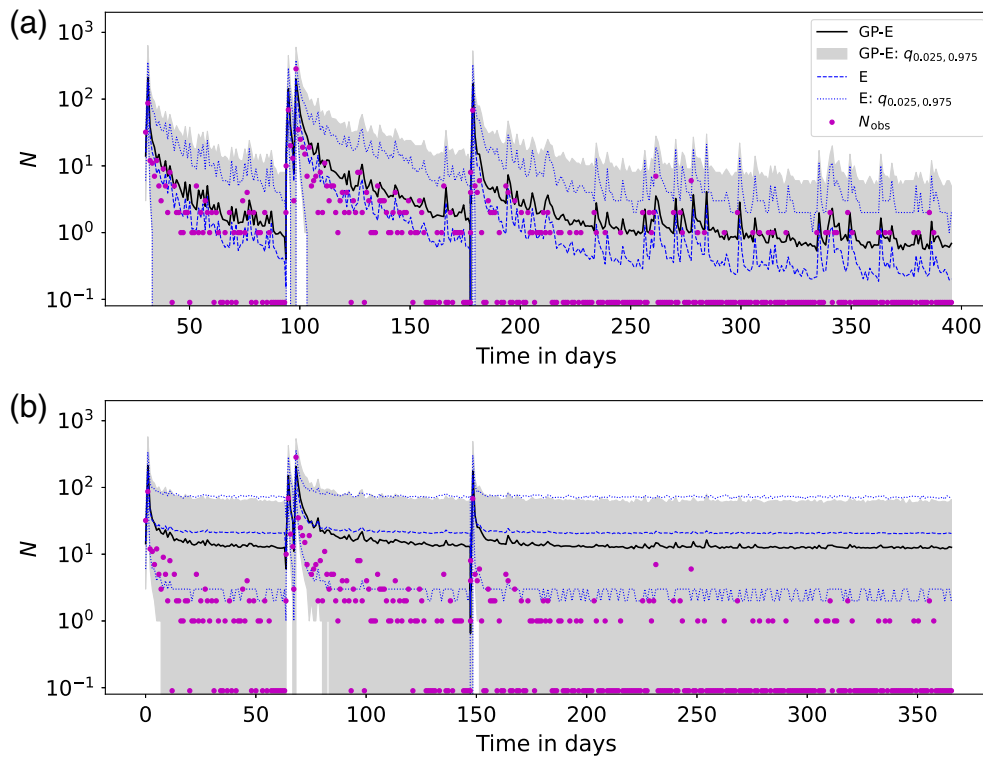GP-ETAS, Gaussian process epidemic-type aftershock sequence.

Figure 3. Forecasts of the total number of events for the entire region: The black solid line and the gray shaded area represent the mean and the 95% credible interval of GP-ETAS (GP-E). The dashed blue and dotted blue lines represent the mean and the 95% confidence interval of the classical ETAS (E). The magenta points are the observed number of earthquakes. Note that observations with zero events are plotted at $N = 0.1$. (a) Total counts for experiment 1 with $K_{\text{sim}} = 5000$ number of forecasts (simulations) and training period $|\mathcal{T}_{\mathcal{D}_1}| = 31$ days. (b) Total counts for experiment 2 under a low data regime with $K_{\text{sim}} = 5000$ number of forecasts (simulations) and training period $|\mathcal{T}_{\mathcal{D}_2}| = 1$ day. The color version of this figure is available only in the electronic edition.

$$p(\mathcal{D}|\mu(\boldsymbol{x}),\boldsymbol{\theta}_{\varphi})$$

$$= \prod_{i=1}^{N_{\mathcal{D}}} \lambda(t_i, \boldsymbol{x}_i | \mu(\boldsymbol{x}_i), \boldsymbol{\theta}_{\varphi})$$

$$\exp\left(-\int_{\mathcal{T}}\int_{\mathcal{X}} \lambda(t,\boldsymbol{x}|\mu(\boldsymbol{x}),\boldsymbol{\theta}_{\varphi})\mathrm{d}\boldsymbol{x}\mathrm{d}t\right), \tag{7}$$

in which the intensity λ is given by the GP-ETAS model or the classical ETAS model, and the dependencies on H_t and H_{t_i} are omitted for notational convenience. For GP-ETAS, ℓ_{test} can be approximated as follows:

$$\ell_{\text{test}} = \ln p(\mathcal{D}^*|\mathcal{D})$$

$$= \ln \int p(\mathcal{D}^*|\mu,\theta_{\varphi})p(\mu,\theta_{\varphi}|\mathcal{D})$$

$$d(\mu,\theta_{\varphi})$$

$$\approx \ln \frac{1}{K}\Sigma_{k=1}^{K} p(\mathcal{D}^*|\mu^{(k)},\theta_{\varphi}^{(k)}), \tag{8}$$

using the posterior samples $\{\mu^{(k)},\theta_{\varphi}^{(k)}\}_{k=1}^{K}$ obtained through the MCMC method (Gibbs sampler). In case of the classical ETAS, we evaluate

posterior distributions and observe significant differences, which is especially important for the low data regime. The stability of parameter estimations is further illustrated in Figures S3–S6. We also provide the classical ETAS model starting values and final ML estimates in Table S1.

The performance evaluation of the two models is based on testing data $\mathcal{D}^*$, which were not seen during the learning period of the models. Testing data $\mathcal{D}^*$ are all events with $t_i > t_1 + 1$ days, which are $N_{\mathcal{D}^*} = 1019$ events, and testing time window is $\mathcal{T}_{\text{test}} = [t_1, 395]$days, the spatial domain is $\mathcal{X}$. The Amatrice data set and its division in training and testing data are depicted in Figure 1.

To evaluate the spatiotemporal predictions of the two models, we use the log predictive probability $\ell_{\text{test}} = \ln p(\mathcal{D}^*|\mathcal{D})$, here interchangeably called the test-likelihood ℓ_{test}, which refers to the joint probability assigned to the observed data by the given statistical model and reflects the predictive power of different modeling approaches. This metric assesses the likelihood for the test data set $\mathcal{D}^*$ (unseen during the inference process), given the inferred model trained on the training data $\mathcal{D}$ using the classical point process likelihood function

$\ell_{\text{test}} = lnp(\mathcal{D}^*|\hat{\mu}_{\text{KDE}},\hat{\theta}_{\varphi})$, in which $\hat{\mu}_{\text{KDE}},\hat{\theta}_{\varphi}$ are the ML estimates. The performance metric ℓ_{test} for the unseen test data set with regard to the two experiments related to scarce data situations, and for the two models, GP-ETAS and classical ETAS are given in Table 2. In both experiments, GP-ETAS and ETAS show an overall reasonable performance with an advantage for GP-ETAS in the scarce data case. This improvement is further highlighted when evaluating the forecasts of total counts for the region. For the daily forecasts of the total number of earthquakes in the entire region, we computed the RPS based on $K_{\text{sim}} = 10,000$ simulations and averaged the values over all considered forecast intervals. The results are also presented in Table 2 for both experiments corresponding to the data-rich and low-data regimes. Specifically, the classical ETAS model exhibits slightly better performance in experiment 1 under the data-rich regime, whereas GP-ETAS outperforms the classical ETAS model in experiment 2 under the low-data regime. It is important that these daily forecasts of total counts for the region completely ignore the spatial distribution of future events.
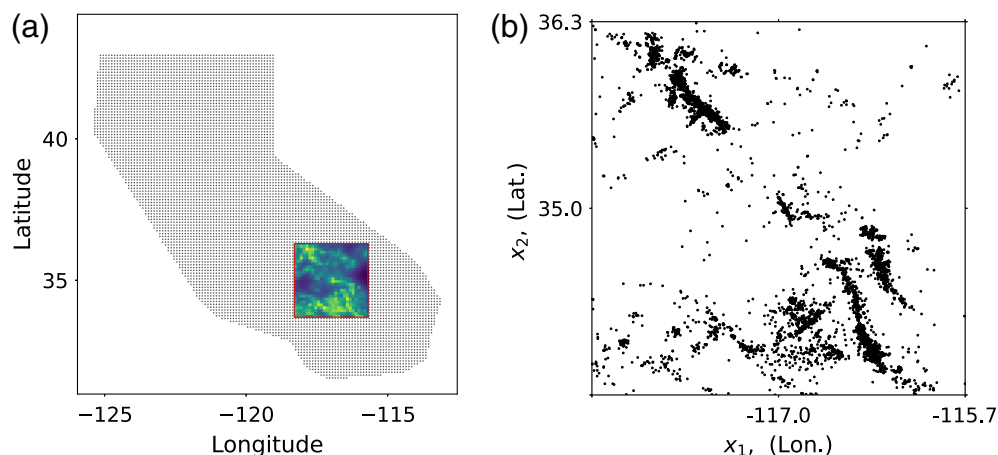
Figure 4. (a) Test site outlined in red (red box) as part of the Collaboratory for the Study of Earthquake Predictability (CSEP) test region in California, with the color-coded Helmstetter *et al.* (2007; referred as HE07) forecast. The remainder of the CSEP test region is shown by the dotted grid. (b) Spatial distribution of the $N = 5647$ $m \geq 3$ earthquakes recorded in the target area since 1981. The color version of this figure is available only in the electronic edition.

Time-independent earthquake forecasts

In addition to its primary use in modeling time-dependent seismicity, ETAS can also be employed for time-independent seismic forecasts. These forecasts can be obtained from the background intensity component $\mu(\boldsymbol{x})$ of the ETAS model or through simulations from the entire ETAS model, as suggested by Ogata (2011) and Zhuang (2011). In the subsequent example, we compare a time-independent forecast generated by GP-ETAS and the established benchmark model of time-independent seismicity for California proposed by Helmstetter *et al.* (2007; hereafter, HE07) in a specific testing location. HE07 is a grid-based forecast giving the time-independent expected seismicity intensity in spatial grid cells of size 0.1° × 0.1° and magnitude bins of 0.1 for $m \geq 4.95$ for the CSEP testing region California. For HE07, it is assumed that earthquakes occur independently in each grid cell following a Poisson process with a mean given by the grid value. Here, we consider the HE07 model for mainshocks and aftershocks, freely available from the pyCSEP webpage (see Data and Resources).
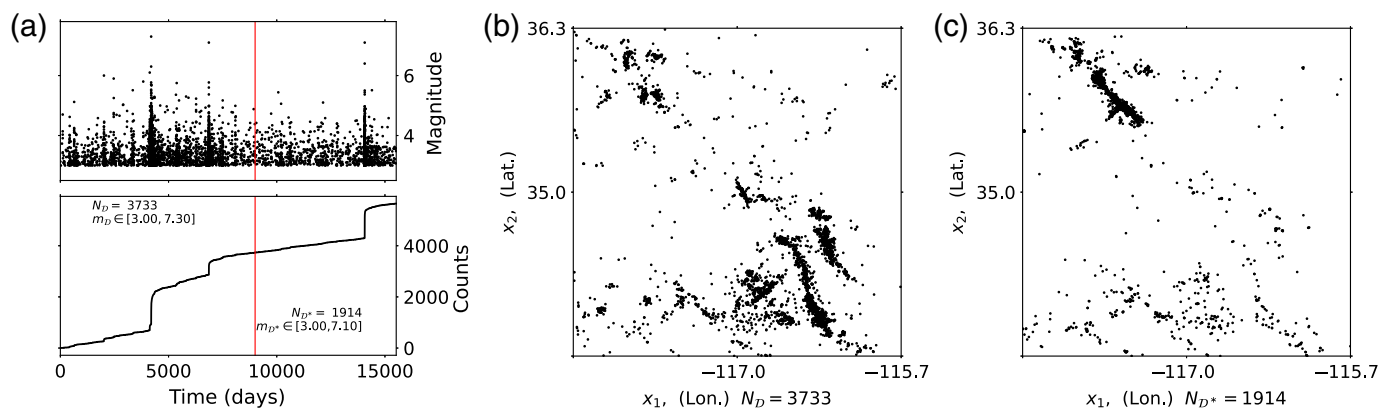
We focus on a test site in California that is of special interest regarding seismicity forecasts. This region, encompassing the spatial domain $\mathcal{X} = [-118.3°\text{E}, -115.7°\text{E}] \times [33.7°\text{N}, 36.3°\text{N}]$, has experienced notable mainshock and aftershock activity, including seismic sequences such as Landers (1992), Hector Mine (1999), and Ridgecrest (2019). Our test site is a part of the CSEP test region, which covers the entire state of California. The earthquake data are obtained from the Advanced National Seismic System Comprehensive Earthquake Catalog website using a tool from the pyCSEP webpage (see Data and Resources). The data set comprises $N = 5674$ events with earthquake magnitudes $3.0 \leq m \leq 7.3$ in $\mathcal{X}$, which occurred in the period from 1 January 1981 to 1 July 2023, that is, in the total time-domain $\mathcal{T} = [0, 15521]$ days since 1 January 1981. The position of our test site $\mathcal{X}$ within the spatial domain of the entire CSEP test region and the spatial distribution of the employed data set are illustrated in Figure 4.

For a fair comparison with HE07, we employ the identical time window $\mathcal{T}_{\mathcal{D}}$ for the training data $\mathcal{D}$ as utilized for HE07: 1 January 1981 to 23 August 2005, corresponding to 9000 days. However, we only use the events above $m \geq 3$, whereas HE07 uses all earthquakes with $m \geq 2$ to train their model. Thus, HE07 utilized approximately 12 times more data than we use for training GP-ETAS at the selected test site. We split the data set into training data $\mathcal{D}$, all events with event times $t_i \leq 9000$ days ($N_{\mathcal{D}} = 3733$ events, $\mathcal{T}_{\mathcal{D}} = [0, 9000]$ days), which corresponds to the training time window of HE07, and test data $\mathcal{D}^*$, all events with $t_i > 9000$ days ($N_{\mathcal{D}^*} = 1914$, $\mathcal{T}_{\mathcal{D}^*} = [9000, 15521]$ days), as shown in Figure 5. Thus the size of the

TABLE 2

Comparison of the Predictive Performance of GP-ETAS and Classical ETAS for Experiments 1 and 2

Experiment	Training	Spatiotemporal ℓ_{test}		Total Counts (RPS)	
		ETAS	GP-ETAS	ETAS	GP-ETAS
1	≈31 days	5520.8	**5722.6**	**1.43**	1.60
2	1 day	3981.0	**5644.7**	11.54	**5.78**

Spatiotemporal forecast evaluated via ℓ_{test} (higher is better, see equations 7 and 8) using $K = 1000$ posterior samples. The daily updated forecasts of the total count of events for the entire region, regardless of its spatial distribution, are evaluated using the ranked probability score (RPS; lower is better, see the Forecast evaluation section). For each case, the best result is highlighted in bold. ETAS, epidemic-type aftershock sequence; and GP-ETAS, Gaussian process epidemic-type aftershock sequence.

training data time window is $|\mathcal{T}_\mathcal{D}| \approx 24.64$ yr, the one of the testing data is $|\mathcal{T}_{\mathcal{D}^*}| \approx 17.85$ yr and the total time window of the data set is $|\mathcal{T}| \approx 42.49$ yr.

Training data $\mathcal{D}$ are utilized to perform Bayesian inference for GP-ETAS, obtaining posterior samples. The analysis follows the standard setup outlined in Table 1, with priors as detailed in the same table. We simulate 500 posterior samples after a burn-in of 5000 iterations and with a thinning of 20. The test data $\mathcal{D}^*$ is used to evaluate the quality of the forecasts of GP-ETAS and HE07. As for the Amatrice case study, we show comparisons between prior and posterior distribution for the model parameters as well as corresponding Markov chains and their autocorrelations in Figures S7–S9. Again, we find robust parameter estimations and clear modifications of the prior distributions by the data.

The time-independent Bayesian forecast of GP-ETAS is obtained via $K_{\mathrm{sim}} = 10{,}000$ simulations from its predictive distribution $p(\mathcal{D}^*|\mathcal{D})$ using the previously obtained posterior samples. Forward simulations are conducted for the forecast time window of $|T_{\mathrm{forecast}}| = |T_{\mathcal{D}^*}| = 6521$ days. It aligns with the duration of available testing data $|T_{\mathcal{D}^*}|$. Earthquake magnitudes are simulated from a Gutenberg–Richter law with $b = 1.01$ obtained from $\mathcal{D}$. A maximum magnitude of 8.5 is assumed during simulations. These simulations do not consider any history data before the start time of the forecast and consist of K_{sim} time series, that is, catalogs. The simulated events contained in the K_{sim} catalogs are then binned according to the spatial grid specified by the HE07 forecast to get the predictive distribution of counts for each spatial grid cell of the HE07 forecast. The HE07 consists of L spatial grid cells.

Figure 6 shows the mean forecasts of GP-ETAS and HE07 together with the $m \geq 3$ earthquakes in test period $\mathcal{T}_{D^*}$. HE07 is scaled using 10^b to obtain a forecast for $m \geq 3$, in which b is the Gutenberg–Richter b-value of $\mathcal{D}$. Although generated from entirely different models, HE07 and the mean GP-ETAS forecast exhibit similar spatial seismicity patterns. However, the HE07 forecast appears more concentrated, whereas the mean GP-ETAS forecast appears smoother. The observed seismicity in the test data $\mathcal{D}^*$ visually aligns with the generated forecast for

Figure 5. Separation of the training and test data set. The entire dataset comprises $N = 5674$ events, visualized as an earthquake sequence alongside its corresponding cumulative counts in panel (a). The temporal separation between the training data $\mathcal{D}$ and the test data $\mathcal{D}^*$ is indicated by a vertical solid red line. The spatial distribution of $\mathcal{D}$ with $N_\mathcal{D} = 3733$ events is shown in panel (b), whereas those of $\mathcal{D}^*$ with $N_{\mathcal{D}^*} = 1914$ unseen test data are shown in panel (c). The color version of this figure is available only in the electronic edition.

both models. We compare the predictive distribution of counts for each grid cell of the GP-ETAS forecast with the one generated by HE07. In the case of HE07, we simulate its predictive distribution of total counts of the lth spatial grid cell from a Poisson process with mean intensity $\lambda_{l,\mathrm{HE07}}|\mathcal{T}^*|/|\mathcal{T}_{\mathrm{HE07}}|$ with K_{sim} realizations. Profiles of GP-ETAS and HE07 forecasts are presented in Figure 7 to illustrate differences or disparities in their forecast uncertainties. This discrepancy may indicate that HE07 provides an uncertainty estimate that appears possibly too narrow, potentially leading to unreliable forecasts that overlook uncertainty in some regions. Overall, uncertainties tend to be greater for GP-ETAS than HE07.

To compare the quality of the probabilistic forecasts, we compute a summary measure based on the test data $\mathcal{D}^*$, employing the RPS (see the Forecast evaluation section). Table 3 shows the resulting average scores for the two models and two magnitude ranges. For both minimal considered forecast magnitudes $m_{\mathrm{f},1} = 3$ and $m_{\mathrm{f},2} = 4.95$, the mean score for the GP-ETAS forecast is lower than that of the HE07 forecast indicating an advantage for GP-ETAS.

We also assessed the reliability of forecasts regarding the total count of events N^* for the entire region $\mathcal{X}$, neglecting the spatial distribution of events. Figure 8 shows the predictive distributions of total counts integrated over the entire spatial domain for both models and the two magnitude ranges. The considered forecast time window is $|\mathcal{T}_{\mathrm{forecast}}| = |\mathcal{T}_{D^*}| = 6521$ days. For both minimal forecast magnitudes, the predictive distribution of GP-ETAS is broader, as expected, because it contains uncertainties originating from the posterior distribution, propagated into the forecast. The probability of
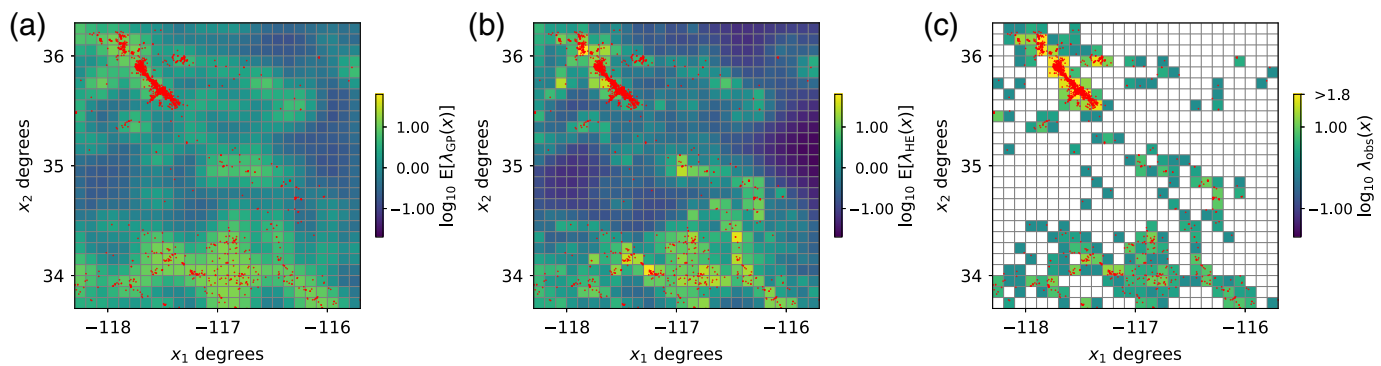
Figure 6. Time-independent forecasts are provided for events with magnitudes $m \geq 3$ and a total forecast duration $|T_{\mathcal{D}^*}| = 6521$ days (≈ 17.85 yr). The mean intensities of (a) GP-ETAS and (b) HE07. (c) The binned distribution of observed test data $\mathcal{D}^*$. Red dots indicate the positions of the test data $\mathcal{D}^*$ with magnitudes $m \geq 3.0$. The color version of this figure is available only in the electronic edition.
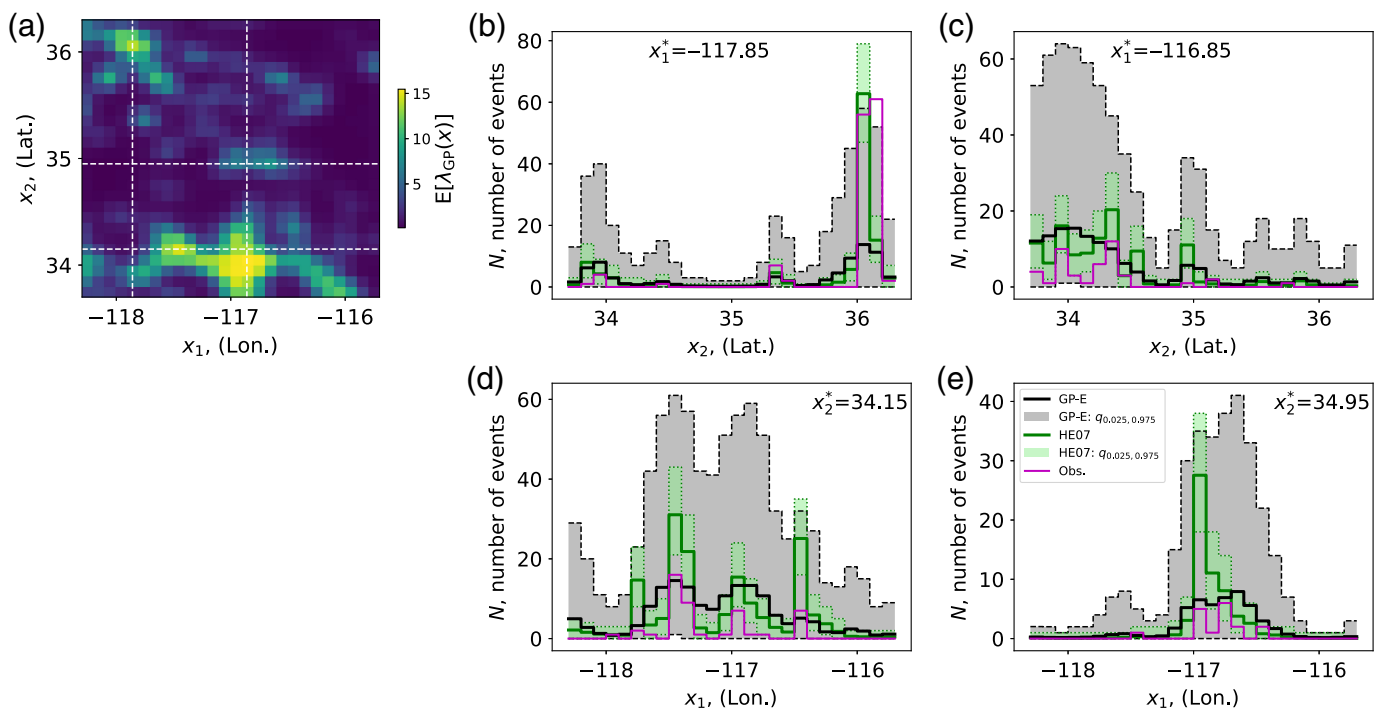

Figure 7. (a) Color-coded GP-ETAS forecast with dashed lines referring to the profiles used in the other panels. (b–e) The predictive mean forecasts with uncertainties of GP-ETAS and HE07 along the profiles indicated in the legend. The color version of this figure is available only in the electronic edition.

the observed test data is higher for the GP-ETAS forecast than for the HE07 forecast. GP-ETAS also performs better in terms of the RPS score. Thus, GP-ETAS shows high forecast skills for the considered test site in California compared to the benchmark model. In addition, forecasts generated by GP-ETAS consider the complete uncertainty encoded in its posterior distribution.

Summary and Conclusions

The semiparametric model GP-ETAS, as introduced in Molkenthin et al. (2022), is based on background activity modeled in a Bayesian nonparametric way using a Gaussian process prior, whereas offspring events (aftershocks) are simulated with the classical ETAS model (Ogata, 1988). Applying various types of data augmentation within a Bayesian framework
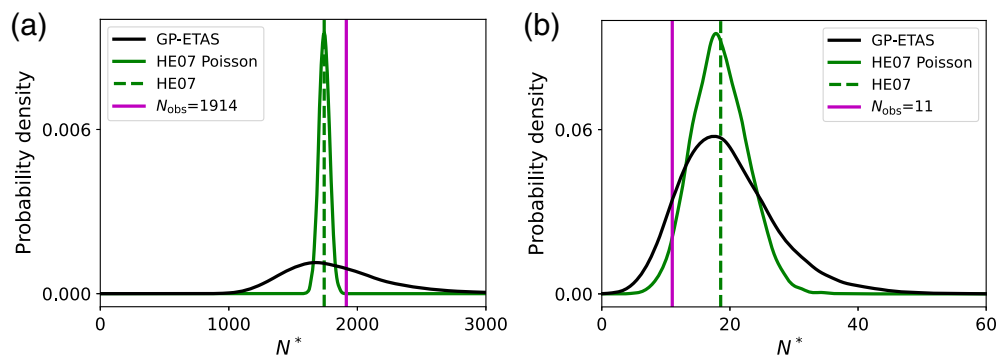
Figure 8. Forecast of total number of events N^* with minimal earthquake magnitudes considered in the forecasts: (a) $m_f = 3$ and (b) $m_f = 4.95$. Median values of the GP-ETAS predictive distribution are 1779 in panel (a) and 18 in panel (b). The color version of this figure is available only in the electronic edition.

allows us to evaluate the posterior distribution of model parameters exactly, resulting in a reasonable evaluation of uncertainties. Synthetic tests have demonstrated that GP-ETAS is more flexible in modeling space–time seismicity than other spatiotemporal extensions of the classical ETAS model, see Molkenthin *et al.* (2022) and references therein. In this study, we have focused on the predictive power of GP-ETAS. For this purpose, the model is applied to two well-known earthquake-prone regions—Amatrice (Italy) and a part of Southern California, including the locations of the Landers (1992), Hector Mine (1999), and Ridgecrest (2019) earthquakes. The first case study refers mainly to time-dependent predictions. In contrast, the latter analyzes the spatial earthquake forecasts over a long period, neglecting the time dependence of earthquake occurrences within this period. For both parts, we have compared GP-ETAS predictions with corresponding results from benchmark models.

An important strength of Bayesian inference is the straightforward uncertainty assessment, even in situations with sparse

TABLE 3
Comparison of the HE07 and GP-ETAS Forecasts for the Selected Region in California Using RPS (Lower Is Better, See the Forecast Evaluation Section), Based on K_{sim}=10,000 Simulations and Either a Minimum Magnitude of m_3 or $m_{4.95}$

Model	Spatial Forecast		Total Counts	
	m_3	$m_{4.95}$	m_3	$m_{4.95}$
HE07	2.85	0.018	148.6	5.17
GP-ETAS	**2.42**	**0.017**	**106.1**	**4.98**

In the case of the spatial forecast, the values represent the average RPS for the spatial cells. For each forecast, the best RPS result is marked in bold. HE07, Helmstetter *et al.* (2007); GP-ETAS, Gaussian process epidemic-type aftershock sequence; and RPS, ranked probability score.

data. For a showcase, we have divided the Amatrice earthquake catalog into a training set, for which the model parameters are estimated, and a testing set, in which the predictions are evaluated. Here, we performed two experiments to study the impact of data richness in the training period: in the "data-rich" situation, the model is trained for 31 days, including one day after the first M6 earthquake, to learn the aftershock behavior. In contrast, the "data-poor" training period includes only the day after the M6 event to fit the model to the data. GP-ETAS outperforms ETAS in both experiments according to the predictive probability as a metric. Although the difference is moderate in the first case, the outperformance is more pronounced in the "data-poor" situation. It must be emphasized that the predictive probability considers the spatial and temporal performance of the predictions. If we only focus on the number of predicted events, the classical ETAS model has a slight advantage for the long training period, whereas GP-ETAS is better for the "data-poor" case. These results confirm that the flexible way of estimating the background activity in GP-ETAS seems beneficial for a realistic model design and, thus, for reasonable forecasts. This advantage becomes particularly visible in the sparse data setting. However, we do not claim that GP-ETAS is superior to other models. Instead, one goal of our study was to understand how GP-ETAS compares with selected benchmark models. Criteria for this selection are the following: the benchmark models should have (1) a similar degree of complexity, (2) should generate a similar output, and (3) should be established in the community. We note that the models of Zhuang *et al.* (2002) and Helmstetter *et al.* (2007) are based on a small number of overall well-established assumptions. Especially KDEs are a simple and frequently used technique to model spatial fields. If now a new model such as GP-ETAS comes into play, it should not perform much worse than the "simple" benchmark models. In this sense, we do not claim the superiority of GP-ETAS but the fulfillment of minimum standards.

Furthermore, it has to be emphasized that the selection of the training and testing period is not a trivial task. If, for example, a short training set is dominated by an aftershock sequence of a former large event, it will be unlikely that the trained model performs well in forecasting long-term background seismicity. High predictive power can be only expected if the model can learn different features of background and triggered seismicity. For this reason, the choice of training and testing interval is to some degree subject to expert knowledge.

The case study related to seismicity in California is close to the CSEP framework (Schorlemmer *et al.*, 2018) and addresses long-term predictions of earthquake rates in a spatial testing region. Again, the framework of training and testing periods is used. The benchmark model of Helmstetter *et al.* (2007) imposes smoothed seismicity in space and Gutenberg–Richter distributed magnitudes with exponential taper for large values. Similar to the Amatrice case, GP-ETAS outperforms the benchmark model for spatially varying predictions in a fixed time frame and also for the total number of events in the entire test region. Interestingly, this result remains stable if predictions are targeted at large events only, that is, for $m \geq 4.95$.

Comparing GP-ETAS with two benchmark models ensures that minimum standards are met. Further comparisons with recently developed models, for example, Fox *et al.* (2016), Ross and Kolev (2022), or Muir and Ross (2023) will provide more insights into the suitability of the different approaches for forecasting purposes and will therefore be part of future activities. Finally, the question, of why a model performs better than another model is still not answered. This will require a deeper study of model features such as physical assumptions, uncertainty assessment, and others, to identify the origin of good or poor performance. Our study has been motivated by practical aspects in the spirit of the CSEP project.

In summary, the results of our retrospective testing experiments demonstrate that GP-ETAS provides a powerful framework for predicting seismicity in time and space. In this context, the assessment of uncertainties of model parameters is a crucial point, especially for sparse data. Point estimators have the disadvantage that uncertainties are usually given in approximation only. Even if Bayesian methods are used, wrong prior assumptions can lead to strongly biased results. So far, all simulations with GP-ETAS, including various tests for synthetic seismicity, have shown a high degree of consistency and robustness in this respect. However, in the present form, the Gibbs sampling of the posterior distribution is computationally very expensive. Therefore, future activities need to focus on reducing the computational burden. Another direction of future research can be the assimilation of additional spatial information as in Bayliss *et al.* (2022) to better constrain the background seismicity. This might include strain maps or fault locations, which could be implemented in the Gaussian process prior.

To assess the accuracy and reliability of probabilistic seismicity forecasts and to enhance the CSEP testing suite, we recommend incorporating the RPS as a usual evaluation metric. RPS is a suitable scoring rule for probabilistic forecasts and offers simplicity in the implementation with simulations. Unlike the logarithmic score frequently used by CSEP, RPS avoids issues when a single future realization has a zero probability under the predictive distribution. This makes RPS a valuable addition to assessing forecast quality and complementing existing evaluation methodologies within the CSEP framework.

Finally, it should be noted that the predictive skills have been analyzed for particular case studies with particular benchmark models in this work. The question of whether GP-ETAS is generally superior to other models in terms of predictive skills remains open. Therefore, it is desirable to adjust GP-ETAS to meet the technical standards of CSEP for automated prediction experiments and then submit the model to the testing centers. Prospective CSEP tests will help to systematically address the previous question.

Data and Resources

Declaration of Competing Interests

The authors acknowledge that there are no conflicts of interest recorded.

Acknowledgments

References

Bayliss, K., M. Naylor, J. Illian, and I. Main (2020). Data-driven optimization of seismicity models using diverse data sets: Generation, evaluation, and ranking using Inlabru, *J. Geophys. Res.* **125,** no. 11, e2020JB020226, doi: 10.1029/2020JB020226.

Bayliss, K., M. Naylor, F. Kamranzad, and I. Main (2022). Pseudo-prospective testing of 5-year earthquake forecasts for California using Inlabru, *Nat. Hazards Earth Syst. Sci.* **22,** no. 10, 3231–3246.

Czado, C., T. Gneiting, and L. Held (2009). Predictive model assessment for count data, *Biometrics* **65,** no. 4, 1254–1261.

Fox, E. W., F. P. Schoenberg, and J. S. Gordon (2016). Spatially inhomogeneous background rate estimators and uncertainty quantification for nonparametric Hawkes point process models of earthquake occurrences, *Ann. Appl. Stat.* **10,** no. 3, 1725–1756.

Gneiting, T., and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation, *J. Am. Stat. Assoc.* **102,** no. 477, 359–378.

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes, *Biometrika* **58,** no. 1, 83–90.

Helmstetter, A., Y. Kagan, and D. Jackson (2007). High-resolution time-independent grid-based forecast for m >= 5 earthquakes in California, *Seismol. Res. Lett.* **78,** no. 1, 78–86.

Krüger, F., S. Lerch, T. Thorarinsdottir, and T. Gneiting (2021). Predictive inference based on Markov Chain Monte Carlo output, *Int. Stat. Rev.* **89,** no. 2, 274–301.

Lippiello, E., F. Giacco, L. de Arcangelis, W. Marzocchi, and C. Godano (2014). Parameter estimation in the ETAS model: Approximations and novel methods, *Bull. Seismol. Soc. Am.* **104,** no. 2, 985–994.

Lombardi, A. M. (2015). Estimation of the parameters of ETAS models by simulated annealing. *Sci. Rep.* **5,** no. 8417, 1–11.

Lombardi, A. M., and W. Marzocchi (2010). The assumption of poisson seismic-rate variability in CSEP/RELM experiments, *Bull. Seismol. Soc. Am.* **100,** 2293–2300.

Marzocchi, W., M. Taroni, and G. Falcone (2017). Earthquake forecasting during the complex Amatrice-Norcia seismic sequence, *Sci. Adv.* **3,** no. 9, e1701239, doi: 10.1126/sciadv.1701239.

Molkenthin, C., C. Donner, S. Reich, G. Zöller, S. Hainzl, M. Holschneider, and M. Opper (2022). GP-ETAS: Semiparametric Bayesian inference for the spatio-temporal epidemic type aftershock sequence model, *Stat. Comput.* **32,** no. 2, 29.

Muir, J., and Z. Ross (2023). A deep Gaussian process model for seismicity background rates, *Geophys. J. Int.* **234,** no. 1, 427–438.

Naylor, M., F. Serafini, F. Lindgren, and I. Main (2023). Bayesian modeling of the temporal evolution of seismicity using the ETAS. Inlabru package, *Front. Appl. Math. Stat.* **9,** 1126759.

Ogata, Y. (1978). Asymptotic behavior of maximum likelihood, *Ann. Inst. Stat. Math.* **30,** 243–261.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes, *J. Am. Stat. Assoc.* **83,** 9–27.

Ogata, Y. (1998). Space-time point-process models for earthquake occurrences, *Ann. Inst. Stat. Math.* **50,** no. 2, 379–402.

Ogata, Y. (2011). Significant improvements of the space-time etas model for forecasting of accurate baseline seismicity, *Earth Planets Space* **63,** 217–229.

Omori, F. (1894). On the after-shocks of earthquakes, *J. Coll. Sci.* **7,** 111–120.

Petrillo, G., and J. Zhuang (2024). Bayesian earthquake forecasting approach based on the epidemic type aftershock sequence model, *Earth Planets Space* **76,** 78.

Rasmussen, J. G. (2013). Bayesian inference for Hawkes processes, *Methodol. Comput. Appl. Probab.* **15,** no. 3, 623–642.

Rathbun, S. L. (1996). Asymptotic properties of the maximum likelihood estimator for spatio-temporal point processes, *J. Stat. Plan. Inference* **51,** no. 1, 55–74.

Reich, S., and C. Cotter (2015). *Probabilistic Forecasting and Bayesian Data Assimilation*, Cambridge University Press, Cambridge, United Kingdom.

Ross, G. J. (2021). Bayesian estimation of the ETAS Model for earthquake occurrences, *Bull. Seismol. Soc. Am.* **111,** no. 3, 1473–1480.

Ross, G. J., and A. K. Kolev (2022). Semiparametric Bayesian forecasting of spatiotemporal earthquake occurrences, *Ann. Appl. Stat.* **16,** no. 4, 2083–2100.

Schoenberg, F. P. (2013). Facilitated estimation of ETAS, *Bull. Seismol. Soc. Am.* **103,** no. 1, 601–605.

Schorlemmer, D., M. Gerstenberger, S. Wiemer, D. Jackson, and D. Rhoades (2007). Earthquake likelihood model testing, *Seismol. Res. Lett.* **78,** 17–29.

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, et al. (2018). The collaboratory for the study of earthquake predictability: achievements and priorities, *Seismol. Res. Lett.* **89,** 1305–1313.

Serafini, F., F. Lindgren, and M. Naylor (2023). Approximation of Bayesian Hawkes process with Inlabru, *Environmetrics* **34,** no. 5, e2798, doi: 10.1002/env.2798.

Serafini, F., M. Naylor, F. Lindgren, M. J. Werner, and I. Main (2022). Ranking earthquake forecasts using proper scoring rules: Binary events in a low probability environment, *Geophys. J. Int.* **230,** 1419–1440.

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

Veen, A., and F. P. Schoenberg (2008). Estimation of space-time branching process models in seismology using an EM-type algorithm, *J. Am. Stat. Assoc.* **103,** no. 482, 614–624.

Vere-Jones, D. (1970). Stochastic models for earthquake occurrence, *J. Roy. Stat. Soc. Ser. B* **32,** no. 2, 1–62.

Wang, Q., F. P. Schoenberg, and D. D. Jackson (2010). Standard errors of parameter estimates in the ETAS model, *Bull. Seismol. Soc. Am.* **100,** no. 5A, 1989–2001.

Zhuang, J. (2011). Next-day earthquake forecasts for the Japan region generated by the etas model, *Earth Planets Space* **63,** no. 3, 207–216.

Zhuang, J., Y. Ogata, and D. Vere-Jones (2002). Stochastic declustering of space-time earthquake occurrences, *J. Am. Stat. Assoc.* **97,** no. 458, 369–380.